A Study of the Co-Evolution of the Genome and Epigenome in Colorectal Cancer Using Multi-Omics Profiling

Timon Heide

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

The Institute of Cancer Research.

Department of Molecular Pathology

The Institute of Cancer Research

University of London

November 10, 2021

I, Timon Heide, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Colorectal cancer is one of the most frequent malignancies in the world and contributes significantly to cancer-related death. While previous studies of the genomic alterations in colorectal cancers (CRCs) have significantly contributed to our understanding of this disease, somatic mutations do not seem to fully explain the evolution of malignant phenotypes. Epigenetic alterations have been suggested to play a crucial role in this context, but they remain insufficiently characterised in CRCs. It is also widely recognised that significant genetic diversity exists in all tumours, but how to interpret this heterogeneity functionally is the subject of significant debate. Even less is known about the role of epigenetic heterogeneity in CRCs.

Here I will present an analysis of the genomic and epigenomic heterogeneity in 30 colorectal carcinoma using a novel multi-omics profiling method. This multi-omics profiling method allows for profiling of somatic mutations with whole-genome sequencing, chromatin accessibility with ATAC-seq and gene expression using RNA-seq concomitantly in single CRC glands. Using data from 1,377 samples from 30 primary cancers and ten adenomas, consisting of 1,212 chromatin accessibility profiles and 527 whole-genomes, I will provide a comprehensive map of genetic and epigenetic heterogeneity in these tumours.

Using an ABC-SMC inference framework based on a spatial tumour growth model, I will also demonstrate how measurements of somatic mutations in multiple single glands can be used to identify subclonal driver mutations undergoing selection. This analysis also suggests that individual CRCs might evolve under different degrees of spatial constrain and that this can be inferred from genomic measurements. Both, the presence of selected subclones and the amount of spatial constraint, might constitute a novel ‘evolutionary biomarker’.

Acknowledgements

First and foremost, I would like to thank my supervisor Andrea Sottoriva as well as Trevor Graham for their incredible support over the last years. Without your mentorship and guidance, this work would not have been possible. I am very grateful for your encouragement, patience and guidance you gave me. The work presented here would also not have been possible without the incredible effort of Inma Spiteri and Chris Kimberley, who conducted the tissue and data collection. Many thanks to both of you!

I would also like to thank all current and previous members of the Sottoriva lab. I want to thank George Cresswell, Nick Trahearn, Daniel Nichol, Luis Zapata, Giulio Caravagna, Claire Lynn and Chela James for the many work-related discussions and sometimes simply random chats over the last years. I will certainly miss those. Many thanks also to the many colleagues from the group of Trevor Graham. Especially to Jacob Househam for his advice and the many discussions over the last couple of years.

Equally, I would also like to thank many of my fellow PhD students at the ICR. Particularly Haider Tari, for the many times he listened to my thoughts and gave me a different perspective on things. In the same way, I would like to thank Rachel Parker, Reda Stankunaite, Julia Hoebart and Kate Chkhaidze. I am very grateful to have met you!

Last but not least, I would like to thank my family and friends for their encouragement, help and never-ending support and keeping me motivated.

Contents

List of Figures

List of Tables

Acronyms

5-FU 5-Fluorouracil.

A adenine.

ABC Approximate Bayesian Computation.

ABC-SMC Approximate Bayesian Computation Sequential Monte Carlo.

ACCTRAN accelerated transformation.

AIC Akaike information criterion.

ATAC-seq assay for transposase-accessible chromatin using sequencing.

AUC area under the ROC curve.

BER base excision repair.

BIC Bayesian information criterion.

C cytosine.

CA chromatin accessibility.

CAA chromatin accessibility alteration.

CCF cancer cell fractions.

ChIP-seq chromatin immunoprecipitation assays with sequencing.

CI credibility interval.

CMG chromatin modifier genes.

CNA copy-number alteration.

CPM counts per million reads mapped.

CRC colorectal cancer.

CSC cancer stem cells.

DNA deoxyribonucleic acid.

DNMT DNA methyltransferase.

EQTL Expression quantitative trait loci.

ESS effective sample size.

FDR false-discovery rate.

G guanine.

GC-content guanine-cytosine content.

GDC Cancer Genomic Data database.

InDel insertion or deletion.

InDel/SNV ratio insertion or deletion to single-nucleotide variant ratio.

ITH intratumor heterogeneity.

LD Luria–Delbrück.

LOH loss of heterozygosity.

LP-WGS low-pass whole-genome sequencing.

MAP maximum a posteriori probability.

MC Monte Carlo.

MCMC Markov chain Monte Carlo.

ML maximum-likelihood.

MLE maximum-likelihood estimation.

MMR DNA mismatch repair.

MMRd DNA mismatch repair deficiency.

MNV multi-nucleotide variant.

MP maximum-parsimony.

MRCA most recent common ancestor.

MSI microsatellite instability.

MSS microsatellite stability.

MT mitochondria.

N non-synonymous mutations.

NGS next-generation sequencing.

NMLL negative marginal log-likelihood.

NNMF non-negative matrix factorization.

ODE ordinary differential equation.

PCR polymerase chain reaction.

PIP3 phosphatidylinositol-3,4,5-triphosphate.

PMC Population Monte Carlo.

RNA ribonucleic acid.

RNA-seq RNA sequencing.

ROC receiver operating characteristics.

S synonymous mutations.

SFS site frequency spectrum.

SL synthetic likelihood.

SMC Sequential Monte Carlo.

SNV single-nucleotide variant.

T thymine.

TF transcription factor.

TSG tumour suppressor gene.

TSS transcription-factor start site.

TSSe transcription-factor start site enrichment.

VAF variant allele frequency.

WES whole-exome sequencing.

WGS whole-genome sequencing.

Chapter 1

General Introduction

1.1 A Historical Perspective

A tumour is composed of billions of cells that grow out of control and do not comply with the normal organisation of a tissue (DeVita, Young, and Canellos 1975; Del Monte 2009). Tumours can therefore arise in virtually all multicellular organisms, including the most basic forms of life. Indeed, tumour-like growth have been observed in *Porifera*[1] (Robert 2010) and *Cnidaria*[2] (Squires 1965; Millane et al. 2011; Domazet-Loso and Tautz 2010). Historical accounts of tumours as pathology in humans date back as far as 3000 BC, and methods of treatment were already described by Hippocrates 400 BC (Hajdu 2011).

Explanations for why tumours arise in some people varied over the centuries, ranging from a contagious disease to imbalances of body humors. One of the first correct identifications of an environmental factor causing the development of cancer was the discovery that the exposure to soot increases the risk to develop scrotal cancer by Pott (1775). Today we know that this is due to the ability of chemicals in the soot to cause mutations of the deoxyribonucleic acid (DNA) — a long polymeric molecule composed of a sequence of the four nucleobases adenine (A), cytosine (C), guanine (G) and thymine (T) — that contains the genomic information. Shortly after the discovery that all living organisms are composed of cells (Schwann and Schleyden 1847), Virchow, maybe inspired by others (Tan and Brown 2006; Wright and Poulsom 2012), established that tumours arise as a disease of cells from a single ancestral cell (Virchow 1860).

Based on the observation of abnormal chromosomes in the nuclei of cancer cells (Hansemann 1890), Boveri suggested that tumours develop due to particular abnormal com-

[1] Sponges

[2] Jellyfish and corals.

binations of stainable particles, the chromosomes[3], that are passed during the division of a cell to the daughter cells (Boveri 1914). While this theory was not universally praised at the time, it is from a modern perspective surprisingly accurate. Almost 50 years later, Nowell and Hungerford (1960) discovered the ‘Philadelphia chromosome’, a chromosomal rearrangement of chromosome 9 and 22 that causes the fusion of the genes *BCR* and *ABL1*. It is the only chromosomal alteration required for the development of chronic myeloid leukaemia and occurs in almost all cases of this disease (Quintás-Cardama and Cortes 2009).

Other research conducted during this period of time had, for example, discovered the cancer-causing effect of x-rays (Marie 1910), tar (Yamagiwa 1915) and various chemical compounds (Friedewald and Rous 1944). Francis Peyton Rous had instead identified a filterable transmissible agent, that was able to induce cancer in birds (Rous 1910; Rous 1911) and it would later be discovered that this effect was caused by a tumour-inducing virus (Claude, Porter, and Pickels 1947). All of these observations lead to the formulation of two, each other apparently contradicting, theories i) that cancer arise due to external factors inducing the growth of cells or ii) that cancer arises due to the spontaneous ‘somatic mutation’ of inherent properties of the genomic information.

After the identification of the DNA as the molecule containing the genomic information (Watson and Crick 1953), it was discovered that the ribonucleic acid (RNA) of the Rous sarcoma virus could also be translated into DNA (Temin and Mizutani 1970; Baltimore 1970). This provided an explanation of how the oncogenic information from the virus would become accessible in infected cells (Huebner and Todaro 1969). Shortly later, research of this genomic information identified almost identical versions of the oncogenic parts of the viral genome in the chicken genome itself (Varmus et al. 1972). This implied that the oncogene of the Rous sarcoma virus was of cellular origin and that a corresponding cellular ‘proto-oncogene’ with a normal physiological function existed. Together with the discovery that many of the previously identified carcinogens were indeed able to change the genomic information (i.e., mutagenic) (Ames et al. 1973), it became clear that cancers did primarily occur because of somatic mutations accumulating in cells and that arguments to the contrary by Rous (1967) and others were erroneous.

[3] It would later be discovered that chromosomes contain the DNA and hence all genomic information.

1.2 Cancer as Evolutionary Process

This somatic mutation theory of cancer was also consistent with statistical observations of cancer incident rates (Nordling 1953; Armitage and Doll 1954) and known familiar predispositions to some cancer types (Knudson 1971). Based on the scaling of incident rates with age, Armitage and Doll (1954) deduced that tumour development could be a multi-stage process requiring around six independent mutations.

Armitage and Doll also derived an alternative explanation for the observed incidence rates (Armitage and Doll 1957). They suggested a two-step process in which a first alteration causes a subset of cells to clonally expand (initiation) with an independent second step causing the transition to malignancy (progression). Further analysis of these two theories concluded that tissue-specific effects and a combination of both might explain the incident rates observed in other tumour entities better (Ashley 1969).

Similarly, Knudson (1971) concluded from the observed number of retinoblastomas in patients with and without familial predisposition that two independent mutations were required for the development of this tumour type. This observation was consistent with the loss of both copies (i.e., alleles) of a single gene, which Knudson suspected to be the explanation for his observation. His prediction would later be proven by the discovery of the gene *RB* that, if lost, causes the formation of retinoblastomas (Friend et al. 1986).

The somatic mutation theory of cancer also implied that — just like for species (Darwin 1859) — evolutionary principles applied (Cairns 1975). In a seminal paper Nowell (1976) outlined the principles of the clonal evolution of cancer, highlighting its equivalence to an asexually reproducing species. In this evolutionary framework, the selection of lineages with advantageous variations, arising due to random mutation, causes the cell population to grow faster than others and hence rise to a higher frequency in the population (Nowell 1976). It also explains why the proliferative capacity of tumours generally increases over time or the ability to grow into the underlying tissue and metastasise arise (Figure 1.1).

1.3 Tumour Biology

The perspective of cancer as a genetic disorder arising due to somatic mutation of the normal cells provided the motivation for the identification of responsible genes. These efforts led to the discovery of p53, as a protein-bound by a viral protein in cells transformed by the tumour inducing virus SV40 (Kress et al. 1979; Chang et al. 1979; Linzer and Levine 1979;

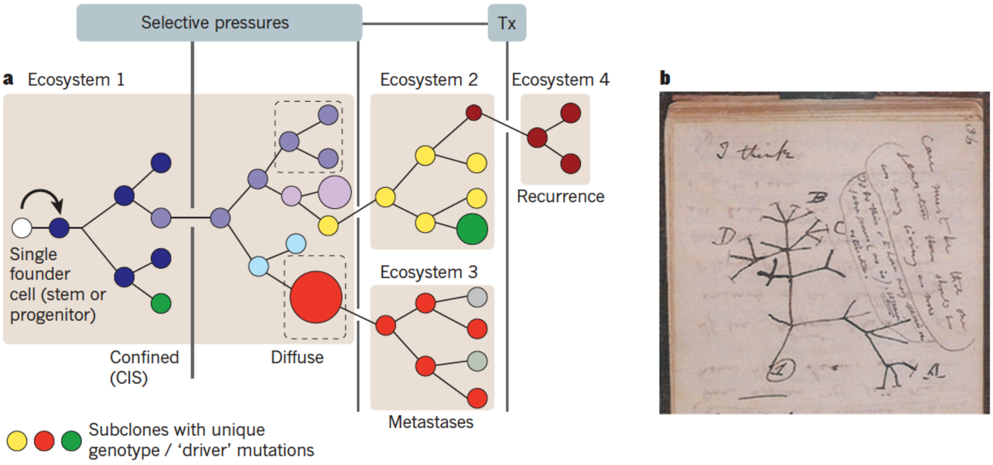


Figure 1.1: Cancer evolution. a) Cancer clones evolve in their micro-environment through an evolutionary process. Vertical lines represent selective pressure and differently coloured circles represent subclones with different phenotypes (i.e., mutations). As already suggested by Nowell (1976) stepwise selection of adapted subclones, might ultimately cause the development of clones able to diffusely infiltrate into the underlying tissue or to metastasis (i.e., to grow in a different ecosystem). Treatment (Tx) introduces a new selective pressure that can cause resistant subclones to arise, hence causing recurrence (dark red clone). b) Darwin's branching evolutionary tree of speciation from his 1837 notebook. (Figure from Greaves and Maley, 2012).

Lane and Crawford 1979) and independently due to its abnormally high expression[4] in some tumours (DeLeo et al. 1979).

The gene *TP53*, which encodes for the protein p53, would later be cloned in mice (Chumakov, Iotsova, and Georgiev 1982) and humans (Matlashewski et al. 1984; Zakut-Houri et al. 1985). Decades of research of the molecular function of this single gene would ultimately lead to the characterisation of its various roles in normal cells and tumours (May and May 1999). Similar research of other tumour-associated genes lead to the discovery of *ERBB2* (King, Kraus, and Aaronson 1985), the RAS gene family (Tsuchida, Ryder, and Ohtsubo 1982; Wong-Staal et al. 1981; Marshall, Hall, and Weiss 1982; Shih and Weinberg 1982) or *APC* (Nishisho et al. 1991).

It was observed that for some of these genes, the mutation of both alleles — equivalent to the 'two-hit' hypothesis proposed by Knudson (1971) for *RB* in glioblastoma — was required. In contrast, other genes only required a single mutation to cause the trans-

[4] The high p53 expression in many tumours (Bártek et al. 1991; Yue et al. 2017), is in contrast to normal cells, in which p53 expression is usually kept at low levels. Non-truncating mutations of p53 are strongly associated with increased expression (Bartek et al. 1990; Alsner et al. 2008), suggesting the change of expression directly arises from higher stability of mutated p53 (Yue et al. 2017). However, the introduction of mutant p53 into mice (i.e., knock-in) does not cause an increase of p53 levels (Lang et al. 2004; Olive et al. 2004), thus suggesting that additional mechanisms are involved (Yue et al. 2017). The negative-dominant phenotype of many p53 mutations would indeed explain why these additional alterations are adaptive and selected for.

formation of cells. These contrasting effects could be explained by the activity these genes have. For genes involved in the active suppression of tumorigenic properties, called tumour suppressor gene (TSG), the second unmutated allele can still perform this activity. Such TSG hence tended to require the loss of both alleles (i.e., they are recessive). In contrast, genes that actively promote tumour-associated traits, like an increased growth rate, typically only required one mutant allele to be present (i.e., they are dominant). These types of cancer-associated genes are also called 'oncogenes' and examples include constitutively active mutants of K-Ras, the most commonly found oncogenes in colorectal cancers (CRCs).

Still, exceptions from this general pattern do exist. An example of this is the previously mentioned p53. While p53 acts primarily as a TSG (Levine and Oren 2009; Vousden and Prives 2009), many of the mutant alleles found in tumours are sufficient to cause a dominant phenotype. This 'dominant-negative' effect is thought to arise from the ability of mutant p53 to disrupt the function of tetrameric p53 complexes. These tetramers are the actual active protein structure able to bind to the DNA and regulate the transcription of target genes. The activity is lost if a single mutant p53 protein is integrated into the complex (Goh, Coffill, and Lane 2011). Since a higher expression of the mutant allele will disrupt an even larger fraction of the p53 complexes (Yue et al. 2017), the 'dominant-negative' phenotype is also able to explain why a high expression of mutant p53 alleles might be advantageous in a tumour.

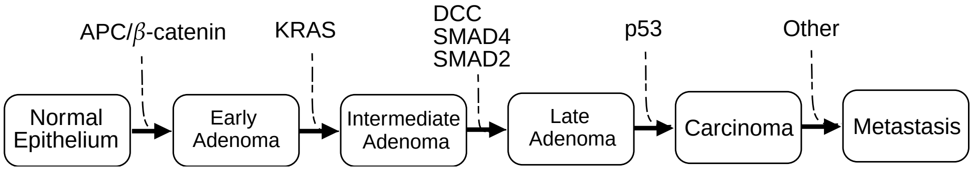


Figure 1.2: Genetic model of colorectal tumorigenesis by Fearon and Vogelstein (1990).

The study of the relative frequency with which tumour-associated genes (i.e., driver genes) were mutated, quickly led to the formulation of sequences able to explain the multi-step nature of cancer. In the context of CRC, the adenoma-carcinoma sequence was described by Fearon and Vogelstein (1990). This simple model of colorectal tumorigenesis suggests that early alterations of APC induce the formation of adenomatous tumours and that the subsequent mutation of K-Ras and the loss of p53 cause the progression to a carcinoma (see Figure 1.2). While simplistic in nature, this model still shapes our understanding of CRC as a disease (e.g., Vogelstein and Kinzler 2015; Lote et al. 2017) and remains at the heart of statistical models (e.g., Paterson, Clevers, and Bozic 2020).

The detailed study of the functional effects mutation of different cancer driver genes had in cells, led to the identification of common characteristics. These ‘Hallmarks of Cancer’ that are acquired during the tumorigenesis, provide a rationale for the interpretation of mutations occurring in individual cancer genomes (Hanahan and Weinberg 2000).

1.4 The Human Genome Project & NGS

The hope that the identification of other cancer-associated genes might reveal a cure for the disease partially motivated large international efforts to sequence the human genome in its entirety. Fierce competition between ‘The Human Genome Project’ (Sinsheimer 1989) and the Celera Corporation lead to the completion of initial drafts of the human genome at the beginning of the 21st century (Venter et al. 2001; Lander et al. 2001). This reference genome provided the basis for the identification of new human genes (e.g., Hubbard et al. 2002) and allowed the study of their evolution in more detail.

Parallel to the sequencing of the human genome, the development of new sequencing techniques allowing high-throughput sequencing took place. These methods are today collectively referred to as next-generation sequencing (NGS). A full description of the various approaches and technologies would certainly be outside of the scope of this simple introduction, but good reviews of NGS methods can be found in Shendure and Ji (2008) or Mardis (2008). The currently most widely used NGS method is Illumina’s sequencing-by-synthesis. Sequencing-by-synthesis allows the parallel sequencing of pools of fragmented DNA with attached primer pairs, so-called libraries.

These libraries are then added onto a glass surface to which complementary primers are covalently bound (Adessi et al. 2000; Fedurco et al. 2006). This causes individual fragments of DNA from the library to bind to the complementary primers. After this, polymerase chain reaction (PCR) based amplification of the bound DNA fragments is performed. At each step of this PCR, the free ends of the DNA fragments form a bridge that binds to a new free primer pair attached to the glass surface. After several rounds, small clusters of nearly[5] identical DNA fragments are formed (see Figure 1.3).

These clusters are then sequenced in a base-by-base fashion using fluorescent nucleotides (Turcatti et al. 2008). At each sequencing step, images of the glass surface are

[5] Errors during the copying of the DNA can arise and are one source of noise that complicates the detection of bona fide mutations from the obtained DNA sequences.

taken. Through the analysis of the images, the sequence of each non-overlapping cluster can be determined. This approach results in the generation of millions of reads, containing the sequence of the DNA in a single cluster and hence the sequence of a single DNA fragment of the library.

In principle, partial matches between reads from overlapping genomic regions can be used to assemble an entire reference genome de novo. Alternatively, reads can be aligned to a known reference genome like the one produced by the Human Genome Project (Schneider et al. 2016). This alignment-based analysis requires far fewer reads and is computationally cheaper. By detecting differences between the reference genome and the sequenced reads, variants present in the library can then be identified. For a variant m at a given locus i the observed variant allele frequency (VAF) f_m of such a variant is given by $f_m = N_m/N_i$, where N_i is the total number of reads covering i and N_m the number of these reads that support the variant m. The observed VAF provides an estimate of the true frequency of the allele in the sequenced sample.

In the absence of a genuine variant, one also expects to see some sporadic mismatches between the reference genome and the generated reads. These mismatches are due to errors that are introduced during the amplification of DNA molecules with PCR or due to random misread bases during the sequencing process itself. The rate at which these errors arise can be locus, library, and sequencing run specific and many different algorithms, so-called mutation or variant callers, have been designed to distinguish bona fide variants from this background noise (see for example Pabinger et al. 2014). With such algorithms, one can readily identify most germline variants a person inherited from their parents.

In order to detect the somatic mutations that are present in a tumour one needs to distinguish these somatic mutations from the millions of germline variants present in all cells of a person. For this, a second normal tissue sample, a so-called reference, is required. Somatic variant callers were developed for the specific purpose of identifying somatic mutations from such paired tumour-normal data. Extensive reviews of the performance of these algorithms have been performed (see for example Wang et al. 2013b; Xu et al. 2014; Xu 2018).

Some variant calling algorithms for both, somatic and germline mutations, are also able to detect mutations that delete or insert a sequence into the DNA. The analysis of these so-called insertion or deletions (InDels), is limited by the length of the available reads. For the sequencing-by-synthesis method, the length of reads that can be obtained is limited to ≈ 150 bp from either end of the fragment. Beyond this, the degradation of the base quality makes sequencing impractical. For this reason, only relatively small alterations can be fully resolved by most currently available NGS data.

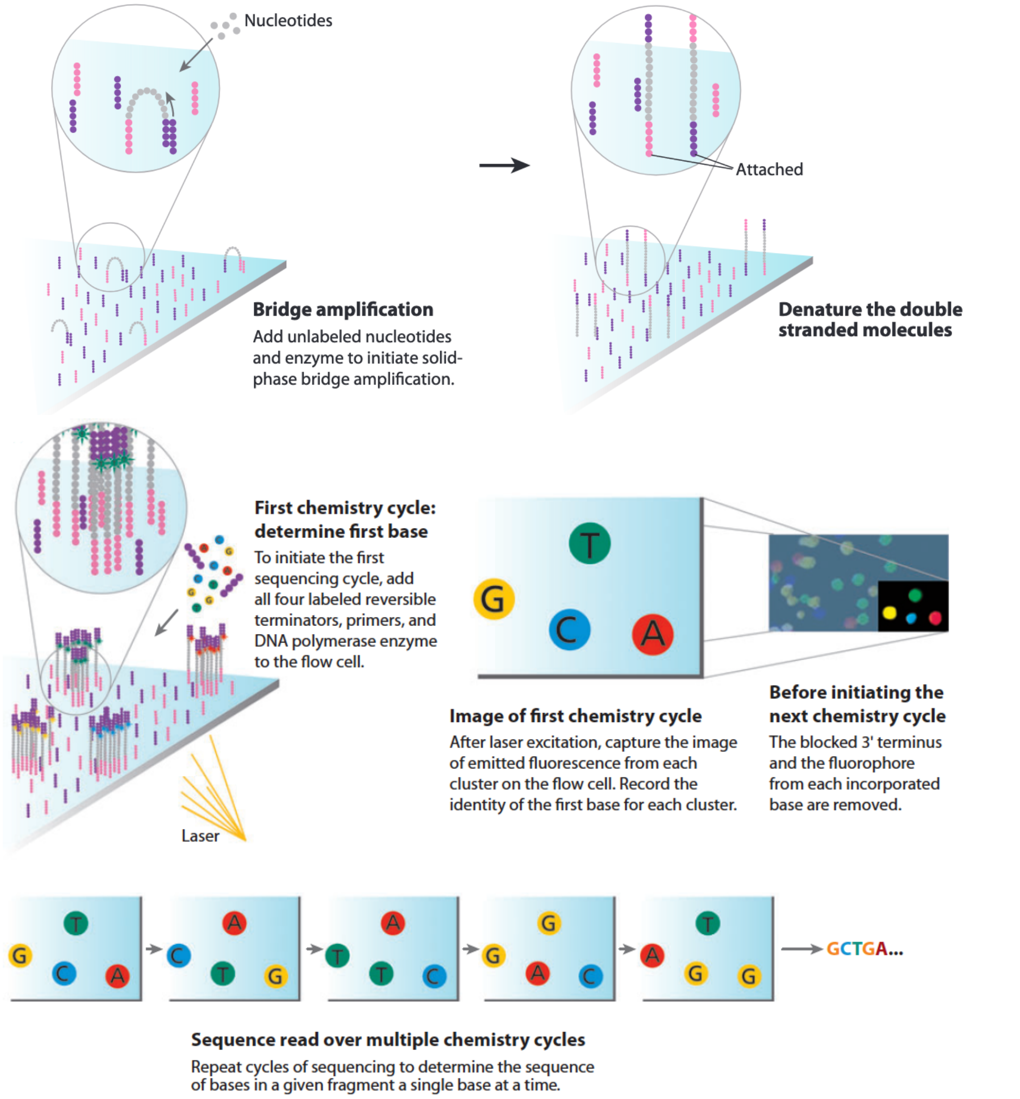

Nucleotides
Attached
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Denature the double stranded molecules
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.
T
G
C
A
Laser
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed.
GCTGA...
Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

Figure 1.3: llumina sequencing-by-synthesis. Clusters of identical fragments are created through ‘Bridge amplification’ on the surface of the flow cell (top row). The sequence of these clusters (bottom left) of reads is then sequenced base-by-base using reversible terminator bases (bottom). During this process, a single base binds to the DNA. The fluorescent signal is picked up using image sensors (bottom right). At this point, the reversible terminator is removed from the base and the process is repeated with the next base (bottom row). (Figure from Mardis, 2008)

A known reference genome and NGS methods now allow to re-sequence the entire human genome within hours. This has provided the technological basis for the comprehensive characterisation of mutations in thousands of cancer genomes as done by The Cancer

Genome Atlas (TCGA, Bailey et al., 2018) or the Pan-Cancer Analysis of Whole Genomes (PCAWG, Campbell and Giocomo, 2019) project. Both studies have also used NGS methods to analyse the transcription of genes in the entire genome or the presence of non-genetic modifications of the genome, which will be explained later.

1.5 Modern Cancer Genomics

1.5.1 Cancer Driver Genes

Large-scale pan-cancer genomic studies have significantly advanced our understanding of the genomic changes underlying carcinogenesis. The analysis of somatic mutations present in individual cancer types has allowed the identification of novel cancer driver genes and the characterisation of the frequency with which these occur in different tumour entities (Kandoth et al. 2013). Similar studies of somatic copy-number alterations (CNAs) of genes have provided significant insight into the recurrence and putative causes of CNAs (Zack et al. 2013).

Due to the complexity of the processes underlying the accumulation of point mutations and their selection, statistical models are required for their analysis. Many approaches to the detection of such recurrently mutated genes exist. Examples of these include methods that analyse an excess of non-synonymous mutations compared to an expected background (Weghorn and Sunyaev 2017; Martincorena 2019; Dietlein et al. 2020), clustering of mutation within protein structures (Arnedo-Pac et al. 2019; Tokheim et al. 2016), the predicted impact of mutations (Mularoni et al. 2016) or a combination of such methods (Lawrence et al. 2014). Dedicated projects for the analysis and curation of such cancer-specific driver genes across datasets and discovery methods have been developed (e.g., Sondka et al. 2018; Martínez-Jiménez et al. 2020).

Despite these efforts little is known about the functional impact the large majority of these driver mutations have *in vivo*. Where such experimental data exist, they often involve mouse models that do not necessarily resemble the effect these have in humans. The longitudinal observation of driver mutations in primary lesions is rarely possible. Longitudinal tracking of somatic driver mutations using liquid biopsies can instead provide a window into disease evolution, but it integrates information over tumour cells from the primary as well as potentially existing metastatic sites (Khan et al. 2018).

Cells in a tumour also do not necessarily shed DNA at a uniform rate from all locations of a tumour. Instead, the rate at which DNA is shed depends on the rate at which tumour cells die, which itself depends on factors like the degree of vascularisation. For this reason, the frequency of mutations in the circulating tumour DNA might not be identical to the frequency of the mutations in all tumour cells. Despite the limited knowledge of the fitness effects of driver mutations *in vivo*, the pan-cancer identification and analyses of driver genes have provided crucial insights into their role in the development of human malignancies and provides the very basis for today's precision oncology and genomic medicine (Vander Velde et al. 2020).

1.5.2 Mutational Signatures

The analysis of somatic mutations across cancer types has also allowed to gain insight into the processes that contribute to their accumulation. In a seminal study Alexandrov et al. (2013b) showed that different mutational processes can be identified based on unique 'fingerprints' from information on the somatic mutations across different tumour entities.

It has in principle been known for a long time that various mutagens affect the DNA in different ways. An example of a well characterised mutational process is the effect of ultraviolet light with a wavelength between 280–315 *nm* (UV-B). The ability of UV-B to induce nucleotide changes and double-strand breaks explains why the exposure to ultraviolet light is a major risk factor for the development of the most common types of skin cancer (Armstrong and Kricker 2001; Narayanan, Saladi, and Fox 2010). The permanent changes of the DNA sequence by ultraviolet light mainly result from the formation of covalent bonds between adjacent pyrimidine bases (i.e., C and T) in the DNA upon exposure to UV-B, causing the formation of cyclobutane-type pyrimidine dimers (Setlow 1966). The incorrect repair of these DNA lesions can then lead to alterations of the DNA sequence itself (Pfeifer, You, and Besaratinia 2005). Depending on which strand of the DNA one considers, this incorrect repair primarily leads to the accumulation of CC>TT/GG>AA and CC>TC/GG>GA (i.e., C>T) mutations.

In line with this, the majority of somatic mutations identified in early sequencing data obtained from a skin-cancer cell line were found to be CC>TT/GG>AA and C>T mutations (Pleasance et al. 2010a). In a small-cell lung cancer cell line sequenced by the same authors, such somatic mutations were in contrast found to be very rare (Pleasance et al. 2010b). This cell line instead mostly showed G>T/C>A, G>A/C>T and A>G/T>C mutations (Pleasance et al. 2010b). This was consistent with earlier observations of mutations in *TP53* obtained through targeted sequencing (Pfeifer et al. 2002) and the mechanism of mutation induction by polycyclic aromatic hydrocarbons, the main mutagenic compounds present in tobacco smoke (Deutsch-Wenzel et al. 1983; Denissenko et al. 1996). Overall, these early studies demonstrated that somatic mutations obtained from sequencing data of individual tumours are a powerful method to characterise the effects of dominant mutational processes.

Alexandrov et al. (2013a) used a similar approach to systematically identify mutational processes based on their induced mutation across patients by using a dimensionality reduction method called non-negative matrix factorization (NNMF) non-negative matrix factorization (Lee and Seung 1999). For the analysis Alexandrov et al. extended the six possible substitutions — i.e., C>A, C>G, C>T, T>A, T>C, and T>G using the opposite strand for sites with a reference G or A base — by the two bases flanking the mutated site (i.e., their 5′ and 3′ context). Since there are four possible bases for the 5′ base, four bases for the 3′ and six substitutions this results in a total of $4 \cdot 4 \cdot 3 = 96$ substitution types. The number m of each of these $K = 96$ substitution types across G patients can be summarised as a matrix

$$M = \begin{bmatrix} m_1^1 & m_2^1 & \dots & m_G^1 \\ m_1^2 & m_2^2 & \dots & m_G^2 \\ \vdots & \vdots & \ddots & \vdots \\ m_1^K & m_2^K & \dots & m_G^K \end{bmatrix}.$$

Alexandrov et al. then used NNMF to factorize this matrix M into a $K \times N$ matrix P and a $N \times G$ matrix E for which $M \approx P \times E$. This approach results in a reduced representation of the data as a linear combination of N 'mutational signatures' stored in the columns of P and the 'exposure' of a patient to each of these stored in the rows of E. They found that the minimum number of N required to factorize the data from total of $7{,}042$ patients from 30 was 21, suggesting that around 21 different mutational process might have been active in various tumours.

It was indeed possible to identify known aetiologies for many of the identified mutational signatures. An example is a signature they primarily identified in lung cancers of smokers, that could be attributed to the mutagenic effects of substances contained in tobacco smoke (Alexandrov et al. 2013b). Another mutational signature, which was only found in sun-exposed skin, matched the known profile of mutations induced by ultraviolet light. A more surprising discovery was that cytidine deaminase from the APOBEC family appeared to contribute substantially to the accumulation of somatic mutations in a subset of tumours from various entities (Nik-Zainal et al. 2012b). Today, similar analyses of 'mutational signatures' in various cancer types have contributed significantly to our understanding of how the exposure to and the activity of mutational processes — that ultimately provide the necessary variation for tumour evolution — change over time and in different disease stages (Alexandrov et al. 2020).

1.6 Intratumor Heterogeneity

As outlined in the above paragraphs, many pan-cancer sequencing studies focused on the analysis of somatic mutations across patients. Doing so, they revealed an extensive pattern of inter-tumour heterogeneity and gained insight into the events involved in the development of the corresponding malignancies. Still, each tumour is composed of 10^8–10^9 cells per gram of tissue (Del Monte 2009) and the dividing cells continue to accumulate mutations as a tumour growths. Since these mutations cause the phenotypic variability that selection can act on, a better understanding of this intratumor heterogeneity (ITH) of mutations is important for the understanding of cancer as a disease (Greaves and Maley 2012; Greaves 2015).

Some mutations can cause subgroups of cells to be better adapted and grow faster, thus causing a positive selection of the corresponding subpopulation. Other mutations might not directly provide a growth advantage, but instead confer resistance to drugs used for the treatment of cancer, these pre-existing resistant subpopulations can then cause the rapid failure of these therapies (Roche-Lestienne et al. 2002; Khan et al. 2018; Shah et al. 2002). Irrespective of their effect on the phenotype, all genetic mutations can serve as naturally arising markers of genetically related cells (i.e., lineage markers) that allow to trace them through time and space. Easily observable markers, like the previously mentioned Philadelphia chromosome, have indeed been used very early to prove that most tumours are clonal and thus arise from a single ancestral cell (Fialkow 1976).

The single cellular nature of cancer makes the observation of most genetic mutations hard (Cairns 1975; Kinzler and Vogelstein 1996), but with technological advancement the detection of other genetic variants became possible.

Examples of this include the usage of microsatellite mutations by Tsao et al. (1998) or gains and losses of chromosomal regions for phylogenetic inference by (Desper et al. 1999). Thanks to the vast improvement of methods over the last decades various other studies (e.g., Siegmund et al. 2009b; Navin et al. 2011; Anderson et al. 2011; Gerlinger et al. 2012; Nik-Zainal et al. 2012a; Sottoriva et al. 2015; Lawson et al. 2020) have enabled similar analyses at ever-increasing level of detail.

The first comprehensive study of the extend of ITH was provided by a seminal study by Gerlinger et al. (2012). In this study, the authors used whole-exome sequencing (WES) — a method that allows the detection of mutations in the majority of the coding genome — of samples from multiple regions obtained from four renal-cell carcinomas. Through this approach, the extensive mutational heterogeneity existing within each tumour was revealed. Similar multi-region sampling combined with NGS based sequencing was applied to many malignancies and, maybe surprisingly, revealed that complex branching patterns and spatial segregation defined the internal clonal structures of all tumours. Likewise, studies of metastasis revealed complex branching patterns that suggested reseeding between sites (Gundem et al. 2015; Yates et al. 2015; Yates et al. 2017; Noorani et al. 2020). While some of these studies identified mutations in previously identified driver genes, how and if these contributed to disease evolution often remained elusive. A notable exception were cases in which multiple independent mutations of the same gene were observed in an independent lineage (e.g., Gerlinger et al. 2012; Gerlinger et al. 2014). This convergent evolution indeed provided strong evidence of context-dependent selection of specific mutations. Nevertheless, convergent evolution of subclonal mutations is fairly rare and the question of whether the observable ITH arises due to pervasive selection of subclones or if it is instead explainable by genetic drift remains unclear.

1.6.1 Phylogenetic Reconstruction

Since the realisation that all existing species arose through the process of evolution from a common ancestor (Darwin and Wallace 1858; Darwin 1859), reconstruction of these ancestral relationships, became a fundamental part of biological research (Haeckel 1866) and is today the subject of the field of phylogenetics. Various methods have been used to reconstruct the relationships between species and between individuals of the same species.

One of the most frequently used methods to reconstruct phylogenetic relationships of N individuals or species is the identification of a tree that requires the smallest number of character changes, a maximum-parsimony (MP) tree. This problem can be split into two sub-problems i) the calculation of the parsimony score S of a given tree T and ii) the exploration of all possible trees. For the calculation of S Fitch's algorithm can be used. Fitch's algorithm labels each internal node with the intersection of the labels of the descendant nodes or if this set is empty, with the union of the labels. The number of changes in sets in the tree are then equal to the minimum number of character changes required. This approach can then be repeated for all analysed characters. The identification of the best tree T is in theory NP-hard and requires to explore the entire tree space. In practice, other approaches, like hill climbing, are often used to reduce the complexity of the problem. Various heuristics, like the parsimony ratchet (Nixon 1999), can be used to ensure that the tree space is explored sufficiently (Felsenstein and Felenstein 2004).

In line with these approaches, many studies of cancer evolution have used methods from phylogenetics to reconstruct phylogenetic trees from mutation data observed in samples. A vast number of algorithms exist for the inference of trees from sequence data and indeed most of these have been used to infer phylogenetic relationships from cancer genome data as well. The options range from simple distance-based methods like unweighted pair-group methods (e.g., Bruin et al. 2014) or neighbour-joining (e.g., Navin et al. 2011; Xu et al. 2012) to maximum parsimony methods (e.g., Bruin et al. 2014; Zhang et al. 2014b). and maximum-likelihood methods (e.g., Jahn, Kuipers, and Beerenwinkel 2016).

The phylogenetic trees reconstructed by these methods are directed and rooted graphs that consist of nodes and edges connecting nodes. In a phylogram, a particular type of phylogenetic tree, each edge has a length proportional to the amount of character change occurring between the two nodes (Santamaría and Therón 2009). Unless otherwise mentioned, phylogenetic trees shown in the following will always be phylograms. Figure 1.4 shows an example of a phylogram and various terms used to describe elements of it are highlighted in it.

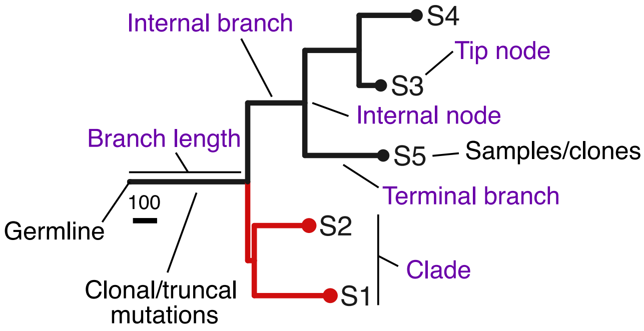


Figure 1.4: Example of a phylogenetic tree. Terms generally used in phylogenomics are shown in blue and those specifically used in cancer evolution are shown in black. Edges associated with the clade formed by the samples S1 and S2 are shown in red.

The tree's root node represents the germline, which can be estimated from appropriate normal tissue (i.e., blood or normal colon bulks). As mentioned before, the length of each edge represents the number of character changes (i.e., mutations) occurring between two nodes. Nodes can either be internal nodes that are connected to two other nodes or a tip node that is only connected to one other node. Tip nodes represent observed samples or clonal entities. Internal nodes instead represent the common ancestor of a set of samples/clones and are unobserved. The subset of a tree containing such a common ancestor and all its descendants are also referred to as a clade (shown in red in Figure 1.4).

Mutations associated with the edge that connects the root of the tree with the most recent common ancestor (MRCA) of the entire tumour are frequently referred to as 'truncal' or 'clonal' mutations and the rest as 'subclonal' mutations. In some contexts, subclonal mutations on terminal edges will also be referred to as 'private' mutations and all the remaining ones as 'shared' mutations.

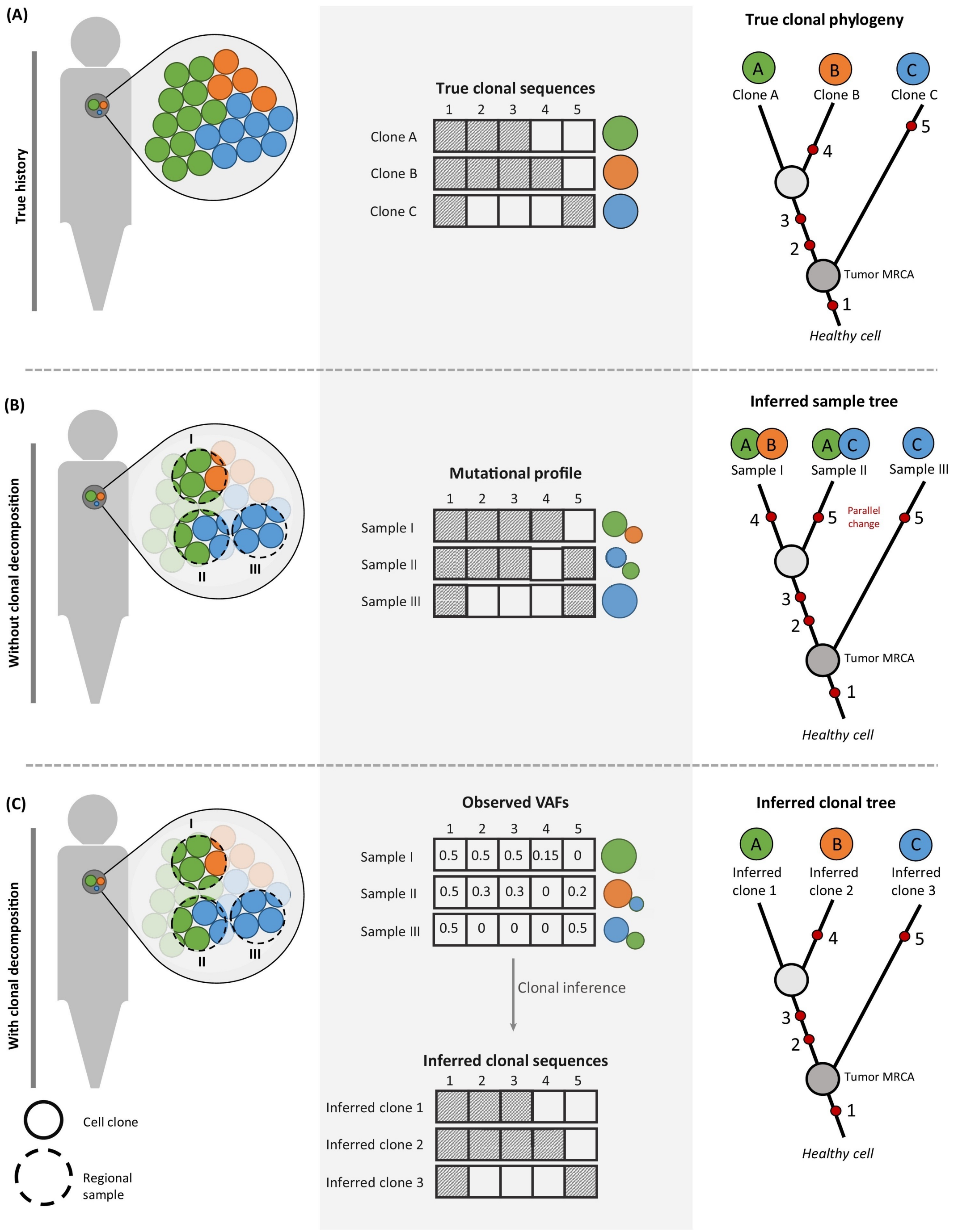
1.6.2 Subclonal Deconvolution

Sample trees are not phylogenies One aspect of cancer genomics data that complicates the application of the described phylogenetic reconstruction methods to multi-region sampling data has to be considered. Each obtained sample consists of an admixture of cells or cell populations that, due to extensive ITH, contain different mutations. When all mutations of each sample are combined, the 'sample trees' reconstructed from these data are not true 'phylogenies' (Alves, Prieto, and Posada 2017).

The reason for this discrepancy is summarised in Figure 1.5A&B. For a tumour with three subpopulations of cells distributed in space (see left of Figure 1.5A), the clonal structure of the tumour can be represented by their three mutation profiles (see middle of Figure 1.5A). Phylogenetic reconstruction methods can be applied to these mutational profiles to infer the ancestral relationships of the subpopulations (see right of Figure 1.5A).

If instead all mutations observed in a spatial sample (see left of Figure 1.5B) are combined and phylogenetic reconstruction conducted on the resulting mutational profiles of the samples (see middle of Figure 1.5B), then a wrong phylogenetic tree might be inferred (see right of Figure 1.5B). This problem can in theory be mitigated through the identification of the 'clonal variants' of each sample, this essentially estimates the mutation state of the MRCA of all cells, but this approach disregards much of the genetic information and can still be problematic if a subclone is present at a high frequency.

Subclonal deconvolution Instead, one should reconstruct the mutational profiles of the present subclonal populations (i.e., clones) from information contained in the VAF of each



True clonal sequences

	1	2	3	4	5
Clone A	▨	▨	▨	□	□
Clone B	▨	▨	▨	▨	□
Clone C	▨	□	□	□	▨

Mutational profile

	1	2	3	4	5
Sample I	▨	▨	▨	▨	□
Sample II	▨	▨	▨	□	▨
Sample III	▨	□	□	□	▨

Observed VAFs

	1	2	3	4	5
Sample I	0.5	0.5	0.5	0.15	0
Sample II	0.5	0.3	0.3	0	0.2
Sample III	0.5	0	0	0	0.5

Inferred clonal sequences

	1	2	3	4	5
Inferred clone 1	▨	▨	▨	□	□
Inferred clone 2	▨	▨	▨	▨	□
Inferred clone 3	▨	□	□	□	▨

Figure 1.5: Phylogenetic Analysis of Bulk Tumour Samples. (A) Left panel: clonal composition of a hypothetical primary tumour. Coloured circles represent the three clones present (Clones A–C). Mid panel: true clonal sequences for five different genomic sites, where the dashed square indicates a somatic mutation. Right panel: true clonal history with red dots depicting the chronological order of mutations. Tumour most recent common ancestor (MRCA) highlighted as an internal node. (B) Left panel: bulk regional samples (I–III), with intermixed clones at different proportions. Mid panel: mutational profile (presence/absence) inferred; dashed square indicates the presence of mutations. Right panel: inferred sample history using maximum parsimony. Red dots depict the chronological order of mutations. (C) Left panel: bulk regional samples (I–III), with intermixed clones at different proportions. Mid panel: variant allele frequency (VAF) estimates for mutation at each sample, and inferred clonal sequences using the Clomial algorithm. (Figure from Alves, Prieto, and Posada, 2017, reproduced under a Creative Commons CC-BY-NC-ND licence.)

of the observed alleles. From the reconstructed mutation profiles, one can in principle infer the correct ancestral relationships (Figure 1.5C). Still, in this approach, the noise associated with NGS becomes a problem.

As described before, NGS generates a set of short reads that can be aligned against the genome. The number of reads N_i covering a genomic site i then defines the ability to resolve the true frequency f_i of the mutation in the population. Ignoring potentially overdispersion, the observed number of mutated alleles y_i follows a Binomial distribution with $X_i \sim B(n_i, f_i)$. This means that if we, for example, assume that two equally sized sets of mutation with $f_1 = 0.4$ and $f_2 = 0.5$ are present in the population and that we sequence this at $\bar{N} = 50$ for only $\approx 10\%$ of the mutations, one can determine to which set they belong at a confidence level of 5%. At $\bar{n} = 100$ this increases to $\approx 31\%$ and at $\bar{n} = 1000 > 90\%$ of mutations could be confidently assigned to either of the two components.

Since sequencing at such high coverage is infeasible in most contexts, statistical methods are often used to instead infer the mixture distributions. Applying these methods to multi-region or single-sample NGS mutation data is called ‘subclonal deconvolution’. These methods have in common that they try to infer the number of mixture components or ‘clones’ and the mixture weight of each. The statistical methods and details surrounding the model vary for any of these, but many are based on Dirichlet Process clustering.

A representative example of these is *DPClust*, which models the VAF distribution as a mixture of n subpopulations of cells, each making up an unknown fraction of tumour cells π_h and contributi ng an unknown fraction of all mutations ω_h. The distribution P of all π_h is modelled as a Dirichlet Process and the number of mutated reads y_i obtained from a variant allele i supported by N_i are then assumed to follow a Binomial distribution. The full model can thus be described as

$$y_i \sim Bin(N_i, \zeta_i(\pi_i)),\ \pi_i \sim DP(P_0, \alpha),$$

where ζ_i is a function that gives the expected VAF of the mutation i if it is present in a fraction π_i of tumour cells. *DPClust* uses the stick-breaking view of the Dirichlet Process

$$P = \sum_{h=1}^{\infty} \omega_h \delta_{\pi_h},\ \omega_h = V_h \prod_{l=1}^{h-1} (1 - V_l),\ \text{with } \pi_h \sim P_0,\ V_h \sim Beta(1, \alpha),$$

where δ_{π_h} represents the indicator function evaluating to one at π_h and ω_h is the weight of cluster h in its implementation.

To obtain samples from the posterior distribution of the model Gibbs sampling is used. The base distribution P_0 is assumed to be $P_0 \sim U(0,1)$, the total number of clusters limited to k and a prior distribution is put on the concentration parameter $\alpha \sim \Gamma(1, \alpha_0))$ with the hyperparameter α_0.

A similar method is used by *PyClone*, which fits a mixture of binomial or overdispersed beta-binomial distributions to cluster mutations (Roth et al. 2014). The Markov chain Monte Carlo (MCMC) step of *DPclust* and *PyClone* is associated with a significant computational cost, which motivated the development of variational Bayesian methods like *SciClone*. *SciClone* can use mixtures of beta, gaussian or binomial distributions to cluster mutation data (Miller et al. 2014). Both *PyClone* and *SciClone* allow the analysis of multiple samples and *PyClone* also allows to conduct the clustering analysis across different copy-number states.

In all cases, the number of reconstructed clusters depends on the available data and importantly is not necessarily equal to the true number of subclonal populations. Mutation clusters present in a very similar fraction of cancer cells, the cancer cell fractions (CCF), are inherently hard to resolve as independent clusters. Low sequencing depth and low tumour purity are other factors that can limit the ability to resolve clusters and can thus cause the underestimation of the number of present subpopulations and their clonal composition. These factors are especially important for single sample sequencing studies. In a multivariate setting, the spatial segregation of mutations often allows to resolve the subclonal structures much better, but the same issues can arise in this context as well.

Reconstruction of phylogenetic relationships Based on inferred subclonal mutation sets identified by clustering methods clone trees — i.e., proper phylogenies — can be reconstructed (Alves, Prieto, and Posada 2017; Dentro, Wedge, and Van Loo 2017; Tarabichi et al. 2018). The methods used for this rely on the ‘pigeonhole principle’, which says that the cellular-prevalence (i.e., the estimated fraction of mutated cells) of a mutation cluster nested into another ancestral cluster must be smaller than the cellular prevalence of the ancestral cluster.

With perfect information on the subclonal structure of a tumour, all trees compatible with the ‘pigeonhole principle’ can be identified. It is important to note that more than one tree can be compatible with the observed subclonal mutation clusters. In this case, the identification of the true tree is then obviously not possible. In principle phasing of mutations located on the same DNA molecule can allow drawing additional inference on the ordering of clusters in a tree, but due to the short read length of readily available NGS, such phasing is often not possible.

Furthermore, the available data themself often do not allow to perfectly reconstruct the subclonal composition of the tumour. Instead, due to the limited sequencing depth, a low number of samples and confounding factors like tumour purity, the subclonal structure can only be resolved imperfectly. This can cause actually separated clusters to be merged and their cellular prevalence to be estimated wrongly. Furthermore, the estimated frequency of mutation clusters is subject to a considerable degree of uncertainty, even if their structure is perfectly resolved.

Due to the imperfect information on the subclonal composition of a tumour, statistical methods that are able to take these errors into account are required. One method that allows the automatic reconstruction of clone trees from the results of subclonal deconvolution has been suggested by Niknafs et al. (2015) Their inference framework combines a fitness function to evaluate the compatibility of a given tree of the inferred subclonal composition with a genetic algorithm to heuristically explore the tree space. An alternative approach that combines both, the subclonal deconvolution and the identification of the clone tree, was suggested by Jiao et al. (2014). This method applies a stick-breaking process that is tree-structured and hence results in tree-compatible cellularity values for mutation clusters.

1.6.3 Definition of a ‘Subclone’

After explaining how statistical deconvolution of NGS data and multi-region sequencing can be used to reconstruct ‘clone trees’, a definition of a subclone is certainly needed. Surprisingly, despite being essential for the interpretation of their results, many publications doing such reconstruction do not define this explicitly (e.g., Dentro et al. 2021). In the following, three possible definitions — all of which are used in cancer genomics studies — will be provided. A more detailed discussion of these and other definitions can be found in Sottoriva, Barnes, and Graham (2017).

The mutation centred perspective When speaking of the results of mutation clustering methods, reconstructed clusters are often referred to as subclones. These can be defined as ‘a set of mutations present in a set of cells due to their shared ancestry’. This mutation centred perspective provides little information on the property of actual tumour cells since mutations from more than one such ‘mutation subclones’ can co-occur in one tumour cell. Given the size of the human genome and the relatively large mutation rates observed in human malignancies, a new mutation cluster is expected to be produced during each cell division of which both created lineages survive. For this reason, there are expected to be more mutation subclones than cells.

The genotype centred perspective When the results of a mutation clustering are instead used to reconstruct a compatible 'clone tree' this tree contains a set of genetic subclones as tip nodes. These can be defined as 'a set of genetically identical cells with common ancestry'. Again, given the size of the human genome and the relatively large mutation rate observed in human malignancies, most cells are expected to accumulate at least one additional mutation during each division. Therefore, the number of such subclones would be almost identical to the total number of cells in the tumour.

The ability to resolve all of these would then primarily be limited by the amount and quality of data obtained. Still, the ancestral relationships of these clones might provide valuable insight into the life history of a tumour. However, the interpretation of these trees is at present still challenging.

Phenotype centred perspective Ultimately, one might be able to use the information contained in reconstructed trees to infer properties of subclones that constitute 'a set of cells with common ancestry with a common phenotype'. This definition allows for the presence of different mutations in cells, but these or other factors must not alter the phenotypic properties of the cells.

The growth dynamics of a tumour containing billions of cells are too complicated to allow an easy interpretation of a reconstructed phylogenetic tree and for this reason, statistical models or simulations that can capture the relevant properties of the process are necessary. Currently, statistical models that allow such inference are lacking. In a tumour significant spatial crowding occurs (Schreck et al. 2019) and how phenotypic properties are altered is not fully understood.

In the following, this definition of a subclone, with the considered phenotype being the replicative potential of cells (i.e., a selected subclone), will be used. How to reliably identify selected subclones from tumour sequencing data is indeed subject of current research and the subject of considerable debate that will be outlined in much more detail in Chapter 2.

It is important to note that all of these possible definitions are fundamentally different in their meaning. This is especially problematic since they are often, at least implicitly, used interchangeably. In the following, a subclone will, unless mentioned otherwise, refer to a subclone that has a selective advantage compared to other cells in the tumour.

1.7 Epigenomics

While the genome provides the ‘blueprint’ for all phenotypes that can be generated by cells in the body, other mechanisms to regulate the expression of these phenotypes must exist. This is obvious as almost all[6] cells of the body contain identical genomic information but express vastly different and stable phenotypes.

Initially introduced as abstract ‘higher level’ or epigenetic control (Waddington 1942), decades of research have revealed a plethora of mechanisms by which this regulation of the expression is archived. By definition, these epigenetic modifications are, like the genome, heritable. Still, unlike the genome, they exhibit much larger flexibility and are controlled by a complex network of regulatory mechanisms. Modification of the epigenome can also occur as a reaction to environmental or cell-intrinsic cues, thus allowing the modification of gene expression (Allis and Jenuwein 2016; Cavalli and Heard 2019; Jung et al. 2020).

1.7.1 Epigenetic Modifications

Various modifications of the chromatin structure — the combination of the DNA and associated proteins — have been identified. Many of these have at some point been implicated in the development of cancer. In the following, the most important epigenetic modifications will be explained in detail.

DNA methylation The most extensively studied epigenetic modification is the methylation of cytosines at the C5 position of CpG dinucleotides (see left of Figure 1.6). Indeed, most CpG dinucleotides are methylated within the genome, and only a small fraction of CpGs in the genome are unmethylated. These are often located in short, $200 - 2,000$ bp long, clusters of CpG rich intervals called ‘CpG islands’ (Suzuki et al. 2007). Such CpG islands frequently occur around the promoter region of genes, and their methylation is associated with reduced expression of the associated genes (Ng and Yu 2015). Similarly, methylation of regulatory elements is associated with a reduction of their activity (Luo et al. 2010). Loss of methylation around retrotransposons — small sequences of DNA that can be removed and inserted in different regions of the genome by special enzymes — is associated with their reactivation and can contribute to genomic instability in CRC (Antelo et al. 2012; Baba et al. 2018).

[6] Physiological somatic recombination is known to occur as part of V(D)J recombination (Market and Papavasiliou 2003), isotype switching of immune cells (Market and Papavasiliou 2003) and brain neurons (Lee et al. 2018).

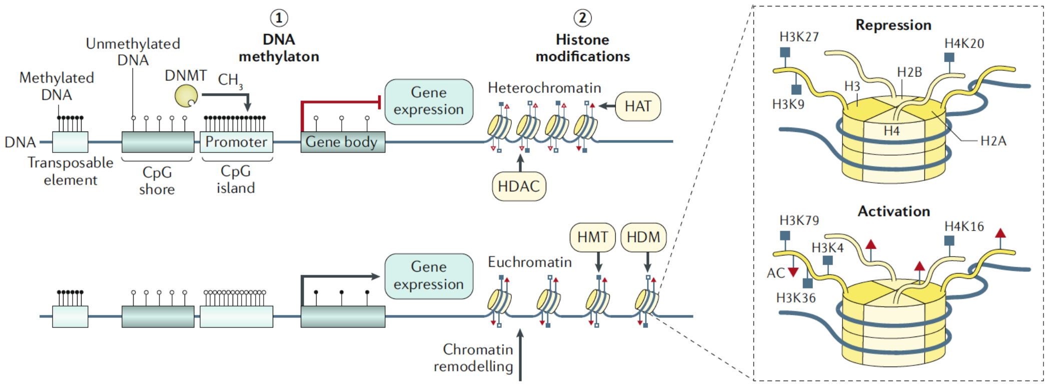


Figure 1.6: Fundamentals of epigenetic modifications. DNA is generally organised in nucleosomes, small segments of DNA wrapped around histones (right inset). These histones are composed of four subunits with tails that can be modified. These histone modifications occur at specific positions of the peptides (e.g., H3K27, meaning histone protein 3, lysine 27). The figure shows the two main chromatin states and their associated chromatin modifications. The first, so-called heterochromatin, is generally compact and less accessible. Heterochromatin is associated with 5-C methylation of CpG dinucleotides transferred by DNA methyltransferases (DNMTs) and methylation (filled blue squares) of histone tails. This type of chromatin is found in the majority of the genome, especially in the promoters of non-expressed genes and transposable elements in the DNA. The second type of chromatin, so-called euchromatin, is less compact and more accessible to protein binding to the DNA. It is associated with low levels of CpG methylation, acetylation of histone tails (filled red squares) and methylation of different peptides in the histone tails (see right inset). Euchromatin can be found around expressed genes, specifically around their promoter regions. Loss of chromatin compaction can lead to the reactivation of transposons in cancer; this can cause them to be reinserted in different genomic regions and contribute to tumorigenesis. In CRC, genome-wide hypomethylation is frequently observed. In a subset of cases, promoter hypermethylation occurs, causing aberrant gene expression. (Figure from Jung et al., 2020)

Histone modifications In normal physiological conditions, the DNA is wrapped around histone proteins, forming the so-called nucleosomes (see right of Figure 1.6). Histones are small proteins composed of eight subunits, which each possess a tail (Chi, Allis, and Wang 2010). A second large group of epigenetic modifications are marks left on peptides at various positions of the histone tails. Common modifications of the histone tails are the addition of methyl and acetyl groups to arginine and lysine peptides in the histone tails (Kouzarides 2007). The absence or presence of these modifications can alter the relative compactness of the chromatin (Struhl 1998), and other proteins can specifically recognise specific histone marks. These DNA binding proteins can cause further modifications of the chromatin structure. Multiple proteins recruited to regions of the chromatin (Jung et al. 2020) can also interact with each other. Together chromatin modifications and proteins binding specific elements of the DNA give rise to a complex and poorly understood regulatory network that ultimately determines how genes are expressed.

Nevertheless, two general histone modification states have been identified. The first one is a generally repressed and compacted state of the chromatin that is associated with H3K27 and H3K9 and H4K20 trimethylation (Peters et al. 2003; Wiles and Selker 2017; Shoaib et al. 2018). Another, generally activated and less compact, chromatin state is instead associated with H3K4, H3K36, and H3K79 trimethylation and the acetylation of histone tails (Kouzarides 2007).

Measurement of chromatin accessibility The presence of histone modifications can be measured using chromatin immunoprecipitation assays with sequencing (ChIP-seq). ChIP-seq isolates small fragments of DNA with modification specific antibodies that can then be profiled through sequencing. Still, ChIP-seq is a very time consuming and complex method. Alternative approaches directly measure the accessibility of the chromatin as a surrogate of the general chromatin states. One such method is called assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al. 2013; Buenrostro et al. 2015). ATAC-seq uses the activity of a modified Tn5 transposase that can nick the DNA and insert short adapter sequences into the flanking regions. The activity of Tn5 occurs primarily in regions of open chromatin and regions deprived of nucleosomes (Figure 1.7a). For this reason, the Tn5 can be used to obtain a high-level surrogate of the general chromatin state. In ATAC-seq promoter regions of actively transcribed genes or active enhancers tend to accumulate many insertions, whereas non-transcript genes do not show an increased number of insertions relative to the background (Figure 1.7c). When two nicks by the Tn5 occur close to each other in the same DNA molecule a barcoded DNA fragment that can be sequenced will be generated. When the NGS reads obtained from these fragments are aligned to the genome, they will reveal characteristic peaks that can be used to profile regions of open-chromatin in a genome-wide fashion. Another important advantage of ATAC-seq compared to ChIP-seq is that it requires very little input material and that it can be conducted with as few as 500 cells (Figure 1.7b).

1.7.2 Epigenetics in Cancer

It is widely recognised that epigenetic alterations play an important role in the development of cancer (Akhtar-Zaidi et al. 2012; Jones and Baylin 2007; Biswas and Rao 2017; Nebbioso et al. 2018). While recurrent genetic alterations detected in tumours clearly demonstrate that they are a major factor contributing to malignant phenotypes, our general understanding of epigenetics and especially its role in human malignancies is still rather limited (Baylin 2011; Lao and Grady 2011; Corces et al. 2018). Much past research of epigenetic alterations in

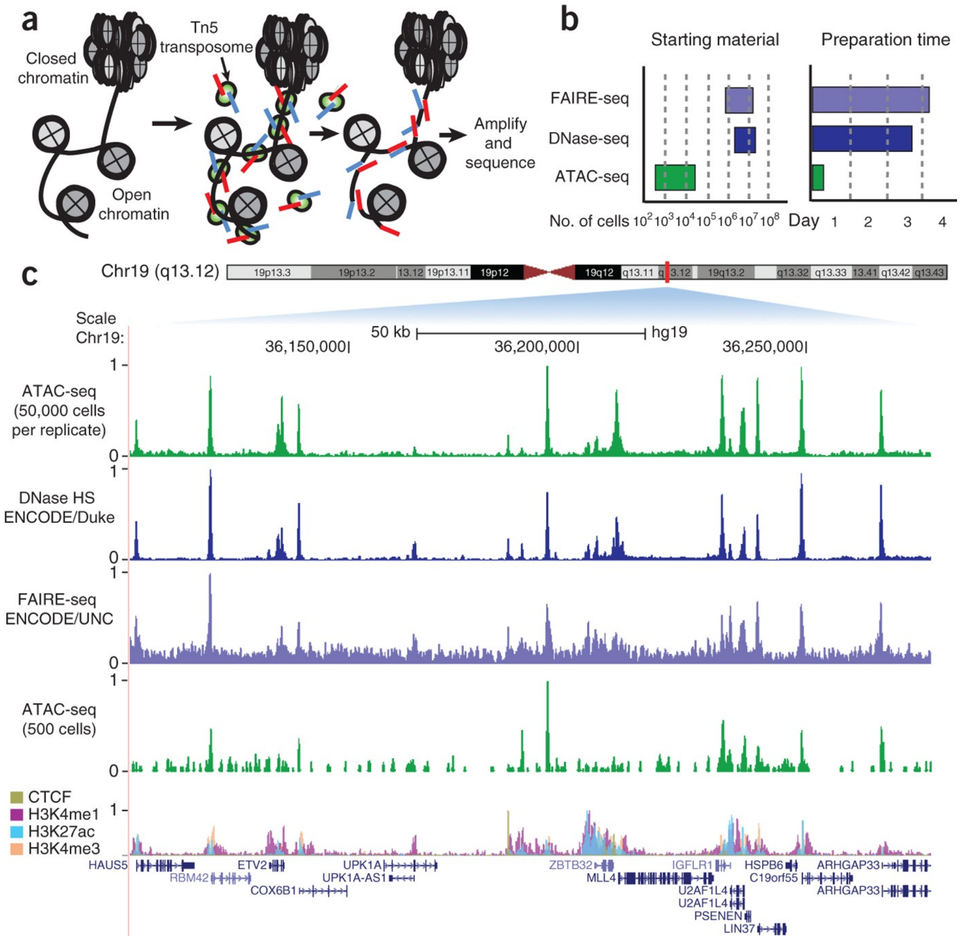


Figure 1.7: The ATAC-seq assay. a) ATAC-seq relies on the activity Tn5 transposon (green) loaded with sequencing adaptors (red and blue). These nick and insert these adaptors preferentially into regions of open chromatin (e.g., between nucleosomes shown in grey). This generates fragments that can be sequenced by conventional NGS. b) Unlike alternative methods ATAC-seq is fast and requires very little input material. c) ATAC-seq generates tracks similar to other chromatin accessibility assays and shows a signal in regions associated with active enhancers and promoters. (Figure from Buenrostro et al., 2015)

CRC has focused on methylation, which is only one of the many known epigenetic modifications (Lao and Grady 2011; Okugawa, Grady, and Goel 2015). This research led to the identification of a CpG island methylator phenotype (Ogino et al. 2008), microsatellite instability (MSI) caused by hypermethylation of DNA mismatch repair (MMR) genes (Herman et al. 1998), a genome-wide hypomethylation phenotype (Suter, Martin, and Ward 2004), and various methylation biomarkers (Okugawa, Grady, and Goel 2015). Studies assessing other epigenetic alterations in CRC are relatively rare, but some have identified recurrent CRC specific alterations of histone modifications (Akhtar-Zaidi et al. 2012). The so-far largest profiling of general chromatin accessibility using ATAC-seq was conducted as part of the TCGA project (Corces et al. 2018) and demonstrated general tissue-specific chromatin states that are able to explain much of the variation in gene expression across tumour types. Still, due to the lack of normal controls, the TCGA study was not able to determine whether the observed chromatin states were a consequence of the tissue of origin or bona fide somatic changes.

Still, very little is known about the ITH of epigenetic alterations and its relationship with the genetic heterogeneity (Black and McGranahan 2021). One seminal study by Roerink et al. characterised the relationship of genomic (WGS), epigenomic (Illumina 450K methylation array) and transcriptomic (RNA-seq) ITH in CRCs (Roerink et al. 2018). For this Roerink et al. analysed single cell-derived organoids from a total of three colorectal cancers along with normal cells from adjacent tissues and showed that heritable and stable subclonal changes occurred in parallel during the expansion of the tumour. Still, clock-like changes of DNA methylation are known to exist (Field et al. 2018; Shibata 2009; Shibata 2011) and a parallel drift of the epigenome and genome seems to be a reasonable assumption. It is of course possible that a subset of the subclonal chromatin state changes observed by Roerink et al. were subject to subclonal selection, but more research is required to elucidate this.

Furthermore, phenotypic changes can also be induced by the microenvironment through pre-existing cellular mechanisms, rather than drift or selection of specific phenotypes (Via and Lande 1985; Price, Qvarnström, and Irwin 2003). This phenotypic plasticity might play a role in the adaptation of cancers to various microenvironments (Anderson et al. 2006; Xue and Leibler 2018; Jolly et al. 2018; Ardaševa et al. 2020). Unfortunately, such microenvironmentally induced differences can be reduced or altered by the *in vitro* cultivation of cells required for the methods used by Roerink et al. For this reason, the concomitant profiling of epigenetic and genetic alterations in primary CRC is required to gain conclusive insight into the prevalence of these.

1.8 Thesis Objective and Outline

The objective of this thesis was to characterise this ITH existing in colorectal carcinoma on the genetic and epigenetic level, as well as the relationship between these. This was done with the goal to derive 'evolutionary biomarkers' that characterise the growth dynamics of individual tumours and to evaluate if they are predictive of the clinical outcome.

In the first two chapters of this thesis, I will illustrate the difficulties of understanding genetic diversity using bulk sequencing of tumour samples. Here, using simulated sequencing data from a stochastic branching-process model of cancer evolution and by reanalysing several large-scale genomic profiling studies, I will show the limitations of simple summary statistics of neutral dynamics, clustering-based methods, and cohort-wide measurements of selection like dN/dS ratios to provide insight into subclonal dynamics from such data.

Following this, I will present results from a novel study on the co-evolution of the genome and epigenome in 30 CRCs at a single-gland level that was motivated by the limitations of bulk whole-genome sequencing (WGS). Here, I will use the multi-omics profiling of individual glands sampled from different regions of the tumours. Using measurements from more than $1,300$ glands of 30 primary cancers and ten concomitant adenomas, consisting of over $1,000$ chromatin accessibility profiles and 500 whole-genomes, I will provide a comprehensive map of genetic and epigenetic heterogeneity in CRCs. I will use these data to identify recurrently altered promoter and enhancer accessibilities and global changes of transcription factor activities.

Finally, I will discuss the observed subclonal architectures of somatic mutations in light of the limited evidence for subclonal selection in most cases. In this context, I will suggest a maximum-likelihood (ML) method to integrate samples subject to WGS and low-pass whole-genome sequencing (LP-WGS) into a single phylogenetic tree. To these trees, I will apply an Approximate Bayesian Computation Sequential Monte Carlo (ABC-SMC) inference framework based on a spatial tumour model that I developed. This provided insight into how competition for space limits expansions on a case-by-case basis and identified subregions likely under selection from driver mutations. The ability to identify such selected driver mutations with this method *in vivo* was also supported by orthogonal dN/dS based methods.

Chapter 2

Neutral Tumour Evolution

2.1 Introduction

Following the previous general introduction into the field of tumour genomics and tumour evolution, I will now provide a more detailed introduction to the current debate on the role and prevalence of selected subclones in tumours. In this context, I will primarily focus on the discussion that followed a seminal paper by Williams et al. (2016), in which the authors suggested that sub-clonal structures observed in a substantial fraction of tumours might also arise in the absence of selection, i.e., under neutral evolution.

Williams et al. based their conclusion on data from the then largest comprehensive study of cancer-genomes, the TCGA project (Bailey et al. 2018). While neutral evolution had been long debated in species evolution, little thought was given to this idea in the context of the somatic evolution of tumours. Maybe curiously, the publication by Williams et al. (2016) was subject to heavy criticism (Tarabichi et al. 2018; Balaparya and De 2018; Noorbakhsh and Chuang 2017; Wu et al. 2016; McDonald, Chakrabarti, and Michor 2018, i.e.,). Others criticised that the test statistic used by Williams et al. (2016) lacked sufficient power to reject the null-hypothesis (i.e., neutrality) and that some models of selection might be practically indistinguishable from the neutral model considered by them.

Interestingly, the general debate of these ideas in the field of cancer genomics (e.g., Bozic, Gerold, and Nowak 2016; Davis, Gao, and Navin 2017; Sun et al. 2017; Turajlic et al. 2019; Williams, Sottoriva, and Graham 2019; Lakatos et al. 2020; Li et al. 2020), resembled the general discussion of idea of neutrality in the field of population genetics and other fields. This eventual even lead some to suggest that there exists a ‘neutral syndrome’ (Leroi et al. 2020), a fascination with the ability of neutral models to give rise to observable patterns. It is certainly true that abundance distributions, like the VAF distribution of alleles obtained from bulk tumour sequencing data, contain only little information on the presence of selection.

Still, this is an important point in itself, given that relatively little attention was given to mechanistic models like the one used by Williams et al. (2016). In the end, two fundamental questions that were raised by Williams et al. and which remain unanswered are: ‘How frequent is subclonal selection within established tumours?’ and ‘Can one use the distribution of somatic variants in a tumour to identify selected subclones’? It is thus not surprising, that the discussion of how to integrate neutral evolution it into the interpretation of cancer genomic data is still on-going (e.g., Caravagna et al. 2020; Edwards, Marusyk, and Basanta 2020; Diamond et al. 2021; Dentro et al. 2021; Black and McGranahan 2021).

In the following a more detailed introduction into the ‘neutral theory’ of Kimura (1968b) in the field of population genetics and dN/dS based methods that can be used to deduce that a population was subject to selection will be provided. After this general introduction of these two relevant topics, I will outline the debate surrounding neutral tumour evolution in the field of tumour evolution. In this context I will present a detailed analysis of the criticism by Tarabichi et al. (2018) that motivated some of the work presented in the following chapters. These were published as reply to Tarabichi et al. (Heide et al. 2018), and in the presentation of them, I will follow the general structure of it. Some of the results were also used in reply to criticism by Balaparya and De (2018) and published separately (Williams et al. 2018a).

2.1.1 Neutral Evolution

In a study in which he tried to reconcile the apparent excess of mutation arising in species evolution (e.g., Zuckerkandl and Pauling 1965; Buettner-Janusch, Buettner-Janusch, and Mason 1969) when compared to the rates expected under theoretical models (Haldane 1957), Kimura (1968b) suggested that a substantial fraction of occurring mutations might be selectively neutral or nearly-neutral. Essentially the same idea was also brought forward by King and Jukes (1969) a year later. While Kimura as well as King and Jukes never questioned the fundamental importance of selection as driving force of evolution, their theories did question whether most variants that fixed in a population — i.e., become present in each individual — did so due to Darwinian selection. Instead they proposed that these fixations could occur due to chance — that is, genetic drift — alone (Kimura 1983; Kimura 1989). Other researcher had also considered dynamics of neutral mutations in populations (i.e., Fisher 1923; Fisher 1958; Wright 1931; Kimura 1955), but by assuming this process to be ubiquitous and studying the implications of this neutral theory, Kimura was able to gain significant insight into the evolution of neutral alleles in a population (Leigh 2007). In this context, Kimura introduced the concept of the infinite sites model in which only unique and novel mutations that are not subject to recombination arise. This model can, for example, be used to make predictions about the number and distribution of alleles present in a population of finite size (Kimura 1969).

Maybe because the theory of neutral evolution was in a stark contrast to the predominant concept of evolution as described by Darwin, the idea of neutral evolution caused immediate criticism and a heated debate (Smith 1968; Langley and Fitch 1974; Gillespie 1984; Kreitman 1996). Today, more and better data as well as improved statistical tests for the detection of selection in sequence data (Tajima 1989; Macdonald and Long 2005) have led to the discovery of striking examples for the selection of adaptive variants in species evolution (Macdonald and Long 2005; Boyko et al. 2008; Halligan et al. 2010; Carneiro et al. 2012; Enard et al. 2016). Discoveries like these have led some to suggest that the neutral theory in itself has outlasted its usefulness and should not be used as a universal basis for hypothesis testing (Kern and Hahn 2018). Nevertheless, doing so still provoked a harsh reaction (Jensen et al. 2019). As reiterated by Jensen et al. (2019) the neutral theory is fundamentally important as most of the genome is not conserved (e.g., many non-coding regions of the genome) and hence only subject to drift. Further signals arising from demographic dynamics, negative selection and hitchhiking of alleles due to genetic linkage can complicate the analysis of genomic data in light of selection (Jensen et al. 2019). Here neutral evolution can serve as a reasonable null model to compare observations again.

While a fascinating debate in itself, there is a key difference in the concept of neutral evolution used in population genetics and its application to cancer genomics. The former is mostly concerned with the evolution of variant alleles in a given, finite or constant, population and the dynamics of such new alleles in the population (e.g., Kimura and Crow 1964; Kimura 1968a; Kimura and Ohta 1969). Opposed to this, cancer is a disease in which one cell expands clonally to an extremely large number (i.e., $\geq 10^8$ cells, Del Monte 2009) of cells through repeated division. In such an expanding population of cell these principles identified by Kimura and others do not apply.

Instead, the allele distribution one expects to observe under neutrality in cancer, is that of the famous Luria–Delbrück (LD) model, which Luria and Delbrück used to demonstrate experimentally that the evolution of resistance to bacteriophages in bacteria arises due to random mutation of sensitive bacteria (Luria and Delbrück 1943). The LD model describes a population of bacteria arising from initially sensitive bacteria through exponential growth and in the absence of any selective pressure. During each division mutation from a sensitive to resistant type are assumed to occur with a given probability. The mutation is assumed to not have any effect on the growth rate in the absence of bacteriophages and mutation back to a resistant state is assumed to never occur.

To distinguish this model from the alternative model, which assumed that a resistant state was only acquired in the presence of the bacteriophages (i.e., induced), Luria and Delbrück (1943) plated solutions of bacteria onto multiple plates. After the bacteria grew to a confluent layer in these, bacteriophages were added and the number of resistant colonies was determined. Luria and Delbrück (1943) showed that the high variability of the number of resistant individuals in the plates was insufficiently explained by the expectation of the alternative model (i.e., a Poisson distribution) and that the LD model provided a better explanation for their observations. Under the LD many resistant individuals arise if a random mutation occurs early in an individual that ultimately gives rise to a large population of daughter cells, thus greatly increasing the variability of the number of resistant individuals per plate.

The elegant experiment they conducted showed that the expected number of pre-existing resistant bacteria in a population grown from a single sensitive bacterium was equivalent to the number they observed in experiments. The mathematical analysis of the birth-death process underlying the LD model has proven challenging, but solutions of the probability distribution of the process have been derived (Antal and Krapivsky 2011; Kessler and Levine 2013). Due to its applicability to cancer, variations of this model have been used to study the evolution of drug resistance (e.g., Coldman and Goldie 1986; Komarova 2006; Iwasa, Nowak, and Michor 2006; Tomasetti and Levy 2010; Kessler, Austin, and Levine 2014), metastasis (e.g., Michor, Nowak, and Iwasa 2006; Dingli et al. 2007; Yachida et al. 2010; Haeno and Michor 2010), and carcinogenesis in general (e.g., Kendall 1960; Moolgavkar 1986; Bozic et al. 2010; Bozic, Gerold, and Nowak 2016; Durrett et al. 2010; Diaz Jr et al. 2012).

2.1.2 dN/dS Ratios

A number of statistics were developed to detect deviations from neutral evolution based on the site frequency spectrum (SFS), the distribution of the allele frequency $f_i = N_i/N$ of alleles i present in N_i individuals of a population of N individuals (Weir and Cockerham 1984; Tajima 1989; Fu and Li 1993; Fay and Wu 2000). Probably the most well know of these statistics is Tajima's D (Tajima 1989). Tajima's D compares the average observed number of pairwise sequence differences between individuals π from a constant effective population size N and given mutation rate μ, against the expected number of divergent sites in a population of effective size N at equilibrium under neutrality ($E[\pi] = 4N\mu$). Still, the power of these statistical tests can be limited (Neuhauser and Krone 1997; Nielsen n.d.). They can also be sensitive to non-selective population dynamics, like temporal changes of the population size (Sano and Tachida 2005; Jensen et al. 2005; Haddrill et al. 2005; Ramírez-Soriano et al. 2008; Simonsen, Churchill, and Aquadro 1995) or spatial dynamics (Ray, Currat, and Excoffier 2003) that are often hard to identify themself.

Here dN/dS methods, which are instead based on the analysis of the effect mutations in protein-coding genes have on the peptide sequence of proteins, provide a valuable orthogonal alternative. dN/dS methods are not based on a specific model explaining the SFS. Instead, dN/dS methods exploit the general property of the genome that only some mutations change the encoded peptide sequence of proteins. This is a property arises from the universal genetic code (Hinegardner and Engelberg 1963; Woese 1964) that translates information from DNA into a sequence of peptides. The genetic code is based on a sequence of trinucleotides, called codons, which each encode for a specific amino acid (Crick et al. 1961; Nirenberg and Matthaei 1961). All possible codons could theoretically translate $4^3 = 64$ amino acids, but only 20 canonical proteinogenic amino acids exist. While three codons cause the termination of the translation into protein sequences (stop codons), the remaining 41 codons encode an amino acid for which at least one other codon exists. From this redundancy of the code, follows that only some mutations, the so-called non-synonymous mutations (N), can change the protein encoded by a gene and that the majority of mutations do not cause a change of protein sequences, hence called synonymous mutations (S).

Since natural selection can only act on the phenotypic differences that arise from structural changes of proteins, S mutations are expected to be selectively neutral. N variants might instead also be under negative or positive selection. The information of non-selected

S variants can hence be used to construct a background model of mutation rates at different sites of the genome. A depletion or an excess of N variants compared to this background model — that is a difference in the rate dN and dS at which these arise — can provide evidence for the selection of a subset of N mutations. Due to their relative simplicity, dN/dS ratios[1] have a long history in population genetics for the detection of selection in sequence data (reviewed in Yang and Bielawski 2000).

Many different methods for calculating dN/dS ratios have been suggested, but in general, these can be grouped based on two properties. First, based on whether they calculate average dN/dS ratios across genomic regions or if they calculate site-specific estimates (Kosakovsky Pond and Frost 2005). Secondly, based on the statistical approach used to calculate the estimates (Yang and Bielawski 2000; Kosakovsky Pond and Frost 2005). simple count-based methods just determine the number of N & S sites, calculate the ratio of the two and then apply a correction factor for biases affecting the ratio in the absence of selective forces. For the analysis of sequence data obtained in the field of population genetics these factors are usually differences in the mutation rate of transitions (i.e., A↔G and C↔T mutations) compared to transversion (i.e., A↔C, A↔T, C↔G, and G↔T mutations) and the codon usage. Some adaptations make the simplistic assumption of equal transition/transversion rates and uniform codon usage (Miyata and Yasunaga 1980; Nei and Gojobori 1986), while others take into account differences of the former (Li, Wu, and Luo 1985; Comeron 1995; Pamilo and Bianchi 1993) or both (Yang and Nielsen 2000). For the analysis of somatic variants detected in tumours, a similar method has been used by Zapata et al. (2018) to assess the prevalence of negative selection. A second class of methods are likelihood-based and directly infer parameters of a substitution model, one of which is the dN/dS ratio (i.e., as a single parameter, often denoted ω) itself (Goldman and Yang 1994; Muse and Gaut 1994; Muse 1996). Both *CBaSE* (Weghorn and Sunyaev 2017) and *dndscv* (Martincorena et al. 2017) are examples of such methods in the context of cancer genomics.

While the various statistical approaches (e.g., summary statistics or likelihood-based methods) tend to obtain similar dN/dS estimates, the assumptions underlying the models themself (e.g., regarding transition/transversion rates or codon usage) tend to have a large influence on the results they obtain (Yang and Bielawski 2000; Kosakovsky Pond and Frost 2005). In the context of somatic mutations and especially in cancer genomes,

[1] Especially in the field of population genetics, the synonymous term K_a/K_s ratio is often used.

several additional factors further complicate the correct estimation of dN/dS-ratios. One major reason for concern are variations of mutation rates across the genome, potentially invalidating models that do not account for these. Regions of closed chromatin, for example, tend to accumulate mutation at a higher rate than those in open chromatin (Polak et al. 2015; Schuster-Böckler and Lehner 2012). Similar effects are caused by differences in gene expression (Fousteri and Mullenders 2008; Pleasance et al. 2010a) and replication timing (Stamatoyannopoulos et al. 2009). Furthermore, complex mutational processes with activity in specific mutational contexts tend to be active in human tumours (Nik-Zainal et al. 2012b; Roberts et al. 2012; Alexandrov et al. 2020). The knowledge of these biases has motivated the development of statistical methods that are able to correct for these sources of variation (Lawrence et al. 2013b).

Martincorena et al. (2017) used such a model to adapt the classic dN/dS methods for somatic variants. Applied to sequence data from cancer genomes, the dN/dS estimates of their model suggested that negative selection, which would be indicated by $dN/dS < 1$, was absent at most genomic sites and that positive selection, indicated by $dN/dS > 1$, acted on sites in known cancer driver genes. This observation was in stark contrast to species evolution, where the majority of variants have a deleterious effect and are quickly removed by purifying selection, thus resulting in global dN/dS ratios $\ll 1$ (Yang et al. 2000).

2.1.3 Neutral Tumour Evolution

In a seminal study Williams et al. (2016) suggested that the observable subclonal structures of many tumours could also arise under neutral evolutionary dynamics. In their publication Williams et al. demonstrate that under the assumption of exponential growth, the cumulative number of subclonal mutations in the VAF spectrum is expected to follow a simple power-law distribution. They showed that the expected number of mutations M in the VAF interval $[f, f_{max}]$ of a tumour with mutation rate μ and a fraction of 'effective divisions' β is then given by

$$M(f, f_{max}) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{max}}\right).$$

Here μ is the number of mutations introduced into the genome of the sister cells during each division and β the fraction of division for which both resulting lineages survive, which might not be the case due to random cell death.

The model analysed by Williams et al. (2016) is equivalent to the previously mentioned LD model and a general solution of the stochastic problem has been derived by Kessler and

Levine (2013), who showed that the number of mutants follows a one-sided Levy α-stable distribution with $\alpha = 1$ in the very large population limit. At high frequencies assessable by currently used next-generation sequencing data the tail of this Landau distribution can be approximated by $1/f^2$ as done by Williams et al. (2016) and others before (e.g., Griffiths and Tavaré 1998; Durrett 2013; Nicholson and Antal 2016).

Williams et al. specifically used the resulting linear relationship between the inverse of the VAF $1/f$ and the cumulative number $M(f)$ of mutations in the interval $[f, f_{max}]$, to determine if the observed VAF data obtained from real tumours followed the distribution expected under neutrality. For this, they fitted a linear model with a fixed intercept to the observed VAF data of each tumour and calculated the coefficient of determination R^2 to measure the obtained goodness-of-fit. Cases with a $R^2 \geq 0.98$ were assumed to be consistent with the neutral model. Williams et al. limited this analysis to mutations with a VAF in the interval $[0.12, 0.24]$ and in diploid regions. The lower bound of this interval was motivated by the general limit of detection of $\approx 10\%$ of the algorithm used for the variant calling (Cibulskis et al. 2013). Since one can typically observe a large number of clonal mutations present in all tumour cells, an upper bound of 0.24 was chosen so that these would not affect $M(f)$.

The expected VAF of clonal mutations at a diploid locus is $\bar{f} = 0.5$. For a mutation at a triploid locus only one out of three alleles are mutated hence resulting in a lower expected VAF of $\bar{f} \approx 0.33$. Another factor that influences the expected VAF is the fraction of normal cells that contaminate the analysed tumour tissue. It is not unusual to see a purity of less than 70% and in this case $\bar{f} \approx 0.35$ for a diploid site. Since the noise associated with NGS can be described by a Binomial distribution, it follows that for a diploid tumour with a purity of 70% less than 1.2% of truly clonal variants are expected to be observed at a VAF ≤ 0.24.

The scaling behaviour of the subclonal VAF and the described influence of clonal variants can easily be seen in simulated sequencing data obtained from a stochastic branching process like the one used in Williams et al. (2018b). Figure 2.1A shows an example of such simulated sequencing data of a neutral tumour obtained from the model used in Heide et al. 2018 (see Methods 2.2.2 for details). In this model a tumour is grown from a single transformed cell with N_{clonal} mutations in an asymmetric branching process using the Gillespie algorithm (Gillespie 1977). Cells are assumed to randomly give birth to two daughter cells with the rate λ and during the division the daughter cells are assumed to die randomly with

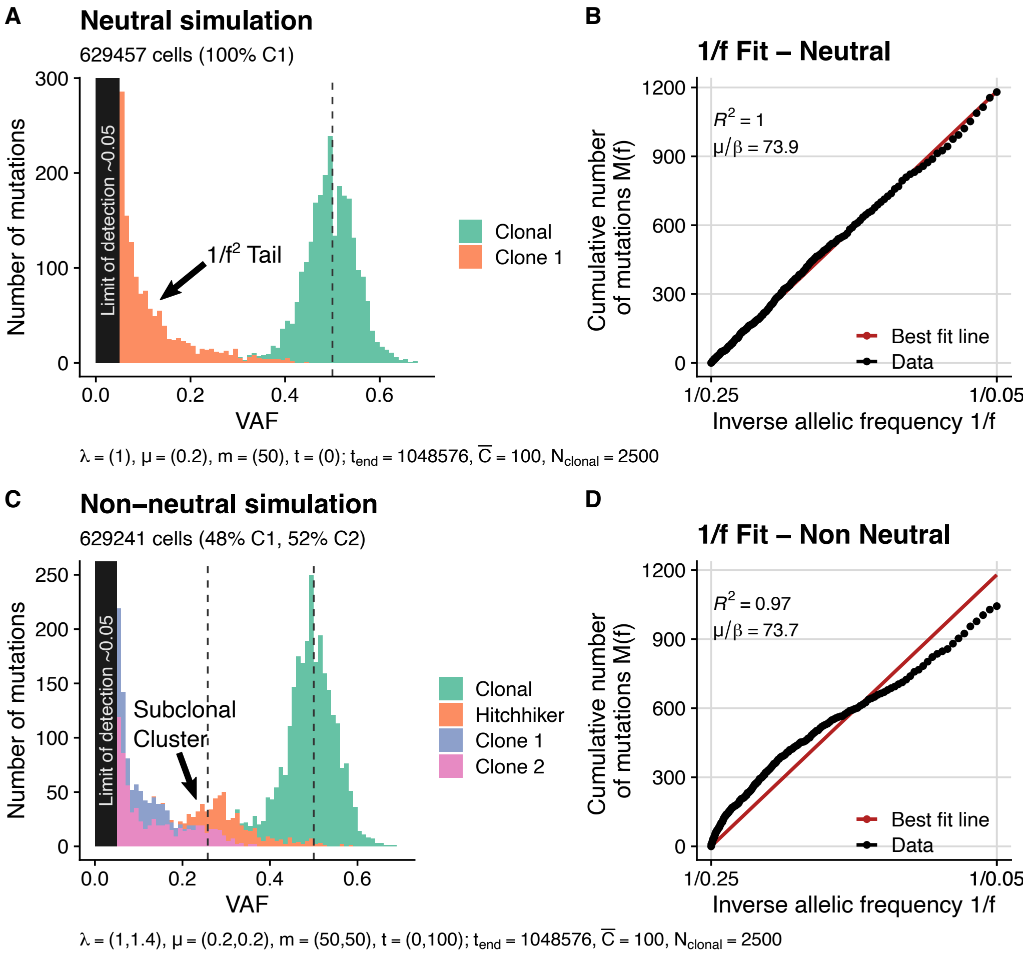


Figure 2.1: VAF spectrum in simulated neutral and non-neutral tumours. Both tumours have 2500 clonal mutations, a mutation rate of 50 mutations per division and a death rate of $\mu = 0.2$. A) The SFS of a simulated neutral tumour. Two structures are visible i) the clonal cluster (green) and a subclonal tail (orange). B) A plot of the cumulative number of mutation against the inverse allelic frequency shows the expected $1/f$ scaling behaviour. C) A simulated sequencing of a tumour with a subclone ($x_{sc} = 51\%$) , showing a subclonal peak at $f \approx 32\%$. Notably, at a higher VAF than 'driver mutations' with $f \approx 26\%$, due to the hitchhiking effect of alleles present in both, the selected and unselected subpopulation. Low-frequency variants are a mixture of two lineages (pink and purple). D) The $1/f$ fit shows a clear deviation from the expected linear scaling between the number of variants with a VAF in the interval $[f, 0.24]$ given by $M(f)$ and the inverse allelic frequency $1/f$. The shown simulations were generated with the simulator used in Heide et al. (2018).

a probability μ. From the generated phylogenetic structure, bulk WGS sequencing data were generated. It was assumed that the number of passenger mutation during each division followed a Poisson distribution with a rate of m.

As expected from the theoretical population genetic model described above, the VAF spectrum of subclonal variants in these neutral simulations tends to follow the expected $1/f^2$ distribution. This can be seen from the linear relationship between the inverse allelic frequency $1/f$ and the cumulative number of mutation $M(f)$ in the interval $[f, 0.24]$ shown in Figure 2.1B. The introduction of a transformed cell with a fitness advantage over the ancestral population (i.e., $\lambda_{sc} > 1$) can instead cause alleles present in the selected subclone to move to a higher frequency. This in turn leads to the presence of a subclonal peak, that causes a clear deviation from the expected $1/f$ scaling behaviour of the VAF spectrum. The VAF distribution of a representative simulated non-neutral tumour, with a selected subclone at a frequency of 52%, is shown in Figure 2.1C. In this simulation, the birth rate of a random cell was increased by 40% when the tumour reached a size of 100 cells. This subclonal cluster causes a clear deviation of the linear $1/f$ scaling as shown in Figure 2.1D.

Since the subclonal structure observed in a substantial fraction of tumours from the TCGA study (Muzny et al. 2012) as well as other bulk-sequencing studies (Wang et al. 2014; Sottoriva et al. 2015) closely resemble that of the neutral theoretical model (i.e., VAF distribution similar to 2.1A), Williams et al. concluded that many tumours evolved effectively neutral. Specifically, by using a high goodness-of-fit as indicated by a R^2 value ≥ 0.98[2] they identify a subset of 32% of tumours in the pan-cancer TCGA cohort that might be evolving neutrally. Still, for the majority of cases (i.e., 68%) the authors identified $R^2 \leq 0.98$ indicating the presence of subclonal selection, or more specifically the deviation from neutral exponential growth.

2.1.4 Criticism of Williams et al. (2016)

Curiously, following the publication of Williams et al. (2016), which brought attention to the concept of neutral evolution in the field of cancer genomics, several authors heavily criticised the methods and conclusions made by them. These followed two main lines of argument i) that alternative models of selection could show patterns identical to neutrality, i.e., questions of identifiability (Balaparya and De 2018; McDonald, Chakrabarti, and Michor 2018) and ii) that the power of the 1/f test is insufficient to reject the null (Tarabichi et al. 2018; Wang et al. 2018a; Noorbakhsh and Chuang 2017).

Unidentifiability of selection in bulk WGS Specifically, McDonald, Chakrabarti, and Michor (2018) argued, using stochastic simulations with a random introduction of subclones, that multiple coexisting subclones could create subclonal structures for which the VAF distribution looks similar to the $1/f^2$ power-law distribution expected under neutrality and that for this reason, the conclusions made by Williams et al. (2016) were logically flawed. Similarly, Balaparya and De (2018) showed that if a significant degree of overdispersion of the

[2] This value was motivated by the observation that none of the simulated neutral tumours showed a $1/f$ fit with a $R^2 < 0.98$ (Williams et al. 2016; Williams 2019).

VAF exists, a single subclone present at a frequency of ≈ 0.15 could cause the right tail of a beta-binomially distributed VAF spectrum to scale very similar to the power-law expected under neutrality. Further Balaparya and De (2018) argued that multiple clones coexisting (i.e., a mixture of binomials) at frequencies between $[0.1, 0.25]$ could likewise cause the mixture distribution to look like neutral $1/f$ tails.

Lack of power of the '1/f test' The remaining criticism primarily focused on the general lack of power to reject neutrality based on observations of the VAF distribution. In this context, Wang et al. (2018a) argued that the narrow window of observability in single bulk-sequencing data severely limits the ability to identify subclonal selection in general. Instead, they suggest that extensive multi-region sequencing methods similar to those conducted by Ling et al. (2015) should be used. The criticism of Noorbakhsh and Chuang (2017) instead focused on the noise of the observational process — reads obtained by NGS methods are approximately binomially distributed — and the consequently limited ability to resolve f_i of individual mutations. They specifically showed that uncertainties in the observed VAF mean that alternative scaling patterns (i.e., $1/f$, $1/\sqrt{f}$ and $1/f^2$) of the subclonal VAF cannot be distinguished at coverage values of $n \approx 100$ available in the TCGA cohort (Muzny et al. 2012).

Critic by Tarabichi et al. (2018) While the criticism by Tarabichi et al. (2018) contained arguments similar to those made by others (i.e., unidentifiability and lack of power), they also provided concrete evidence for the presence of subclonal selection in the tumours Williams et al. (2016) classified as 'neutral'. For this, the authors used a dN/dS based method (Martincorena et al. 2017) that is similar to those commonly used to analyse sequence data in population genetics. As described above, these methods analyse if an excess or depletion of non-synonymous variants relative to synonymous mutations exists.

A depletion of non-synonymous variants (i.e., $dN/dS < 1$) would suggest their removal through negative selection and an excess of non-synonymous (i.e., $dN/dS > 1$) would instead suggest that these were positively selected. In their letter Tarabichi et al. showed dN/dS estimates > 1 for subclonal mutation in TCGA cases for which the '$1/f$ test' did not reject neutrality, demonstrating the presence of selection in these 'neutral' tumours.

They further used simulations across a wide range of subclonal selection rates λ and subclonal mutation rates μ to suggest that the '$1/f$' classifier used by Williams et al. (2016) performs worse than random. Tarabichi et al. based this argument on the receiver operating characteristics (ROC) of the $1/f$ statistic for various thresholds of the R^2 value. The ROC is the curve that shows the relationship between the false positive rate (i.e., neutral tumours classified as non-neutral) and the false negative rate (i.e., non-neutral tumours classified as neutral). A point on the ROC curve below the diagonal of the plot represents a classification that is worse than random for a given discrimination threshold. The area under the ROC curve (AUC) can be used as a summary statistic of a classifier.

For the $1/f$ statistic Tarabichi et al. report a AUC of 42% and a behaviour that is worse than random across a wide range of classification thresholds. While certainly a curious suggestion, this seems to contradict previous theoretical work (see above for details). As Tarabichi et al. provided no explanation for these observations and to address their criticism in general, a detailed analysis of a similar setup was performed. The results of this work, which will be presented below, showed that Tarabichi et al. nonexplicitly made arguments similar to that of other authors: i) some models without selection can lead to (consistent) rejection of neutrality and other models with selection can look like neutral simulations (i.e., unidentifiability) and ii) that the $1/f$ test lacks power in some areas of the parameter space. While these arguments are undoubtedly valid, they certainly apply, as shown below, to other methods as well.

In addition to this analysis, the behaviour of a commonly used Dirichlet Process based clustering method *DPclust* was analysed. This analysis showed that such clustering methods were unable to accurately cluster the mutations of the selected subclones.

2.2 Methods

2.2.1 Analysis of TCGA Data

The dN/dS analysis Tarabichi et al. (2018) used to assess the discriminatory power of the $1/f$ test statistic, was based on *CAVEMAN* (Jones et al. 2016) variant calls from the analysis of TCGA samples by Martincorena and Campbell (2015) . These variant calls were unfortunately not publicly available and *Mutect2* (Cibulskis et al. 2013) variant calls from the Cancer Genomic Data database (GDC) were used instead (Grossman et al. 2016).

2.2.1.1 Pan-Cancer Classification

In order to reproduce the results of the analysis conducted by Tarabichi et al. (2018), somatic variant calls and copy-number array data (log-R ratios) of $8,455$ TCGA tumours were downloaded through the GDC data portal (`https://portal.gdc.cancer.gov/`).

Annotations of sample purities were obtained from a separate study of pan-cancer purities in the TCGA cohort (Aran, Sirota, and Butte 2015). Next, diploid regions in the log-R ratios obtained through GDC were identified and the VAF of somatic mutations adjusted for purity estimates.

As in the original publication by Williams et al. (2016) samples with a purity below 70% or less than 12 diploid subclonal variants with a purity adjusted VAF f within the integration range $[0.12, 0.24]$ were removed from the analysis[3].

The $1/f$ test statistic — i.e., the R^2 value of a linear model with a fixed intercept fitted to $1/f$ and $M(f)$ in the VAF interval $[0.12, 0.24]$ — was calculated for each case on the mutations in diploid regions. In line with the previous analysis by Williams (2018) cases with a $R^2 < 0.98$ were classified as 'non-neutral' and those with $R^2 \geq 0.98$ as 'neutral'. Of the total of $8,455$ tumours analysed 724 satisfied all the filtering criteria (see Figure 2.2A). Of these cases, $1,021$ were already available during the original analysis conducted by Williams et al. (2016) and 117 of them passed the filtering criteria in the analysis presented here (see Figure 2.2B).

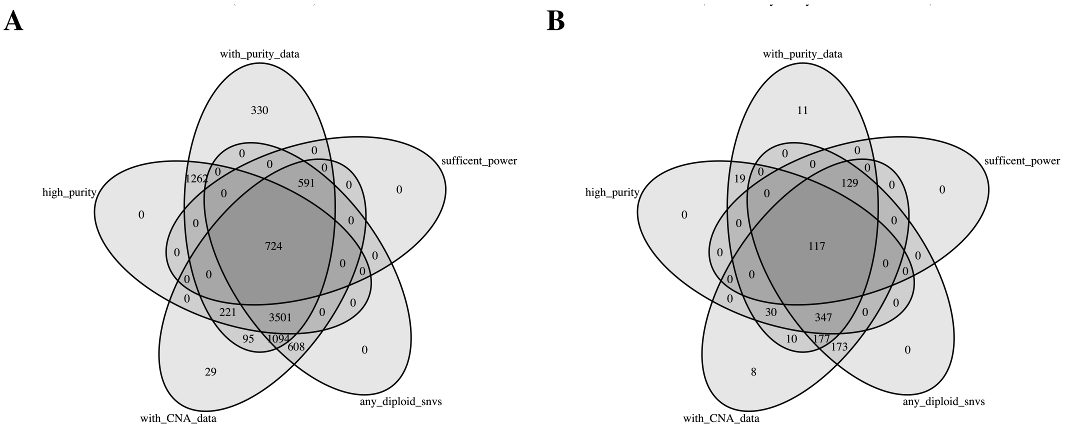


Figure 2.2: Reason for exclusion of samples from reanalysis of TCGA data. All samples were required to have purity data available ('*with_purity_data*'), a sample purity > 70% ('*high_purity*'), matched copy-number data ('*with_cna_data*'), any diploid regions ('*any_diploid*') and at least 12 variants in the interval $[0.12, 0.24]$ ('*sufficient_power*'). A) Annotation of all $8,455$ TCGA samples obtained from GDC. B) Annotation of the subset of TCGA samples analysed by Williams et al. (2016).

[3] No specific reason for the value of 12 subclonal variants, apart from the need to remove cases with too few subclonal variants, was given by Williams et al. (2016), but for the sake of consistency this value was also used here.

2.2.1.2 dN/dS Analysis

For the dN/dS analysis, variants of each case were split — in line with the $1/f$-test integration range of $[0.12, 0.24]$ — into a set of clonal variants with a purity adjusted VAF $f > 0.24$ and a set of subclonal variants with $f \leq 0.24$. The clonal and subclonal mutations of cases were then grouped based on the classification of the $1/f$ test statistic. This resulted in a total of four sets of somatic variants (neutral clonal, non-neutral clonal, neutral subclonal and non-neutral subclonal) on which dN/dS estimates were calculated.

The estimation of these dN/dS values was done with the *dndscv* model developed by Martincorena et al. (2017). To increase the power of this analysis to detect positive dN/dS values, only coding regions of previously identified cancer driver genes were considered. For the pan-cancer analysis of the TCGA cohort a set of 198 previously identified genes reported Martincorena et al. (2017) was used. For the analysis of the 169 colorectal cancers previously analysed in Williams et al. (2016) a set of 369 driver genes from (Martincorena et al. 2017) was used instead. Default parameters were used for the model, this especially uses the included covariate model and the '192r_3w' substitution model.

After the estimation of dN/dS values for cancer driver genes, dN/dS values of genes that are likely not under selection were calculated as reference. For this, a gene set composed of all $\approx 19,000$ genes used by Martincorena et al. (2017) excluding the 198 driver genes, a set of genes that were identified as neutral (i.e., top 25% of the highest p-values) by an orthogonal dN/dS method (Zapata et al. 2018), and third a set of genes reported as neutral by (Martincorena et al. 2017) was used.

To each of these sets of genes, a bootstrap procedure (Efron 1992) was applied to calculate null distributions of dN/dS values to which the point estimates of cancer driver genes could be compared. For this, a random set of genes with a size equivalent to that of the driver genes (i.e., 198) was sampled $1,000$ times with repetition from all genes and dN/dS estimates in each of the four variant sets were calculated for these. For subclonal variants, p-values were calculated by comparing the dN/dS points estimates against the distribution of the three neutral-background sets. For subclonal nonsense variants, dN/dS and p-values were recalculated after the removal of $1/57$ (1.7%) and $11/290$ (3.8%) cases with ≥ 3 subclonal nonsense variants from the gastric and pan-cancer cohort respectively.

Table 2.1: $1/f$ classification results of the TCGA cohort per tumour type.

Tumour type	$R^2 \geq 0.98$	$R^2 < 0.98$	Fraction $R^2 < 0.98$
Adrenocortical carcinoma	1	2	67%
Bladder Urothelial Carcinoma	7	13	65%
Breast invasive carcinoma	21	33	61%
Cervical squamous cell carcinoma	6	20	77%
Colon adenocarcinoma	24	48	67%
Glioblastoma multiforme	30	15	33%
Head and Neck squamous cell carcinoma	6	30	83%
Kidney renal clear cell carcinoma	3	5	62%
Brain Lower Grade Glioma	8	14	64%
Liver hepatocellular carcinoma	4	11	73%
Lung adenocarcinoma	5	31	86%
Lung squamous cell carcinoma	12	41	77%
Ovarian serous cystadenocarcinoma	41	21	34%
Prostate adenocarcinoma	17	30	64%
Rectum adenocarcinoma	11	2	15%
Skin Cutaneous Melanoma	0	18	100%
Thyroid carcinoma	8	10	56%
Uterine Corpus Endometrial Carcinoma	85	89	51%
Uterine Carcinosarcoma	1	1	50%

2.2.2 Stochastic Simulations

An in-depth analysis of the stochastic simulations performed by Tarabichi et al. (2018) was conducted to explain the apparent mismatch between the deterministic model (see Figure 1a, Tarabichi et al., 2018) and the stochastic simulations (see Figure 1b, Tarabichi et al., 2018).

Generation of simulations I assumed that the parameter space explored by Tarabichi et al. (2018) was indeed realistic and explored the behaviour of the model under these parameters in more detail. A stochastic branching process model using the Gillespie algorithm (Gillespie 1976), equivalent to the one used by Tarabichi et al. (2018) implemented in C++ was used for this purpose.

The tumour model was initiated with a single cell of the ancestral cell type. This cell was assumed to carry no mutations (i.e., $N_{clonal} = 0$). The birthrate of all cells of the ancestral cell type were assumed to be $\lambda = 1$ and the doubling of cells in the tumour were simulated using the Gillespie algorithm. During each division either daughter cells was assumed to die with a probability determined by the deathrate μ. At a given population size t_{sc} a random cell of the ancestral type was selected and converted to a subclone with altered birthrate $\lambda_{sc} = 1 + a_{sc}$, where a_{sc} is the relative growth advantage of these cells over the ancestral type. The simulation was terminated once the tumour reached a given size t_{end}. To prevent an entire cell type from dying out the last member of a cell type was assumed to never die.[4] Synthetic sequencing data were then generated from the recorded ancestral history of cells. For each mutation Poisson distributed coverage $N_i \sim Pois(\lambda = \bar{N})$ and a Binomial distributed number of mutated reads $n \sim Bin(N, p_i)$ with success probability $p_i = f_i/2$ being determined by the fraction of cells f_i carrying the mutation i.

A number of parameters were set to fixed values identical to the ones used by Tarabichi et al. (2018). The deathrate was fixed at $\mu = 0.2$ per division and the mutation rate of the ancestral clone was assumed to be $\mu = 16$ mutations per division. All simulations were terminated at a tumour size of $t_{end} = 2^{20} = 1,048,576$ cells. Subclones were always introduced at a population size of $t_{sc} = 2^8 = 256$ cells. Simulated sequencing data were generated with an average sequencing coverage of $\bar{N} = 100$.

For each combination of the subclones selective advantages $a \in \{0, 0.01, 0.02, ..., 1\}$ and mutation rates $\mu_{sc} \in \{1^2, 1.5^2, ..., 10^2\}$ a total of 200 realisations were generated. This resulted in $17 \times 101 \times 200 = 383,800$ simulation for further analysis (Figure 2.3).

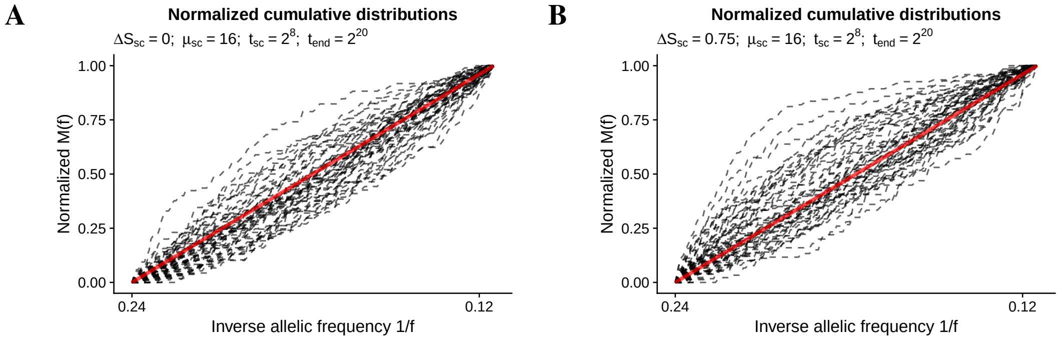


Figure 2.3: Examples of $1/f$ plots in simulated sequencing data. A) Random realisations of neutral simulations ($t_{sc} = 2^8$, $a = 0$). B) Random realisations of simulations with selection ($t_{sc} = 2^8$, $a = 0.75$). The 'normalised $M(f)$' is the cumulative number of alleles $M(f)$ in the interval $[0.25, f]$ divided by the maximum of $M(f)$.

$1/f$ classification Equivalent to the analysis of the TCGA dataset the $1/f$ test statistic (Williams 2018) with the integration range $[0.12, 0.24]$ was applied to the simulated sequencing data. Cases with a $R^2 < 0.98$ were classified as 'non-neutral' and those with $R^2 \geq 0.98$ as 'neutral'.

[4] It should be noted that the choice to not let cell types to die out introduces some biases. Specifically, this leads to a prolonged duration of drift around a low number of cells. For the ancestral clone this would lead to the presence of additional clonal mutations. For a subclone this would likewise cause presence of additional mutations in the clonal peak and a more variable clone size for a given set of parameters. A better approach would be to reject simulations in which either population did die out. Still, the amount of simulated death was low and for this reason differences between both approaches should be relatively small.

Cluster analysis of simulated data Clustering of the simulated sequencing data was done with the Bayesian Dirichlet Process (Dunson 2010) based clustering method implemented in the *DPClust* package for R (Nik-Zainal et al. 2012a; Dentro, Wedge, and Van Loo 2017). *DPClust* and similar methods are commonly used to interpret the VAF spectrum observed in tumour sequencing data (Tarabichi et al. 2021) and to better understand the behaviour of these methods when applied to data from the considered branching-process model was considered to be important.

Bayesian dirichlet process model of *DPClust* *DPClust* models the VAF distribution as a mixture of n subpopulations of cells, each making up an unknown fraction of tumour cells π_h and contributing an unknown fraction of all mutations ω_h. The distribution P of all π_h is modelled as a Dirichlet Process and the number of mutated reads y_i obtained from a variant allele i supported by N_i are assumed to follow a Binomial distribution. The full model is hence described by

$$y_i \sim Bin(N_i, \zeta_i \pi_i),\ \pi_i \sim DP(P_0, \alpha),$$

where ζ_i is the expected VAF of the site if the mutation is present in all tumour cells and π_i the fraction of tumour cells containing i. *DPClust* uses Gibbs sampling to obtain samples from the posterior distribution with priors of $P_0 \sim U(0,1)$ and $\alpha \sim \Gamma(1, \alpha_0))$, where α_0 is a hyper-parameter. The total number of clusters is unusually limited to k.

To characterise the behaviour of *DPClust* on the simulated WGS data, a subset of $3,780$ tumours consisting of 20 simulations in which the subclone made up more than 5% of the total number of cells were selected for each combination of $a \in \{0, 0.05, 0.1, ..., 1\}$ and $\mu_{sc} \in \{2^1, 2^2, ..., 2^9\}$. *DPClust* was then run with the default parameters $k = 20$ and $\alpha_0 = 0.01$ for a total of $10,000$ iterations. Since samples from the beginning of a MCMC chain may not accurately represent the posterior distribution, the first $5,000$ samples were treated as burn-in period and discarded. *DPClust* uses the samples from the posterior distribution to determine the position π_h and weight ω_h of clusters to which mutations can be assigned. Posterior clusters with no assigned mutations were discarded.

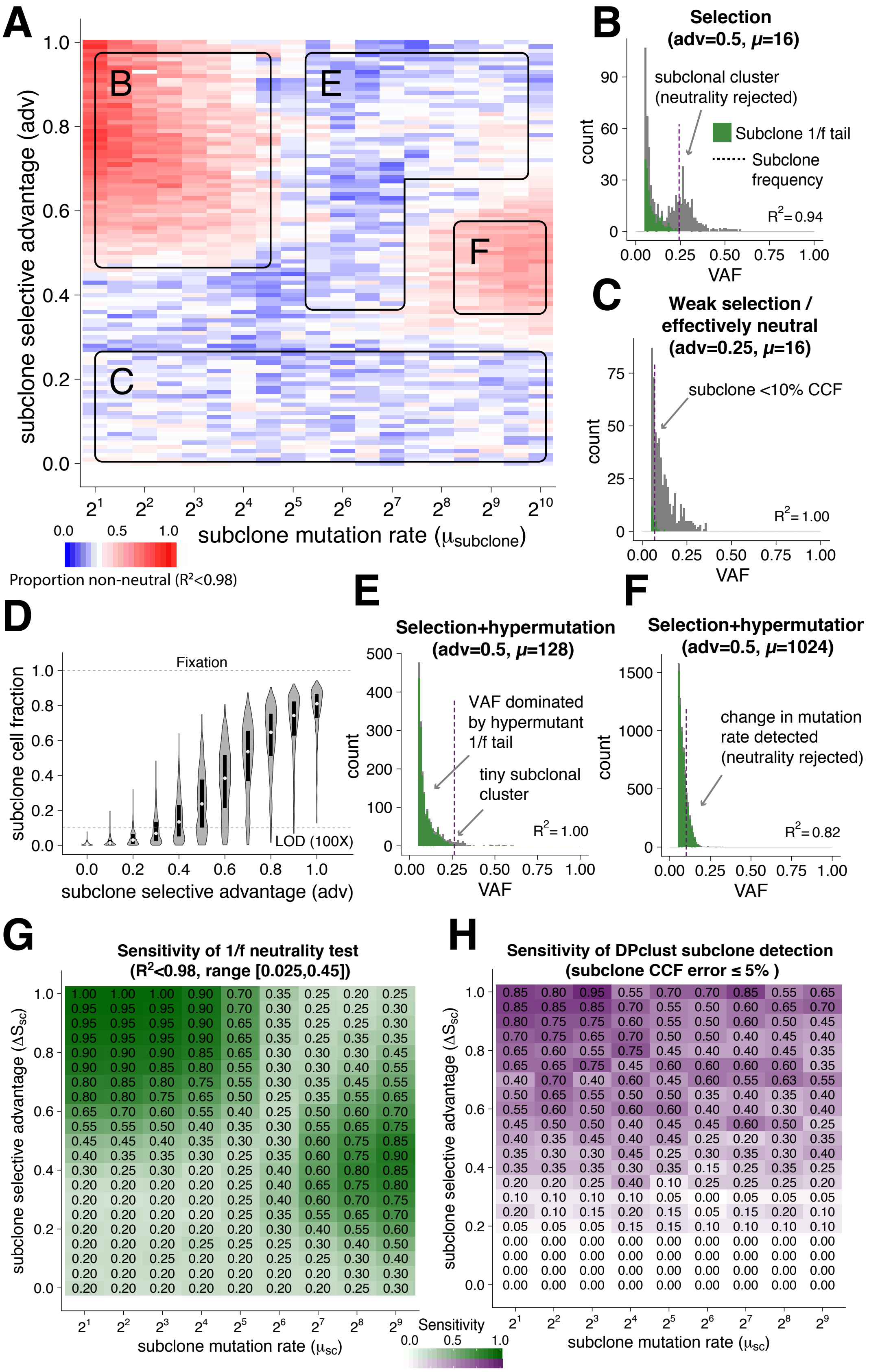
2.3 Results

2.3.1 Insights From Simulated Tumours

In their letter, Tarabichi et al. (2018) used an analytical solution of the tumour growth model (their Figure 1a) and a small number of stochastic simulations (their Figure 1b) to argue that the $1/f$ test used by Williams et al. (2016) leads to the arbitrary classification of tumours (their Figure 1c). As an explanation for this Tarabichi et al. (2018) suggested that the biological noise caused by the stochasticity of the process (i.e., genetic drift) lead to the arbitrary classifications by the $1/f$ test. This seems at odds with previous theoretical work on this (Williams et al. 2016; Bozic, Gerold, and Nowak 2016; Kessler and Levine 2013; Durrett 2013). It is worth noting, that subclones were introduced at fixed time points in all simulations and instead of testing the classification at different time points of subclone introduction t_{sc} and selective advantages a_{sc}, the effect of subclonal selection a_{sc} and changes of mutation rate μ_{sc} were assessed.

More specifically it was assumed that the change of μ_{sc} co-occurred with a change of a_{sc}. Whether such a change of the properties of a subclone in a tumour is realistic might be questionable in itself. Further the relative changes considered by Tarabichi et al. (2018) seem rather extreme. The change of the mutation rate per division Tarabichi et al. (2018) they tested ranged from a decrease from 16 mutations per division to ≈ 1 mutation per division — a decrease by a factor of 16 — to an increase to $\approx 2,000$ mutations per division — a decrease by a factor of more than 100 (x-axis of Figure 1a, in Tarabichi et al.). Indeed, such a substantial increase of mutation rate can likely only be explained by cases with a subclonal defect of the DNA mismatch repair machinery or a POLE/POLD alteration (Billingsley et al. 2015). Studies of subclonal MMR gene defects indicate that these are very rare in endometrial cancer (Stelloo et al. 2017) or colorectal cancer (Joost et al. 2014). An analysis of POLE mutated subclones arising within a POLE wild-type background also indicated that these are very infrequent events (Temko et al. 2018). Together these initial observations indeed suggest that the extremities of the parameters considered by Tarabichi et al. should be interpreted with caution.

To address their criticism and explain the observed results, I tried to reproduce the analysis shown in Tarabichi et al. Figure 1b. For this I generated multiple realisations of simulations for parameter sets across the range considered by them (Figure 2.3 and Methods). This analysis showed that when the selected subclone was present at $f_{sc} \geq 10\%$ (i.e., when $a_{sc} \geq 0.2$), the $1/f$ test correctly rejected neutrality in the majority of cases if no simultaneous increase of the mutation rate occurred (top left quadrant of Figure 2.4A). An example of such a simulation is shown in Figure 2.4B. Here the presence of a subclone at a frequency of ≈ 0.5 (i.e., 0.25 in the VAF spectrum) leads to a clear deviation from the $1/f$ distribution ($R^2 = 0.94$) similar to some realisation shown in Figure 2.3B.



2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9
1.00	1.00	1.00	0.90	0.70	0.35	0.25	0.20	0.25
0.95	0.95	0.95	0.90	0.70	0.30	0.25	0.25	0.30
0.95	0.95	0.95	0.90	0.65	0.35	0.25	0.25	0.30
0.95	0.95	0.95	0.90	0.65	0.35	0.25	0.35	0.35
0.90	0.90	0.90	0.85	0.65	0.30	0.30	0.30	0.45
0.90	0.90	0.85	0.80	0.55	0.30	0.30	0.40	0.55
0.80	0.85	0.80	0.75	0.55	0.30	0.35	0.45	0.55
0.80	0.80	0.75	0.65	0.50	0.25	0.35	0.55	0.65
0.65	0.70	0.60	0.55	0.40	0.25	0.50	0.60	0.70
0.55	0.55	0.50	0.40	0.35	0.30	0.55	0.65	0.75
0.45	0.45	0.40	0.35	0.30	0.30	0.60	0.75	0.85
0.40	0.35	0.35	0.30	0.25	0.35	0.60	0.75	0.90
0.30	0.25	0.30	0.20	0.25	0.40	0.60	0.80	0.85
0.20	0.20	0.20	0.20	0.25	0.40	0.65	0.75	0.80
0.20	0.20	0.20	0.20	0.25	0.40	0.60	0.70	0.75
0.20	0.20	0.20	0.20	0.25	0.35	0.55	0.65	0.70
0.20	0.20	0.20	0.20	0.20	0.30	0.40	0.55	0.60
0.20	0.20	0.20	0.25	0.25	0.25	0.30	0.40	0.50
0.20	0.20	0.25	0.20	0.20	0.25	0.25	0.30	0.40
0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.30	0.30
0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.25	0.30

2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9
0.85	0.80	0.95	0.55	0.70	0.70	0.85	0.55	0.65
0.85	0.85	0.85	0.70	0.55	0.50	0.60	0.65	0.70
0.80	0.75	0.75	0.60	0.55	0.50	0.60	0.50	0.45
0.70	0.75	0.65	0.70	0.50	0.50	0.40	0.45	0.40
0.65	0.60	0.55	0.75	0.45	0.40	0.40	0.45	0.35
0.65	0.65	0.75	0.45	0.60	0.60	0.60	0.60	0.35
0.40	0.70	0.40	0.60	0.45	0.60	0.55	0.63	0.55
0.50	0.65	0.55	0.50	0.50	0.35	0.40	0.35	0.40
0.55	0.60	0.50	0.60	0.60	0.40	0.40	0.30	0.40
0.45	0.50	0.50	0.40	0.45	0.45	0.60	0.50	0.25
0.40	0.35	0.45	0.40	0.45	0.25	0.20	0.30	0.35
0.35	0.30	0.30	0.45	0.25	0.30	0.35	0.30	0.40
0.35	0.35	0.35	0.30	0.35	0.15	0.25	0.35	0.25
0.20	0.20	0.25	0.40	0.10	0.25	0.25	0.25	0.20
0.10	0.10	0.10	0.10	0.05	0.00	0.05	0.05	0.05
0.20	0.10	0.15	0.20	0.15	0.05	0.15	0.20	0.10
0.05	0.05	0.05	0.15	0.15	0.10	0.10	0.10	0.10
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 2.4: Insights from stochastic simulations of cancer growth. A) Heat map recapitulating Tarabichi et al.'s 2018 Figure 1b with the same parameter set and showing the proportion of simulations in which neutrality was rejected (200 cases per parameter combination).

Figure 2.4: (Continued) B) Example VAF distribution with a detectable subclonal cluster (dashed line indicates subclone frequency). The $1/f$ test rejects neutrality in favour of selection (R2 reported). C) Example VAF distribution with a weakly selected subclone that remains below the limit of detection (100x depth). D) Subclone cell fraction in the final tumour as a function of fitness advantage; for $a < 0.5$, the subclone rarely reaches a detectable size of $\approx 10\%$ cell fraction (assuming 100× depth). LOD, the limit of detectability. E) Example VAF distribution for a subclone with a selective advantage and, at the same time, a high mutation rate. F) Example VAF distribution for a selected and extreme mutator subclone. G) Sensitivity of the $1/f$ test applied to subclonal mutations in the extended range of VAF $f = [0.025, 0.45]$ from the simulations in a. Numbers report the proportion of cases in which neutrality was rejected ($R^2 < 0.98$). H) AUC values of the $1/f$ test in various regions of the tested parameter space. (Modified version of the figure presented in Heide et al. 2018)

As expected a relationship between the selective advantage a_{sc} and the subclone cell fraction f_{sc} in the final tumour was observed (Figure 2.4D). Notably, subclones in simulations with $a_{sc} \leq 0.3$ rarely reached a size $f_{sc} \geq 0.1$. Variants at such a low VAF are basically undetectable at a depth of $\approx 100x$ commonly used for sequencing. This highlighted again the issue of the limit of detectability in currently used methods (Williams et al. 2018b), a point that was later confirmed by a more rigours analysis (Caravagna et al. 2020). Since selected clones at such a low f do not significantly change the observable clonal composition of the tumour, the signature of neutral growth (i.e., the '$1/f$ tail') does still dominate the detectable VAF spectrum (Figure 2.4C) and neutrality was not rejected (bottom part of Figure 2.4A). Importantly, this is not an issue of the test statistic itself, but rather seen as a general limitation of NGS.

Notably, for hypermutant subclones with strong selective advantage (i.e., $\mu_{sc} \geq 64$ and $a_{sc} \geq 0.4$, top right of Figure 2.4A), the analysis indicated that the method consistently failed to reject neutrality. Examination of realisations of these simulations showed that a massive $1/f$ tail containing thousands of the subclone's private mutations was frequently observed. These effectively masked the comparatively small cluster of mutations that were present in the selected subpopulation of cells and hence overrepresented at a high VAF (example in Figure 2.4E). These mutations dominated the entire VAF distribution and obscure the underlying subclonal structure. Unsurprisingly, the $1/f$ test and likely any other similar test would struggle to detect any subclonal cluster or deviation from the expected power-law distribution in these cases.

Curiously, for moderate values of selection $a_{sc} \approx 0.5$ and very high mutation rates $\mu_{sc} \geq 2^8$, a change in mutation rate from normal to hypermutant was detected, thus leading to consistent rejection of neutrality (mid-right area in Figure 2.4A; example in 2.4F). A similar example was indeed shown in Supplementary Figure 11h of the original paper in Williams et al. (2016). In cases with weak selection and hypermutation, subclones did not reach a detectable size, and therefore neutrality was not rejected (bottom right of Figure 2.4A)

As discussed in the original paper (Williams et al. 2016) and the reply (Heide et al. 2018), the main motivation for the narrow integration range of $0.12 \leq f \leq 0.24$ were concerns of mutations from triploid sites or impure samples affecting the power-law tail and for this reason an upper threshold of $f \approx 0.25$ was used by them. A larger integration range would potentially allow to detect the presence of subclones outside of this fairly narrow window. Since the simulated tumours were all diploid and did not contain such subclonal mutations, it was possible to test this hypothesis. For this reason, the $1/f$ test was used with an extended integration range and this did indeed demonstrate that the $1/f$ test is more accurate when applied to the entire VAF spectrum (Figure 2.4G). Under these conditions, neutrality was consistently rejected (i.e., $\geq 75\%$) for non-neutral simulations at background mutation rates and sufficiently large subclones.

I further suspected that the lack of discriminatory power in the peculiar scenarios considered by Tarabichi et al. did not depend on the method per se but was largely due to minimal signal in the data. To demonstrate this, the $1/f$ test using the extended integration range (Figure 2.4G) was compared to results from *DPclust* (Nik-Zainal et al. 2012a), a method often used to detect subclones on the basis of Dirichlet Process clustering. Indeed, the sensitivity of *DPclust* was suboptimal in most cases (Figure 2.4H), even in the presence of strong selection. This despite the fairly consistent number of 3–5 clusters inferred to be present from the simulated VAF data (see Figure S.4, page 263). The clusters inferred by *DPclust* were often located at similar positions of the VAF distribution and independent of the true subclone frequency (see Figure S.4, page 263). Importantly, the positions of the cluster also implied that mutations in them occurred in independent lineages, thus raising question on how one should interpret the result from such clustering methods in general.

Still, this observation did still not explain why the $1/f$ classifier might have performed for than random, as Tarabichi et al. (2018) reported. For this reason, I calculated conducted a similar analysis on the stochastic simulations generated as described above. For these additional neutral simulations with $a_{sc} = 0$ and $\mu_{sc} = 16$ were generated and used as a comparison for the classifier. ROC curved and the AUC of these were calculated for each parameter combination. This analysis confirmed that the AUC was substantially larger than 0.5 in some areas and at least 0.5 across the entire parameter space (Figure S.3A, page 263). In light of the observed ROC across the entire parameter range for cases with a putatively detectable subclone ($0.25 < f_{sc} < 0.75$) shown in Figure S.3B (page 263), it seems likely that Tarabichi et al. (2018) swapped the false-positive and false-negative rates when they conducted a similar analysis. In summary the $1/f$ statistic is an imperfect interpretation of the VAF spectrum, but certainly not worse than a random classifier.

In summary, the detailed analysis of stochastic simulations described above confirmed the initial concern that Tarabichi et al. (2018) failed to perform a fair test of the $1/f$ statistic. Instead of considering the behaviour of simulations in the parameter range considered (Figure 2.4), the authors appeared to instead integrate over a wide range of parameters. This likely lead them to underestimate the strength of the test under more realistic scenarios of subclonal selection. Further, the analysis of a commonly used clustering approach (Figure 2.4H) demonstrated that applying these methods to data of somatic mutations detected from cancer bulks might be problematic.

2.3.2 Analysis of Subclonal Selection Using dN/dS Ratios

In the second part of their letter Tarabichi et al. used a test inspired by the classical dN/dS method to demonstrate evidence for the presence of selected subclonal variants in tumours classified as neutral. Specifically, the authors pooled subclonal mutations in known cancer genes from multiple patients and calculated dN/dS ratios for the neutral and non-neutral groups. Tarabichi et al. argue that for cases in which the $1/f$ test failed to reject the null hypothesis, subclonal mutations should lack evidence of selection (i.e., $dN/dS \approx 1$). While this is a sound argument if one assumes that there is no classification error whatsoever, it is incorrect to draw conclusions about individual samples from such a population-level statistic. Instead, the observation of $dN/dS > 1$ for mutations from the subclonal mutations of all samples might simply indicate that the $1/f$ test misclassified one or more patients.

To investigate this possibility, I repeated the dN/dS analysis conducted by Tarabichi et al. with the same method. Summarised, global dN/dS estimates for 369 the driver genes reported by Martincorena et al. (2017) were calculated for the colorectal and gastric cancers analysed in the original publication (Williams et al. 2016). Since the TCGA CAVEMAN calls Tarabichi et al. (2018) used were not available publicly, I instead reanalysed the pan-cancer TCGA variant calls that were available through GDC. Due to the criticism by Tarabichi et al. regarding the presence of tetraploid tumours in the original analysis conducted by Williams et al., which could cause the false rejection of the null hypothesis, I restricted the analysis to diploid regions and samples with high purity (see Methods for details). The usage of the newly published ploidy and purity estimates for the TCGA samples should generally have improved the classification. Curiously, this new analysis found that $290/724$ (40%) of cases compared to the 31% in the original analysis were consistent with neutrality (Table 2.1), thus confirming the findings by Williams et al. (2016).

Consistent with the results by Tarabichi et al. (2018) the dN/dS estimates of missense[5] and nonsense[6] mutations were significantly above one for the clonal variants of the pan-cancer TCGA cases classified as neutral and non-neutral (Figure 2.5C). Equivalent results were also observed for the 101 CRCs that were also analysed by Williams et al. (2016). As shown in Figure 2.5A, $dN/dS > 1$ was observed for clonal missense and nonsense mutation of cases classified as neutral ($34/101$) and non-neutral ($67/101$). For the 68 gastric tumours Williams et al. (2016) analysed only the clonal nonsense mutations showed a $dN/dS > 1$ (Figure 2.5B).

In contrast to clonal mutations, which should have a $dN/dS > 1$, subclonal mutations might in principle only have $dN/dS > 1$ in cases classified as non-neutral, but not in those classified as neutral. Consistent with this expectation the dN/dS ratios of subclonal missense mutations of tumours from all three cohorts were found to not be significantly different from 1 (Figure 2.5A–C, missense mutations at left, blue bars). Likewise, dN/dS estimates of subclonal nonsense mutations from the colorectal and gastric cohort were not significantly above one either Figure 2.5A–C, missense mutations at right, blue bars).

In contrast the analysis of subclonal nonsense mutations for neutral cases of the TCGA cohort suggested $dN/dS > 1$. This observation is of course in conflict with the classification of these cases as ‘neutral’. Still, a more detailed analysis of the cases showed that a small subset of patients classified as neutral showed a high number of subclonal nonsense mutations in putative driver genes. Specifically, $1/57$ cases (1.7%) of the gastric cancers and $11/290$ (3.8%) cases of the pan-cancer cohort classified as neutral contained ≥ 3 subclonal nonsense mutations.

Manual examination of these patients (Figure S.5-S.14, page 263-266) suggested that

[5] Non-synonymous mutation of the DNA that cause the replace of one encode amino acid by another.

[6] Non-synonymous mutations of the DNA that cause the premature termination of the translation and hence the expression of a shorter, unfinished protein product.

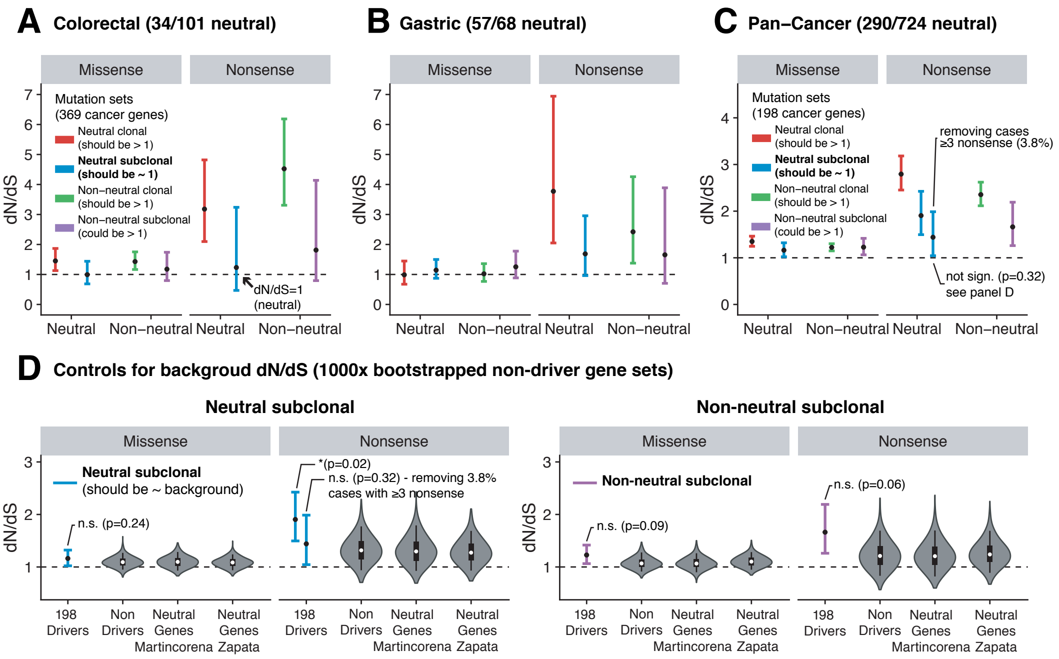


Figure 2.5: dN/dS analysis with the method of Martincorena et al. (2017) applied to colorectal cancers. A) Gastric cancers from Wang et al. (2014) analysed in Williams et al. (2016) B) TCGA pan-cancer cases analysed by using newly available GDC calls to reproduce Tarabichi et al.'s 2018 dN/dS analysis. C) Cancers were classified as neutral or non-neutral with the $1/f$ test, and the dN/dS values of were calculated over pooled variants from each group (split between clonal/subclonal and missense/nonsense). D) Comparison of the dN/dS estimates obtained for the 198 driver genes (black dots, point estimates; error bars, 95% confidence-intervals) with the distribution of 1,000 random subsets from three control sets of non-driver genes, demonstrating a general positive bias of estimated dN/dS values (white dots, median; box, interquartile range; whiskers, 90% prediction interval). After removal of 3.8% of pan-cancer cases with three or more subclonal nonsense mutations in driver genes, both missense and nonsense dN/dS in neutral cancers were not significantly different from the neutral expectation. 'Martincorena' refers to Martincorena et al. (2017), 'Zapata' refers to Zapata et al. (2018). (Figure as presented in Heide et al. 2018)

some clonal mutations were 'bleeding' into the subclonal integration range. Since clonal mutations are expected to have a $dN/dS > 1$, this would explain the elevated dN/dS value of subclonal mutations in 'neutral' cases. In other cases, a misclassification caused by erroneous ploidy estimates or the presence of a selected subclones underneath a power-law tail seemed possible. Regardless of the exact reason, after the removal of the 3.8% of cases with ≥ 3 subclonal nonsense mutations from the analysis, the dN/dS values of subclonal nonsense mutations were found to not be significantly different from that of the neutral background (Figure 2.5C; $dN/dS = 1.44$, $p = 0.32$). For the calculation of this background, dN/dS values of known passenger genes was generated using a bootstrap method of $1,000$ random sets of 198 non-drivers as described in the Methods (Figure 2.5D). This showed a systematic positive bias for the estimation of dN/dS, possibly due to publicly available somatic GDC calls being filtered for common human germline variants present in dbSNP. Since germline mutations are composed of more synonymous than non-synonymous variants, estimates of dN/dS ratios generated from such data are skewed upward (Martincorena et al. 2017). While not significant, dN/dS values were consistently higher in non-neutral versus neutral cases (Figure 2.5D).

2.4 Discussion

Summarised, the analysis of the simulations conducted by Tarabichi et al. (2018) explained the apparent mismatch between the stochastic simulations conducted by the authors and the previous mathematical theory on the convergent solution of the continuous-time stochastic branching process (Durrett 2013; Kessler and Levine 2013; Kessler and Levine 2015; Williams et al. 2016; Bozic, Gerold, and Nowak 2016). Simulations based on the Gillespie algorithm, which explicitly model asynchronous cell divisions, did agree with the solutions of the stochastic branching process and, as shown by others such stochastic neutral models (Durrett 2013; Kessler and Levine 2013; Kessler and Levine 2015) do generally scale according to the expected $1/f^2$. This general scaling behaviour even holds in the presence of stochastic cell death (Kessler and Levine 2013).

While Tarabichi et al. (2018) appear to, at least implicitly, acknowledge that simulations of tumour expansion as a branching process (i.e., Bozic, Gerold, and Nowak 2016) provide a reasonable model of tumour evolution, they seem to have missed why the observed structures (i.e., the $1/f^2$ distribution) arise. They instead allude to classic studies of neutral evolution in population genetics like that of Kimura and Ohta (1969) by suggesting that ‘drift can drive novel variants to high frequencies’. These studies are concerned with the drift of novel variants arising in $1/N$ individuals within a population of constant size. However, the argument by Williams et al. (2016) was on the site frequency spectrum arising in an exponentially expanding population, which also arise in the absence of selection and drift. At least at sufficiently high mutation rates, neutral tails, similar to those observable in the cancer genomic data analysed by Williams et al. (2016), are simply a consequence of the mutations that arise with each cell division during the clonal expansion of a tumour (Williams et al. 2016; Williams et al. 2018b). Drift can obviously also arise due to the stochastic events in exponentially expanding tumours, but this would emulate the properties of selection revealed by the $1/f$ test. Importantly, the presence of such subclonal structures is also not taken into account by the clustering-based methods some of these authors suggest to use for the analysis of subclonal mutations observed in bulk sequencing data (Tarabichi et al. 2021).

Tarabichi et al. (2018) also seem to ignore that the limitations they highlight for the $1/f$ test, namely that it is 'neither a necessary nor a sufficient' method to detect selection, apply in the same way to commonly used clustering methods (e.g., Dentro, Wedge, and Van Loo 2017). The analysis of one such clustering method (Figure 2.4H) showed that the application of these to data of somatic mutations detected from cancer bulks might be problematic in general. Most importantly, if the assertions made by Williams et al. are correct, variants at a subclonal frequency would often be present in different lineages (e.g., Sottoriva, Barnes, and Graham 2017). The way in which some of the authors suggest to interpret results from clustering methods would then be inherently flawed (Tarabichi et al. 2021; Dentro et al. 2021).

Curiously, the application of the dN/dS methods Tarabichi et al. (2018) used to criticise the $1/f$ test statistic has demonstrated that most mutations detected in individual tumours are selectively neutral (Martincorena et al. 2017). This observation is entirely consistent with the premise of neutral tumour evolution, which is that the majority of genetic variation arising through mutation are selectively neutral (Kimura 1968b; Kimura 1991). The presence of a positive dN/dS ratio for subclonal mutations in known cancer genes, as described by Dentro et al. (2021) in a recent pan-cancer analysis of subclonal drivers, is not at odds with this. The detection of a positive dN/dS ratio in a set of patients does not imply that all of these are non-neutral, but only means that at least some have a subclone arising through selection. Indeed, the analysis of dN/dS ratios in the pan-cancer TCGA cohort shown above identified a subset of tumours with multiple subclonal non-synonymous variants. The removal of this small subset (3.8%) of cases reduced the dN/dS ratio to a level at which it was not significantly above one. In theory, a single misclassified patient carrying multiple nonsense mutations in driver genes could significantly alter the dN/dS value of an entire cohort. This highlights that, since dN/dS analysis at the cohort level combines mutations from different patients, it cannot easily evaluate the performance of statistical tests that aim to detect neutrality at the patient level.

Last but not least, the point that the 'failure to reject the null hypothesis is not the same as proving it true; made by Tarabichi et al. (2018) is certainly correct. Still, it somewhat misses the main point made by Williams et al. in the original 2016 paper. Here neutrality is explicitly formulated as the null model for a frequentist approach. This null hypothesis is rejected by the proposed test statistic in most cases, suggesting the widespread presence of subclonal selection. The fact that the remaining cases are referred to as ‘neutral’ in the publication does not change the setup of the test itself.

Summarised, the critique by Tarabichi et al. (2018), did not invalidate the conclusions made by Williams et al. (2016). Neutral evolution provides an adequate null model for the pattern of ITH that can be observed in many tumours. Ignoring this risks to misinterpret existing cancer genomic data or, even worse, to conduct ill-equipped experiments. This will, in turn, delay potential clinical improvements that could be archived from a better understanding of the dynamics driving late-stage cancer evolution.

Chapter 3

Modelling Cancer Evolution in Space

The simple $1/f$ summary statistic described in Williams et al. (2016) and the Approximate Bayesian Computation (ABC) inference able to detect selection using the entire VAF spectrum Williams et al. (2018b) developed later use single bulk WGS sequencing data as the basis for statistical inference. Due to the abundance of such datasets, generated as part of several large-scale cancer sequencing projects like TCGA (Bailey et al. 2018) or PCAWG (Campbell and Giocomo 2019), the development of such methods was crucial. Despite this, the methods were heavily criticised for their lack of discriminatory power. One example of this is the criticism by Tarabichi et al. (2018) and in the previous chapter, I have presented results from the reply to this criticism. While the analysis showed that Tarabichi et al. overstated the severity of these problems, detection of selected subclones from bulk sequencing data is inherently challenging.

One of the most significant drawbacks of bulk WGS is that information on which alleles co-occur in individual cells is lost. Especially, at the commonly used sequencing depth, it is thus not possible to confidently demine if mutations with a similar VAF occur in the same lineage or not. A number of studies assume that only a few genetically identical subpopulations of cells are present at frequencies that are detectable by NGS. These populations are assumed to be co-existing subclones that have expanded to a significant size. Clustering methods, like the previously mentioned *DPclust*, could in this case be applied to reconstruct phylogenetic relationships among subclones (see Section 1.6 for details).

If instead, as suggested by Williams et al. (2016), subclonal structures primarily arise as a consequence of the clonal expansion itself, then these assumptions do not hold. In this case, one would instead expect that many clusters of mutations are identifiable at a CCFs between 10% and 100%. Many of these would be present at such similar frequencies that they could not be resolved by currently used sequencing approaches.

This issue, was already demonstrated in the previous chapter, where I applied *DPclust* to simulated sequencing data of neutral and non-neutral tumours that were obtained from a branching process-based tumour model. The results of this analysis showed that *DPclust* was often unable to identify the 'peak' of mutations that the selected subclone carried to a higher frequency. Instead, the clustering results often appeared to be dominated by the power-law tail, which is itself composed of many mutation clusters generated by multiple parallelly expanding lineages. As expected *DPclust* was also unable to resolve the mutations of these parallel lineages as independent clusters and instead suggested the presence of a small number of large clusters. Still, tumours usually expand as a mass of cells and different lineages are thus expected to variegate in space. For this reason, one can in principle use multiple WGS samples obtained from different areas of a tumour to resolve lineages much better. This formed the basic motivation for multi-region sequencing studies like the one conducted by Gerlinger et al. (2012).

Multi-region sequencing experiments do allow a much more accurate reconstruction of ancestral relationships for the dominant cell populations (Tarabichi et al. 2021). Still, it is not entirely clear if the detection of subclonal selection in such phylogenies would easily be possible or not. Similar to the previous interpretation of single bulk samples, various issues arising from neutral dynamics might exist. Specifically with regard to biases arising from spatial sampling in a tumour relatively little is known. The behaviour of commonly applied multivariate clustering methods, when applied to simulated multi-region sequencing data arising under neutrality, was previously also uncharacterised. To better understand these key questions and to gain insight into if bulk sequencing can easily be used gain insight into tumour growth dynamics and especially the presence of subclonal selection, a model that could generate artificial multi-region sequencing data was required. For this reason, I developed a spatial tumour simulator together with Ketevan Chkhaidze.[1] Given the general interest in the field, I aimed to make this method as easily accessible to others as possible. The code was implemented in C++ and then integrated into a package for the R programming language (R Core Team 2020). R is very commonly used in the field, and as an interpreted language, it is a suitable option for this purpose. We also integrate

[1] Details of the model will be provided below. Ketevan Chkhaidze implemented a first version of the model in Python. This first version assumed random 'pushing' in space and non-boundary driven growth. I implemented a version in C++ of this model and integrated it into an R package. I also modified the model to consider boundary driven growth and let cells in the tumour 'push' to the closest edge.

additional code into the package, allowing the simulation of various sampling schemas and plotting of simulated datasets. Our method allows the simulation of exponentially growing and boundary-driven tumours. This simulator was used to describe the general properties of spatially growing tumours and demonstrate that ABC inference can infer model parameters from genomic measurements. The results of this were published in Chkhaidze et al. (2019).

I also expanded on some of the work presented in the previous chapter. Especially, I tried to characterise the behaviour of commonly used clustering methods (e.g., Roth et al. 2014) when applied to simulated sequencing data of spatial and non-spatial tumours. The ability of these methods to distinguish mutations present in individual 'mutation clones' was, as expected, very poor for all commonly used coverage values. This provided some important insight into the results expected from real experiments, especially with regard to the effect of purity and coverage. They also provided some important insights into the usability of multi-region bulk WGS sequencing data for statistical inference of selection. The results suggest that some of the identified issues might be hard to mitigate in practice. These results were added to the publication of a statistical method *MOBSTER* developed by Giulio Caravagna (Caravagna et al. 2020).

Combined, these extended analysis of simulated spatial and non-spatial WGS datasets showed that sequencing of individual clonal units (i.e., cells or glands) might be better suited for the inference of selection from sequencing data and that great care has to be taken in the interpretation of results obtained from clustering of mutation calls. Even extreme divergence of samples obtained very closely in space can arise from neutral dynamics. Overall, this provided a rationale for the multi-region single-gland sequencing study of CRCs called EPICC, which I will present in the next chapter.

3.1 Methods

3.1.1 Spatial simulator

Due to the limitations of single bulk WGS data, I decided to explore how commonly conducted multi-region sequencing data could be used for improved detection of selected subclones in individual tumours. For this purpose, we developed a simple spatial tumour simulator with which spatial dynamics could be simulated and to which different spatial sampling strategies could be applied. The spatial simulator we developed for this purpose, models tumour growth using a stochastic spatial model of cells that incorporates cell division, cell death, random mutations, clonal selection and effects of spatial crowding. This simple model is described in detail below.

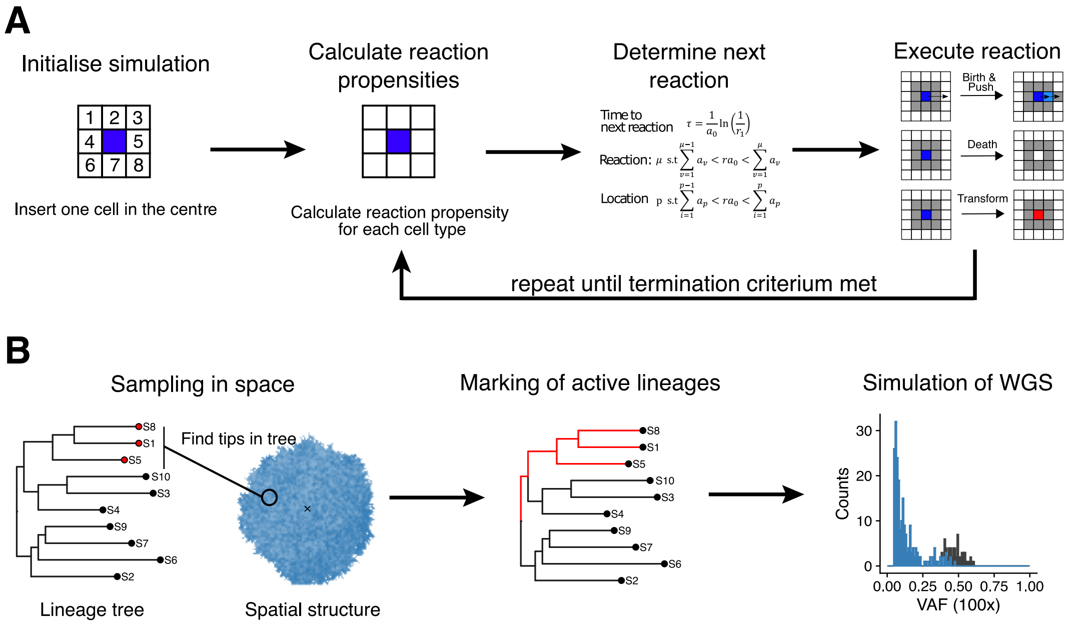

A
Initialise simulation
1 2 3
4 5
6 7 8
Insert one cell in the centre
Calculate reaction propensities
Calculate reaction propensity for each cell type
Determine next reaction
Time to next reaction $\tau = \frac{1}{a_0}\ln\left(\frac{1}{r_1}\right)$
Reaction: μ s.t $\sum_{v=1}^{\mu-1} a_v < ra_0 < \sum_{v=1}^{\mu} a_v$
Location p s.t $\sum_{i=1}^{p-1} a_p < ra_0 < \sum_{i=1}^{p} a_p$
Execute reaction
Birth & Push
Death
Transform
repeat until termination criterium met
B
Sampling in space
Find tips in tree
S8
S1
S5
S10
S3
S4
S9
S7
S6
S2
Lineage tree
Spatial structure
Marking of active lineages
Simulation of WGS
Counts
30
20
10
0
0.00 0.25 0.50 0.75 1.00
VAF (100x)

Figure 3.1: Schema of the spatial simulator. Simulations involve two steps: A) The spatial simulation of the tumour using the Gillespie algorithm. B) The simulation of WGS sequencing data. For this, a subset of cells is selected in space. Then active lineages — that is, edges that connect any of the sequenced cells to the root of the tree – were annotated with the number of sequenced cells below them. Finally, the active part of the tree was traversed from the root to simulate WGS data.

3.1.1.1 Simulation of the Tumours

The birth-death process The growth of a tumour is simulated on a 2D or 3D lattice with Moore neighbourhood (see left panel in Figure 3.1A). Each simulation is initialised with a single cell placed at the centre of the space at the time point $t_g = 0$. All cells A are assumed to be able to undergo two reactions, birth and death. A cell that undergoes death is removed from the simulation and frees up the occupied space ($A \xrightarrow{\mu} \cancel{A}$). This reaction is assumed to occur with the birth rate μ. A cell that undergoes a birth event is assumed to give rise to a second, identical daughter cell A' that it tries to place into a location in its neighbourhood ($A \xrightarrow{\lambda} A + A'$). This reaction occurs with the birth rate λ.

Pushing of cells If during a birth event empty grid points (i.e., in the Moore neighbourhood) exist next to A, then A' is simply placed into one of these at random. If instead all neighbouring grid points are occupied by other cells, then the cell A tries to 'make room' for the cell A' by trying to 'push' other cells away. This pushing in space is done along a vector $\mathbf{v}$ up to maximum distance d_{push} and if it is possible all cells along this vector are moved one position forward. If instead, the pushing was unsuccessful, then the division is considered to have failed and the cell A' dies. For the choice of $\mathbf{v}$ two options were considered: i) pushing into a random direction of the space and ii) pushing towards the closest edge of the tumour.[2]

Gillespie Algorithm As outlined above, cells are assumed to be able to undergo two reactions, birth according to the birth rate λ and death according to the death rate μ. The actual rate with which cells undergo these two reactions is determined by their 'cell type' i. As daughter cells are assumed to be identical copies of their parents, their cell type is also identical to that of the parent. This means that for any cell type i a set of member cells M_i exists.

Since each cell type is assumed to undergo the same reaction types (i.e., birth and death) with a different birth and death rate λ_i and μ_i respectively, one can consider these to be different reactions. For this reason, the total number of reactions is twice the number of cells types. To simulate a trajectory of reactions one can use to the Gillespie algorithm (Gillespie 1977; Gillespie 1977). The Gillespie algorithm allows sampling of both, the time to the next reaction τ and the index of the corresponding reaction j. Several variations of the Gillespie algorithm exist, of which the so-called 'first-reaction method' (Gillespie 2007) was used. For a set of uni-molecular, like the one considered here, one can sample

$$\tau_j = \frac{1}{\alpha_j N_j} log(\frac{1}{r_j}),\ with\ r_j \sim U(0,1),$$

where $N_j = |M_i|$ are the number of cells of the species i taking part in the reaction and α_j is the rate of the reaction. The next reaction j and the time to it τ is then given by

$$\tau = \min_{j'} \ \tau_{j'},\ j = \arg\min_{j'} \ \tau_{j'}.$$

From the species i taking part in the reaction j, a random element k was chosen. The reaction j was then executed on k as described above and the Gillespie time updated: $t_g = t_g + \tau$.

Introduction of subclones & Termination The transformation of one cell of type A to another one B was assumed to take place at a specific population size t_{sc}. For this, a random member of A was chosen and its reaction rates were updated with those of B. To prevent the random disappearance of the cell type A, transformations were delayed until A had more than one member. All simulations were stopped once the simulated tumour reached a predetermined size.

[2] I will later comment on the effects of these two choices in more detail (see Section 3.2.1).

3.1.1.2 Generation of Simulated Sequencing Data

During simulations, the ancestral relationship of all cells was recorded in form of a phylogenetic tree to allow the generation of simulated sequencing data (Figure 3.1B). Different ways to simulate such datasets were considered. These can be distinguished by how cells were sampled in space and by the parameters that described the model used to simulate the sequencing data (see Table 3.1).

In all cases, simulated sequencing data were obtained through a traversal of the recorded phylogeny (Figure 3.1B). The generation of simulated sequencing data for a set of cells C can be done by applying the following three steps to recorded ancestors of any of the cells: i) determine the number of passenger mutations that occurred during the ancestors' division, ii) determine the expected frequency of these passenger mutations, and iii) generate simulated sequencing data for each of these passenger mutations.

Number of mutation per division The number of mutations that occurred during each division were assumed to be Poisson distributed with $\Delta m \sim Pois(m_i)$, where m_i is the mutation rate of the ancestors' cell type i. For divisions that failed due to lack of space, mutations were still assumed to have occurred in the corresponding ancestral cell. This behaviour makes sense if one assumes that one of the two cells resulting from such a division into insufficient space immediately dies from overcrowding. Other options were separately considered. Specifically, a setup in which mutations are only accumulated during 'successful' divisions and where a second process, simulated as a reaction in the Gillespie algorithm, causes the continuous accumulation of mutations.

Determination of the frequency of mutations The expected frequency f_N of mutations that occurred in a ancestor N depends on the number of descendants of N that are in C. The frequency of mutated cells is

$$f_i = \frac{1}{|C|} \sum_{c \in C} \mathbb{I}_{c \in desc(N)},$$

where $\mathbb{I}_{c \in desc(N)}$ indicates if the cell c is a descendant of N.

Generation of simulated sequencing data The generation of simulated sequencing data for each mutation i can be broken down into the simulation of the coverage n_i and the simulation of mutant reads y_i. Three model M_{seq} for the simulation of n_i under a given average sequencing depth $\bar{n}$ were considered: i) Poisson distributed sequencing depth: $n_i \sim Pois(\bar{n})$, ii) overdispersed sequencing depth with $n_i \sim Bin(\frac{\bar{n}}{0.6}, \pi)$, where $\pi \sim Beta(\frac{0.6}{d} - 1, \frac{(0.6-1)(d-1)}{d})$ and d is a constant dispersion parameter $d = 0.08$, and iii) constant sequencing depth $n_i = \bar{n}$.

Table 3.1: Sequencing model parameters of the spatial simulator. Variables with a subclone index i are set individually for each subclone. All other variables are assumed the be constant for the whole tumour. The default values were chosen to represent an ideal sample (i.e., 100% purity) with a coverage similar to sequencing data from the TCGA project that was filtered with commonly used filters.

Symbol	**Description**	**Values**	**Default**
$\bar{n}$	Average sequencing depth	$[0,\infty]$	100
M_{seq}	Sequencing depth model[3]	$\{1,2,3\}$	1
ρ	Sample purity	$(0,1]$	1.0
f_{min}	Minimum VAF for detection	$[0,1)$	0.05
y_{min}	Minimum reads for detection	$[0,\infty)$	2

The number of mutated alleles y_i was in all cases assumed to follow a Binomial distribution $y_i \sim Bin(n_i, p_i)$ with the expected VAF being $p_i = \frac{f\rho m_i}{\rho c_i + 2 - 2\rho_s}$, where $c_i = 2$ is the copy-number of the mutated site in the tumour, $m_i = 1$ the multiplicity of the mutated allele in the tumour and f_N the fraction of mutated cells in the sample.

The generated mutation data were then filtered to only retain those with a minimum number of mutated reads $y_i \geq y_{min}$ and a minimum VAF $y_i/n_i \geq f_{min}$. Unless otherwise mentioned values of $f_{min} = 0.05$ and $y_{min} = 2$ were used. The filtering based on y_i and y_i/n_i was motivated by their common use as filtering criteria for the reduction of spurious false-positive mutations in NGS experiments (e.g., Williams et al. 2016; Cross et al. 2018). An overview of all parameters of the model used for the generation of sequencing data can be found in Table 3.1.

3.1.1.3 R Package - CHESS

The spatial simulator described above was implemented in the C++ general-purpose programming language and integrated into a package for the R statistical programming language (R Core Team 2020). For this, methods from the Rcpp package (Eddelbuettel and Francois 2011; Eddelbuettel 2013), which allows seamless integration of R and C++, were used. The code of the package is available on GitHub: `https://github.com/T-Heide/CHESS.cpp`. This package also contains the code for the ABC-SMC algorithm described in Chapter 6. Some additional notes on the implementation of the model can be found in Section S.2.1 (page 267).

3.1.2 Tree Statistics

Three tree balancing methods and one statistic that describes the distribution of the relative branching times were used to assess deviations in the tree shapes introduced by the subclonal selection.

Sackin Index The first and probably the most commonly used statistic to describe the balance of a tree is the so-called Sackin Index S (Shao 1990). This is a modification of a similar index proposed by Sackin (1972) and used for the first time by Shao (1990). While there are several definitions of the Sackin index, these can be shown to be equivalent (Fischer 2020). Here S was calculated as

$$S(T) \sum_{t \in V^1(T)} \delta_{\rho,t},$$

where $\delta_{\rho,j}$ denotes the number of edges that have to be traversed to reach the node j from the root ρ of the tree (i.e., the depth of j) and $V^1(T)$ is the set of all leave nodes in T.

Colless' index The second one, another commonly used index, is the so-called Colless' index C (Colless 1982; Mir, Rotger, and Rosselló 2018; Coronado et al. 2020) and defined as

$$C(T) = \sum_{t \in V^{2,3}} bal_T(t) = \sum_{t \in V^{2,3}} |K(c_1) - K(c2)|,$$

where c_1 and c_2 are the two children of the node t and $K(s)$ is the number of leaves that are part of the descendants of s.

Total Cophenetic Index Third, the Total Cophenetic Index Φ proposed by Mir, Rosselló, and Rotger (2013) was assessed. The statistic is given by

$$\Phi(T) = \sum_{s \in V^1} \sum_{t \in V^1 \setminus \{s\}} \delta_\rho(LCA(s,t)),$$

where $LCA(s,t)$ denotes the last common ancestor of s and t.

The γ statistic Most of these classic indices disregard information about branch lengths (Mooers and Heard 1997). I hence also calculated Pybus and Harvey (2000) γ statistic. This statistic has well defined properties extensively described in the literature (Pybus and Harvey 2000) and defined as

$$\gamma(T) = \frac{\left(\frac{1}{|V|-2} \sum_{i=2}^{|V|-1} \left(\sum_{k=2}^{i} k(\delta_{\rho,i} - \delta_{\rho,i-1})\right)\right) - \frac{T}{2}}{T\sqrt{\frac{1}{12(|V|-2)}}}, \text{ with } T = \sum_{j=2}^{|V|} j(\delta_{\rho,i} - \delta_{\rho,i-1}),$$

where nodes are ordered by their distance from the node $\delta_{\rho,i}$ and $|V|$ denotes the number of nodes.

Intermixing statistic Another summary statistic I was used to calculate the degree of intermixing within a simulated tree. For this, a number of cells were sampled from the tumour and labelled by a lineage marker m_i to create a set of cells C. The intermixing within the reconstructed tree was then measured as

$$I(T) = \frac{1}{|C|}\sum_{s \in C}\left(\frac{1}{|D_s|}\sum_{t \in D_s} \mathbb{1}_{m_s \neq m_t}\right),\ D_s = \{t \in V^1 | t \in desc(pa(s))\},$$

where V^1 are all tip nodes of the tree, $desc(s)$ the descendants of s, $pa(s)$ the parent node of s and $\mathbb{1}_{m_s \neq m_t}$ an indicator function that indicates if s and t had different labels m.

3.2 Results

3.2.1 Artefacts Arising From ‘Random Pushing’

One of the main aspects of the spatial simulator described here was to consider the effect of spatial crowding on tumour growth dynamics. This aspect of the simulator was controlled by the d_{push} parameter, which describes up to which distance cells can push other cells away to make room for a daughter cell. In the initial implementation of the simulator, if no empty grid point in the Moore neighbour existed, a random vector $\mathbf{v}$ was generated, and a push was initialised along this vector. Upon reaching the maximum allowed distance, the tried push was aborted, and the division skipped.

While this heuristic might appeared to be a reasonable and computationally cheap approach, upon closer inspection artefacts arising from it were identified. Considering a mass of N cells, only a subset at a distance $r_g = d_{push}$ from the outer edge should be able to grow in a spatially constrained tumour. By simply assuming that growth occurs in form of a disc or sphere, the expected growth dynamics in 2-3 dimensions can be described by the following ordinary differential equations (ODEs):

$$2D: \frac{dN}{dt} = N - \pi\ max(0, r - r_g)^2; r = \sqrt{N/\pi},$$

$$3D: \frac{dN}{dt} = N - \frac{4}{3}\pi\ max(0, r - r_g)^3;\ r = \left(\frac{3\pi N}{4}\right)^{\frac{1}{3}}.$$

Comparison of the growth curves expected from these ODEs to those obtained from the spatial simulator revealed an obvious discrepancy between the two (compare red and black lines in Figure 3.2). For both extreme parametrisations of the model (i.e., $d_{push} = 1$ and $d_{push} = \infty$), no deviations from the expected behaviour existed, but for intermediate degrees of constraint, the simulations started to deviate from the expectation at some point.

To show that this resulted from the pushing of cells into a random direction, an alternative ODE model that would take the effects of this into account was constructed. In two

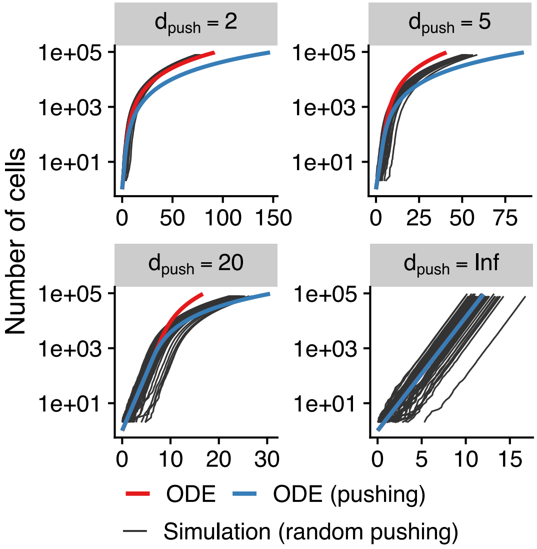


Figure 3.2: Simulated vs expected growth curves using random pushing. The black lines show 50 random realisations obtained from the simulator.

dimensions the pushing distance forms a radius r_g around the location of the dividing cell located at a distance r_i from the centre of the tumour (Figure S.17B, page 269). Given the size of the tumour N, its radius can be assumed to be $r = \sqrt{N/\pi}$. If $r < r_i + r_g$, then the pushing radius r_g around the position r_i and the outer edge of the tumour r will intersect in two points $(x, \pm a)$ given by

$$x = \frac{r_i^2 - r_g^2 + r^2}{2r_i}, \; a = \frac{1}{2r_i}\sqrt{(-r_i + r_g - r)(-r_i - r_g + r)(-r_i + r_g + r)(r_i + r_g + r)}.$$

The angle between these points is proportional to the likelihood that a random push is successful. Taking the special cases of a cell being on the edge and those in which no intersection with the edge exists into account, we have:

$$p_{push}(r_i|r, r_g) = \begin{cases} 0, & \text{if } r_g + r_i \leq r, \\ 1, & \text{if } r_g - r_i > r \vee r - r_i \leq 1, \\ \frac{1}{\pi}\cos^{-2}\left(\frac{x - r_i}{r_g}\right), & \text{otherwise.} \end{cases}$$

Using this likelihood, the effective population size N_{eff} can be calculated and used in the ODE model instead of N to factor in random pushing, thus giving:

$$\frac{dN}{dt} = N_{eff} - \pi(\sqrt{N/pi} - r_g)^2, N_{eff} = \int_0^r 2\pi r_i \; p_{push}(r_i|r, r_g)\, \mathrm{d}r_i.$$

Indeed, this ODE explained the observed behaviour of the simulator better for higher values of d_{push} (see blue lines in Figure 3.2). Upon inspection of the function $p_{push}(r_i|r, r_g)$ for different parameter values, it also became apparent that the random pushing leads to unwanted behaviour if the radius of the tumour is $r \leq d_{push}$. In this case, cells at the centre of the tumour are more likely to divide than cells on the outer edge (Figure S.17A, page 269). This behaviour arises since cells in the centre will be able to push in all directions, whereas cells on the periphery can only push towards the edge of the tumour, although with a lower likelihood (Figure S.17B, page 269). Due to this and the increased intermixing arising as a consequence of this method, an alternative heuristic that identifies the closest edge was added to the simulator (see Section 3.1.1.1 of Methods, page 78). With this alternative method, the simulated growth curves (Figure 3.3A or Figure S.18A, page 269) and the number of generations required to reach a specific tumour size for different values of d_{push} (Figure 3.3B) matched the expected ones almost exactly.

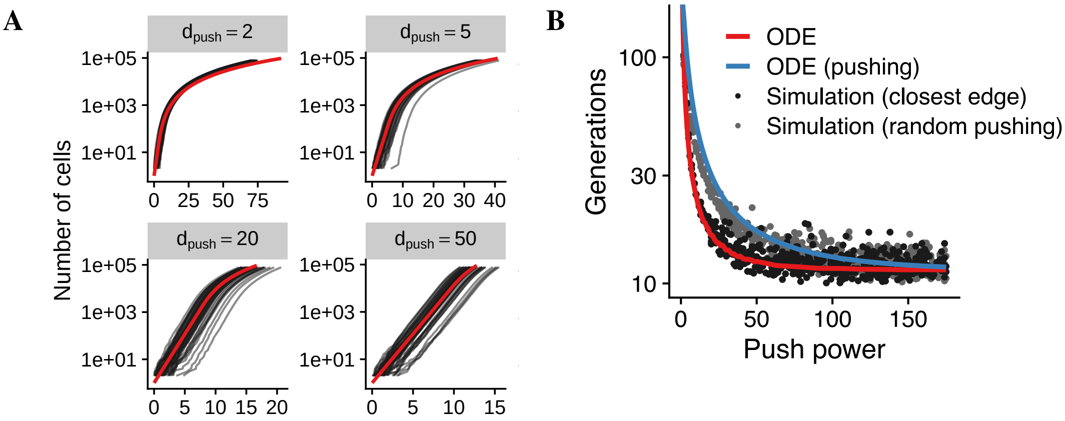


Figure 3.3: Expected vs simulated generation times for different degrees of boundary-driven growth. A) Random realisations of growth curves obtained from the spatial simulator for different diameters of the outer growing edge (d_{push}) when cells 'push' to the closest edge. The red line shows the expected distribution obtained from a set of ODEs. B) The time required for a neutrally growing tumour to reach a radius of $r_{end} = 175$ in two dimensions. The red line shows the expected growth dynamics according to a simple ODE model, and the blue line shows the same for a set of ODEs that accounted for the effect of pushing into a random direction instead of to the closest edge during cell divisions. The black and grey dots show random realisations obtained from the stochastic simulator, where cells push approximately to the closest edge or into a random direction respectively. It is evident that the random pushing causes artefacts leading to a deviation from the expected behaviour (grey dots vs red line). No such deviation is visible for the simulations in which pushing occurs approximately towards the closest edge (black dots vs the red line).

3.2.2 General Insights Into Spatial Tumour Growth

In Figure 3.4A-C, three examples of neutral and non-neutral spatial simulations, obtained using a small degree of boundary-driven growth ($d_{push} = 20$), are shown. It is important to note that the results significantly depend on this parameter. Staining Ki67, a marker for the replicative activity of cells, in tumours have shown a higher replication rate on the edge of tumours, thus supporting that tumour growth is primarily driven by cells on the outer edge of a tumour. For this reason, boundary driven growth was assumed to be the most realistic model in the following. In the following the behaviour of the model under various degrees of boundary-driven growth will be described and in Chapter 6 I will apply an ABC-SMC algorithm to fit this parameter to the trees of individual patients.

The first example shown in Figure 3.4A is an entirely neutral simulation. At a tumour size of 10 cells, each existing cell was ‘marked’. In reality this might correspond to a random passenger mutation or a lentiviral barcode (Lamprecht et al. 2017). Following this, the simulated tumour was grown to a final size of 10^5 cells under neutral dynamics. The top of Figure 3.4A shows the distribution of the marked lineages in space, and at the bottom, each clone (i.e., cells with identical fitness due to common ancestry) are marked in different colours. Equivalent plots for a tumour with one selected subclone and two selected subclones (branching) are shown in Figure 3.4B and 3.4C respectively.

From the neutral simulation, it can be seen that the relative size of each marked sub-lineage (i.e., the descendants of the marked single cells) differ substantially. This effect arises from drift, which is amplified due to the competition for space under boundary-driven growth. Under sufficiently significant selective advantages, deviations of the relative clone sizes can, of course, be observed (yellow lineage in Figure 3.4B&C). Still, due to the potential effect of strong drift, deviations of relative sub-lineage sizes observed under selective advantages can be hard to distinguish from neutrality (e.g., the dark blue lineage in Figure 3.4C).

Single-cell sequencing can be used to detect selection Single-cell sequencing provides information on which mutations co-occure in groups of cells. The information encoded in the somatic mutations can easily be used to reconstruct the ancestral history of cells. In Figure 3.4D, two simulated neutral phylograms are shown. As seen here, relatively balanced trees are obtained under both boundary-driven (top) and non-boundary-driven (bottom) growth. Since mutations were also assumed to be accumulated in non-dividing cells in the simulations, most cells have a relatively similar mutation burden. Still, one difference that can be seen between boundary-driven and non-boundary-driven growth, which will be discussed in more detail below, is that the relative branching-times (i.e., the relative position of internal nodes between the root and the tip) differ. This effect arises due to cells going ‘practically extinct’ once they fall behind the growing edge.

In a case of relatively late arising selection, phylogenetic trees reconstructed from ran-

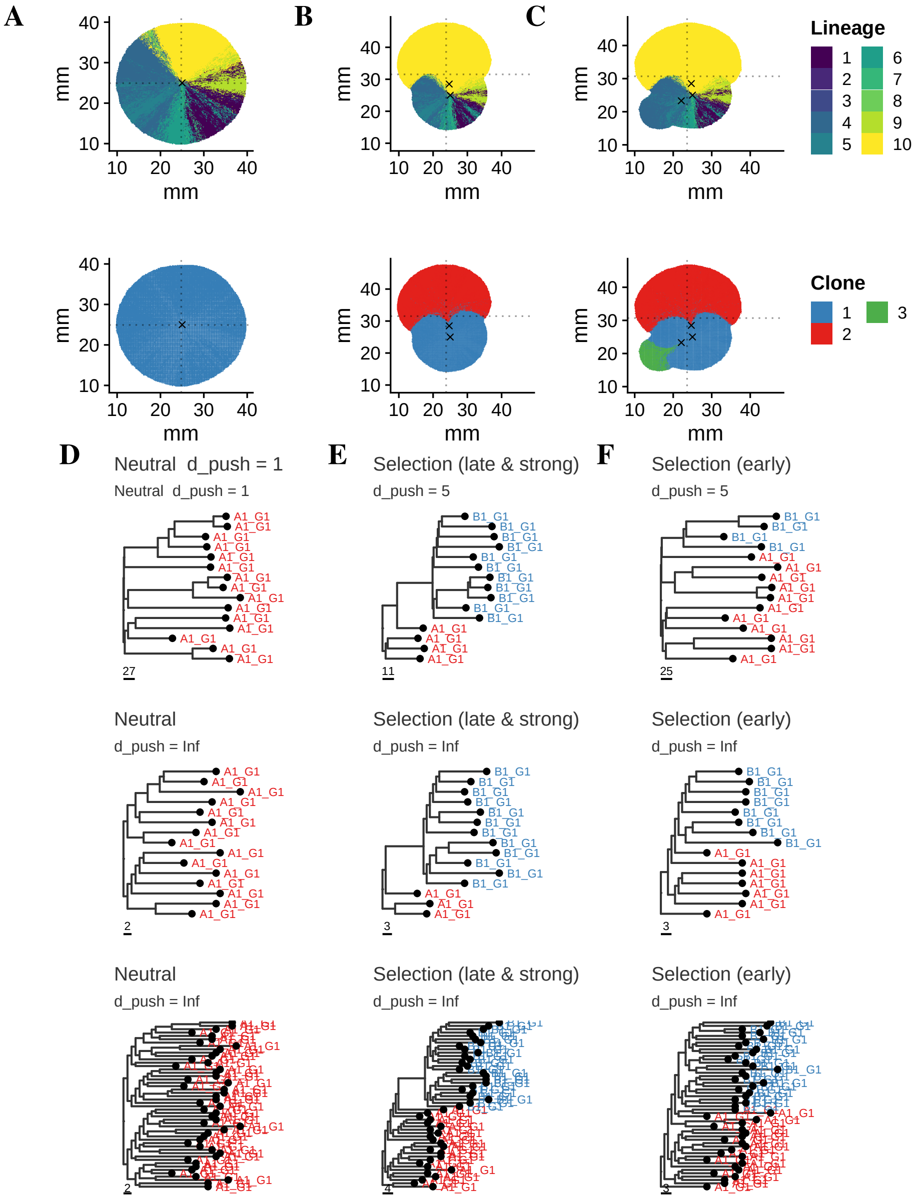


Figure 3.4: Examples of spatial simulations and single-cell trees. A) Representative simulations of a neutral tumour. B) A tumour with one subclone. C) A tumour with two subclones. The top plots show lineages of cells marked at a tumour size of 10 cells. D) Two representative single-cell trees of random cells from a neutral tumour under boundary-drive (top) and exponential growth(bottom). E) Equivalent single-cell trees from tumours with a late arsing subclone with a strong selective advantage, similar to the ones in B. Here a clear deviation from the balanced tree in D is visible, indicating the presence of a selected subclone. F) If a subclone arises early with only moderate selective advantage and hence does not sweep through the whole population, deviations are less clear and again hard to distinguish from selection.

domly taken single-cell samples revealed elongated internal edges and a subset of cells with a higher mutation burden than others (Figure 3.4E). From these data, an obvious deviation from neutrality is evident. This is something that would have been hard to resolve from the size of randomly marked lineages alone. Still, if a selected subclone arises very early —

i.e., under weak selection so that the subclone ultimately coexist with the ancestral clone — these patterns are less pronounced and again, especially from relatively sparsely sampled data, hard to distinguish from neutrality (Figure 3.4E).

3.2.2.1 Neutral Boundary-Driven Growth

I next assessed the effect of boundary-driven growth on the growth dynamics of individual tumours. For this, spatial simulations with marked lineages, similar to the ones described above, were generated and used to quantify the amount of spatial intermixing of different lineages. Examples of these simulations are shown in Figure 3.5A.

From this figure, it can be seen that more intermixing of lineages occurs for lower degrees of boundary-driven growth (left to right). In the case of fully exponential growth, scattering is widespread. Increased death (top to bottom) only has a relatively minor influence on the amount of intermixing observed. These observations are also summarised by the statistics shown in Figure 3.5B&C. Variable strength of boundary-driven growth in individual tumours might explain the different rates of spatial variegation observed by Sottoriva et al. (2015) between carcinomas and adenomas.

Phylogenies can be used to resolve boundary-driven growth The differences in the growth dynamics implied by the different intermixing and scattering of cells in space implied by the data summarised in Figure 3.5 should also be encoded within genomic measurements obtained from single-cells. To test this hypothesis, a spatial sampling layout similar to the one used by Sottoriva et al. (2015) was used. In brief, random single-cells were obtained from four regions with diameters of $\approx 50 \times 50$ grid located on the outer edges of the tumour ($350x350$ grid points) with a 90° offset from each other (i.e., at a 12, 3, 6, and 9 o'clock position) were subjected to simulated sequencing and phylogenies were reconstructed from these data using a maximum-parsimony method. Similar to the example using random sampling of single-cells mentioned previously (top tree in Figure 3.4D), a clear difference in the relative distribution of branching-times could be seen in these simulated trees (Figure 3.6A). Under strict boundary-driven growth (left side of Figure 3.6B), 'palm-tree' shaped phylogenies can be observed, and the strength of this effect is only moderately reduced if the death rate is high (left bottom left of Figure 3.6B). Each clade in the corresponding trees was formed by samples from one region, which are indicated by the colour of the added labels (left site of Figure 3.6B).

For purely exponential growth, branching occurs instead at a relatively early position

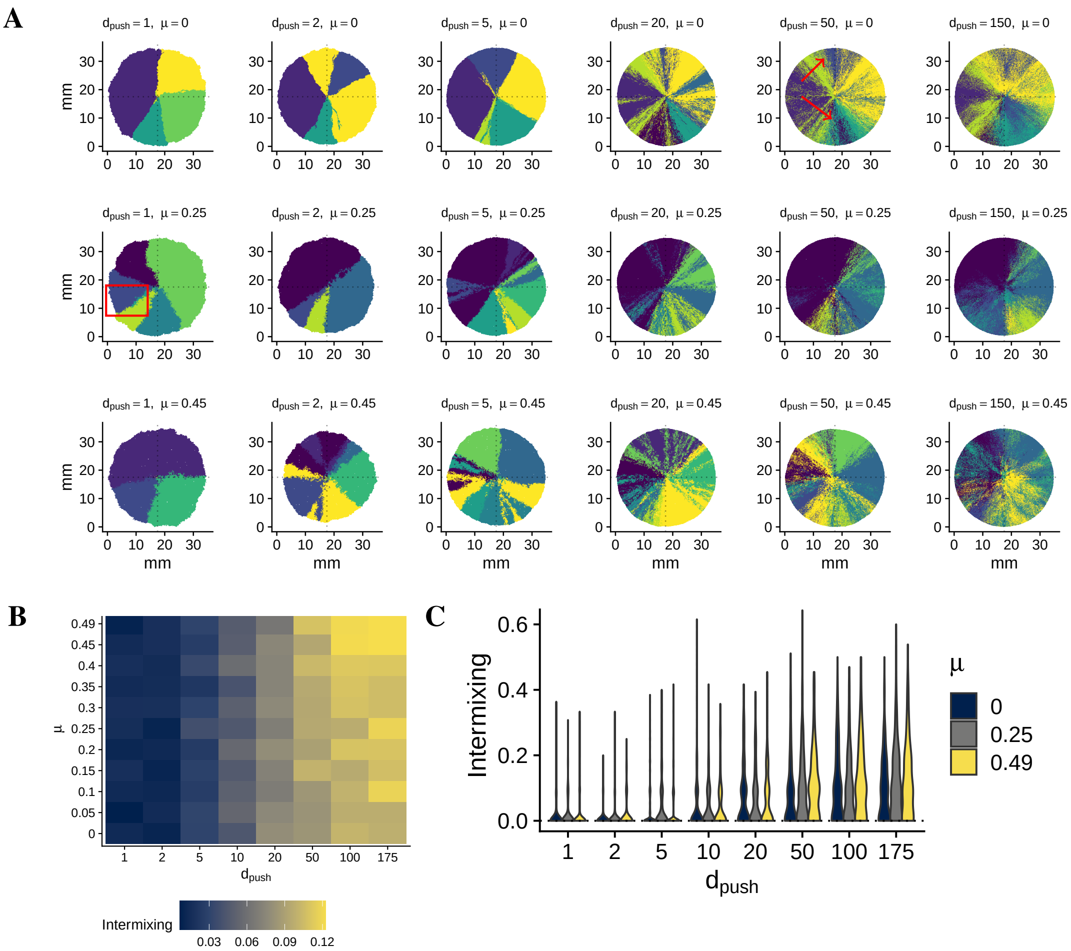


Figure 3.5: Illustration of mixing effects due to boundary and non-boundary-driven growth. A) Neutral simulations obtain for combinations of the push distance (d_{push}) and death rate (mu) parameters. Cells and their descendants were marked as distinct lineages after reaching a population size of ten cells. Red box: Stochastic out-competition of a lineage on the growing edge by surrounding cells. This occurs more frequent under boundary-driven growth. Red arrows: Spatial segregation arising due to early intermixing of lineages commonly observed under non-boundary-driven growth. B-C) Summary statistics of the intermixing rate show how intermixing rates increase with a larger width of the growing edge (d_{push}) and slightly with increasing death rates μ. The intermixing rates were calculated with the tree statistic I (see Methods section) on 100 randomly samples cells from a tumour with a diameter of 350 points.

in the trees (right side of Figure 3.6B). Within the tree, samples from the same region were frequently less distant to each other than those from different regions, but the formation of clades by all samples obtained from one single region of the tumour occurred very rarely. At the intermediate parametrisations of d_{push}, a transition between the patterns seen in the two extreme cases became apparent (middle of Figure 3.6B). These observations suggest that through the analysis of the shape of single-cell phylogenetic trees, an accurate estimation of the strength of boundary-driven growth might be possible.

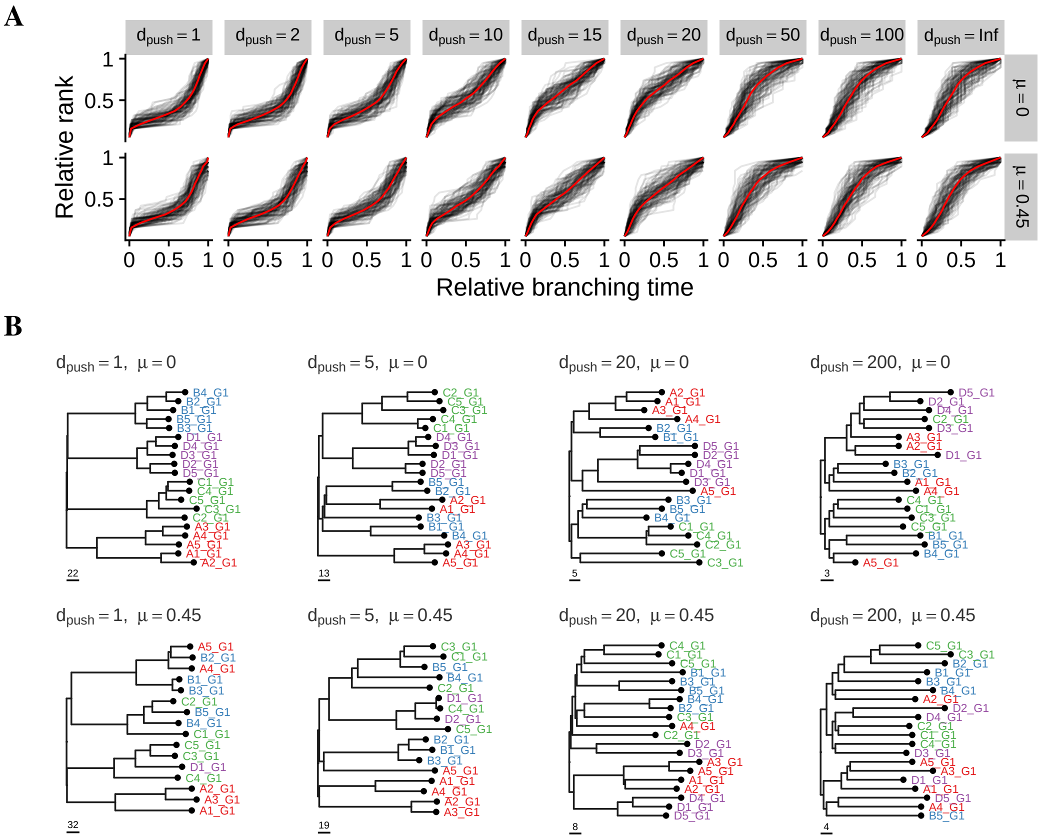


Figure 3.6: Relative branching-time and structured sampling should allow the recovery of growth laws. A) Relative branching-times of trees obtained from spatial sampling in four regions from the outer edge of a tumour. Black lines show estimates of random realisations of trees containing 20 single cells each. Little difference with regard to the death rate μ is visible, but changes in the width of the growing edge d_{push} cause clear deviations. B) Examples for different parameter values from A.

These patterns were also evident in the distribution of the lineages-through-time plots shown in Figure 3.6A. In this context, it is important to note that a combination of two effects is at play i) the effect of the boundary-driven growth itself and ii) biases introduced due to non-random spatially sampling. If the latter effect did not exist, one would expect to see a uniform branching across lineages under exponential growth, but due to the effect of non-random sampling, we instead have to compare the relative distribution of branch times for different values of d_{push} to assess its effect. However, the analysis showed that branching in reconstructed phylogenies consistently occurred later (concave up) in boundary-driven growth. For non-boundary-driven growth, branching occurred instead earlier (convex up).

Mutational processes can reveal boundary-driven growth Another possible way to distinguish boundary and non-boundary-driven growth might be the activity of different mutational processes. In this context, two simple mutation processes can be conceived i) those that are continuously active and cause damage to the DNA (i.e., a ‘non-mitotic’ process) and ii) those that are only active during cell division when a second copy of the DNA is created (i.e., a ‘mitotic’ process). Given that cells within the centre of a boundary-driven tumour have a reduced mitotic turnover and assuming that these two mutation types can be distinguished from each other, differences in the mutation rate of these might reveal a pattern that is indicative of the growth law in bulk WGS data.

To explore this hypothesis, I integrated both of these mutational processes into the spatial simulator and generated simulated bulk WGS datasets from the tumour as a whole. For the implementation of the two different processes, the default behaviour of the model was slightly modified. Cells that did not manage to divide successfully were assumed to not accumulate mutations to represent the ‘mitotic’ mutational process. The ‘non-mitotic’ mutation process was included as an additional reaction, which added one mutation to a random cell, in the Gillespie algorithm. The rate of this process was set to the mutation rate per division of the ‘mitotic’ process to ensure that mutations from both processes were present at equal proportions under exponential growth.

The results of a representative simulated boundary-driven tumour are shown in Figure 3.7. These data revealed, as hypothesised, a pattern that could potentially distinguish the presence of boundary-driven growth, namely an excess of mitotic mutations compared to non-mitotic ones at a low VAF .

It might, in principle, be possible to distinguish such processes based on the analysis of mutational signatures. Indeed, a previous study found that the number of mutations assigned to individual mutational signatures was only weakly correlated with the age of a person at the time of tumour diagnosis (Alexandrov et al. 2015). Only a single mutational signature (S1) showed a strong correlation with age across tumour entities. S1 is associated with the spontaneous deamination of methylated CpG dinucleotides (Alexandrov et al. 2013b) and is therefore also expected to occur in non-dividing tissue (i.e., non-mitotic). Other mutational signatures with a known aetiology are in contrast associated with defects introduced during the duplication of the DNA (i.e., mitotic). One example of this would be signature S10 from COSMIC, which is associated with mutations of POLE, causing error-prone DNA replication during division (Heitzer and Tomlinson 2014). Depending on how widespread such mitotic signatures are, the observations made here might partially explain why shifts in the activity of different mutational processes occur so frequently between clonal and

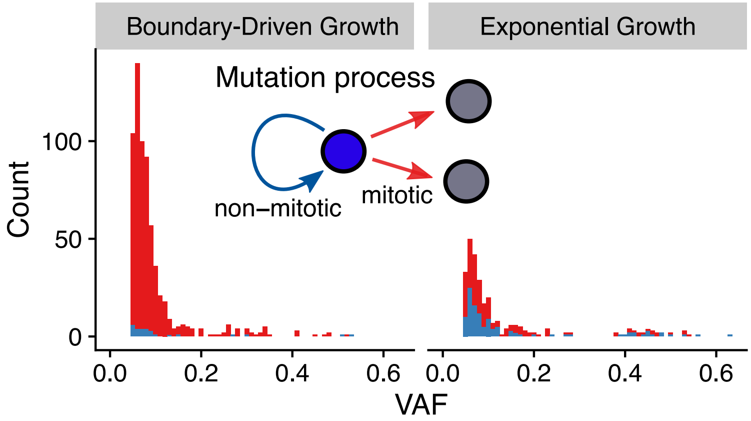


Figure 3.7: Mutational processes can reveal boundary-driven growth. The histograms show the global VAF distribution of mutations generated by two different mutational processes (colours): i) mutations generated by a continuously active non-mitotic process (red) and ii) a mutational process that is only active during division (i.e., mitotic). A clear difference in the VAF distribution of mutations generated from these two processes can be seen between boundary-driven tumours (i.e., cells growing only on the outer edge, $d_{push} = 1$) and exponentially growing tumours ($d_{push} = \infty$) can be seen. Tumours with boundary-driven growth show an excess of low-frequency mutations generated by the mitotic process compared to exponentially growing tumours. Identification of such processes from WGS data might allow discriminating between these two modes of growth in tumours.

subclonal mutations observable in tumour sequencing data.

3.2.2.2 Non-Neutral Boundary-Driven Growth

Boundary-driven growth dampens selection Similar to the previous analysis, an assessment of what effects selected subclones have on the structure of reconstructed phylogenetic trees in combination with boundary and non-boundary-driven growth was conducted. For this subclones with a given selective advantage λ_{sc} were introduced into the simulation at a given population size t_{sc}. The simulated tumours (2D) were grown to a total population size of $N_{end} = 10^5$ and the relative size of the subclone f_{sc} was determined. A range of parameter combinations, were tested with this setup, specifically all combinations of $t_{sc} = \{\lfloor 2^{(n/2)} \rceil \mid n \in \{0,...,30\}\}$, $\lambda_{sc} = \{1 + x/4 \mid n \in \{0,...,36\}\}$ and $d_{push} \in \{1,5,20,\infty\}$ with 25 realisations each. A mutation rate of $m = 50$ and a death rate of $\mu = 0$ was used in all cases.

For the introduction of the selected subclone, a random cell was chosen from the population and modified. Given that for some of the parameter simulations, a large number of cells were already present at this point, it is expected that some of the transformed cells were located behind the growing edge. For this reason, simulations in which the introduced subclone did not expand were rejected. The rate of this rejection is shown in Figure S.20 (page 270). On simulations in which the subclone was able to expand (i.e., the non-rejected ones), the average size of the subclone after reaching a total size of 10^5 cells was calculated. These numbers are summarised in Figure 3.8A. From the same setup of simulations, the fraction of simulations in which the subclone made up between 10% and 90% of cells in the simulation were also calculated. These are shown in Figure 3.8B.

The observations under boundary-driven growth, shown in the bottom right corner of the two figures, can be used as a reference. As seen here, in the majority of the tested parameter combinations the subclone effectively swept through the population (yellow in Figure 3.8A and grey in Figure 3.8B). In this parameter range, it would be relatively unlikely to sample from the ancestral clone. Likewise, for a number of parameter combinations the subclone did not have enough time to grow to a sufficient size (dark blue in Figure 3.8A and grey in Figure 3.8B). In these, it would be unlikely to sample from the subclone. In neither of these two sets, we would expect to ever observe any evidence for selection in the global VAF spectrum if we were to sequence the tumour as a whole. The issue of this relatively narrow range in which subclones could potentially be detected in such data, the 'wedge of selection', was also described in Williams et al. (2016).

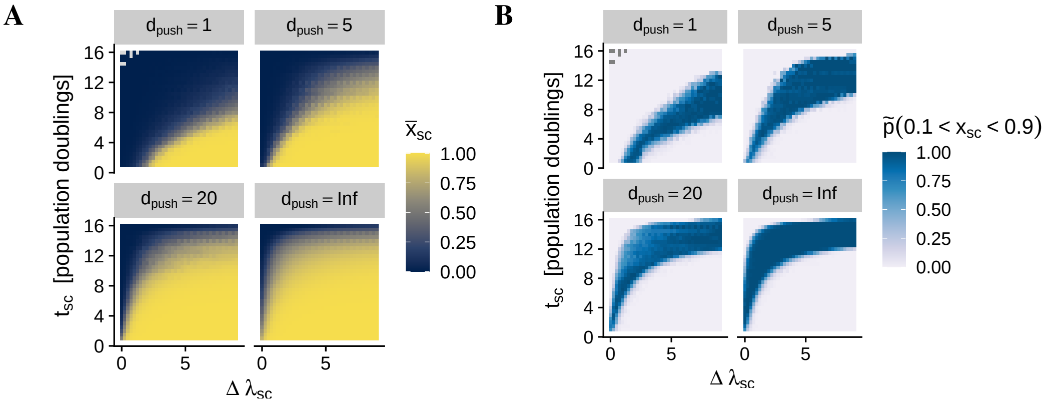


Figure 3.8: Effect of boundary-driven growth on subclone sizes reached for different values of the pushing distance d_{push}, the subclone start time t_{sc} and the selective advantage of the subclone $\Delta\lambda_{sc} = \lambda_{sc} - \lambda_{ac}$, where $\lambda_{ac} = 1$ and λ_{sc} are the birthrates of the ancestral clone and the subclone respectively.

Comparison of the observations under fully exponential growth to those observed under various degrees of boundary-driven growth showed that the efficiency of selection (i.e., the ability of the subclone to grow to a very large size) was reduced under boundary-driven growth. Given that the growth of subclonal cells is restricted to the growing surface, this is certainly expected. Still, depending on the actual growth law applying to human malignancies (i.e., boundary vs non-boundary-driven growth), this effect would have to be considered to estimate correct parameters from bulk or multi-region WGS.

Boundary-driven growth is detectable in single-cell WGS data After the characterisation of the parameter range in which subclones in spatial simulations could potentially be observed, random spatial sampling of single-cells, followed by simulated sequencing of these, was conducted to determine the ability to detect subclonal selection from single-cell sequencing data. For this, three tree balancing metrics and a metric that describes the distribution of branching within the trees were assessed (see Methods for details). A simulation setup identical to the one used to analyse clone sizes under selection, fully described in the previous paragraph, was used. In each case, 20 random cells were obtained from the simulated tumour, subjected to simulated sequencing and maximum-parsimony phylogenetic reconstruction. From the reconstructed phylogenies, the four summary statistics were calculated. These statistics are summarised in Figure 3.9. From the results shown here, it is evident that in those intervals in which samples from both subclones can in principle be obtained (see Figure 3.8), an apparent deviation from the typical tree balance expected under neutrality can be observed (dark blue colours in Figure 3.9). This suggests that even a moderate amount of single-cells subjected to WGS sequencing should be sufficient to detect subclonal selection. A larger number of cells should in principle even allow the detection of selected subclones at a frequency far below the limit of detection in bulk WGS sequencing data (i.e., $< 10\%$).

3.3 Contributions to MOBSTER

In the previous chapter, some issues of commonly used clustering methods, when applied to bulk simulated WGS sequencing data obtained from a neutral branching process model of cancer evolution, were described (see Figure 2.3H, page 62). In short, these methods were found unable to explain the expected power-law distribution of subclonal variants expected under neutrality (see Figure 2.1A-B, page 55), causing these clustering methods to include several subclonal clusters, almost irrespective of the true number and position of subclones (Figure S.4, page 263 and Figure S.2, page 262). Importantly these subclonal clusters are, as they are expected to be composed of multiple lineages present at a similar VAF, almost uninterpretable.

These observations motivated Giulio Caravagna, a colleague in Andrea Sottoriva's group, to create an alternative clustering method called 'model-based tumour subclonal reconstruction' (*MOBSTER*). *MOBSTER* can fit a mixture of a Pareto distribution (i.e., the

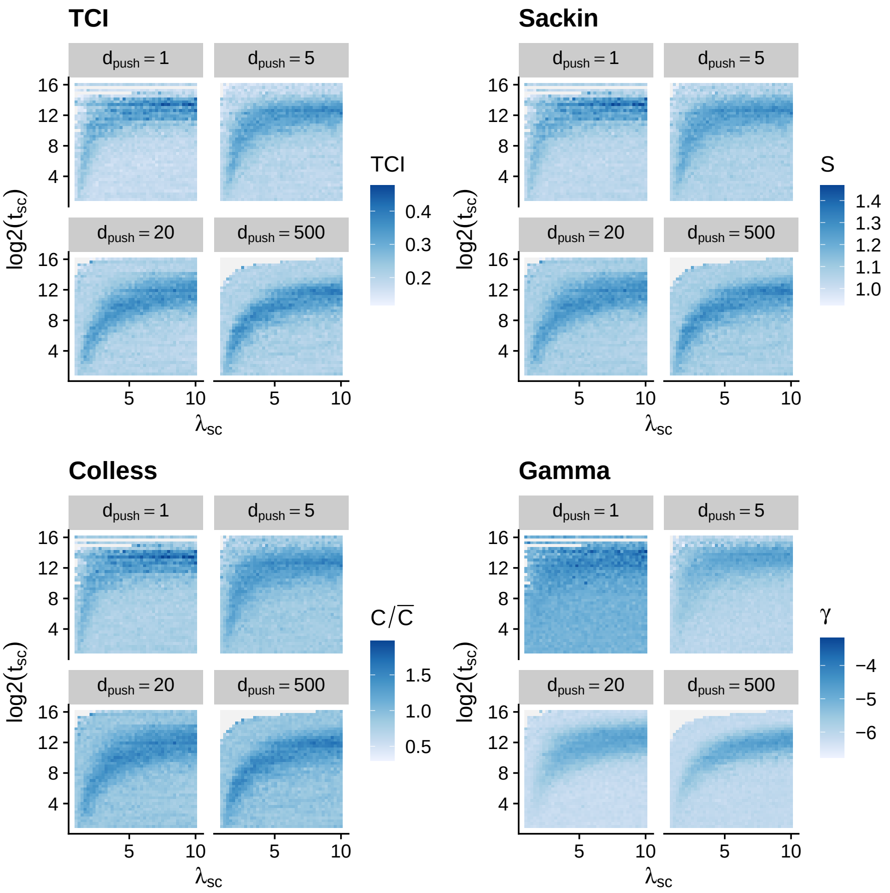


Figure 3.9: Detection of subclonal selection in single-cell sequencing data. The presence of selected subclones arising under various parametrisations causes detectable deviations in tree balances.

power-law ‘$1/f$’ tail) and multiple Beta distributions to the sequencing data and is hence a method that should in principle be able to account for the structures expected to arise under neutrality (Figure 3.10).

For the validation of the *MOBSTER* clustering method, two reference datasets composed of simulated non-spatial (univariate) and spatial (multivariate) sequencing datasets were generated. Each of these was composed of neutral simulations and multiple non-neutral simulations with subclones present at different frequencies. Both of these datasets were used to characterise the ability of *MOBSTER* to detect selected subclonal clusters in comparison to other commonly used methods.

3.3.1 Non-Spatial simulations (Univariate dataset)

Generation of simulations The first dataset was composed of simulated sequencing data obtained from a non-spatial (i.e., univariate) tumour model. For this, the simulator described in Chapter 2 (see Section 2.2.2 on page 61 for details) was used. Instead of a Poisson distributed coverage C of mutant alleles x_i, an over-dispersed Beta-Binomial distribution was added to the model (see Section 3.1.1.2 on page 80). For each simulation, one ancestral population and a single mutant subclone, introduced at a fixed time-step t_s, were generated.

Constant parameter values were used for the mutation rate $m = 16$ (mutations per doubling), the death rate $\mu = 0.2$, the total number of reactions $t_{end} = 179,782,830$ [4], and the total number of clonal mutations $N_c = 500$. For the initial dataset, an average sequencing depth of $\bar{C} = 120$ and sample purity $a = 1$ were used. Nine random realisations for each combination of subclone birthrates $\lambda_s \in \{1+0.1i \mid i \in \mathbb{N} \wedge 1 \leq i \leq 13\}$ and number reactions prior to initiation of a subclonal expansion $t_s \in \{2^i \mid i \in \mathbb{N} \wedge 4 \leq i \leq 14\}$ were simulated.

Selection of 150 datasets for testing All simulations in which the subclone accumulated less than 50 mutations before its transformation (i.e., less than 4-5 divisions) were removed, and three datasets with a specific fraction of mutated cells in the population (x_s, the CCF of the subclone) were generated by randomly selecting from the remaining simulations as follows: i) 20 effectively neutral cases where $x_s < 5\%$, ii) 20 effectively neutral cases with $x_s > 90\%$, and iii) 110 cases with a potentially detectable subclone, with $20\% < x_s < 80\%$. These cases represent tumours with minor, almost undetectable subclones (e.g., Figure 3.10C), tumours where the subclone has swept through the entire population and cases where the subclone is detectable within the VAF spectrum (e.g., Figure 3.10D).

Analysis Mutations from each of these 150 WGS sequencing datasets were clustered with *DPclust* (Nik-Zainal et al. 2012a), *PyClone* (Roth et al. 2014) and *SciClone* (Miller et al. 2014) before and after the removal of subclonal tails with *MOBSTER* (Figure 3.10B). The number of inferred clusters relative to the true number of 'clone cluster' (i.e., $k = 1$ for neutral and $k = 2$ for non-neutral cases) is summarised in Figure 3.10E. These data demonstrate that for all of the four tested methods a similar number of additional clusters were inserted due to subclonal $1/f$ tail present in the simulated data (yellow colour in Figure 3.10E). This was the case in the same way for both, neutral, and non-neutral WGS data. After removal of the subclonal tail with *MOBSTER* the number of additional clusters was significantly reduced (green colour in Figure 3.10E).

Representative fits for a simulated tumour with one subclone and an effectively neutral case are shown in Figure S.22A (page 271) and Figure S.22B (page 271) respectively. Still,

[4] It is not entirely clear to me why I chose this somewhat arbitrary value, but ultimately one obtains very similar data for a weakly selected subclone in a tumour grown to a larger size and a strongly selected subclone in tumour grown to a smaller size (Williams et al. 2018b) . The size of the simulated tumours was $\approx 107,800,000$ cells.

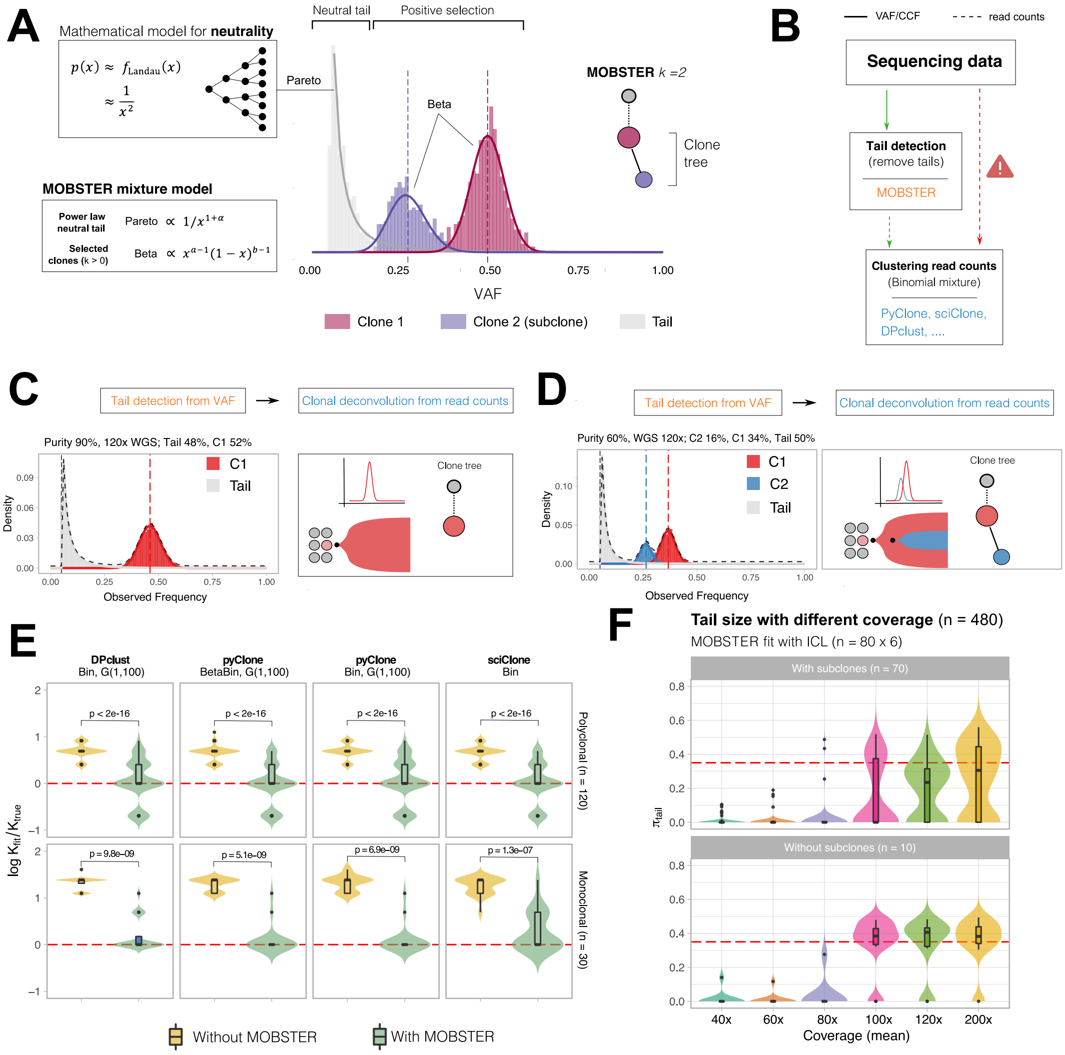


Figure 3.10: Validation of *MOBSTER* using synthetic tumour bulk sequencing data. A) Principle of the *MOBSTER* method. Details can be found in Caravagna et al. (2020). B) Various clustering methods were either applied directly to simulated sequencing data or after the removal of the subclonal power-law tails with *MOBSTER*. C) A representative neutral simulation. D) A representative non-neutral simulation with a subclonal cluster present at a VAF of ≈ 0.25. E) Summary of the number of clusters inferred by *DPclust*, *PyClone* and *SciClone* before (yellow) and after (green) removal of the subclonal tails with *MOBSTER*. F) The effect of reduced purity on the mixture weight of the tail component.

as seen in Figure 3.10E misclassification did occur by *MOBSTER*. Analysis of these cases identified three modes of failure.

Failure modes First, in ≈ 70% of misclassified cases, the subclone was present at a very high frequency (Figure S.22C, page 271). Here the additional variability of the beta components fitted the mixture of two Binomial distributions sufficiently well. In these cases, an approach that removed mutations assigned to the tail and clustered the remaining ones with a method that fits a mixture of binomials on the raw count data (e.g., *BMix*) was typically able to identify the subclone correctly. The *BMix* package for R was used to fit maximum likelihood Binomial mixtures to the data since the clustering of mutations was not the objective of this analysis. Still, in principle Dirichlet Process based clustering using Binomial distributions like *DPClust*, would be expected to obtain similar results if clustering of mutations would be required. Secondly, $\approx 17\%$ of misclassified cases the subclone was ‘hidden’ below the power-law tail (Figure S.22D, page 271). This problem especially arose when the subclonal cluster was small (i.e., small t_{sc}). While this is a genuine error of the method, it is inherently hard to resolve. The remaining $\approx 13\%$ of misclassified cases had a low frequency subclone with no fitted tail (Figure S.22E, page 271). In these cases, the low-frequency mutations of the tail were assigned to the subclonal cluster instead. While incorrect, this might, in practice, be irrelevant.

Notably, relatively high coverage is required to detect the subclonal tails (Figure 3.10F). At $C \leq 100$, reliably detecting variants at a low frequency is compromised, and subclonal tails are often not detected in these data. This can, in turn, lead to over-calling of subclonal selection in low-coverage WGS data. More extensive tests of this behaviour are shown in Caravagna et al. (2020), but generally, a minimum of 100x sequencing coverage appears to be required for subclonal reconstruction from single-bulk WGS data. A conclusion that was also supported by the simulated synthetic tumour datasets obtained from non-spatial simulations.

3.3.2 Spatial Simulations (Multivariate dataset)

Generation of synthetic datasets A second multivariate dataset composed of tumours with one ($n = 50$), two ($n = 10$) and three ($n = 10$) selected subclones at a detectable frequency were created. Between two to nine simulated biopsies were obtained from these synthetic tumours. Each tumour was grown on a 800×800 2D lattice until one of the cells reached the edge of the space. This results in tumours containing roughly $\approx 5 \cdot 10^5$ cells. New subclones with a birth rate $\lambda = [1, 1.6, 2.4]$ were introduced at time points $[0, 4, 6.7]$ respectively. These were chosen to allow coexistence of each subpopulation at approximately equal abundance at the termination of the simulation. The remaining parameters, equivalent to the non-spatial simulations, were kept constant: $m = 10$, $N_c = 100$, $\bar{C} = 100$, $\mu = 0$, $a = 1$, and $d = 100$ (see methods above). Biopsies of $10,000$ cells (i.e., 100×100 grid points) were taken along the outer perimeter with an equal angular distance relative to the centre between them. Representative examples of a simulated neutral tumour with two biopsies are shown in Figure 3.11A&B. Figure 3.12A&B shows a representative example of a tumour with two subclones and a total of three biopsies.

Fitting of multivariate datasets The multivariate datasets were fitted with the multivariate variational Binomial clustering method *VIBER*[5] on the raw read counts. The same analysis was run after removal on tail variants with *MOBSTER* along the marginals of samples (e.g., Figure 3.11B and Figure 3.12B).

Observations on neutral tumours In order to show how the reconstructed subclonal clusters after and before the removal of tails with *MOBSTER* were related to the spatial distribution of variants within the tumour, a virtual *in situ* staining was applied to the simulations. The results of this method for the neutral case shown as example above before and after removal of tails with MOBSTER are shown in Figure 3.11C and Figure 3.11D respectively. In these plots, a perfectly resolved 'mutation cluster' will have non-transparent colours in the entire tumour (i.e., all mutations are present or absent). Imperfectly resolved 'mutation cluster' will, in contrast, have a variable amount of staining within the tumour. This means that all mutations are present in some cells, whereas others only contain a subset of the mutations from the cluster. These imperfectly resolved clusters could still be identified through more extensive sampling in space but should be removed for subclonal reconstruction. Comparing Figure 3.11C and Figure 3.11D shows that removing tails with *MOBSTER* can help to reduce the amount of spurious unresolved subclonal clusters. This was supported by the observations in the remaining case (see Figure S.21, page 270).

Observations on non-neutral tumours Equivalent *in situ* staining data for a non-neutral tumour with two subclones are shown in Figure 3.12C. Here staining of all clusters detected after the removal of tails along the marginals is shown. First, the clonal cluster (C1) was identified correctly, and these mutations are present in the entire tumour (dark green staining). A second large cluster (C3) were those mutations that formed the MRCA of the subclones #2 and #3. While this cluster also contains a small number of non-neutral mutations present in some, notably unsampled, cells of the background clone #1 (note the red cells in purple staining), this cluster would also be expected to contain the first 'driver' mutation(s) responsible for the selection of the first subclone #2.

Similarly, cluster C5 contains the simulated 'driver' mutations responsible for the se-

[5] In principle, any other mixture model or clustering method could be used instead of *VIBER* in the same way. Due to the absence of overdispersion in the simulated data, the choice of a Binomial mixture was considered to be reasonable. If a significant degree of overdispersion might be present, another distribution, like a Beta distribution, able to capture this should be fitted instead.

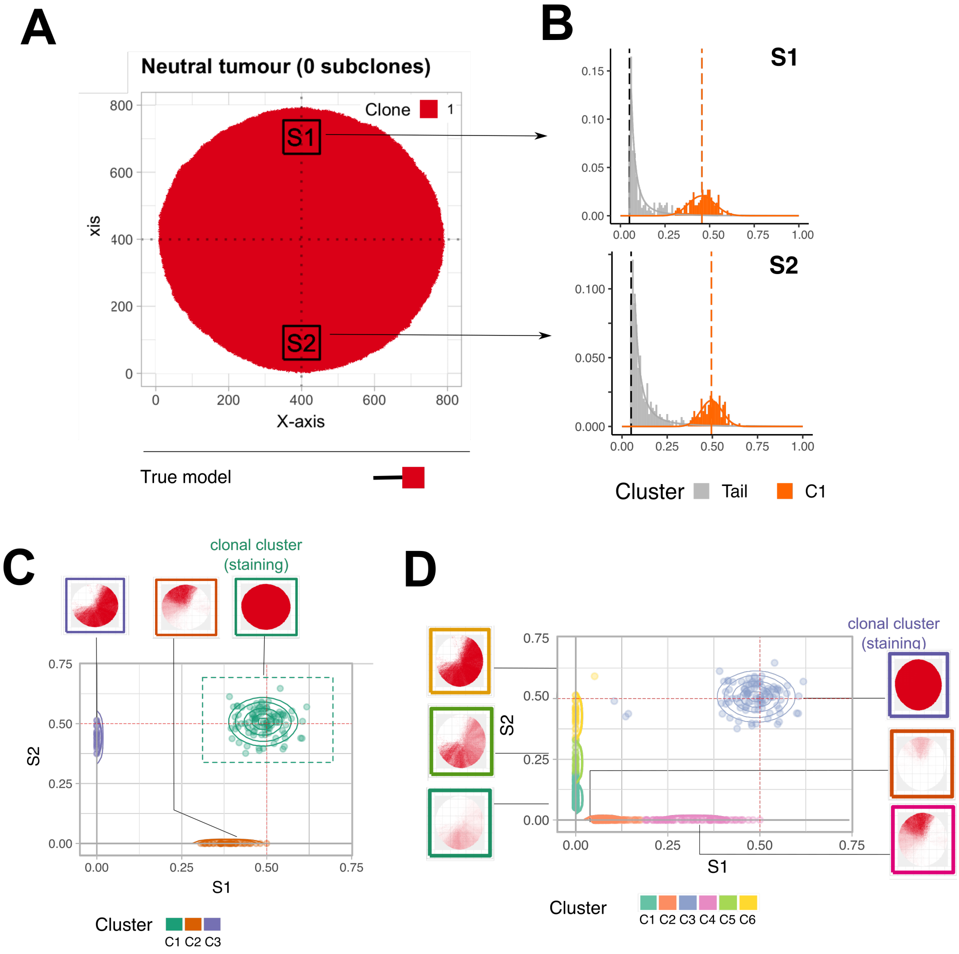


Figure 3.11: Example of a spatial simulation with no subclone. A) Spatial layout of cells coloured by the clone they belong to within the tumour. The location of two 100 × 100 bulk samples (S1 and S2) are shown as black boxes. B) Histograms of simulated NGS mutation data obtained from the two bulk samples. Colours highlight the cluster mutations were assigned to during a univariate clustering, i.e., along the marginals, with *MOBSTER*. C) Result of multivariate cluster analysis with *VIBER*, a variational Bayesian able to fit multi-variate Binomial mixtures, after mutations assigned to '$1/f$ tails' identified by *MOBSTER* (highlighted in grey in B) in all samples were removed. The scatter plot shows the VAF of mutations in both samples, and mutations (dots) are coloured by the cluster they were assigned to. A 'virtual staining' of these mutations within the tumour is shown in insets. D) Like C without the removal of mutations in tails identified by *MOBSTER*.

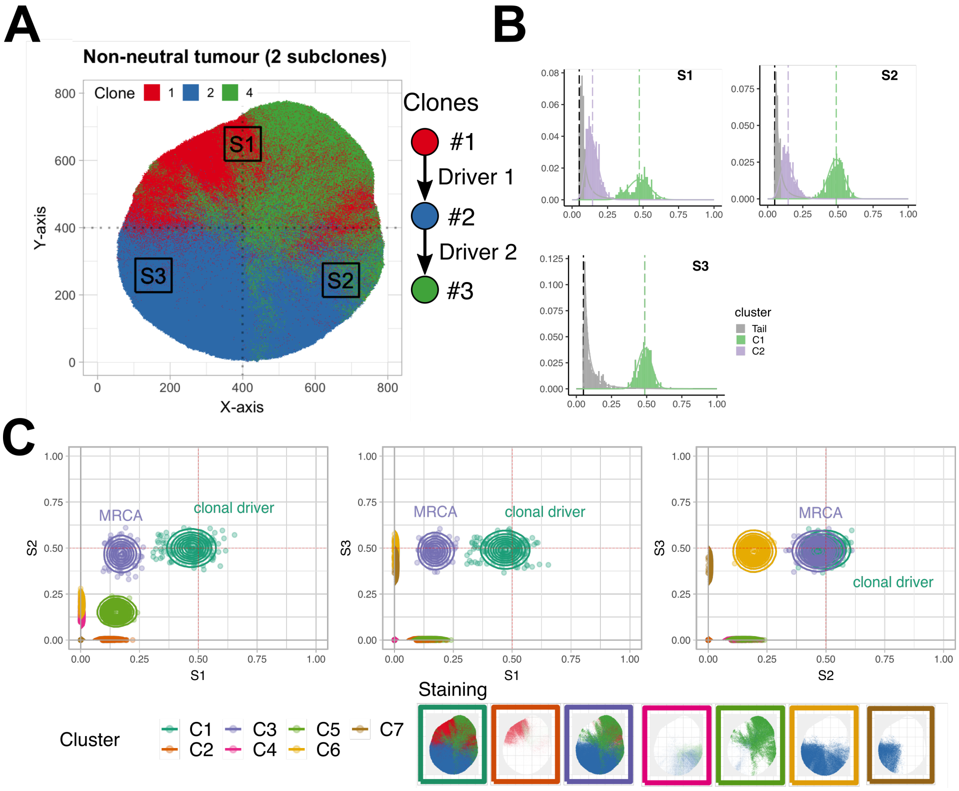


Figure 3.12: Example of a spatial simulation with two subclones. A) Spatial layout of cells coloured by the clone they belong to within the tumour. The location of three 100 × 100 bulk samples (S1, S2 and S3) are shown as black boxes. B) Histograms of simulated NGS mutation data obtained from the three bulk samples. Colours highlight the cluster mutations were assigned to during a univariate clustering, i.e., along the marginals, with *MOBSTER*. C) Result of a multivariate cluster analysis with *VIBER*, a variational Bayesian model able to fit multi-variate Binomial mixtures, after mutations assigned to '$1/f$ tails' identified by *MOBSTER* (highlighted in grey in B) in all samples were removed. The scatter plot shows the VAF of mutations in both samples, and mutations (dots) are coloured by the cluster they were assigned to. A 'virtual staining' of these mutations within the tumour is shown in insets.

lection of the subclone #3. All remaining identified subclonal clusters contained mutations that formed (unselected) MRCAs of subpopulations of cells present in the biopsies. Taken together, these results show how *MOBSTER* can be used to simplify the reconstruction of 'clone trees' by removing a large number of clones otherwise added to subclonal tails arising in the marginals (compare Figure S.21, page 270). Still, inferred clusters and 'clone trees' reconstructed from these have to be interpreted carefully. Not all detected clusters arise due to subclonal selection, and additional spatial effects, two of which will be described in more detail in the following, can heavily influence the shape of reconstructed trees.

3.3.3 Spatial Effects in Bulk WGS

Based on the spatial simulations generated for the validation of *MOBSTER* several issues that can arise from the usage of bulk WGS sequencing samples taken in space were identified. Two of these with particular importance for the usage of multi-region bulk samples for the detection of selection in such data were identified. As described in Section 3.2.2 one of the ‘hallmarks’ of a selected subclone in multi-region sequencing data is the presence of an elongated edge in reconstructed phylogenies (see Figure 3.4E).

Spatial structures in neutral tumours Now, the question is whether this property is sufficient to identify subclonal selection in such a reconstructed phylogenetic tree. Here two relevant spatial effects that can complicate such an analysis were identified. Both of these can easily be demonstrated on simulated sequencing data taken from a simulated spatial tumour. In Figure 3.13A, the growth curve of a neutral tumour simulation is shown. Once the tumour reached a size of 6 cells, each of these was ‘marked’ as an individual lineage. This is similar to what one would expect a real tumour to look like if lentiviral barcoding of individual cells was conducted *in vivo* (Heijden et al. 2019; Lamprecht et al. 2017).

In Figure 3.13A two general properties of spatial simulations can be seen: i) imprisonment of one lineage of the tumour (lineage #4) and ii) strong spatial drift on the edge (lineage #5). Both of these result from the ‘competition for space’ on the growing edge of the tumour. In the case of lineage, #4 this competition was unsuccessful, and in the case of lineage #5 descendants of the cell were ultimately able to survive. These behaviours are ultimately driven by a small number of cells that are able to ‘surf’ on the growing edge (Schreck et al. 2019).

This effect of gene surfing can also be seen in the radial patterns in the spatial staining shown in Figure 3.13B. The resulting distribution of clones within the tumour are mainly confined to specific regions of the tumour. By sampling and sequencing single cells within the tumour, we can see that the majority of these contain a roughly similar number of mutations (Figure 3.13C). From the global VAF spectrum shown in Figure 3.13D, it becomes obvious that some drift, seen by subclonal mutations at a high frequency (light grey colour), did occur, but generally, the distribution is consistent with the expected $1/f$ tails.

Admixture effect If now a sample is taken in space at a position at which it overlaps a lineage boundary, of which some a marked by colours in the spatial plot, a curious effect might arise (Figure 3.13E). The simulated sequencing data of the bulk sample B4, taken

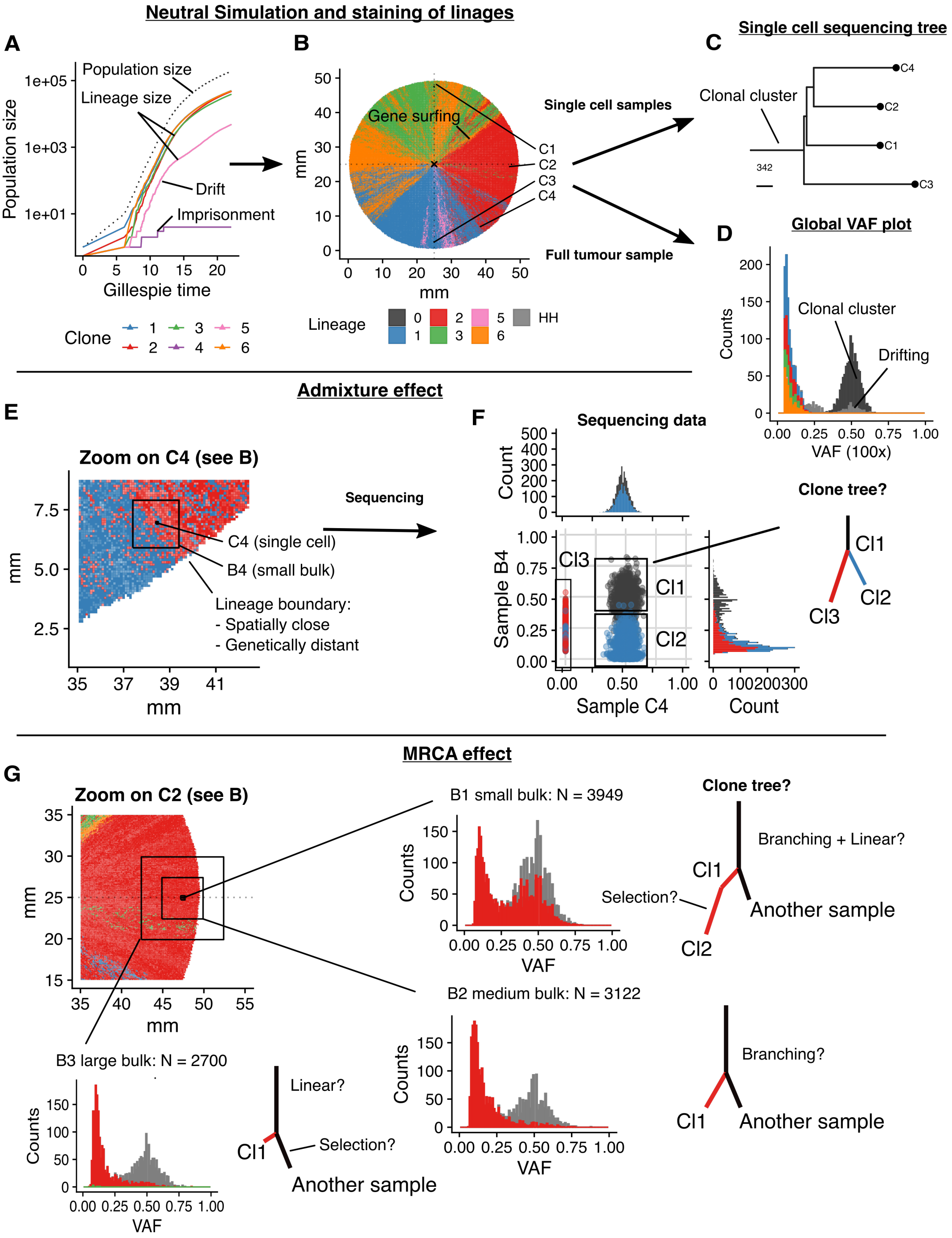


Figure 3.13: Spatial effects that arise in bulk samples. A) Growth dynamics of lineages over time. An example of imprisonment arising from the weak boundary-driven growth, lineage #4 and random drift on the growing edge restricting the expansion of lineage #5 are visible. B) From the spatial staining, radial lineage boundaries arising from gene surfing are visible. C) Simulated single gland sequencing reveals a very balanced subclonal structure indicative of neutral evolution. D) Similarly, a single clonal cluster and a $1/f$ tail are visible. Nested lineages within the tail and excess clonal mutations arising from drift (see A) are visible. E) Demonstrates the admixture effect in bulk samples crossing a lineage boundary (here red & blue). F) Sequencing and combination with an adjacent sample can reveal clusters that might be interpreted to indicate the selection of two subclones (i.e., branching). A single cell was used as a second sample, but the same effect can arise with bulks. It is important to note that spatial distance does not necessarily relate to genetic distance. G) The MRCA effect arising from different sizes of bulk samples.

across the boundary of the tumour, are shown on the right side of the scatter plot in Figure 3.13F, the single-cell sample from the blue sample is shown on top of the plot. In the case that a separate sample taken from the blue or red lineage is available, the mutations would reveal the presence of three clusters Cl1-3, as shown in the scatter plot. From this, one would be able to reconstruct a 'clone tree' similar to the one shown on the right-hand side of the plot. While technically correct, this clear 'branching pattern' would still not imply the presence of subclonal selection. It has instead to be considered that cells close in space can be genetically extremely distant. An even more concerning problem arises when we do not happen to sample either of the two lineages separately (i.e., in the absence of C4). Then subclonal mutations from both lineages (red and blue) are present at identical frequencies in the VAF spectrum. Using standard subclonal or *MOBSTER* would likely split these into a subset, leaving a very small subset of mutations considered clonal and altering the reconstructed trees.

MRCA effect A related effect arises when the size of individual bulk samples is varied (Figure 3.13G). This effect arises from the fact that only mutations of the MRCA of all cells in the sample will be present at a clonal frequency. All other mutations will instead be at a subclonal frequency. As the number of cells within the sample increases, more mutations of the $1/f$ tail of these subclonal mutations will fall below the level at which they will be detectable. In a small sample (e.g., B1), a relatively large number of clonal mutations and an additional number of subclonal mutations can be detected. As the size of the sample increases (e.g., B2 or B3), fewer variants will be found to be clonal or within the tail. Therefore, no matter whether the subclonal mutations are removed (e.g., using MOBSTER) or kept, various tree shapes can be obtained from the roughly same region of a tumour, just by altering the size of the bulk samples taken.

Clinical samples used for large-scale WGS studies like TCGA (Bailey et al. 2018) or PCAWG (Dentro et al. 2021) are normally obtained as part of diagnostic or therapeutic procedures. Normally, little control over how samples are obtained from the tumour is possible. Even if the obtained specimens are very large, these are not representative of the whole tumour. Therefore, bulk sequencing done on such samples can suffer from the spatial effects described here without the ability to resolve these through the extensive sampling of other regions. It is important to note that these patterns can emulate the hallmarks of selection, that is, the elongation of individual edges, making the interpretation of clustering results obtained from such bulk WGS sequencing data more challenging. Especially, the relative length of edges from single bulk samples has little meaning.

3.4 Summary

Through the use of a spatial tumour simulator, which includes effects from spatial crowding and boundary-driven growth, some general insights into the ability to detect i) the strength of boundary-driven growth and ii) subclonal selection was obtained. These data suggested that a relatively small number of sequenced single-cells should be sufficient to detect reasonable large subclones arising under selection and that single-cell phylogenies should in principle reveal whether tumours grew exponentially or under boundary-driven growth.

The spatial simulator was also used to characterise the behaviour of commonly used clustering methods in a univariate and multivariate setting. Specifically, it was possible to show that the new method *MOBSTER* (Caravagna et al. 2020), which Giulio Caravagna created, was able to resolve the issues arising from the '$1/f$ tail' in single-bulk WGS data. The method can remove subclonal tails expected under neutrality, which would otherwise be fitted by many binomial clusters. However, here issues remain and using simulation-based methods like ABC might, at least for the time being, be the best option for the interpretation of such data.

Last but not least, two general issues with bulk WGS data were identified: first that even small changes in the size of the tissue pieces used for the generation of libraries can significantly alter the length of edges in reconstructed phylogenies. Secondly, sampling across 'lineage boundaries' can also cause a miss-ordering of somatic variants when commonly used clustering methods are applied. Together these results highlight the importance of single-gland or single-cell sampling methods.

Chapter 4

Analysis of the EPICC Cohort

4.1 Introduction

Motivated by the results of previous studies that inferred the strength of subclonal selection from single-bulk WGS samples (Williams et al. 2016; Williams et al. 2018b) and the identification of general issues arising from spatial sampling effects in bulk WGS datasets (Caravagna et al. 2020), a multi-region single-gland sequencing study, called 'Evolutionary Predictions in Colorectal Cancer' (EPICC), was set up. The ultimate goal of this study was to obtain measurements that would allow inference and prediction of evolutionary dynamics in individual CRCs. The secondary aims of the project were i) to study the relationship of genomic and epigenetic intra-tumour heterogeneity, ii) to identify somatic epigenetic alterations with a potential role in CRC development, iii) and characterise the prevalence of subclonal driver alterations in this disease.

4.1.1 Clonal Architecture of Colorectal Cancers

For this, the property of colorectal adenomas and well to moderately differentiated carcinomas to contain tumour glands was exploited (Hamilton, Aaltonen, et al. 2000). These tumour glands are a structure that resembles colorectal crypts, small finger-like invaginations into the underlying tissue that normally form the epithelium of the colon (Humphries and Wright 2008). Normal crypts typically contain five to six stem cells located at the bottom of the crypt (Lopez-Garcia et al. 2010; Snippert et al. 2010; Baker et al. 2014). The stem cells continuously give rise to transient cells, which migrate to the top of the crypt and are shed after several days into the colon lumen (Wright, Alison, et al. 1984). As mentioned before, CRCs are composed of glands and are believed to contain the same structure (Merlos-Suárez et al. 2011). The clonal expansion of CRCs is thought to take place due to bifurcation or fission of crypts/glands (Garcia et al. 1999; Humphries et al. 2013; Baker et al. 2014; Bruens et al. 2017). Experiments have further shown that the lineages within a gland undergo frequent sweeps (Graham et al. 2011; Baker et al. 2014). This implies that all cells within a gland share a recent common ancestor and that they are only a few cell divisions apart. Taken together, these properties make glands the fundamental 'clonal unit' of CRCs.

While the full genomic profiling of individual cells would in principle be optimal, due to errors and artefacts arising from the necessary whole-genome amplification, the direct measurement of mutations in single-cells remains elusive (Leung et al. 2017). Alternative approaches use the ability of cells to copy their genome with high fidelity *in vitro*. This ability has been used in immortalised cell-lines (Meyer et al. 2015) or patient-derived organoids (Sato et al. 2011; Wetering et al. 2015; Blokzijl et al. 2016; Roerink et al. 2018). Still, these techniques are time consuming, expensive, and might introduce a selection bias for a subpopulation of cells. Further, the specific micro-environment required for the *in vitro* cultivation might not reflect the one encountered in the primary tissue and could consequently introduce artefacts in assays (e.g., gene expression, methylation or chromatin accessibility assays).

Conveniently, the tissue architecture and replication machinery of cells produces exactly what these *in vitro* methods do — i.e., the creation of a large number of genetically similar cells from a recent common ancestor — in the micro-environment of the analysed tumour *in vivo*. As each colorectal cancer gland contains between $2,000$ to $10,000$ individual cells (Siegmund et al. 2009a), a sufficient amount of genetic-material for various assays can be obtained from these. Indeed, the ability to conduct genomic profiling at essentially single clone resolution has been exploited in previous studies to elucidate clonal dynamics in normal colorectal epithelium (e.g., Yatabe, Tavaré, and Shibata 2001; Nicolas et al. 2007; Shibata 2009) as well as in cancer (e.g., Tsao, Grisham, and Nelson 1985; Tsao et al. 1998; Tsao et al. 2000; Siegmund et al. 2009b; Humphries et al. 2013; Sottoriva et al. 2015; Cross et al. 2018; Baker et al. 2019; Cross et al. 2020).

4.1.2 Multi-Omics Profiling

For the EPICC project a novel multi-omics profiling method was developed[1], which combines the single-crypt isolation methods described by Martinez et al. (2018) with three genomic assays i) mutation profiling using WGS, expression profiling with RNA sequencing

[1] This work was primarily done by Inmaculada Spiteri.

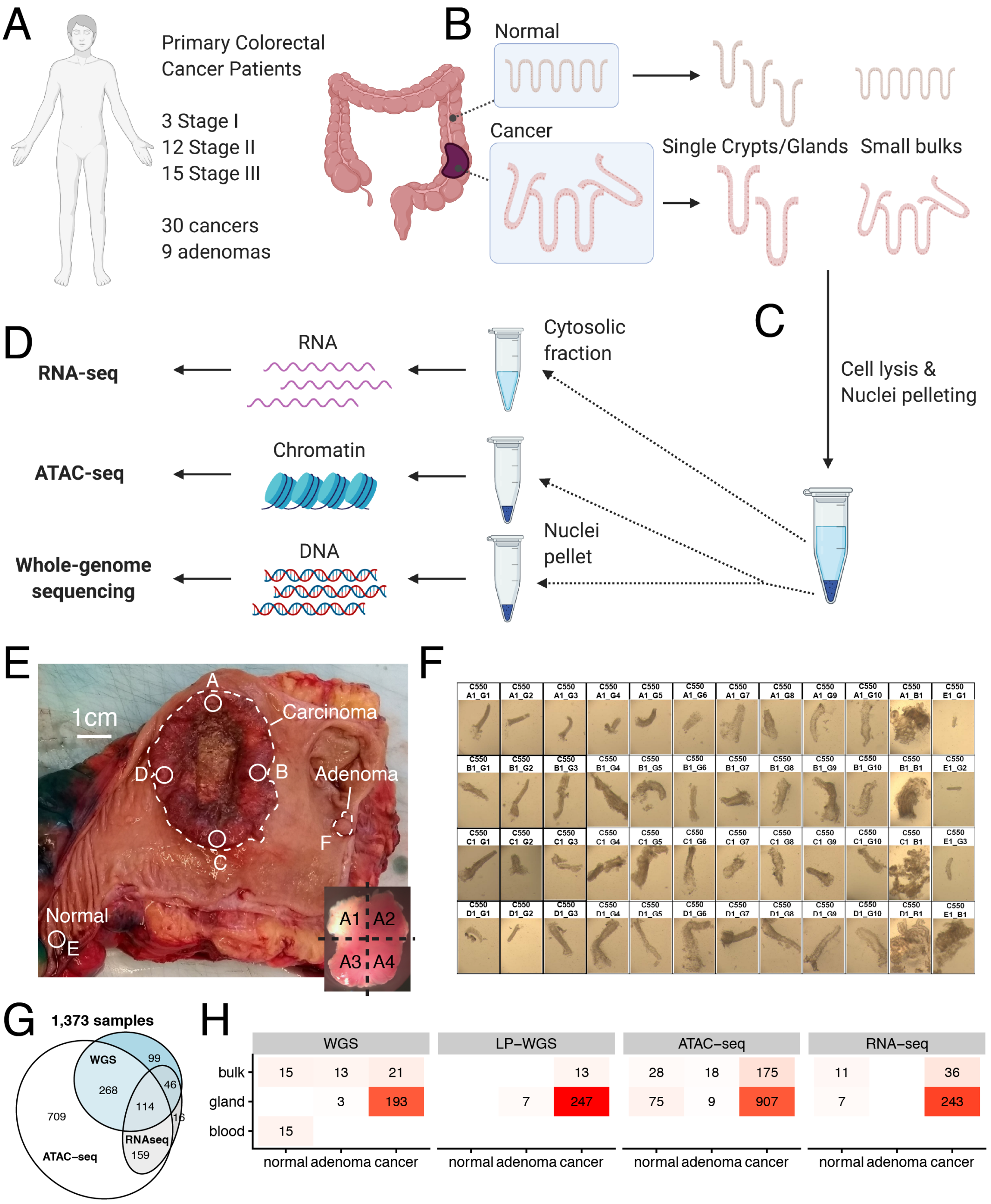


Figure 4.1: Sample collection scheme and sample numbers. A) A total of 30 cancers and 9 adenomas obtained from colectomy specimens from 30 patients were obtained. B) For each case normal reference and cancer glands and bulks were taken from different regions. C) Cells of each sample were lysed and split into a cytosolic and nucleic fraction. D) RNA-seq data were generated from the cytosolic fractions. ATAC-seq and WGS data from the nucleic fraction. E) A macroscopic image of a tumour with the layout of the four tumour regions (A-D) An adjacent normal sample region (E) An adenoma (F). F) Microscopic images of individual glands obtained from the four regions. G) The overlap of the different assays. H) The total number of samples obtained for each data and sample type.

(RNA-seq) (Schuierer 2017) and chromatin accessibility measurements using ATAC-seq (Buenrostro et al. 2013; Buenrostro et al. 2015).

The multi-omics profiling method was applied to a total of 30 stage I–III primary colorectal carcinoma and nine concomitant adenomas from a total of 30 patients that underwent a colectomy at the University College London Hospital (Figure 4.1A and Table S.1).

From these primary specimens, individual crypts as well as tiny bulk samples — in the following referred to as 'minibulks' — were obtained from normal and tumour tissue (Figure 4.1B). The cells obtained from each of these tissue samples were then lysed and, split into a cytosolic fraction and nucleic fraction through centrifugation (Figure 4.1C). ATAC-seq and WGS libraries were then created from the nuclei fraction and RNA-seq libraries were created from the cytosolic fraction[2] (Figure 4.1D).

Some of the generated libraries were then sequenced on an Illumina sequencer and the total number of samples for which sequencing data were generated are shown in Figure 4.1H. Overall WGS data at sequencing coverage of $\approx 30\times$ were obtained for a total of 214 samples (tissues pieces and single-glands) from carcinomas and for 16 samples from adenomas. A large number of additional LP-WGS sequencing was conducted on 260 and 7 samples obtained from cancers and adenomas respectively. For most of the tissue samples — i.e., 1082 obtained from carcinomas, 27 from adenomas and 103 healthy normal tissue — ATAC-seq data were generated. Additionally, RNA-seq was obtained for a subset of samples[3]. The overlap of the different measurements is summarised in Figure 4.1G. As shown here, matched WGS (including LP-WGS) and ATAC-seq was available for 268 samples. In a total of 114 samples, all three measurements (i.e., WGS, ATAC-seq and RNA-seq) were available.

In this chapter, I will present an analysis of the WGS and ATAC-seq datasets generated as part of this project. Using these, I will provide a comprehensive overview of the subclonal architecture of the analysed CRCs. Furthermore, I will present an analysis of the somatic chromatin accessibility profiles and identify a number of functional recurrent focal alterations of the chromatin accessibility as well as the general deregulation of TF activity as putative non-genetic drivers of CRCs.

Sample barcodes Each obtained sample was given a unique identifier that allowed to identify its specific properties. The barcodes are a series of alphanumeric identifiers like the

[2] Tissue collection and sample preparation were done by Inmaculada Spiteri and Chris Kimberley.

[3] These RNA-seq data were analysed by Jacob Househam.

following: 'EPICC_C501_A1_B1_D1'. The elements of these labels separated by '_' consist of i) the project code (i.e., EPICC), ii) the patient identifier (e.g., C501), iii) the region identifier (i.e., A-G followed by a number), iv) the sample type (i.e., B — bulk or G — gland followed by a number), v) the analyte type (i.e., D — WGS, C — ATAC-seq, R — RNA-seq or L — LP-WGS) followed by a number. Especially the case identifier (e.g., C516) will frequently be referenced in the following.

4.2 Methods

ATAC-seq and WGS data were generated on an Illumina NextSeq 500 (ATAC-seq of C516) and an Illumina NovaSeq sequencer. The primary analysis pipeline for WGS, LP-WGS and ATAC-seq sequencing datasets described below was implemented using the Snakemake workflow engine (Köster and Rahmann 2012).

4.2.1 Alignment

4.2.1.1 WGS

Contaminating adapter sequences were first removed with *Skewer* version 0.2.2 (Jiang et al. 2014) using the adapter sequences 'AGATCGGAAGAGC' and 'ACGCTCTTCCGATCT', a maximum error rate of 0.1, minimum mean quality value of 10, and a minimum read length of 35 bp after trimming using options '-l 35 -r 0.1 -Q 10 -n'. The trimmed and filtered reads from each sequencing run and library were separately aligned to the GRCh38 reference assembly of the human genome (Schneider et al. 2016) with version 0.7.17 of the BWA-MEM algorithm (Li 2013).

Using the GATK version 4.1.4.1 and following GATKs best practices (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013) reads were then sorted by coordinates with *GATK SortSam*, merged across independent sequencing runs, and libraries generated from the same tissue and duplicated reads marked using *GATK MarkDuplicates*. The structure of the final BAM files was verified using *GATK ValidateSamFile*.

4.2.1.2 ATAC-seq

Adapter sequences were removed with *Skewer* version 0.2.2 (Jiang et al. 2014) using the full-length adapter sequences (see Table 4.1) with the option '-m any' set.

The reads of each sequencing run and library were aligned to the GRCh38 reference genome using *Bowtie2* (Langmead and Salzberg 2012; Langmead et al. 2019, version 2.3.4.3) with the options '–very-sensitive -X 2000' set. After sorting the reads with *sam-*

Table 4.1: ATAC-seq adapter sequences.

Adapter	Sequence
Pair 1	CTGTCTCTTATACACATCTCCGAGCCCACGAGACNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
Pair 2	CTGTCTCTTATACACATCTGACGCTGCCGACGANNNNGTGTAGATCTCGGTGGTCGCCGTATCATT

tools sort version 1.9 (Li et al. 2009), reads mapping to non-canonical chromosomes and mitochondria (i.e., chrM) were removed (*GATK PrintReads* followed by *GATK RevertSam* and *GATK SortSam*). After merging the independent libraries of each sample, duplicated reads were removed using *GATK MarkDuplicates*. Likewise, all reads mapping to multiple locations (i.e., multi-mappers) were excluded using *samtools view*. The final bam files were validated with *GATK ValidateSamFile*.

4.2.2 Detection of Germline Variants

The GATK HaplotypeCaller (Poplin et al. 2018) was used to identify germline variants from the reference normal samples in each patient (buffycoats or adjacent normal tissue). In short, candidate germline variants were identified using the HaplotypeCaller with known germline variant annotations from the build 146 of the dbSNP database (Sherry, Ward, and Sirotkin 1999; Sherry et al. 2001). This was conducted separately for each chromosome and VCF files were merged later using GATKs MergeVcfs. Following the methods by Poplin et al. (2018) variant recalibration data were then calculated for single-nucleotide variants (SNVs) using *GATK VariantRecalibrator* with the options:

```
--resource hapmap,known=false,training=true,truth=true,prior=15.0:hapmap_3.3.hg38.vcf.gz
--resource omni,known=false,training=true,truth=true,prior=12.0:1000G_omni2.5.hg38.vcf.gz
--resource 1000G,known=false,training=true,truth=false,\
prior=10.0:1000G_phase1.snps.high_confidence.hg38.vcf.gz}
--resource dbsnp,known=true,training=false,truth=false,prior=2.0:dbsnp_146.hg38.vcf.gz
--max-gaussians 6 -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0
-an DP -an QD -an FS -an SOR -an MQ -an MQRankSum -an ReadPosRankSum -mode SNP
```

and for InDels with the options

```
--resource mills,known=false,training=true,truth=true,\
prior=12.0:Mills_and_1000G_gold_standard.indels.hg38.vcf.gz
--resource dbsnp,known=true,training=false,truth=false,prior=2.0:dbsnp_146.hg38.vcf.gz
-mode INDEL --max-gaussians 4 -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0
-an DP -an QD -an FS -an SOR -an MQRankSum -an ReadPosRankSum
```

following suggestions by Frazer et al. (2007), Auton et al. (2015), Sherry et al. (2001), and Mills et al. (2006). Recalibration data were then applied to the VCF files using *GATK*

ApplyVQSR with the options '-mode SNP -ts-filter-level 99.0' and '-mode INDEL -ts-filter-level 99.0' respectively. Only germline variants with the filter flag 'PASS' were retained.

4.2.3 Verification of Sample-Patient Matches

It was verified for each sample that they matched the expected patient identity by using the germline variants identified in normal tissue samples of patients. Reads of each read-group were extracted from bam files with *'samtools view'* using options '-bh {input_bam} -r {read_group_id}' and then used *GATK CheckFingerprint* tool to extract statistics on sample-patient matches (Javed et al. 2020).

For all but a couple of high-purity samples with extensive loss of heterozygosity, this analysis confirmed that the samples were obtained from the expected patient. For the latter group, copy-number profiles were manually reviewed to confirm that they matched the remaining samples.

4.2.4 Copy-Number Analysis

4.2.4.1 WGS

For the analysis of CNA from deep WGS samples, coverage of genomic loci relative to matched normal tissue samples (buffycoats or adjacent normal) were extracted and binned in non-overlapping windows of 10^6 bp B-allele frequencies of germline mutations determined as outlined above for each patient were added to these binned files.

Joined segmentation on all samples obtained from a given tumour was performed using the heterozygous B-allele frequencies to determine a set of breakpoints for subsequent analysis. Biases introduced by differences in guanine-cytosine content (GC-content) as well as mappability were corrected, and piecewise constant curves fitted to data from all samples using the multipcf function from the *copynumber* package version 1.22.0 for R (Nilsen et al. 2012).

Using the per-patient set of breakpoints, binned depth-ratio and B-allele frequency data, the *sequenza* algorithm (version 2.1.2) was used to determine allele-specific copy-numbers, ploidy (Ψ), and purity (ρ) estimates from these data (Favero et al. 2015). The initial parameter space searched was restricted to $0.1 \leq \rho \leq 1$ and $1 \leq \Psi \leq 7$. After reviewing the results, several samples with unrealistic fits (e.g., extremely variable ploidy values across samples) were identified. For these samples, alternative solutions consistent with the other samples of a patient and somatic variant calls were manually selected.

4.2.4.2 LP-WGS

Reads from LP-WGS samples were extracted from bam files with methods from *QDNAseq* (Scheinin et al. 2014), binned in windows of 500 kb across the autosomes and converted to log2-ratios. The log2-ratios were normalised according to the workflow described by Scheinin et al. (2014) apart from the outlier smoothing steps. Next, log2-ratios were normalised by subtraction of the median log2 ratio in a given sample, segmented with the *multipcf* method from the *copynumber* package for R (Nilsen et al. 2012) using $\gamma = 10$ and summarised by the average log2-ratio across identified segments.

For the estimation of absolute copy-number values from the average log2-ratios of each segment a tool, similar to ASCAT (Loo et al. 2010), which was developed by George C. Cresswell and is briefly described in the following was used[4]. As described by Loo et al. (2010) the expected log2-ratio r_i of a genomic locus i present at a copy-number of c_i in a sample with purity ρ is given by

$$r_i = \gamma \, log_2 \left(\frac{2 - 2\rho + \rho c_i}{2 - 2\rho + \rho \Psi_t} \right),$$

where Ψ_t is the average copy-number across the entire genome in the tumour population (i.e., the tumour ploidy) and γ a correction factor accounting for dampening of the signal resulting from a specific assay. For sequencing data $\gamma = 1$ is used. The average ploidy was calculated from the WGS copy-number analysis described and used as a plug-in estimate for Ψ_t.

Expected log2-ratios for a range of purity values $\{\rho \in \mathbb{R} \mid 0.1 \leq \rho \leq 1\}$ were calculated and compared to the observed log2-ratios using the L2-norm as described in the ASCAT paper (Loo et al. 2010) to identify a value for ρ minimising this distance. The closest absolute copy number $\{c_i \in \mathbb{N} \mid 0 \leq c_i \leq 20\}$ of each segment for this value of ρ was then calculated for each segment.

4.2.4.3 ATAC-seq

For ATAC-seq data, reads in the vicinity of peaks (open-chromatin) and those in regions of closed-chromatin were analysed separately. The former was defined as reads mapping to intervals of the filtered, extended (100 bp) and merged (distance of 2000 bp) peak set (see above). The latter were defined as intervals in a distance of 1000 bp from the open region with a minimal size of 10000 bp. Coverage for all intervals relative to normal colorectal

[4] R functions for the estimation of allele-specific copy-numbers from the log2-ratios were provided by George C. Cresswell.

ATAC-seq datasets was determined, normalised for GC-content as well as genomic repeats (Talevich et al. 2016), averaged in windows of 10^6 bp, and segmented with a circular binary segmentation algorithm (Olshen et al. 2004; Seshan, Olshen, et al. 2015).

4.2.5 Somatic Variant Detection

4.2.5.1 Calling

Somatic mutations were called for each tumour sample separately against matched blood-derived or adjacent normal tissue samples with *Mutect2* version 4.1.4.1 using the options '–af-of-alleles-not-in-resource 0.0000025 –germline-resource af-only-gnomad.hg38.vcf.gz' (Cibulskis et al. 2013; Poplin et al. 2018). Variants detected in any tumour sample marked PASS, coverage AD 10 in both normal and tumour, ≥ 3 variant reads in the tumour, 0 variant reads in the normal, reference genotype in normal, and non-reference genotype in cancer) were jointly summarised with *Platypus* version 0.8.1.1 (Rimmer et al. 2014) in all samples of a patient.

This set of joined variant calls was then filtered to keep high-quality variants with flags 'PASS', 'alleleBias', 'QD' or 'Q20', in canonical chromosomes (i.e., not in a decoy), a minimum number of reads $NR \geq 5$ in all samples, a genotyping quality $GQ \geq 10$ in all samples, a reference genotype (i.e., 0/0) in the matched normal reference, and a non-reference genotype (i.e., 0/1 or 1/1) in at least one tumour sample. Due to concerns to filter out important driver mutations, a second set of variant calls to which the second filtering step was not applied was generated to identify mutations in known driver genes and the dN/dS analysis (see details below).

4.2.5.2 Annotation

Somatic variants were annotated and candidate driver genes of colorectal cancers reported by TCGA (Muzny et al. 2012), Cross et al. (2018) and IntOGen (Gonzalez-Perez et al. 2013; Martínez-Jiménez et al. 2020) as well as pan-cancer driver genes reported by Martincorena et al. (2017) and Tarabichi et al. (2018) filtered with the Variant Effect Predictor toolkit version 93.2 (McLaren et al. 2016).

4.2.5.3 MSI Status Detection

The identification of MSI colorectal cancers was performed with version v0.2 of the *MSIsensor* C++ program developed by Niu et al. (2014). The position of microsatellites sites was first determined by applying the *msisensor scan* command to the GRCh38 reference assembly. These were then subset to the first chromosome to speed up the subsequent analysis. In a second step, the fraction of mutated microsatellites in each sample were determined with the *msisensor msi* command using default options. Generally, in known MSI affected cases, more than 30% of microsatellites were mutated and this was used as a critical value to classify cases as with microsatellite stability (MSS) and MSI.

4.2.5.4 Extraction of Reads Supporting Variants

Using the VCF files from both somatic and germline variant calling (see above), the number of reads supporting the reference and alternate alleles as well as the total number of reads covering the sites from WGS, LP-WGS and ATAC-seq samples were extracted using python and the version 0.15.2 of pysam library with samtools version 1.9 (Li et al. 2009; Andreas Heger et al. 2021).

4.2.5.5 dN/dS Analysis

The dN/dS analysis was conducted using the *dndscv* package for R (Martincorena et al. 2017). Per-patient variant calls were obtained from the VCF files (Obenchain et al. 2014) and lifted to the hg19 reference genome using the *rtracklayer* package for R (Lawrence, Gentleman, and Carey 2009). Variants were split into clonal (i.e., present in all samples) and subclonal mutations (i.e., present in a subset of samples) present in the cancers as well as a set of mutations present in any of the adenomas. Patients were further split into MSI and MSS cases and the *dndscv* model was fitted for each of the four sets (MSI/MSS & clonal/subclonal) separately. For this default parameters apart from deactivated removal of cases due to the number of variants were used. Global dN/dS values for a set of 167 chromatin modifier genes and the previously described CRC driver genes were obtained.

4.2.5.6 Mutational Signature Analysis

The analysis of mutational signatures[5] was conducted with the *deconstructSigs* package for R (Rosenthal 2016) based on the mutational signatures reported in version 2 of the COSMIC database (Tate et al. 2019). For a given set of mutations, the trinucleotide contexts were obtained with methods from the *GRanges* (Lawrence et al. 2013a) and *BSgenome* (Pagès 2021) package for R. All substitutions were annotated with their specific context (e.g., A[C>A]A) and those with a central A or G base were replaced by their reverse complements. The number of mutation types were tabulated as vectors

[5] The results presented here are from an initial analysis of the mutational signature activity in normal colorectal crypts as well as colorectal cancer glands. For the purpose of this analysis, the activity of previously identified mutational signatures with an activity in CRC reported in the COSMIC database was assessed. A de novo determination and analysis of mutational signatures using *SparseSignatures* (Lal et al. 2021) was conducted by Daniele Ramazzotti as part of the EPICC project, which corroborated the results reported here.

$\mathbf{m} = [m_{[}\mathrm{A[C{>}A]A}, ..., m_{\mathrm{T[T>G]T}}]$, where m_i are the number of mutations of type i. Selected mutational signatures $\mathbf{s_i} = [s_{\mathrm{A[C>A]A}}, ..., s_{\mathrm{T[T>G]T}}]$, with $\sum_j \mathbf{s}_{i,j} = 1$ were then obtained from the COSMIC database and arranged as a matrix S with the signatures as column vectors. Next, *deconstructSigs* was used to estimate vectors of exposures to each selected signature $\mathbf{e} = [e_1, e_2, ..., e_n]$ of each sample. The default option of *deconstructSigs* to remove signatures with a relative contribution of $< 6\%$ was set deactivated by setting the corresponding option 'signature.cutoff=1' to 0% to ensure that all selected signatures were considered. Residuals were calculated as $\mathbf{r} = \mathbf{m} - \mathbf{S}\,\mathbf{e}$ for each sample. The results of the analysis were plotted with methods from the *deconstructSigs* package.

In order to determine the stability of the obtained results — this is especially important in cases where few variant loci are used for the analysis — a non-parametric bootstrap was used. For this, a number of mutations equal to the actual number in the sample were sampled with replacement from the entire set and analysed as described above. The procedure was repeated 100 times and the average exposure was calculated as the arithmetic means across all replicates. The results were plotted with basic R methods (R Core Team 2020) as shown in Figure S.159 (page 337). The panels in Figure S.159 (page 337) show from top to bottom i) the mutation spectrum $\mathbf{m}$, ii) the estimated mutation spectrum $\mathbf{S}\,\mathbf{e}$, iii) the exposure $\mathbf{e}$ and iv) the residuals $\mathbf{r}$ across replicates, with the error bars indicating the 95% confidence interval estimated from the bootstrap.

4.2.6 ATAC-seq Peak Analysis

4.2.6.1 Peak Calling

Extraction of cut-sites To detect peaks, bed files of ATAC-seq cut-sites were extracted. These were obtained from the bam files by first sorting the reads by their read names using 'samtools sort -n {bam}', isolating all proper reads pairs (i.e., reads mapped to the same chromosome and with correct read orientation) using 'samtools view -bf 0x2', and finally converting these paired reads to the bed format using 'bedtools bamtobed -bedpe -mate1 -i {input}' (Li et al. 2009; Quinlan and Hall 2010).

Equivalent to Buenrostro et al. (2013) the start site of reads aligned to the forward strand were shifted by four bases and those aligned to the reverse strand by five bases to obtain positions of cut-sites during transposition.

ATAC-seq reads spanning nucleosomes have an insertion size periodicity of multiples of $\approx 200\,\mathrm{bp}$, and reads in regions of open-chromatin have insertion sizes smaller than $100\,\mathrm{bp}$

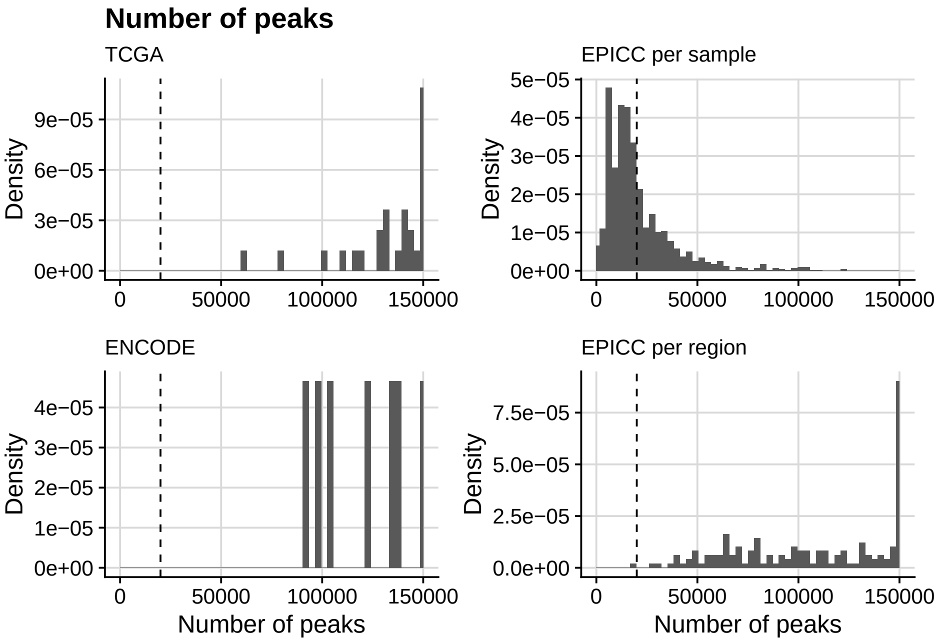


Figure 4.2: Number of peaks called by *MACS2* in samples from the TCGA, ENCODE and EPICC cohort at an FDR of 0.1%.

(Buenrostro et al. 2013). For this reason and in line with previous studies, the ATAC-seq reads were split into a set of nucleosome-free reads (insertion size $\leq$ 100) and nucleosome associated reads (180 $\leq$ insertion size $\leq$ 620).

Peak detection Likely due to the low library complexity of the single-gland ATAC-seq libraries, exhaustive identification of regions of open chromatin (i.e., peaks) with data from individual glands at a high level of confidence was problematic (see Figure 4.2). While an imperfect substitute, when pooling reads from single-glands obtained from a single region of a tumour, statistical power was sufficient to call a number of peaks similar to high-quality large bulk samples from the TCGA and ENCODE cohort (Figure 4.2).

For this reason, peaks were called per tumour region using *MACS2* (Zhang et al. 2008, version 2.12) with ‘macs2 callpeak -f BED -g hs –shift -75 –extsize 150 –nomodel –call-summits –keep-dup all -p 0.01’ with the concatenated and sorted bed read files of nucleosome-free cut-sites of all samples as input. A set of normal peaks (pan-patient) was called separately on the concatenated normal sample bed files (i.e., region E) and per adenoma peak calls using all adenoma bulk samples as input.

Filtering and concatenation of peaks Strict filtering of per-region peak calls extended by 250 bp was applied. Only variants with a minimum q-value of $\leq$ 0.1%, enrichment of $\geq$ 4.0 and a maximum number of peaks 20,000 were kept. Iterative merging, equivalent to that

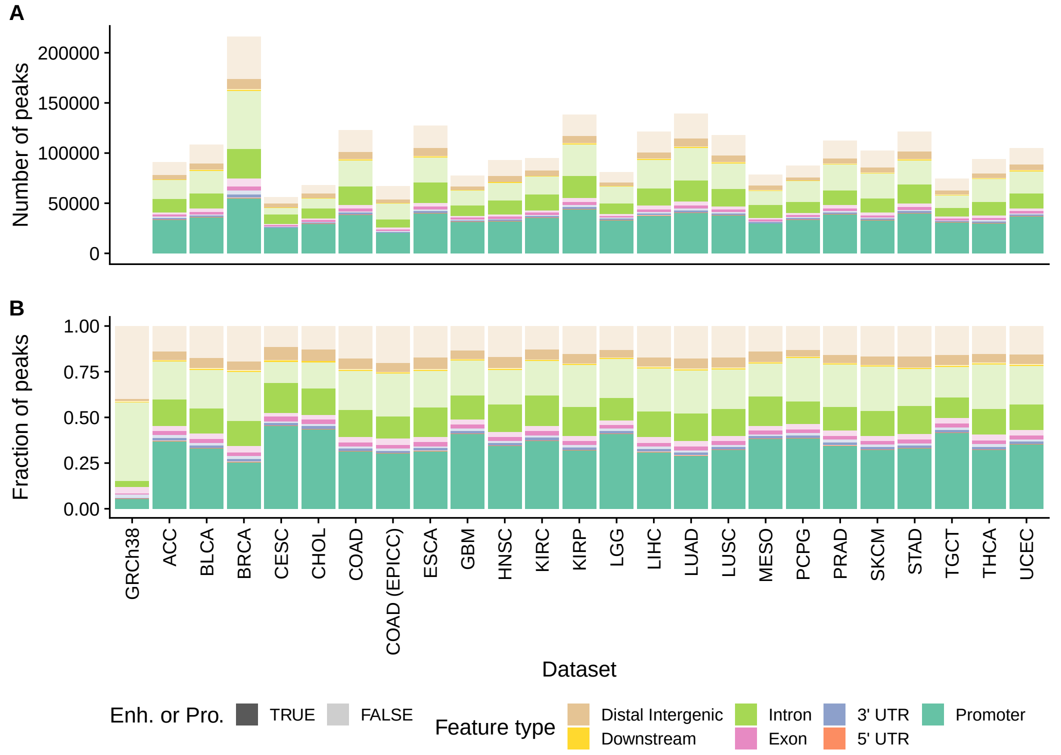


Figure 4.3: Distribution of filtered peak calls in the genome. Shown the peak call sets from various cancer types in the TCGA cohort and the merged peak calls from samples in the EPICC cohort labelled COAD (EPICC). A) Total number of peaks overlapping with the corresponding genomic features (see legend at the bottom). B) Fraction of peaks overlapping with the genomic features. The first column, labelled 'GRCh38', shows the distribution of features within the genome. Enhancer elements reported in the GenHancer database and Promoters are shown in dark, those without overlap in light colours.

used by Corces et al. (2018), was then performed on per-region peak calls of patients (per-tumour peaks set) as well as across all cancer samples, and pan-patient normal peak calls (pan-patient peak set). This procedure resulted in a total of $N = 343,240$ peaks, of which $N = 67,215$ peaks were called in ≥ 2 tumour regions or in the panel of normal samples.

The *ChIPseeker* package for R (Yu, Wang, and He 2015, version 1.24.0) was used to annotate peaks based on their genomic location. For peaks that were not proximal to known promoter regions (± 1000 bp), overlaps with known enhancer elements reported in the double-elite annotations of the GeneHancer database (Fishilevich et al. 2017) were determined. The overlaps for both of these with the pan-patient peak set is shown in Figure 4.3 in comparison to the distribution of annotations from Corces et al. (2018) and the general distribution of these features in the genome.

Extraction of reads in peaks For the final set of peaks, the number of shifted cut-sites overlapping each peak were counted separately for individual samples using *bedtools* (Quinlan and Hall 2010) as follows: 'bedtools coverage -a bed_peaks -b bed_cut_sites -split -counts -sorted'.

Purity estimation for ATAC-seq samples To obtain purity estimates for the ATAC-seq samples of the study, clonal variants identified as part of the WGS sequencing (i.e., those present in all samples from cancer) were used. It was assumed that these variants would also be clonal in the ATAC-seq samples obtained from the same regions of cases and were hence used to obtain purity estimation of ATAC-seq samples. First variants in intervals with identical copy-number states (i.e., A/B states) in all WGS samples were identified. From these variants within regions of open chromatin (i.e., peaks) and copy-number values > 4 were then excluded. For each variant i, allele copy-number values c_i and mutation multiplicity m_i were estimated using the WGS data. An example of these estimates in one case, C539, are shown in Figure S.27 (page 274).

For a mutation at site i covered by $n_{s,i}$ reads in sample s the number of reads k_i containing the alternate allele is expected to follow a binomial distribution with the likelihood

$$B(k_i|p_{s,i},n_{s,i}) = \binom{n_{s,i}}{k_i} p_{s,i}^{k_i}(1-p_{s,i})^{n_{s,i}-k_i},$$

where the expected success probability $p_{s,i}$ is a function of the samples purity a_s, the number of mutated alleles in the tumour cells $m_{s,i}$, the total copy-number of the mutated site in the tumour cells $c_{s,i}$ and the copy-number in contaminating normal cells $c_n = 2$ is

$$p_{s,i} = \frac{\rho_s m_{s,i}}{\rho_s c_{s,i} + (1-\rho_s)c_n} = \frac{\rho_s m_{s,i}}{\rho_s c_{s,i} + 2 - 2\rho_s}.$$

The negative-log-likelihood across N mutated sites is then

$$l(\rho_s) = \sum_{i=0}^{N} -log\left(B(k_i|p_{s,i},n_{s,i})\right),$$

which was minimised to obtain a ML estimate of the sample purity ρ_s:

$$\hat{\rho}_s = \underset{\rho_s:\ 0\leq\rho_s\leq 1}{arg\,min}\ l(\rho_s).$$

Identification of recurrently altered peaks across patients Next, events for which data indicated general chromatin structure changes within a given cancer were identified. For this, samples with purity $\rho \geq 0.4$ were determined, data merged, and the counts of reads per peak were obtained as described above. Based on the assumption that peaks proximal ($\leq$ 1000 bp) to a transcription start site (i.e., promoters) and those more distant to a transcription-factor start site (i.e., putative enhancers) might have a different amount of dispersion, peaks were split into these two groups.

An overdispersed Poisson model was then split to each of these datasets using *edgeR* version 3.30.3 (Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012). Per sample set normalisation factors were calculated using the TMMwsp method (Robinson and Oshlack 2010; Robinson, McCarthy, and Smyth 2010) and a global dispersion estimate were estimated across all 'cancer' samples.

In order to identify somatic chromatin accessibility alterations, each 'cancer pool' of pure glands was compared against a large 'pool-of-normal' composed of normal tissue ATAC-seq samples. These were independently filtered for events with a minimum counts per million reads mapped (CPM) of 15 in the tumour or normal and a minimum fold change of 2. From the remaining tests, peaks significantly altered at a level of $p \leq 0.01$ in at least $5/25$ (i.e., 20%) of cases were identified.

Identification of associated gene expression changes A subset of $27,731$ peaks that were either adjacent to a known transcription-factor start site (TSS) of a gene (Team and Maintainer 2019; Haeussler et al. 2019) or overlapped a previously characterised enhancer element described in the GenHancer database (Fishilevich et al. 2017) were identified. Of these, $944/27731 (\approx 3.40\%)$ were recurrently altered. To test whether any of these alterations were associated with changes in gene expression, the *results* method from *DESeq2* (Love, Huber, and Anders 2014) was used to compare coefficients of the fitted beta-binomial regression model (design: $\sim Patient$, with all normal samples as 'Normal') with the *contrast* argument being a list of vectors containing the significant and non-significant patient sets.

For promoters, a one-tailed hypothesis test was conducted by setting the *altHypothesis* argument to 'less' (for closed peaks) or 'greater' (for opened peaks). For enhancers, a two-tailed hypothesis test was instead conducted on all associated genes by setting the *altHypothesis* argument to 'greaterAbs'. All p-values were adjusted for multiple hypothesis testing using the false-discovery rate (FDR) method (Benjamini and Hochberg 1995) and reported if the $FDR < 0.1\%$. For the visualisation of gene-expression values, the average gene expression was calculated on variance stabilised (log-transformed) CPM values across all samples from a given cancer or across all normal samples.

Identification of subclonal changes is recurrently altered peaks To explore putatively subclonal epigenetic events, a subset including recurrent events affecting known drivers of CRC, and the top 20 most recurrent events in each of the following four categories were selected: i) gained promoter, ii) lost promoter, iii) gained enhancer and iv) lost enhancer. To determine the significance of the spatial region whilst controlling for purity, a log-ratio test from *DESeq2* (Love, Huber, and Anders 2014) was used to compare the full model $\sim purity + region$ to a reduced model $\sim purity$. Samples from the same region of a patient were used as biological replicates. Events were considered to be putatively subclonal if the adjusted p-value was ≤ 0.05 and if the direction of log fold change from bulk analysis matched the expected change. In the case of gained events, subclonal events were filtered out if no peaks were called within 500 bp. Log ratio tests for the effects of region and purity were also performed with single parameter models. For visualisation of peaks, coverage per region was calculated 1kb upstream and 1kb downstream from the centre of the peak. Coverage was normalised per million reads in peaks and was plotted using functions from *GenomicRanges* (Lawrence et al. 2013a) and *Gviz* (Hahne and Ivanek 2016).

4.2.7 TF Binding Site Analysis

Binding site prediction The *motifmatchr* package for R (Schep 2020), a reimplementation of the C++ library *MOODS* (Korhonen et al. 2009; Pizzi, Rastas, and Ukkonen 2011), was used to identify binding sites for all human transcription factor (TF) motifs defined in a curated version of the CIS-BP database (Weirauch et al. 2014). The list of predicted binding sites was filtered using a minimum significance value of $p \leq 1 \times 10^{-6}$, followed by removal of binding sites in centromeric regions (Schneider et al. 2016) and non-autosomal (i.e., sex and non-canonical) chromosomes.

After this initial filtering, predicted binding sites were split into six distinct groups based on their distance to the next TSS (proximal: $d \leq 2000\,\text{bp}$, close: $2000\,\text{bp} < d \leq 10,000\,\text{bp}$, distal $d > 10,000\,\text{bp}$) and ii) whether they overlapped with a peak observed in the ATAC-seq data. For a number of TF, clustering of binding sites of the same TF motif in specific genomic regions was observed. For this reason, binding sites that were closer than $d \leq 1000\,\text{bp}$ to the next predicted binding site of the same TF were removed.

Extraction of signal values For each of the TF sets described above, the counts of insertions around the centre of the TF binding site ($\pm 1000\,\text{bp}$) as well as the insertion size of the read pair (i.e., the distance to the second nick) for each sample (Lawrence et al. 2013a) were tabulated. The insertion sizes (rows) were binned into intervals of 5 bp and divided by total count of reads with an equivalent size in the entire genome. After this, the background signal was estimated to be the average number of insertions $1000-750$ bp from the centre of TF binding site per insertion size and subtracted from the counts. The difference between these 'normalised and background corrected TF signals' in each sample and a pool of normal samples was calculated and integrated across the central region of the TF binding sites — specifically, insertion sizes in $[25,120]$ and distances in $[-100\,\text{bp}, 100\,\text{bp}]$ — as a summary statistic.

Regression analysis Linear regression was used to identify associations with purity estimates and in this context, signals were found to correlate with transcription-factor start site enrichment (TSSe) for both nucleosome-free and all reads. For this reason, an additional term was added to the regression model of each TF to correct for this effect: $signal \sim tsse * tsse\_nf + purity : patient$, where $tsse$ and $tsse\_nf$ are the TSSe differences of the sample and the pooled-normal samples. Each observation was weighted by the square root of the number of reads in the sample. A second linear model in which a region-specific effect of the purity: $signal \sim tsse * tsse\_nf + purity : tumour\_region$ was also fitted to the data. For both models, the significance of the 'purity' coefficient was determined and estimates of the coefficients were used as a patient specific summary for subsequent analysis.

Cluster analysis The analysis was focused on the 150 TF for which a significant association with the tumour cell content and TF signal was most frequently observed. With the aim to identify general patterns in these data, a clustering analysis was conducted using hierarchical clustering with Euclidean distance and complete linkage. This method identified three major groups of TFs. Each of these were analysed with String-DB (Szklarczyk et al. 2019) to identify significantly overrepresented pathways.

4.2.8 Reconstruction of Phylogenetic Trees

An MP method was used to reconstruct phylogenetic trees from the mutation data. This method requires the definition of a set of mutations that are present in a given sample. Due to the various purity values and copy-number states of mutations, estimates of CCFs were calculated from the VAF as

$$CCF_s = \frac{VAF_s(\rho_s\, c_i + 2 - 2\rho_s)}{\rho_s\, VAF_s\, m_i},$$

where m_i is the multiplicity of the variant assumed to be $m = 1$, ρ the purity of the sample s, and c_i the copy-number of the variant estimated from the WGS data as described above.

Variants with an estimated $CCF \geq 0.25$ were assumed to be present in a given sample. With this set of sequence data, MP trees were then inferred with the Parsimony Ratchet method (Nixon 1999) implemented in the *phangorn* package for R (Schliep 2011). A minimum of 100 iterations, a maximum of 10^6 iterations, and termination of the ratchet after 100 rounds without improvement were used.

4.3 Results

4.3.1 Sample Purity and Coverage

All sequencing data from all obtained samples (see Figure 4.1H) were aligned as described in the Method section above. CNA profiles of samples were manually reviewed and curated where necessary. For this curation, CNA fits of other samples from the same patient and the VAF distribution of SNVs were taken into account. The latter was especially helpful for the identification of nearly tetraploid tumours (e.g., Figure S.25, page 273). After this, somatic mutations were identified in all samples as outlined above.

Identification of normal crypts Initial inspection of the called SNVs revealed the presence of a small number of samples that showed private mutations at a high VAF and did not contain any of the variants present in the majority of other samples (e.g., A1_G3 and A1_G6 in Figure S.26, page 274). The analysis of their CNA profiles showed that these samples were diploid and an analysis of their mutational signatures revealed a strikingly different signature profile. Instead of MMR associated signatures in cases with MSI, these samples primarily showed the presence of S1, associated with ageing and S5, a general background process. These observations suggested that these spurious 'odd samples' were most likely healthy normal crypts that were interspersed with or at least closely adjacent to the tumour. A manual review of the obtained H&E slides confirmed this initial suspicion. Samples that were suspected to be normal samples were removed from the subsequent analysis of cancer samples, but below, a separate analysis of these healthy normal crypts will be presented.

Sample coverage After the exclusion of normal crypts, the coverage of all samples was determined. The archived median coverage of each sample type is summarised in Figure 4.4 and split by cases in Figure S.30 (page 276).

As shown in these two plots, the target coverage of ≥ 30 for deep WGS was archived for the majority (89.1%) of tumour samples (adenoma & cancer), with the median coverage being 35.0×. Later conducted LP-WGS archived a relatively variable coverage with a me-

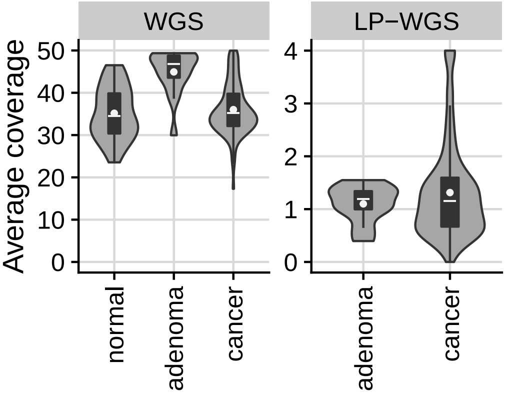


Figure 4.4: Mean coverage of whole-genome (WGS) and low-pass WGS (LP-WGS) samples from the EPICC cohort.

dian of 1.15× in samples from cancers. More than 86.9% of LP-WGS samples had average coverage of 0.5×.

WGS sample purities As part of the CNA analysis, purity estimates of each WGS and LP-WGS sample were obtained. As shown in Figure 4.5A, purity values depended on the type of sample they were obtained from. Generally, single-gland WGS and LP-WGS samples had a median purity of 81%, with a subset of these showing much lower values. Bulk WGS samples had in contrast significantly lower (median of 73%) and more variable purities. This is overall consistent with the expectation of single CRC glands being composed of a small set of genetically closely related tumour cells.

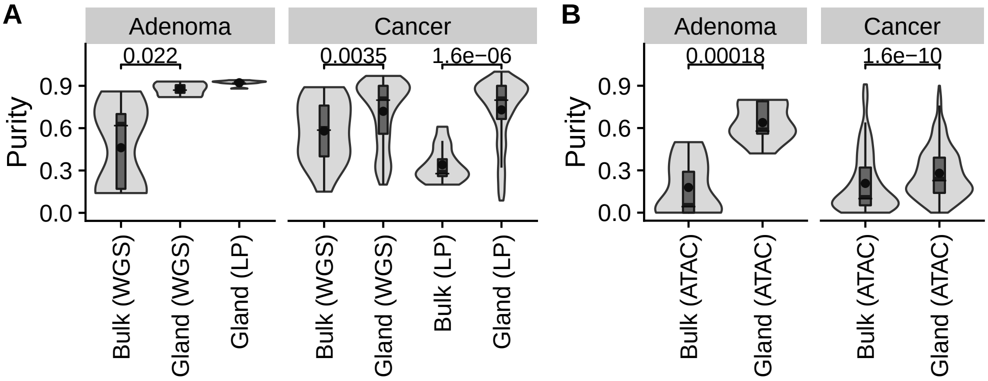


Figure 4.5: WGS and ATAC-seq purity estimates in the EPICC cohort. A) Distribution of estimated purities of deeply whole-genome sequenced (WGS) and low-pass whole-genome sequenced samples (LP). B) ML purity estimates for ATAC-seq samples (ATAC).

Some ploidy states can be hard to distinguish from each other (e.g., fully diploid vs tetraploid). In these cases, a higher ploidy can often fit the data if the purity of the sample is reduced. No association of purity and ploidy values was identified, thus confirming the overall correctness of the estimates ploidy values (Figure S.28, page 274).

The purity estimates obtained from the analysis of LP-WGS sequenced samples were confirmed by purity estimates derived from clonal somatic SNVs (see methods section for details). A scatter plot showing the correlation of CNA and SNV derived purity estimates can be found in Figure S.32A (page 277).

ATAC-seq sample purities Estimates of ATAC-seq sample purities were obtained using a simple ML method based on the information of clonal SNVs in the samples (see methods for details). As shown in Figure 4.5 these purity estimates were, similar to WGS data, significantly higher in single-glands compared to bulk samples. Oddly, purity estimates were much lower than those of WGS samples. This was also the case for samples in which matched WGS and ATAC-seq samples were available (Figure S.31, page 276). Still, a significant correlation between both measurements did exist ($r = 0.551$, $p \leq 10^{-8}$), suggesting that the estimates obtained for ATAC-seq samples might indeed be correct.

To rule out the possibility that the ML method itself produced biased estimates, it was also applied to all deep WGS samples. This approach confirmed that the estimates were accurate for the WGS samples themself (Figure S.32B, page 277) and the matched LP-WGS samples (Figure S.32C, page 277), hence ruling out issues with the method per se. Further, the analysis of the local chromatin accessibility measurements obtained from the ATAC-seq data (i.e., using principle-component analysis) and the analysis of CNA obtained using the ATAC-seq ‘background signal’ confirmed this low ‘apparent purity’.

While the exact reason for this lower ATAC-seq purity is unclear, the most likely explanation appears to be a bias in the speed of tissue/chromatin degradation following ischemia or a difference in the nuclear stability between tumour and normal cells.

In the following, the ML estimates of ATAC-seq samples will be used to identify samples with low amounts of tumour-associated signals.

4.3.2 Analysis of SNVs

4.3.2.1 MSI Status Detection

I next assessed the somatic mutations identified through WGS sequencing. In $6/30$ patients (C516, C518, C536, C548, C552, and C562), DNA mismatch repair deficiency (MMRd) was identified clinically through immunohistochemical staining and reported in the pathology report. For the case C516 loss of MSH2 expression and for the remaining four cases loss of MLH1, PMS2 expression was reported. MMRd arises from defects of a group of enzymes that orchestrate the repair of mismatched nucleotides in the genome (Peltomäki 2003). Loss of these proteins, MLH1, PMS2, MSH2, and MSH6 in specific, is frequently tested as part of routine cancer diagnostics since MMRd is associated with a more favourable diagnosis (Popat, Hubner, and Houlston 2005) and Lynch syndrome, a hereditary condition causing susceptibility to some cancer types (Evrard et al. 2019).

The separate evaluation of the number of unstable microsatellites with *msisensor* applied to the WGS data showed a variable degree of MSI in the cohort (Figure 4.6A). As expected, cases with clinically identified MMRd showed a clear increase in the fraction of mutated microsatellites (i.e., $\geq 30\%$ of all microsatellites). This value is consistent with previous studies (Dietmaier et al. 1997; Jass, Young, and Leggett 2001). While no evidence of MSI was identified in case C562, this was likely due to the low purity of the obtained samples.

4.3.2.2 Clonal Mutation Burden

Consistent with previous findings (Sia et al. 1997; Fujimoto et al. 2020), a substantial increase in the mutational burden in cases with MSI was identified. The average mutation burden was $200,250$ (range: $31,047 - 357,160$, median $197,092$) and $7,570$ (range: $2,154 - 11,907$, median $7,967$) in MSI and MSS cases respectively (Figure 4.6B). Likewise, an increased insertion or deletion to single-nucleotide variant ratio (InDel/SNV ratio) was identified for MSI cases, with an average of 1.82 (range: $1.56 - 2.18$, median: 1.79) and 0.13 ($0.06 - 0.22$, median: 0.13) in MSI and MSS cases respectively (Figure 4.6B).

As expected in light of the clinically detected MMRd, an increased InDel/SNV ratio of 1.56 was also observed in case C562. Overall, the observed mutational burden was similar to those reported in other studies of CRCs (Muzny et al. 2012; Liu et al. 2018), hence corroborating the ability to detect variants in this study.

4.3.2.3 Mitochondrial Variants

Curiously, it is still debated if MMR activity exists in mitochondria (MT) and whether it is independent of the nuclear MMR pathways (Mason et al. 2003; Alexeyev et al. 2013; Fontana and Gahlon 2020). Reports of mitochondrial MSI in MMRd colorectal cancers do exist (Habano, Nakamura, and Sugai 1998). However, experimental data indicate that crucial proteins (i.e., MSH2, MSH3, MSH6 and MLH1) of the nuclear MMR pathways are either not present or do not contribute to MMR repair activity in MTs (Souza-Pinto et al. 2009). It seems that an independent MMR system, potentially involving the protein YB-

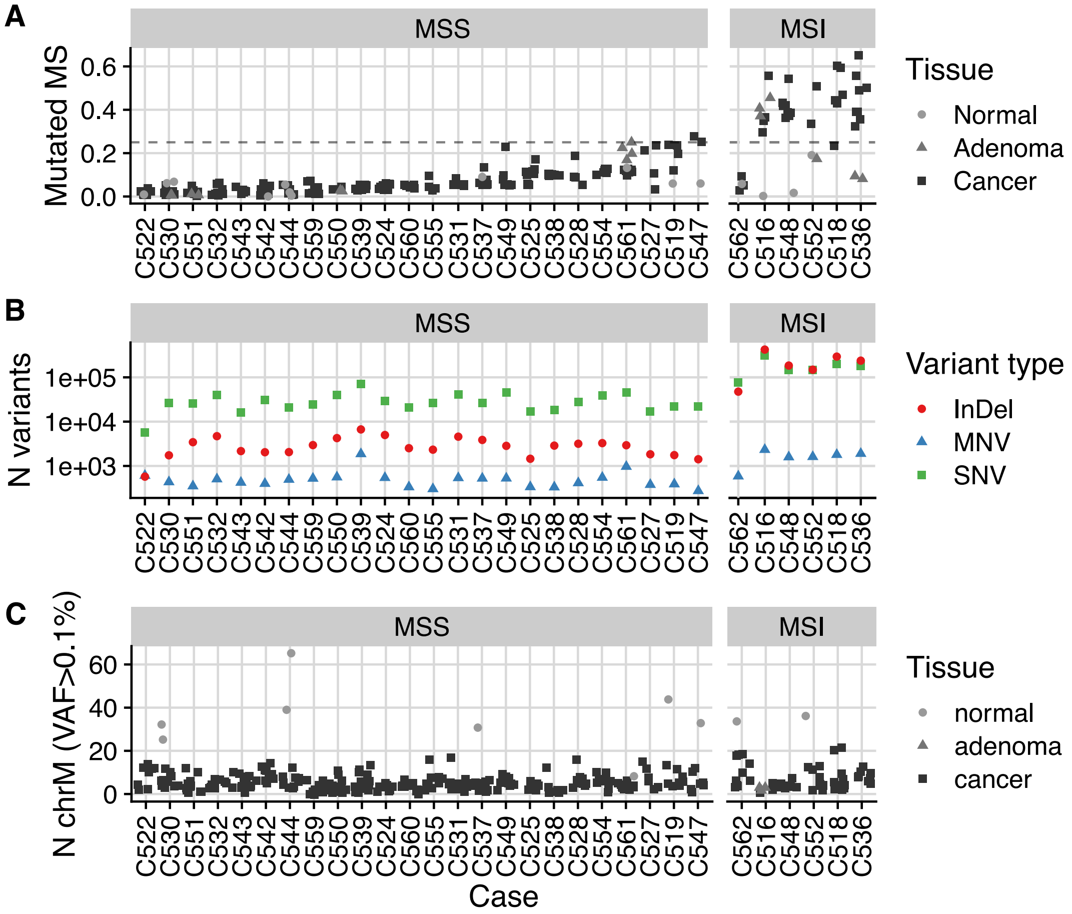


Figure 4.6: Driver events and recurrent copy-number alterations. A) Percentage of altered microsatellites in MSS and MSI samples. The shape and colour of points show the tissue type (legend on the right). B) Total number of somatic mutations of different types (legend on the right) identified in the tumours. C) Number of somatic mitochondrial mutations with a VAF $> 0.1\%$ in different tissue types of cases with MSI and MSS (legend on the right).

1 (YBX1), might be responsible for the experimentally detectable MMR activity in MTs (Souza-Pinto et al. 2009). Other studies described a protective effect of MLH1 — this protein is part of the nuclear MMR pathway and often lost in MSI colorectal cancers — on the genomic integrity of MT in retinal endothelial cells (Mishra and Kowluru 2014).

Absence of increased number of MT variants in MSI CRC Since the single-gland WGS sequenced samples provided excellent coverage of the MT genome (average coverage: $11,300$ in cancers with MSI, $13,600$ cancers with MSS, $9,790$ in normal and $5,900$ in adenoma glands), the data were used to test if a significant increase in SNVs in the MSI cases existed. Variants were called against a beta-binomial background model with *deepSNV* (Gerstung et al. 2012). To exclude non-somatic variants, any variants detected at a posterior probability of being mutated $PP \leq 0.1$ in the normal reference samples of a patient were removed from all other samples of this case. Variants present at a minimum VAF of $f \geq 0.1\%$ with a $PP \leq 0.05$ were considered to be present in a given sample (Figure 4.6C). As seen in Figure 4.6C, no difference in the average number of point mutations in single glands of MSS and MSI cases were evident and statistical testing using a permutation method did show that the small observed difference of 4.9 in MSS versus 6.4 in MSI cancer glands was not statistically significant ($p = 0.16$).

Significant increase in MT variants in normal crypts Interestingly, normal crypts appeared to contain a much larger number of MT SNVs than cancer glands, with an average of 33.3 vs 5.2 variants respectively (Figure 4.6C). For these, the permutation test did indeed indicate a highly significant p-value $\leq 10^{-5}$. Curiously, no significant differences were observed when only variants present at a VAF $f \geq 1\%$ were considered. A potential explanation for this observation could be the presence of a more divergent stem-cell population in the normal colorectal crypts in general. Variants detected in the tumour often had at a high VAF (e.g., Figure S.29, page 275) and for this reason, the more suitable explanation appears to be that an equivalently diverse set of MT variants was lost when a subset of MT moved to a higher frequency (i.e., swept within cells). These sweeps could either be explained by selection acting on a subset of MTs or alternatively by a reduction of the number of MT in the ancestor of the tumour causing higher rates of genetic drift. Unfortunately, a conclusive answer would require more extensive analysis and modelling, which was considered outside of the scope of this thesis.

4.3.2.4 Normal Gland Mutation Rates

As mentioned above, a relatively large number of single glands sequenced as part of the study were later identified to be normal crypts. Some of these were located within surrounding tumour tissue (51 tumour ‘adjacent normal crypts’) and others even had low-frequency tumour contamination (17 ‘mixture crypts’). While these crypts were excluded from the analysis of the cancer’s genomes, they still provided an opportunity to gain insight into underlying stem cell dynamics and the process of ageing in normal CRC epithelium. As a control, 10 additional ‘normal crypts’ from a total of 8 patients were isolated from tumour distant regions and subjected to WGS sequencing.

Normal crypt mutation burden For the analysis of all normal crypts, variants detected in the matched cancers were removed to avoid the effect of tumour contamination in tumour adjacent glands. Through WGS on average 1440, 2100 and 1620 SNVs, 66, 98 and 82

InDels and 28, 29 and 24 multi-nucleotide variants (MNVs) were detected in the healthy normal, tumour adjacent normal, and tumour-normal mixture glands respectively. The InDel/SNV ratio was with ≈ 0.06 (range: $0.02 - 0.11$, median: 0.053) remarkably similar to that of MSS colorectal glands (range: $0.06 - 0.22$). This demonstrates that no change of the relative mutation rate of InDels compared to SNVs occurs during the early development of MSS cancers. The numbers reported here are generally consistent with the findings by Lee-Six et al. (2019), who were able to identify on average $2,599$ somatic SNVs and 226 InDels in $2,035$ single colorectal crypts isolated by laser-capture microdissection from the colon epithelium of 42 individuals.

I hypothesised that the micro-environmental effects of the tumour might alter the type of mutations or rate with which these are acquired in the normal crypts. Such an effect could, for example, be explained by reactive oxygen species or local inflammation. Both of these mechanisms were previously described to promote tumour development and cause mutations (Waris and Ahsan 2006; Grivennikov, Greten, and Karin 2010; Costa, Scholer-Dahirel, and Mechta-Grigoriou 2014; Canli et al. 2017; El-Kenawi and Ruffell 2017).

Analysis of mutational signatures For this reason, the burden of SNVs, InDels and MNVs, as well as the relative burden attributed to individual mutational signatures was determined (Alexandrov et al. 2013a; Alexandrov and Stratton 2014). For the analysis of signatures with the *deconstructSigs* R package (Rosenthal 2016) four signatures Alexandrov et al. (2013b) previously reported to be active in colorectal cancers were considered: S1 (CpG associated de-methylation), S5 (unknown aetiology), S6 (mismatch repair defects) and S10 (POLE defects).

Examination of the residuals of *deconstructSigs* suggested the presence of an unexplained mutational signature characterised by an excess of T>C mutations (Figure S.37, page 278). This signature resembled the *pks*$^+$ *E. Coli* signatures described by Pleguezuelos-Manzano et al. (2020), which was also added to the assessed signatures.

Representative examples of this analysis are shown for a normal crypt in Figure S.34 (page 277), for a cancer adjacent crypt in Figure S.35 (page 278), and for an intermixed normal crypt in Figure S.36 (page 278). The analysis revealed a substantial contribution $\geq 25\%$ of the *pks*$^+$ associated mutational signature to the SNV burden in 2 of the 18 cases with analysed normal crypts (C537 and C547, compare Figure 4.8A-C). Almost identical contributions of the *pks*$^+$ signature were observed in normal and cancer adjacent crypts of

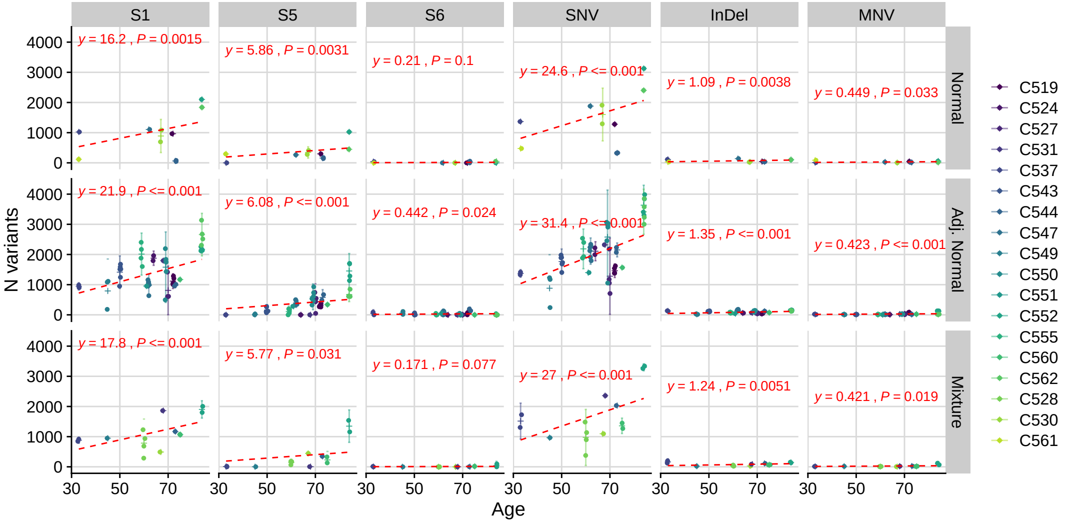


Figure 4.7: Mutation rates in normal colorectal crypts. Shown are the average mutation burden in different types of normal crypts (y-axis panel) isolated from patients with colorectal cancer. Linear regression (red lines and annotations) revealed an increase of the average mutational burden in crypts of a given patient for different mutation types (y-axis panel) with age. S1, S5 and S6 (S10 not sig.) show the mutation burden associated with the corresponding mutational signatures in the COSMIC database estimated with *deconstructSigs* (Rosenthal 2016).

the same cases. This is consistent with the proposed colon wide genotoxic effect of the *pks*$^+$ *E. Coli* strains.

Mutation rates in normal crypts In general, a significant association of age and the mutational burden attributed to each signature and the three different mutation types was found to be present (Figure 4.7). The only exception from this rule was S10. Given the absence of POLE mutations in any of the crypts, this is unsurprising. Especially the positive correlation of age and S1 is consistent with findings by others.

The amount of S1 associated mutations detected in colorectal tumours was previously reported to be associated with the age at diagnosis by Alexandrov et al. (2015). Similar results were also obtained by Lee-Six et al. (2019) who described a positive association of age and the number of mutations associated with the equivalent of S1 and S5 in normal colorectal crypts isolated with laser-capture micro-dissection.

The coefficients of a conducted linear regression were relatively similar for each of the three types of glands considered. On initial examination, no difference between the mutation rate of crypts with a significant tumour cell contamination and healthy normal crypts was found. A slightly larger mutation rate was estimated for tumour-adjacent crypts compared to normal crypts (Figure 4.7). Overall, the data suggested that approximately

24 SNVs and 1.6 InDels accumulate in normal colorectal crypts every year. The relative contribution of S1 to the total mutation burden in normal crypts was consistently higher compared to cancer glands (Figure 4.8), with about 16.5 of the SNVs ($\approx 66\%$) assumed to be caused by the associated deamination of 5-methylcytosine at CpG sites. Most of the remaining variants (≈ 5.8, 23%) were instead attributed to S5.

In order to identify whether the observed mutation rates of tumour adjacent normal crypts were significantly different between the types of crypts, a generalised linear mixed model with crypt type-specific coefficients for age ($y \sim age : type + (1 \mid patient)$) was compared against a reduced model ($y \sim age + (1 \mid patient)$). Since a patient-specific variation of signatures was evident (p$\leq 10^{-12}$, Akaike information criterion (AIC) $65,011$ vs $11,954$ in a binomial regression on S1), the patient variables were added as a random effect to the analysis. A significant difference between the two considered models was found for both SNVs ($p \leq 0.0057$ and AIC 985.38 vs 979.07) and InDels ($p = 0.002$ and AIC 609.26 vs 601.26) and the coefficients suggested that the mutation rate in tumour adjacent crypts was $\approx 50\%$ higher than that in normal crypts.

A non-parametric test (Wilcoxon signed-rank exact test) applied to the average number of SNVs in normal and adjacent normal crypts of cases with paired data ($N = 6$) also suggested a significantly higher ($p = 0.0156$) number of SNVs in the adjacent crypts. Obviously, the specific mechanism for the observed differences remains elusive, but care was taken to avoid tumour contamination confounding the analysis. Among others, microenvironmental effects (Reynolds, Rockwell, and Glazer 1996) or increased mutation rates in a surrounding field defect (Bernstein et al. 2008; Shen et al. 2005) could explain this phenomenon.

4.3.2.5 Driver Mutations

The large number of WGS samples obtained from different regions of each profiled cancer in this study (median: 7, range: $2-11$) allowed for the accurate identification of clonal and subclonal mutations (Werner et al. 2017; Opasic et al. 2019). I used this ability to systematically identify subclonal mutations in known cancer driver genes. For this analysis, previously reported cancer-related genes were considered, specifically the 369 pan-cancer driver genes described by Martincorena et al. (2017) and 69 colorectal cancer-specific drivers

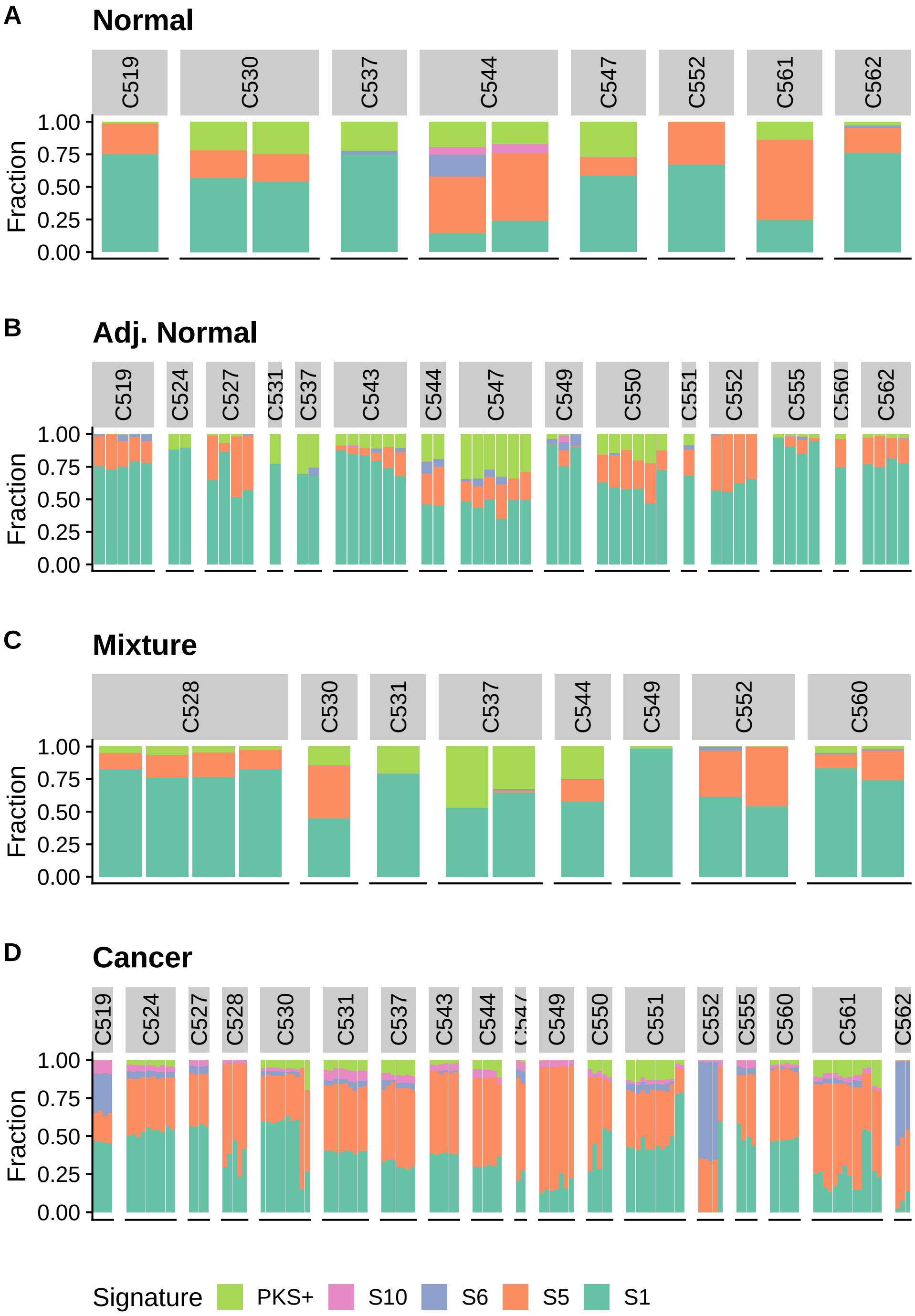


Figure 4.8: Signature contribution in normal colorectal crypts. Relative contribution of mutational signatures S1, S5, S6 and S10 reported in the COSMIC database and the *pks*$^{+}$ signature described by Pleguezuelos-Manzano et al. (2020) were estimated with *deconstructSigs* (Rosenthal 2016).

genes from IntOGen[6] (Martínez-Jiménez et al. 2020).

The analysis of all non-synonymous somatic mutations of the 69 colorectal cancer driver genes revealed a plethora of mutations in MSI and MSS cases (Figure 4.9A). As expected (Priestley et al. 2019; Zhang et al. 2011; Muzny et al. 2012; Campbell et al. 2020), point mutations and InDels of APC (23/30, 77%), damaging mutations of p53 (16/30, 53%) and activating hotspot mutations of K-Ras (13/30, 43%) were found to be the three most common driver mutations observed. Other recurrently mutated genes identified were PIK3CA (11/30, 37%), FAT4 (8/30, 27%) and SOX-9 (10/30). All of these are commonly observed to be mutated in CRCs, albeit at a somewhat lower frequency of $\approx 17\%$ for PIK3CA, 6.5% for FAT4 and 5.4% for SOX-9 (Martínez-Jiménez et al. 2020).

In line with existing literature (Rowan et al. 2000), cases with APC mutations frequently showed a loss of the unmutated allele through loss of heterozygosity (LOH)[7] or an additional truncating APC mutation[8]. These events were observed in 10/23 (43%) and 9/23 (39%) of the carcinomas respectively. In four cases[9] only one APC mutation was identified, potentially indicating the presence of additional undetected mutations (i.e., false negative calls). In C560 a subclonal LOH affecting the *APC* locus was observed in one sample (B1_G1_D1) (Figure S.48, page 282). Still, as both *APC* alleles were deleted in either of the two states, this event was likely selectively neutral. In contrast, for *TP53* LOH events[10] were the most frequent 10/16 (62%) alteration leading to the loss of the second allele. Only one example of a secondary truncating p53 mutation was observed in C552 (6%). In the remaining four cases a single p53 mutation[11] was identified. For the p.R175H in C536 the literature suggests a negative-dominant effect (Marutani et al. 1999; Willis et al. 2004; Boettcher et al. 2019). In summary, the available data indicate the full loss of p53 function in 12/30 (54%) of the analysed cases.

While the consistency of the results obtained in this cohort with previously published studies (e.g., Waris and Ahsan 2006; Grivennikov, Greten, and Karin 2010; Costa, Scholer-

[6] As part of manual curation three of the 73 reported genes, were identified as likely false positives and removed. These were: LRP1B — which is even recognised as a potentially spurious driver by Martínez-Jiménez et al. (2020) — KMT2C and PARP4. These genes contained many repetitive regions with low mappability or passenger hotspots and had a dN/dS of ≈ 1 in the TCGA colorectal cancer cohort (L. Zapata, personal communication).

[7] Observed in C561, C547, C539, C549, C538, C544, C554, C543, C530, and C516.

[8] Observed in C560, C542, C527, C555, C531, C524, C531, C524, C550, C537, C548, and C516.

[9] Observed in C519, C525, C559, and C536.

[10] Observed in C561, C547, C527, C539, C554, C549, C544, C560, C538, and C528.

[11] Observed in C519 (p.R158H), C542 (p.R213*), C536 (p.R175H), C543 (p.87Rfs*63), C516 (p.A138V).

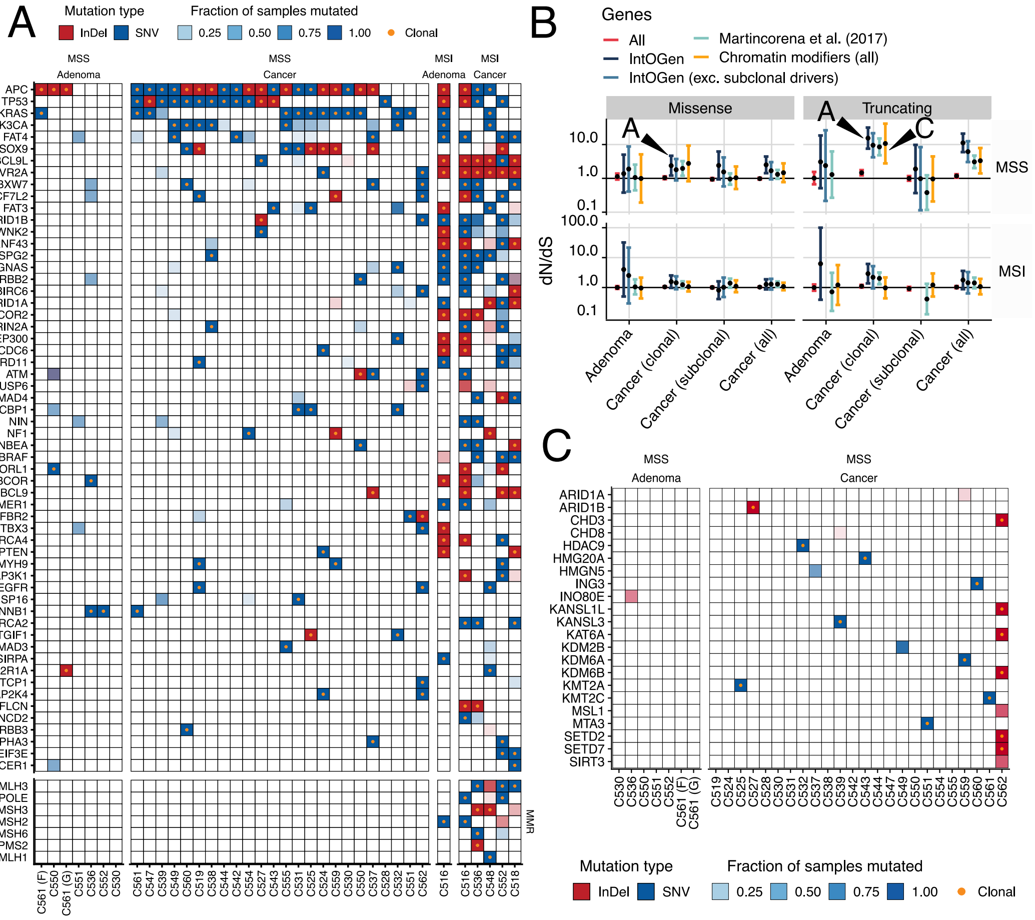


Figure 4.9: Driver mutations and dN/dS analysis of somatic variants in the EPICC cohort. A) Somatic driver mutations in previously identified colorectal driver genes (Martínez-Jiménez et al. 2020) in adenomas and carcinomas of the EPICC cohort. B) Results from the dN/dS analysis with *dndscv* demonstrate evidence of positive selection for missense and truncating mutations in CRC driver genes from IntOGen (A) and pan-cancer drivers from Martincorena et al. (2017) for clonal variants, but not for subclonal variants. Additionally, an excess of clonal truncating mutations in chromatin modifier genes for MSS CRCs was apparent (C). C) Truncating chromatin modifier mutations in MSS CRCs for which the dN/dS analysis indicated positive selection.

Dahirel, and Mechta-Grigoriou 2014; Canli et al. 2017; El-Kenawi and Ruffell 2017) is reassuring, one of the primary objectives was to characterise the frequency of subclonal selection. This should in principle be revealed by subclonal mutations of known cancer driver genes. Some previously conducted analyses of single bulk sequencing data suggest that subclonal driver mutations are very frequent (Tarabichi et al. 2018; Dentro et al. 2021). In contrast to these studies, mutations of the three classic colorectal cancer driver genes APC, p53 and K-Ras were only identified in one case, specifically an activating K-Ras p.G12C mutation (Bos et al. 1987) in C539 (Figure S.40, page 279). Likewise, other less frequently mutated colorectal driver genes like FBXW7 (Iwatsuki et al. 2010; Yeh, Bellon, and Nicot 2018), truncating SOX-9 mutations (Javier et al. 2016), EGFR (Barber et al. 2004), PTEN (Molinari and Frattini 2014) or TCF7L2 (Tang et al. 2008; Wenzel et al. 2020) were clonal in almost all cases (Figure 4.9A).

A notable exception from this general pattern were subclonal mutations of PIK3CA and FAT4 which were observed in more than one case. Subclonal PIK3CA mutation were found in 5/30 (17%) cases: C544 (Figure S.41, page 280), C531 (Figure S.42, page 280), C525 (Figure S.43, page 280), C524 (Figure S.44, page 281) and C537 (Figure S.45, page 281). In two cases, C531 and C524, even more than one subclonal PIK3CA mutation was present. This resembles the frequent parallel evolution of PIK3CA mutations in clear cell renal carcinomas reported by Gerlinger et al. (2014) and has not been described in similar studies of CRC (e.g., Uchi et al. 2016; Cross et al. 2018).

As expected for bona fide driver mutations of PIK3CA, the majority of these mutations occurred in previously identified 'hotspots' around the RAS binding, helical or kinase domain of the protein (Samuels et al. 2004). The activation of the enzymatic activity conferred by these alterations and consequently, the downstream signalling cascades was previously shown to promote cell growth and invasive capabilities in cell lines (Samuels et al. 2005). It would therefore be reasonable to assume that PIK3CA could also confer such a growth advantage in vivo either at primary or metastatic sites. The prognostic value of PIK3CA mutations is still unclear (Mei et al. 2016), but increased resistance to chemotherapy in PIK3CA mutated CRCs (Wang et al. 2018b) and an effect of non-steroidal anti-inflammatory drugs on survival in PIK3CA mutated CRCs (Liao et al. 2012) has been previously reported. These studies suggest that PIK3CA might play a fundamental role in the evolution of CRCs.

Similar results were obtained by Sottoriva et al. (2015), Kim et al. (2015), Uchi et al. (2016), and Cross et al. (2018) who observed subclonal PIK3CA mutations in 1/6, 1/5, 4/9 and 1/10 CRCs profiled by multi-region WES respectively. No subclonal PIK3CA mutation was identified in a similar multi-region WES study of 4 CRCs by Suzuki et al. (2017). While the used methods and the design of the studies differed substantially, taken together they suggest that around 19% (95% CI: 10%–30%) of CRC harbour subclonal PIK3CA mutations at a frequency at which they can be detected with multi-region WES and WGS. A meta-analysis of all five studies is shown in Figure S.49A (page 283), showing that the proportion of cases with a subclonal PIK3CA mutation are consistent across these.

The observation of subclonal activating PIK3CA mutations in many different studies, including this one, suggests that PIK3CA mutations are a genuine driver alteration in CRC and that it is consistently acquired late. This also suggests that a substantial fraction of tumour cells in those CRC grew in the absence of activation of the PI3K pathway. This alone would suggest that PI3K inhibition in CRC might only elicit a partial response, at least if the PIK3CA wild-type cells do not already rely on the activation of the PI3K pathway.

The only other CRC driver gene for which more than one MSS case with subclonal mutations was observed in the EPICC cohort was FAT4, which was found to be mutated in C561 (Figure S.46, page 282) and C554 (Figure S.47, page 282). A meta-analysis identical to that conducted for PIK3CA is shown in Figure S.49B (page 283), indicating that subclonal mutation of FAT4 are relatively rare events occurring in only 3.4% (95% CI: 0.4%-11.9%) of CRCs.

In summary, the analysis of the 30 CRCs by extensive single-gland WGS from multiple regions of the tumour revealed only a few recurrent subclonal driver alterations: a single subclonal K-Ras mutation, seven subclonal PIK3CA mutations from five cases and two FAT4 mutations (Table 4.2). This appears to be at odds with the results obtained by Dentro et al. (2021) from single-bulk WGS sequencing data in the PCAWG cohort (Campbell et al. 2020). This study suggests that subclonal driver mutations are widespread in CRCs. From Figure 6 in Dentro et al. (2021) it can be seen that subclonal driver mutations of K-Ras were identified in $\approx 10\%$, of p53 in $\approx 5\%$, and of APC in $\approx 5\%$ of cases. Assuming that these alterations are independent $\approx 20\%$ of cases were found to have evidence of subclonal drivers in these three classic driver genes alone and additional subclonal mutations in a large number of other putative driver genes appear to suggest pervasive subclonal selection driven by somatic events.

4.3.2.6 dN/dS Analysis

To potentially resolve this discrepancy, I assessed if an excess of non-synonymous driver mutations in all driver genes might indicate widespread positive selection of subclonal mutations in the 30 CRC analysed by us. For this, two previously published tools for the analysis of dN/dS ratios in cancer genomic data were applied to the data (Martincorena et al. 2017; Zapata et al. 2018).

In line with previous studies (Martincorena et al. 2017), the cohort was stratified into MSS and MSI cases to account for the large difference of mutation rates in these two CRC

Table 4.2: Potential subclonal driver mutations identified in the EPICC cohort.

Case	Gene	Mutation	Hotspot?
C539	K-Ras	p.G12C (Region A & B)	Yes (Bos et al. 1987)
C544	PIK3CA	p.H1047Q (D1_G3)	Yes (Zaidi et al. 2020)
C531	PIK3CA	p.Q546K (Region C)	Yes (Zaidi et al. 2020)
C531	PIK3CA	p.G118D (Region D)	No, but activating (Masoodi et al. 2019)
C524	PIK3CA	p.C378R (Region B)	No, but activating (Samuels et al. 2004)
C524	PIK3CA	p.R88Q (Region C & D)	No, but activating (Oda et al. 2008)
C525	PIK3CA	p.Q546P (Region A & C)	Yes (Zaidi et al. 2020)
C537	PIK3CA	p.E545K (Region C)	Yes (Bader, Kang, and Vogt 2006)
C561	FAT4	p.L2617V (Region A)	-
C554	FAT4	p.M1825K (Region A)	-

subgroups. The somatic mutations were then split into sets of subclonal and clonal mutations. For each of these sets, dN/dS values were determined for the entire coding genome, pan-cancer driver genes reported in Martincorena et al. (2017), all chromatin modifier genes (Yates et al. 2020), and the IntOGen colorectal cancer driver genes shown in Figure 4.9A.

As expected, clear evidence for selection of clonal missense and truncating mutations in MSS cancers were found (i.e., dN/dS $\geq$ 1) in both lists of putative cancer driver genes (Arrow A in Figure 4.9B). Likely due to the higher fraction of CRC specific driver genes in the IntOGen list, dN/dS estimates were slightly higher for this set of genes when compared to the pan-cancer gene list. Likewise, the dN/dS analysis revealed a clear excess of clonal truncating driver alterations in MSI cancers for both gene lists and clonal missense mutations in the IntOGen genes. The dN/dS estimates obtained for the MSI cases were generally lower than that of the MSS cases. This is consistent with the depletion of signal due to the excess of non-selected variants arising due to the higher mutation rates compared to a small set of driver mutations.

In contrast, no significant evidence for subclonal selection of truncating variants was identified in MSI or MSS cases. The dN/dS estimates of subclonal missense mutations for the IntOGen gene list were slightly higher than one, indicating subclonal selection of a subset of mutation in these CRC driver genes. This is consistent with the previously described K-Ras and PIK3CA hotspot mutations (see Table 4.2). After the removal of these two genes from the IntOGen list ('IntOGen (excl. subclonal drivers)'), observed dN/dS values were indeed not significantly above one as shown in Figure 4.9B, thus supporting the conclusions made above.

Interestingly, the dN/dS analysis of chromatin modifier genes (CMG) revealed an excess of truncating clonal mutations in them (Figure 4.9B). The distribution of these truncating SNVs, which is shown in Figure 4.9C, did not reveal any apparent genes with an increased number of somatic variants. Due to the complex function of CMGs, it appears plausible that mutation of different CMG might produce a similar phenotype. In analogy to MSI, the mutation of different CMGs might for example cause 'epigenetic hypermutation' and the frequent selection for truncating alterations of various CMG could occur if the resulting epigenetic mutations themself are subject to strong positive selection.

Overall, these findings, that is $\mathrm{dN/dS} \approx 1$ for subclonal mutations excluding K-Ras and PIK3CA, are consistent with the conclusions made based on the assessment of recurrently mutated CRC driver genes. In general, subclonal selection due to somatic mutations — with the exception of a few specific genes identified in a subset of cases — appears to be rare in CRC. This is in agreement with conclusions from the previous analysis of single-bulk WES data by Williams et al. (2016) and Williams et al. (2018b), which found that subclonal mutation spectra resembled that of a neutral null model in 38/108 (35%) and 55/70 (79%) of CRCs respectively. The 'Big-Bang' model of tumour growth (Sottoriva et al. 2015) might explain this apparent lack of subclonal selection.

4.3.3 Reconstruction of Phylogenetic Relationships

I next sought to explore the ancestral relationships of glands as revealed by somatic variants accumulated during the clonal expansion of the tumour. For this I used, like many other studies before (e.g., Gerlinger et al. 2014; Bruin et al. 2014; Ling et al. 2015; Sottoriva et al. 2015; Cross et al. 2018), a simple maximum-parsimony reconstruction of phylogenies from mutations detected in individual samples.

The application of such methods to bulk WGS sequencing data has been criticised by Alves, Prieto, and Posada (2017), since mutations detected in bulk WGS sequence data (i.e., 'sample trees') do not — or at least not necessarily — inform on the mutations present in individual 'evolutionary units'. In the context of this study, the mutation profiles were obtained from single colorectal cancer glands. These are indeed assumed to be the evolutionary units of CRC (Humphries et al. 2013) and the direct application of classic phylogenetic methods to mutations identified in samples seems for this reason less controversial.

The decision of using a MP method was primarily based on their simplicity and the advantage of not requiring the formulation of an explicit model with potentially questionable applicability to cancer genomic data. For example, commonly used substitution models do not account for the complexity of mutational processes known to be active in cancer (Alexandrov et al. 2013b) and readily available models of population structures might be violated due to spatial dynamics. For this reason, MP trees were reconstructed as a simple 'model-free summary statistic' of the data. In Chapter 6 (page 195 ff.) I will present some results from a simulation-based inference in which I used these trees as a data basis.

The subclonal structure of MP trees reconstructed from mutations identified by WGS are shown in Figure S.51 (page 284). Bootstrapping of the mutation data (Figure S.50, page 283), confirmed that the topologies of the reconstructed trees were robust. At a later stage, LP-WGS was applied to a larger number of single-gland obtained from the same 30 cases. These additional LP-WGS samples were assigned to edges of the tree using a maximum-likelihood method, which will be described in detail in Chapter 5 (page 171 ff.). In the following, these LP-WGS trees will be used to discuss the general patterns observed. As shown in Figures 4.10A, C & D, the macroscopic locations of the sampling regions were recorded during sample collection. The macroscopic structure of most carcinomas was similar to that of the two examples shown in Figure 4.10 C & D, i.e., a relatively flat and round crater-like erosion that is typical for colorectal adenocarcinoma (Nagtegaal et al. 2020).

In a subset of cases[12], one or more synchronous adenomatous polyps were identified. The proportion of cases with synchronous adenomatous polyps (6/30, 20%) was consistent with the proportion of ≈ 30% cases reported elsewhere (Kim and Park 2007). In another case (C516) a tumour adjacent polypoid precursor lesion was found to be present. Additional WGS of samples from these lesions were obtained in three cases (C516, C551, and C561).

As shown in Figure 4.10A&B, the concomitant polypoid lesion in C516 (region C&D) and the carcinoma (A&B) shared a large number of variants, including a mutation of APC (single-hit) and p53 (single-hit). This clearly demonstrates that the two cell masses arose from a common, likely precancerous, MSI ancestor.

In contrast to this, the adenomatous polyp (region F) of C551 (Figure 4.10 C&D) and the two sampled polyps (region F&G) of the case C561 (Figure 4.10 E&F) demonstrated that these distant polyps arose independently from the healthy colorectal epithelium. The 109 variants shared by the main tumour and the two adenomas, as well as the 38 variants

[12] C530, C547, C550, C551, C552, and C561

shared by the two adenomas with each other, could then have arisen during the embryonal development and the formation of the colon. A similar number of mutations 154 was shared between the tumour and the independent adenoma in C551. In all three cases, no known non-synonymous mutations in any colorectal driver gene were identified.

I annotated the tumour phylogenies reconstructed from single-gland and small bulk samples in the remaining 27 cases with putative driver mutation identified to assess if the selection of these variants might have induced any obvious distortion of the general tree structure. The 27 reconstructed sample trees, including the assigned LP-WGS samples, are shown in Figure 4.11 and the MP trees excluding the added LP-WGS samples can be found in Figure S.51 (page 284).

Structures in tree topologies The trees reconstructed from the majority (17/23, 74%) of tumours in which the evaluation of such patterns was possible showed a clear formation of clades containing samples from the same regions. This pattern is best summarised by the example C548 shown in Figure 4.12B and a list of all cases in which this lack of intermixing was observed can be found in Table 4.3.

Notably, in a smaller set of cases (6/23, 26%) some degree of inter-region mixing of samples was detected. These patterns were consistent with the spatial variegation observed by Sottoriva et al. (2015). A clear example of this was found to be present in case C559 (Figure 4.12D). For the remaining cases, summarised in Table 4.3, different sources of evidence were available i) the WGS sequencing data, ii) the ML estimates of the LP-WGS sample positions in the tree, and iii) the CNA analysis of LP-WGS samples consistent with alteration found to be present in the majority of samples from a different region. Considering i) or a combination of ii) and iii) clear evidence, spatial variegation was found in $\approx 22\%$ (5/23) of analysed cases. This is considerably less than the 6/6 (Fisher's Exact Test: $p = 0.001$, OR: 0, 95% CI: $[0 - 0.35]$) of carcinomas with such patterns in Sottoriva et al. (2015) based on the analysis of WES data and 9/11 (Fisher's Exact Test: $p = 0.002$, OR: 0.069, 95% CI: $[0.0055 - 0.47]$) based on CNA analysis alone, but supports the presence of this pattern in general. As noted above, a high frequency of spatial intermixing might be explained by nearly exponential growth and the lack of such intermixing by boundary driven growth. Based on these observations, it appears reasonable to suggest that the presence of intermixing might be a surrogate of a fast and potentially aggressively growing tumour, indicative of a worse outcome (i.e., a potential prognostic marker).

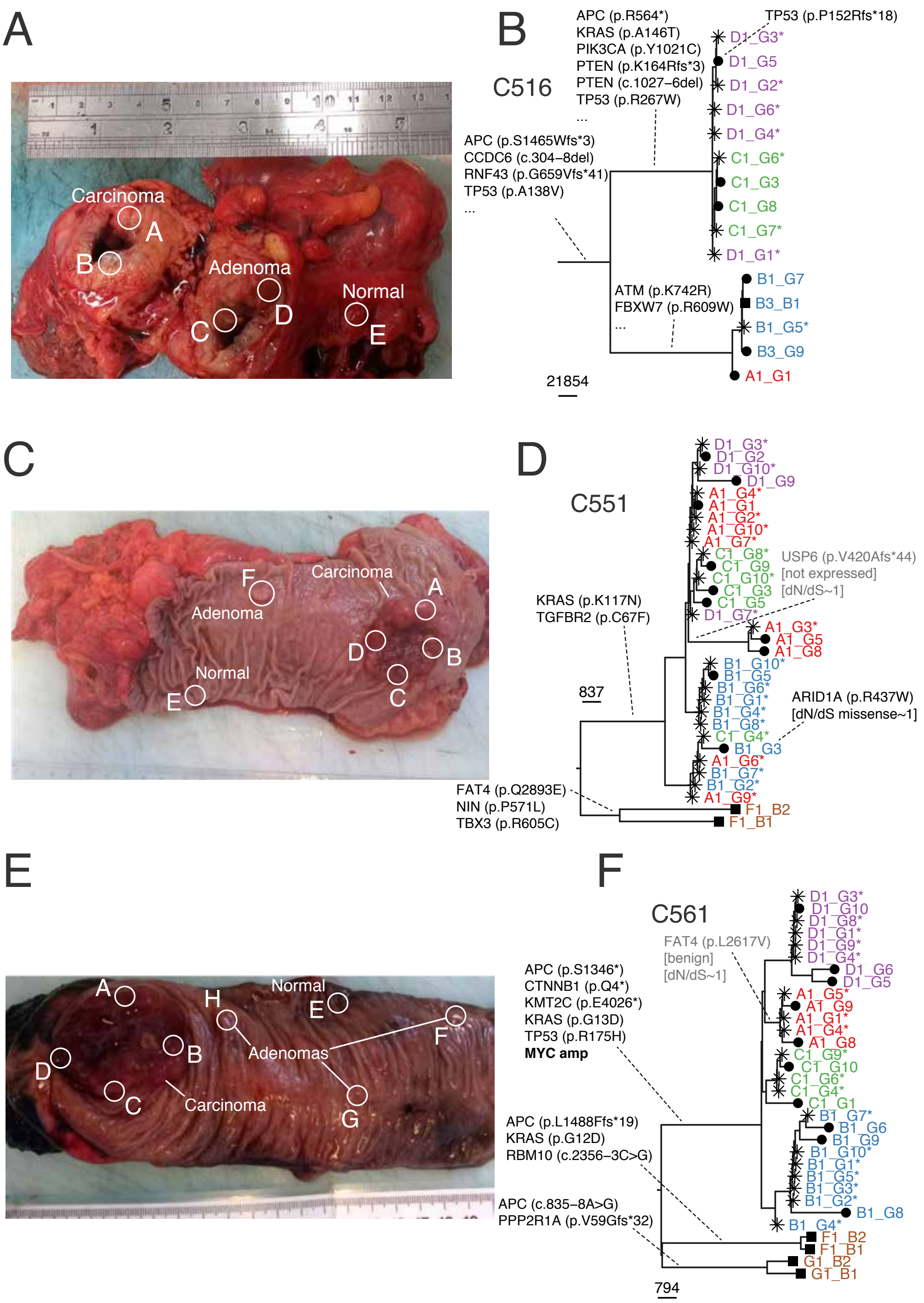


Figure 4.10: Spatial sampling and phylogenetic trees reconstructed for carcinomas with associated adenomas. Arrows indicate putative driver mutations in known colorectal driver genes reported in IntOGen (Martínez-Jiménez et al. 2020). A) Shows a macroscopic image of the tumour of C516. Here a polypoid adjacent adenoma (C&D) was adjacent to the carcinoma (B%A). B) Shows the phylogenetic tree reconstructed for somatic mutations in C516. This tree indicates that the adenoma and the tumour evolved from a common precursor lesion. C) Shows a case (C551) in which a synchronous adenoma (F) was located several centimetres away from the main carcinoma (A-D). D) In this case the reconstructed phylogenetic relationships suggested that both tumours evolved independently from the healthy colorectal epithelium. E) Shows a macroscopic image of the tumours found in C561. A carcinoma (A-D) and several adenomas (G-F) were found to be present in this case. F) The phylogenetic relationships suggested, like for C551, that the adenomas arose independently from the carcinoma.

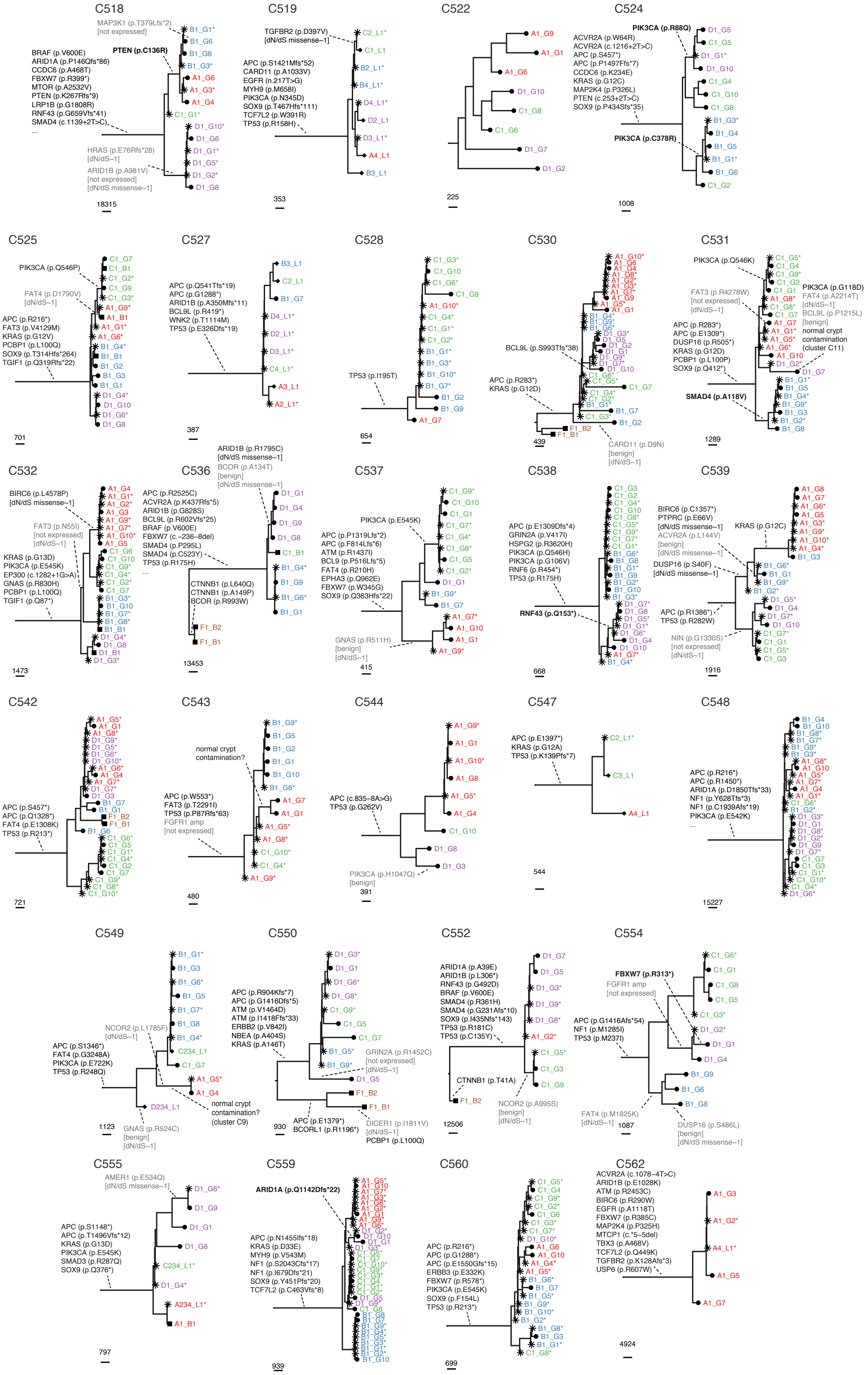


Figure 4.11: Maximum-parsimony phylogenies reconstructed from mutations identified in samples of 27 colorectal carcinomas. Arrows indicate putative driver mutations in known colorectal driver genes reported in IntOGen(Martínez-Jiménez et al. 2020).

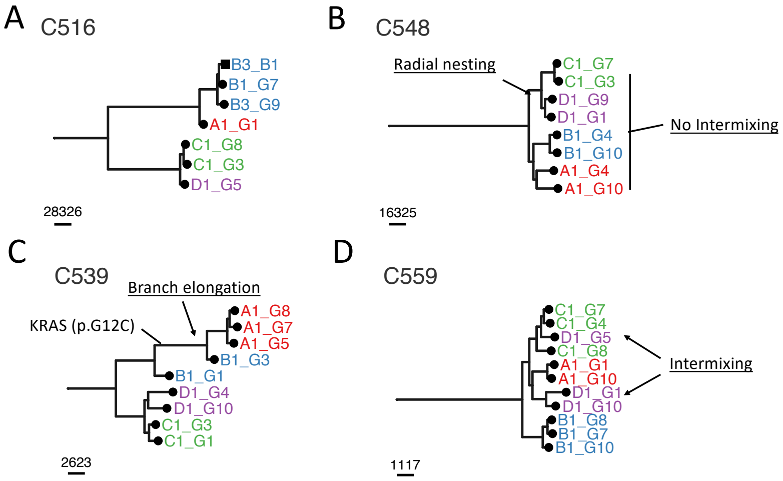


Figure 4.12: Examples of observed tree structures. A) Shows a clearly ‘branching’ tree indicating parallel evolution of distinct regions. B) Shows a tree with clear spatial segregation, indicated by the lack of intermixing and radial nesting of lineages. C) Shows an example of a case with an elongated internal edge. This elongation suggests the presence of a selected subclone, which in this case was explainable by a p.G12C K-Ras mutation. D) Shows an example in which the intermixing of samples from different regions suggested the presence of spatial variegation.

Across cases, perfectly star-shaped topologies were, except for C536 and C561 (2/19, 11%), not observed. Instead, a pattern consistent with sampling from ‘radial clones’ (compare Figure 3.13B, page 103), manifesting itself as clades formed by samples from two adjacent regions (e.g., A+B or D+A, but not A+C, compare Figure 4.10E) was found to be present (13/19, 68%, see Table 4.3). An example of this pattern can be seen in case C548 (Figure 4.12B). Maybe surprisingly, in the remaining 4/19 (21%) cases, clades formed by samples from opposing sites of the tumour (e.g., A&C or B&D) were found to be present. Given that the proportion of adjacent regions is itself lower, this appears to indicate frequent scattering or intermixing of glands during the early tumour development. While the reasons for this might not be entirely clear, the presence of such a feature could also be associated with a specific growth pattern of individual tumours and potentially be prognostic.

Last but not least, in a small set of cases, elongated internal branches were observable. In a single case, C539, this pattern was obvious and indeed a subclonal activating K-Ras mutation (p.G12C) — one of the most common colorectal driver alterations with a well-established effect on cells growth rates (Konishi et al. 2007; Platt et al. 2014) — mapped exactly to this branch of the tree (Figure 4.12C). In all other cases, listed in Table 4.3, the presence of such a pattern, if existent, was more subtle. In some but not all cases, previously identified driver mutations (e.g., PIK3CA) mapped to these edges. Examples of this can be found in C524, in which a subclonal PIK3CA p.C378R mutation was present or C538 in

Table 4.3: Summary of features identified in tumour phylogenies. Trees from C519, C522, C527, C547, C555 and C562 were considered not evaluable based on the data obtained.

Case	**Mixing?**	**Subclades?** [1]	**Branch elongation?**
C516	No	-	No
C518	No	A&B	No
C525	No	**A&C**	Maybe (PIK3CA p.C378R)
C528	No	B&C	No
C530	No	A&D	No
C532	No	**A&C**	No
C536	No	Star shaped	No
C537	No	B,C,D	Maybe (PIK3CA p.E545K)
C539	No	C&D + A&B	**Yes (KRAS p.G12C)**
C542	Maybe	A,B,D	No
C543	No	-	No
C544	No	**A,C**	Maybe
C548	Maybe	C&D	No
C549	No	B&C	Maybe
C552	No	-	No
C554	No	C&D	Maybe
C561	No	Star shaped	Maybe
C524	**Yes**[2]	C&D + B&B	Maybe (PIK3CA p.C378R)
C531	**Yes**[3]	A,C,D	Maybe (PIK3CA p.G118D)
C538	**Yes**[3,4]	B&C	Maybe (RNF43 p.Q153*)
C551	**Yes**[2,3,4]	-	Maybe
C559	**Yes**[2,3]	C&D	No
C560	**Yes**[3,4]	**A&C**	No

[1] Non-adjacent regions highlighted in bold.
[2] Based on WGS samples, [3] Based on LP-WGS samples
[4] LP-WGS position supported by LP-WGS CNA analysis.

which a subclonal RNF48 p.Q153* mutation was identified (see Figure 4.10). In other examples, like C543 or C549, no known driver mutations were found to be present. The analysis of the VAF spectrum suggested that contamination from individual adjacent normal crypts, which according to the analysis presented above contain $\approx$ 1500 SNV each, could have contributed to this phenomenon in some cases. However, the question remains to what extent these patterns could be attributed to spatial drift instead of selection. Further, what tree structure to expect in general from spatially sampled tumours is unclear.

Estimated clone size vs lineage age Disregarding these spatial effects, one could use the information of lineages contained in the trees to derive estimates of the relative clone[13] sizes, which is equivalent to the VAF in bulk samples, and the approximate time t_{MRCA} at which variants in these clones arose.

Under the assumption of neutral evolution and in the absence of drift, the relative

[13] In this context a clone refers to a set of cells with a MRCA identifiable by their shared mutations.

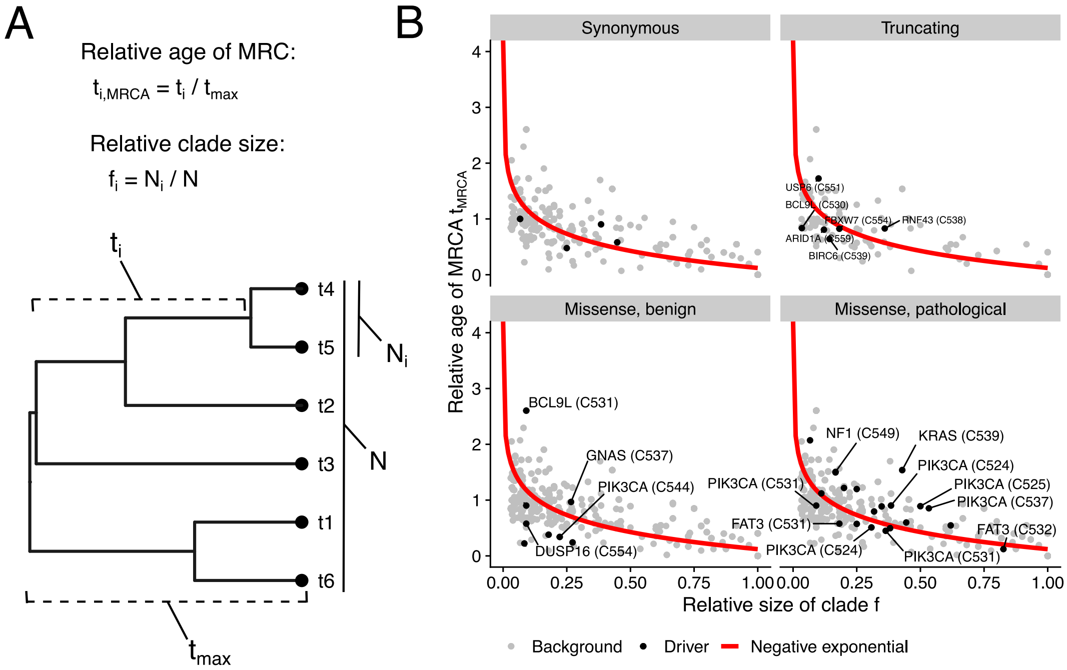


Figure 4.13: Relationship of clone size and MRCA age. A) Illustration showing how estimates for the relative clone size $f = \frac{N_m}{N}$ and relative MRCA age $t_{MRCA} = \frac{t}{t_{max}}$ can be obtained from the tree. B) The distribution of f and t_{MRCA} obtained from the trees across the entire cohort. The red line shows a fit of the negative exponential relationship expected under neutrality. Black dots highlight lineages with mutations of given impact (panels) in an IntOGen CRC driver gene.

clone size $f = N_m/N$, where N_m is the number of mutated and N total number of cells, in an exponentially growing population, is $\approx 1/N_t$, where N_t is the total population size at the time t at which the mutation arose. The number of cells in such an exponentially growing population at time t is given by $N(t) = e^{\lambda t}$ with the growth rate λ. For this reason, the relationship between f and t_m is expected to follow $t_m = \frac{-ln(f)}{\lambda}$. Positive selection should instead increase the clone size relative to the age of the MRCA, causing a deviation from this relationship. In other words, even mutations present in a recent MRCA could reach a significant size due to the expansion of the selected subclone. Estimates of t_m and f can be obtained from reconstructed phylogenetic trees as illustrated in Figure 4.13A. The observed relationship between $t_{MRCA} \approx t_m$ and f across all trees is shown in Figure 4.13B.

As expected, the relationship of these two measures from most resolved lineages followed the expected exponential distribution derived above (red line in 4.13). One has to consider that this analysis does ignore the biases introduced by the spatial sampling or other effects arising from growth in space and for this reason, direct testing for the significance of deviations is likely not warranted. Still, in order to gain some insight into whether subclonal CRC driver mutations (IntOGen, Martínez-Jiménez et al. 2020) might explain any of the observed deviations, I added annotations of these for i) synonymous mutations and ii) benign missense mutation, both of which are expected to have no effect as well as iii) truncating variants and iv) pathological, missense mutation for which some might induce selection, thus causing deviations from the expected exponential fit. In line with the conclusions drawn from both the analysis of dN/dS and the general topology of trees, the majority of mutations, including those in CRC driver genes, seemed to follow the expected fit. Especially, synonymous and CRC driver gene mutations predicted to be benign appeared to follow the distribution of the background. A few notable exceptions of this general pattern existed for missense mutations in CRC driver genes predicted to be pathological existed (bottom left corner of Figure 4.13B). Again, the clearest example of this was the K-Ras p.G12C mutation, which was also identifiable through manual examination of the tree described above.

Assessment of the PIK3CA mutations listed in Table 4.2 provided some evidence of subclonal selection of these in C525 and C537, but for example, not in C544. For no other driver mutation like FAT3, FAT4 or RNF43 deviations from the exponential fit was evident, thus suggesting that these genes might only be weak CRC drivers or only act in a specific genetic/environmental background.

While it is entirely reasonable to assume that only the small subset of the CRC driver gene mutations caused measurable fitness effects, the analysis disregarded spatial effects as well as biases resulting from non-random sampling in space. For this reason, the described findings must be considered to be anecdotal. The interpretation of the information in the analysed single-gland multi-region sequencing data is not straightforward. To account for these concerns, I will apply an ABC-SMC inference framework that uses a spatial tumour simulator to these data. This approach allows to explicitly model the spatio-temporal dynamics (i.e., tumour phylodynamics). In contrast to the model-free maximum parsimony method described here, it can explicitly take into account the spatial sampling performed on a patient-specific basis and infer the populations' dynamics (i.e., mutation, selection, death and effects of spatial crowding), that could have created the population structure and model the process used to obtain the trees (Stadler, Pybus, and Stumpf 2021). I will present the results of this analysis in Chapter 6 (page 195 ff.).

4.3.4 Analysis of Mutational Signatures in Tumour Glands

I next assessed mutational signatures, to understand if and how changes of active mutational processes might have altered the rate at which mutations were acquired. While a de novo reconstruction of mutational signatures from the available data themself can discover novel signatures and might for this reason generally be the preferred approach, I instead based the analysis on previously identified mutational signatures.[14]

Some mutational signatures are known to be very similar to a linear combination of other ones and this can lead to problems of identifiability. Due to this, it is generally advisable to instead aim to identify the contribution of a relatively small number of relevant signatures. For these reasons, three signatures — S1, S6 and S10 from the COSMIC database (Tate et al. 2019) — originally identified by Alexandrov et al. (2013b) in CRCs and signature S5 — a mutational signature that was previously found to correlate with age at diagnosis in many tumour types (Alexandrov et al. 2015) and suggested to represent a general background process (Lal et al. 2021) — were included in the initial analysis. Data on each signature were obtained from the COSMIC database (Tate et al. 2019). The analysis itself was conducted with the *deconstructSigs* package for R (Rosenthal 2016), a simple and frequently used tool that identifies each patients' exposures to provided signatures by fitting the mutation counts as a non-negative combination of these.

After the initial analysis of the four signatures (S1, S5, S6 and S10) residuals very similar to signatures S2 and S13 (see Figure 4.15) attributed to the activity of cytidine deaminases of the AID/APOBEC family (Nik-Zainal et al. 2012b), S17 (see Figure 4.16) that was also identified in a similar study by Cross et al. (2018) and with proposed aetiology of 5-Fluorouracil (5-FU) chemotherapy-related DNA damage (Christensen et al. 2019), and a *pks*$^+$ *E. coli* associated signature (i.e., SBS88 in COSMIC v3) described by Pleguezuelos-Manzano et al. (2020), were observed in some cases. For this reason, I also included these four additional signatures and repeated the analysis for all cases.

To resolve temporal changes of the active processes, the mutations of each case were split into sets of clonal (i.e., in all samples), shared (i.e., in multiple samples), and private (i.e., a single sample) mutations. Those present in concomitant adenomas were added as an additional group. This analysis demonstrated that in virtually all cases, a much larger fraction of clonal mutations were attributed to S1 and S5 when compared to subclonal variants

[14] A de novo analysis of mutational signatures with *SparseSignatures* (Lal et al. 2021) was later conducted in collaboration with Daniele Ramazzotti. The results of this analysis corroborated the results described here.

(Figure 4.14).

As such, the mutational signature activity reconstructed from clonal variants was much more similar to that of normal crypts (compare Figure 4.8), which, given that majority of the CRC evolution occurred in phenotypic normal crypts, is unsurprising. The differences between ‘Shared’ and ‘Private’ mutations were generally much less pronounced. The separately conducted analysis of the mutations of each edge of the phylogenetic trees confirmed this observation. Overall, little variation of active mutational processes appeared to have occurred during the clonal expansion of the tumours themself.

In MSI cases, a substantial fraction of clonal variants and virtually all subclonal variants were attributed to the MMR associated signature S6. This confirmed the clinically reported MMR in these cases. The relative contribution of S6 to the clonal mutation burden in MSI cases varied substantially. The relative contribution of S6 (i.e., S6 compared to S1+S6) ranged from 73% (C518) to $\approx$ 100% (C552), suggesting that MSI arose at various time points during the evolution of the MRCA of the CRCs. Assuming the 15-fold increase of mutation rates in MSI cases reported by Williams et al. (2016), this suggests that MSI arose around three times earlier than the MRCA of the tumour (i.e., relatively early). Consistent with the absence of cases with Lynch syndrome (i.e., hereditary defects of MMR) in the cohort, none of the concomitant adenomas showed evidence of signature S6. Instead, a signature spectrum extremely similar to that of normal crypts was generally found to be present in the adenomas, demonstrating that abnormal mutational processes normally contribute little to the early evolution of them. Apart from these general patterns, two noteworthy observations were made.

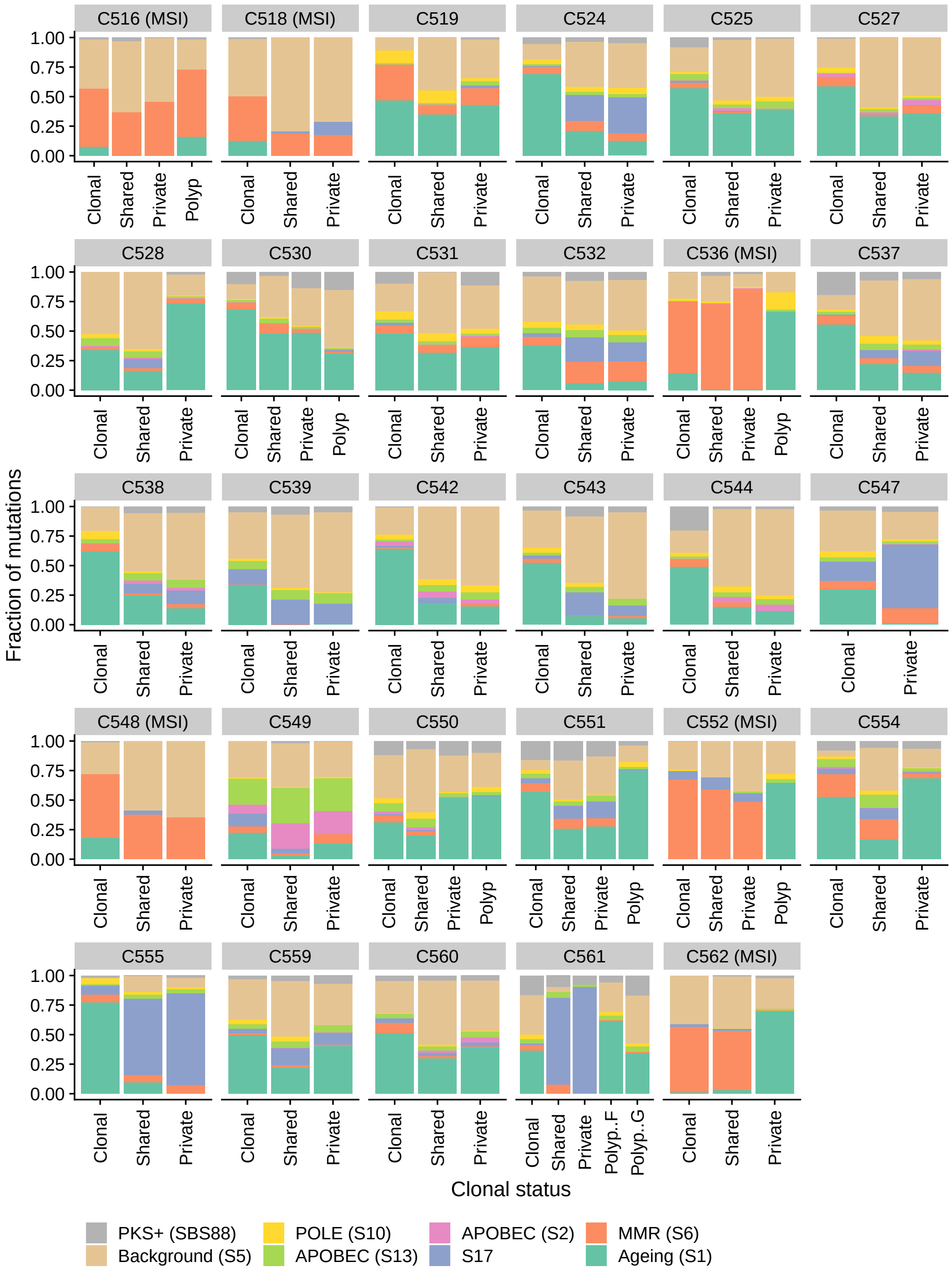


Figure 4.14: Signature contribution in individual colorectal adenocarcinoma and concomitant adenoma (‘Polyps’) of the EPICC cohort. Relative contribution of mutational signatures S1, S2, S5, S6, S10, S13, S17 reported in the COSMIC database (Tate et al. 2019) and the *pks*$^+$ signature described by Pleguezuelos-Manzano et al. (2020) were estimated with *deconstructSigs* (Rosenthal 2016). Mutations were split into groups based on the number of samples they were found to be present in: ‘Clonal’ mutations were present in all samples, ‘Shared’ mutations in more than one, but not all samples, and ‘Private’ mutation in only one sample.

APOBEC associated mutagenesis in a CRC First, one case (C549) was found to exhibited a clear and dominant APOBEC associated mutational signature (see Figure 4.14). APOBEC mediated mutagenesis is frequently found in other tumour types, like bladder or cervical cancers, but not in CRCs (Roberts et al. 2013). In the case of C549 however, two APOBEC associated signatures, S2 and S13 (Figure 4.15B), were found to be present. The total fraction of variants attributed to these was over 40% for clonal and substantially higher for subclonal variants. To further dissect the temporal dynamics of the observed APOBEC mediated mutagenesis, the mutational signatures of each edge of the reconstructed phylogenetic tree were analysed separately. The results of this analysis are shown in Figure 4.15.

As seen here, a signature of APOBEC mediated mutagenesis was identified in the clonal variants of this case ($\approx 40\%$). The contribution to mutations associated with intermediate and terminal edges of the tree was even higher, with $> 90\%$ of variants on some internal edges being attributed to S2 and S13. These observations suggest that prior to the formation of the MRCA of the observable part of the tumour, a stable activation of the associated mutational process occurred. The precise reasons for this are elusive, and no indication of a potential explanation was found in the pathology report of this case. A recent paper by Roufas, Georgakopoulos-Soares, and Zaravinos (2021) suggested that an elevated level of APOBEC associated substitutions can be identified in CRC with high antitumoral

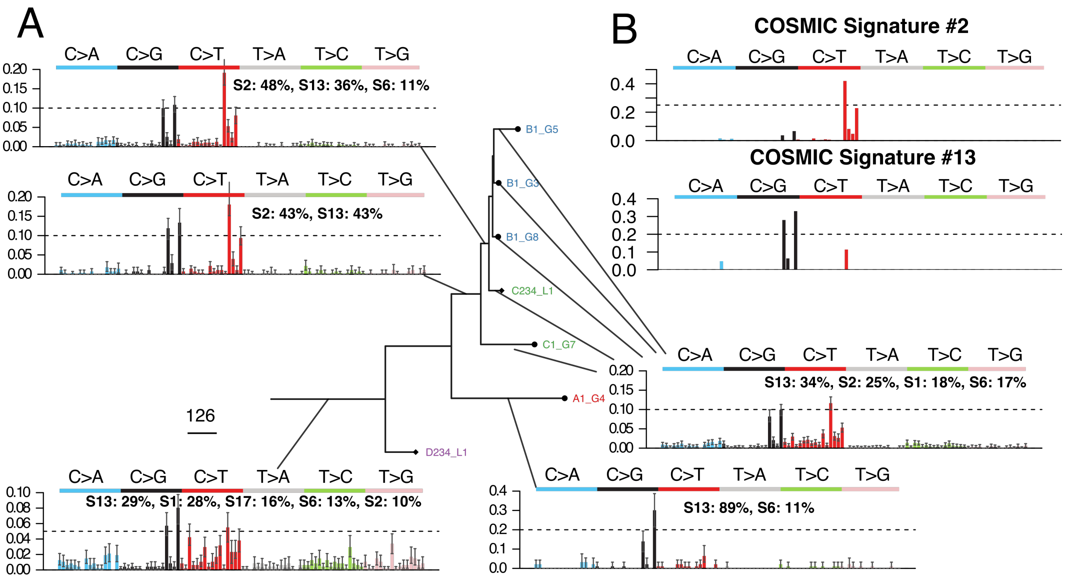


Figure 4.15: Activity of APOBEC associated mutational signature identified in C549. A) The trinucleotide distribution of mutations associated with each edge of the tree. It can be seen that the entire subclonal mutations are dominated by the two APOBEC associated signatures S2+S13. B) The two APOBEC associated COSMIC signatures S2+S13.

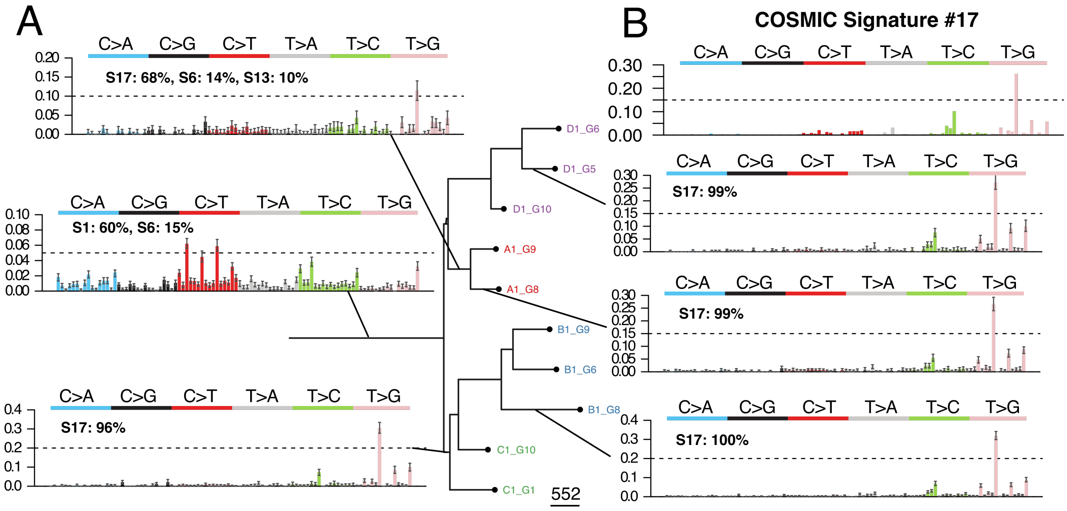


Figure 4.16: S17 mutational signature identified in C561. A) The trinucleotide distribution of mutations associated with each edge of the tree. It can be seen that all subclonal mutations are associated with S17. B) COSMIC signatures S17.

immune cytolytic activity. However, the difference observed by these authors were much less pronounced and it is unclear if an immunoreaction would provide an adequate explanation for the observations made here. In summary, the analysis of the mutational processes active in this CRC provides, to my knowledge, the first example of APOBEC mediated mutagenesis in a CRC, suggesting that it might indeed occur in a very small subset of CRCs.

Subclonal activity of S17 in multiple CRCs Consistent with previous findings by Cross et al. (2018), evidence for a subclonal increase of signature S17 activity in a subset of cases was identified. Such a pattern was observed in $11/23$ MSS cases[15] (see Figure 4.14). Interestingly, S17 appeared to be the dominant mutational process in two of these cases (C555 and C561) and more than 75% of variants were attributed to the activity of S17 in these. In C555, a small but significant contribution of S17 to clonal variants was also observed, whereas no such variants were found in C561. This suggests that a stable activation of an underlying mutational process occurred around or shortly before the MRCA of these tumours arose. The presence of S17 associated mutations in different regions and at different time-points of the tumour evolution is demonstrated by the tree-based analysis of mutational signatures for C561 shown in Figure 4.16A. Here a clear pattern of T>G substitutions associated with S17 (Figure 4.16B) is dominant on various edges of the reconstructed phylogenetic tree.

Interestingly, the aetiology of S17 is not perfectly understood. Christensen et al. (2019)

[15] C524, C532, C537, C539, C543, C547, C551, C552, C555, and C561

have shown that treatment with 5-FU chemotherapy induces DNA damage similar to S17 in vitro and in vivo. However, according to the available information on the treatment of these patients, no adjuvant or neoadjuvant chemotherapy with 5-FU or indeed any other drug took place. For this reason, the alternative explanation that the presence of oxidised deoxyguanosine triphosphate nucleotides in the nucleotide pool, as suggested by Tomkova et al. (2018), seems to be a more likely explanation. Other publications have also suggested that base excision repair (BER) might play a role in the repair of mutations caused by S17 (Pich et al. 2018). I evaluated this hypothesis but found no somatic mutations causing defects of the BER pathway in any of the MSS cases assessed here. I further tested if somatic mutation of K-Ras, p53 or PI3KCA were associated with the presence of subclonal S17. This indicated a weak association with K-Ras ($p = 0.036$, Fisher's Exact Test) but not with p53 ($p = 0.42$, Fisher's Exact Test) or PI3KCA mutations ($p = 0.66$, Fisher's Exact Test). Still, after adjusting these p-values for multiple hypothesis tests, no significant effects remained.

In summary, no clear explanation for the presence of S17 in a subset of the CRC could be found. Through the usage of extensive information contained in the multi-region sequencing data, it was however possible to demonstrate that the presence of this process is stable in time and space. For this reason, it appears more likely that the underlying cause is a cell-intrinsic property and not a local or transient process (i.e., chemotherapy or local micro-environmental effects).

4.3.5 Analysis of Chromatin Accessibility Using ATAC-seq

Alterations of the chromatin structure have been suggested to play an important role as non-genetic drivers in carcinogenesis (Flavahan, Gaskell, and Bernstein 2017) and the development of metastasis (McDonald et al. 2017). Nevertheless, such epigenetic driver alterations have not been studied extensively in colorectal cancers. Studies of cancer cell lines (Akhtar-Zaidi et al. 2012) and primary tissues (Johnstone et al. 2020; Corces et al. 2018) have provided some insight into the role chromatin alterations might have. Still, little is known about the epigenetic heterogeneity existing within human malignancies (Black and McGranahan 2021) and its relationship with the genetic diversity that occurs during the expansion of tumours. The largest pan-cancer study of chromatin alterations in human cancers as part of the TCGA project (Corces et al. 2018) lacked normal controls. The results of this study are for this reason likely dominated by the signal of the 'tissue of origin' and unable to unveil somatic changes of the epigenome.

As part of this project, $1,109$ chromatin accessibility profiling using ATAC-seq of single-glands and bulks obtained from 8 adenoma and 24 carcinoma was performed (Figure 4.1G–H). Additional ATAC-seq on patient-matched normal bulk tissue and normal crypts was used to generate normal reference data to distinguish actual somatic alterations from signals of the tissue of origin (Figure 4.1H). During the initial analysis of somatic mutation detected in the 30 CRCs of the EPICC cohort, an excess of truncating mutations of chromatin-modifier genes in MSS CRC was found (Figure 4.9B–C). These recurrent mutations of CMG hint that epigenetic alterations might have an essential role in the evolution of CRCs. A relationship between these epigenetic alterations and the selection of somatic CMG defects might exist. Using the chromatin accessibility profiling data from the EPICC cohort, I identified recurrent somatic alterations of chromatin-accessibility and demonstrated that these were primarily late clonal events.

4.3.5.1 Recurrent Changes of Chromatin Accessibility

To identify these recurrent alterations of chromatin accessibility in the analysed CRCs, a reference set of open chromatin regions (i.e., ATAC-seq peaks) was created first. In brief, peaks were called in the ATAC-seq data of individual tumour regions using *MACS2* (Zhang et al. 2008) and merged across regions and patients using an approach similar to the one used by Corces et al. (2018). The number of reads covering each of these peaks was obtained for all samples. Due to the insufficient number of reads obtained from the low complexity single-gland ATAC-seq libraries (see Figure 4.2), statistical analysis was challenging and data from glands were combined instead. Summarised, reads obtained from all glands with a purity $\geq 40\%$ of each tumour were pooled to generate synthetic ‘megabulks’. Likewise, reads from all normal reference samples were merged to create a ‘pool of normals’. A clear association of independently measured gene expression values in normal tissue with the promoter accessibility in this ‘pool of normals’ was evident (Figure S.54, page 285), supporting that the generated reference was indeed representative of the chromatin accessibility of healthy colorectal crypts.

Statistical analysis of the number of reads in the megabulks compared to the pool of normals allowed the identification of peaks with significantly altered accessibility (see Figure 4.17A, see Methods section for details). I conducted this analysis separately for peaks overlapping putative enhancers ($9,706$) and promoters ($17,885$) across the carcino-

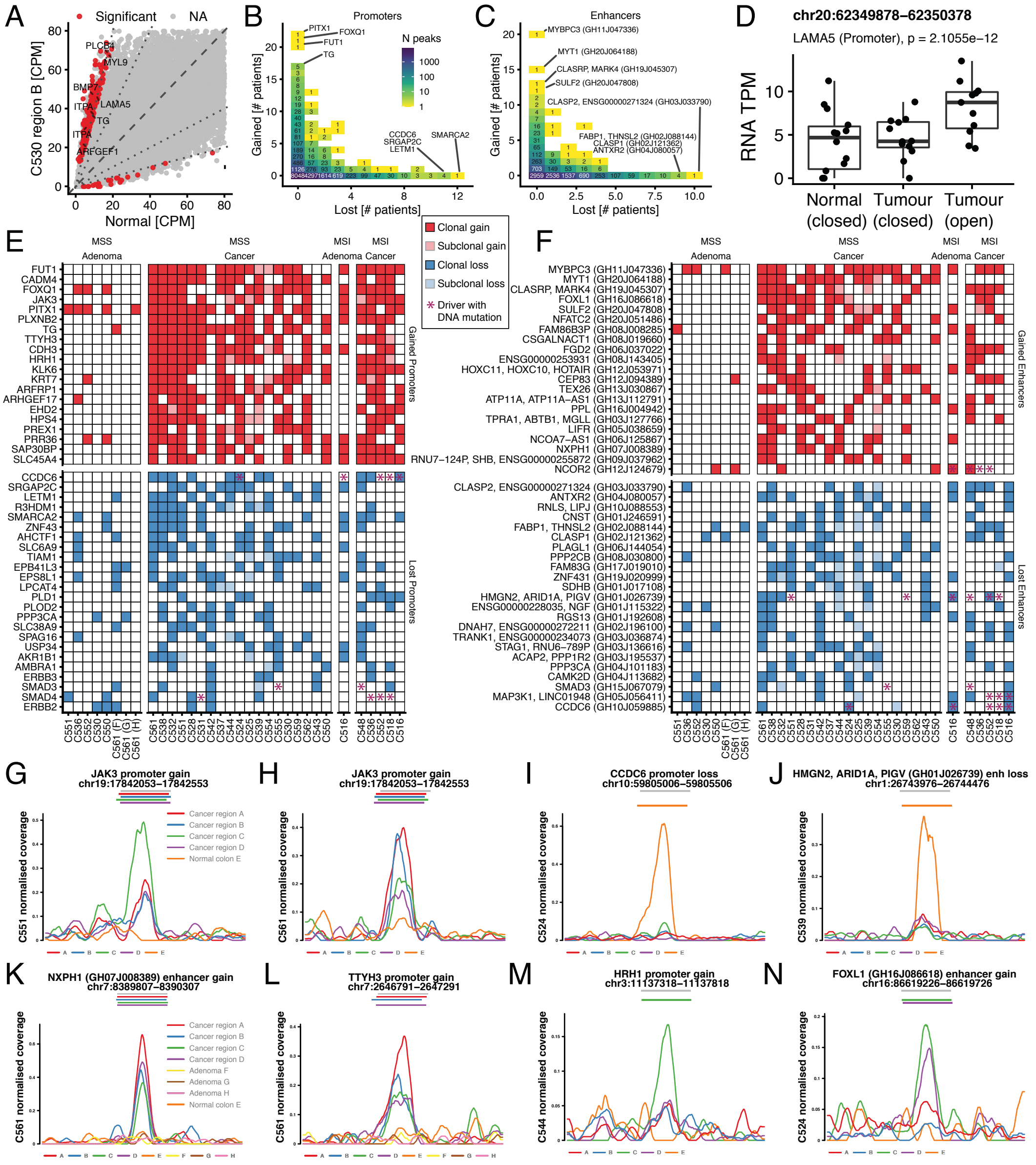


Figure 4.17: Recurrent somatic chromatin accessibility changes identified in the EPICC cohort. A) Recurrent changes were identified for each cancer and adenoma by comparing pooled reads from pure single-glands ATAC-seq data against a pool of normal glands. Significant differences are highlighted in red. Values shown are CPM to normalise for the total number of reads in the samples. Comparison of these somatic chromatin accessibility alterations (CAAs) across patients identified recurrent CAA at promoters (B) and enhancer regions (C). D) Gene expression analysis of matched RNA-seq data from the same cases showed concordant changes in ≈ 16% of promoter CAA and ≈ 12% of enhancer CAA. E) and F) The 20 most recurrently gained and lost CAA at promoter and enhancer regions respectively. G–N) Tracks of representative somatic CAAs.

mas and adenomas in the cohort. I then assessed how frequent losses (e.g, closing) and gains (opening) of chromatin accessibility (CA) occurred for each peak. Here a subset of promoters (Figure 4.17B) and enhancers (Figure 4.17C) were identified that showed recurrently altered CA in the carcinomas. Notably, losses of CA were more frequently observed than gains ($p < 10^{-12}$ for promoter and enhancer), but the recurrence of these CA losses was lower compared to CA gains. A total of 93 gained vs 8 lost promoter-associated CAs ($\chi^2 = 70.1$, $p < 10^{-12}$) and 8 vs 1 lost enhancers associated CA ($\chi^2 = 4.00$, $p = 0.0455$) occurred in ≥ 10 cases. The excess of promoter losses might, at least partially, be explained by the problem of identifying rarely opened regions (i.e., peaks) in the impure ATAC-seq signal obtained from the tumours. Despite this, the data suggested that a pattern of increased chromatin accessibility in specific genomic regions exists in CRCs compared to normal colorectal tissue.

Still, the effect such differential CA might have, are especially for distant enhancer elements unclear. For this reason, I also explored whether changes in CA were associated with altered gene expression. In brief, matched RNA-seq data[16] were used to identify concordant expression changes of promoter adjacent genes and previously identified enhancer-gene pairs reported in the GeneHancer database (Fishilevich et al. 2017). This analysis demonstrated that $\approx$ 16% (92/586) of recurrently ($\geq$ 20%) altered promoters and $\approx$ 12% (29/244) of recurrently altered enhancers were accompanied by corresponding gene expression changes at a FDR of 1%. A representative example of such a recurrent CAA associated with differential gene expression, an opening of a *LAMA5* promoter, is shown in Figure 4.17D. Additional examples can be found in Figure S.55 (page 286).

Despite the limited power to detect changes in chromatin accessibility and gene expression, this analysis revealed a fairly large number of recurrent CAAs, many of which have been previously identified to have a role in the development of CRC. Examples of this include LAMA5, which was suggested to be a potential biomarker with prognostic value due to its ability to promote growth of liver metastasis and induce angiogenesis (e.g., Pyke et al. 1994; Hlubek et al. 2001; Bartolini et al. 2016; Galatenko et al. 2018; Gordon-Weeks et al. 2019) or TNNT1, which appears to be implicated in the induction of increased proliferative and invasive capabilities of CRC cells (e.g., Chen et al. 2020; Hao et al. 2020).

Together these observations could be explained by a model of epigenome evolution in

[16] This part of the EPICC project was led by Jacob Househam, who shared the processed RNA-seq data with me. I used these to identify differentially expressed genes associated with identified CAA.

which a relaxed control of the cellular state causes a drift away from the epigenomic status of the tissue of origin. In this case, the selection of specific alteration or changes to cell-intrinsic regulation networks could be responsible for the recurrent changes of chromatin accessibility. The difference between the number of gains and losses might arise due to a small number of mechanisms that have to be lost to release cells from proliferation control. Many more mechanisms might instead exist that modify the interaction with the cellular environment and are ultimately beneficial to the growth of the tumour.

Validation with data from ENCODE and TCGA The ATAC-seq peaks called in the dataset showed a clear overlap with regions of open chromatin identified in CRCs (Corces et al. 2018) and healthy normal colon epithelium (ENCODE Project Consortium 2012). The data from both of these projects were obtained and reprocessed with the same analysis pipeline used here to reduce potential biases. Since no matched normal tissue samples were available from the TCGA study (Corces et al. 2018), I instead determined whether the average CA across cases were differenced for the peaks I identified here (see Figure S.57, page 288).

Reassuringly, this analysis confirmed that a strong correlation between the average CPM of peaks in the normal tissue samples from the EPICC cohort and those from the ENCODE project existed (top left, Figure S.57, page 288). This correlation existed for all as well as recurrently altered CAAs. This demonstrates the consistency of the single-crypt ATAC-seq profiles analysed here with the normal bulk colon tissue CA data from ENCODE. Likewise, a similarly strong correlation was observed between CA data from the cancer samples of the EPICC cohort and the TCGA project (bottom right, Figure S.57, page 288). In contrast, significant differences in the average CPM between different sample types (i.e., between normals and tumours) did (exist bottom left, Figure S.57, page 288). Together these observations support that the analysis described above was able to identify genuine recurrent somatic CAAs from the EPICC ATAC-seq data alone.

Assessment of the impact of CNAs During the detection of differentially accessible genomic regions, changes of copy-number states were not taken into account. Upon review of the CA data, a weak, but significant relationship between CNAs and the number of gained or lost CAAs was observed. Of the recurrent CAAs reported in Figure 4.17E&F, the majority of CAAs (95.5%) did not show a significant association with the relative copy-number of the locus across patients (see Figure S.66, page 293). In light of the relative change in coverage ratios caused by CNAs at the given sample purities compared to the expected change in read numbers due to CAAs, this is generally not surprising.

Identification of subclonal chromatin accessibility alterations The above analysis of somatic CAAs in primary CRCs and the associated gene expression changes demonstrated the general importance of epigenetic alterations in CRC. I next sought to determine if these alterations tended to be clonal or subclonal. This analysis was significantly complicated by the generally low complexity of the obtained ATAC-seq libraries, the very variable sample purities, and the differences of TSSe of samples.

Only after realising that, for unknown reasons, the signal of cancer cells was underrepresented in the signal obtained from ATAC-seq samples (i.e., the ‘apparent sample purity’) the variability of signal across samples was explainable. To measure this apparent purity of the ATAC-seq data, clonal SNVs identified during the WGS of single-glands were determined, and their frequency in the ATAC-seq libraries was used to obtain ML estimates of their purity. In order to estimate the amount of overdispersion associated with the measurements, samples obtained from the same region of the tumour were treated as ‘biological replicates’ in *DESeq2* (Love, Huber, and Anders 2014) to fit a regression to account for the identified confounding factors.[17] Based on this regression, sites with a significantly different ATAC-seq signal in individual tumour regions were identified. The analysis of these subclonal variants was focussed on the 20 most recurrent CAAs of each type — i.e., gained and lost enhancer and promoters, shown in Figure 4.17E (promoter) and 4.17F (enhancer) as well as promoter and enhancers of CRC driver genes from the IntOGen database (Martínez-Jiménez et al. 2020).

This analysis demonstrated that $\approx 92\%$ ($782/854$) of the recurrent CAA were consistently altered in all analysed regions of affected tumours. In general, this observation was consistent with the assumption that most of these CAA are bona fide somatic alterations arising early during tumour evolution and not, for example, alterations arising due to plasticity from different local microenvironments. Very similar observations are common for somatic driver mutations, of which the overwhelming majority are present clonally. Representative examples of somatic CAAs are shown in Figures 4.17G–N, these specifically show examples of clonal promoter gains (Figure 4.17G,H,L) and a loss (Figure 4.17I), a clonal enhancer gain (Figure 4.17K) and loss (Figure 4.17J) as well as a subclonal promoter

[17] This regression analysis was performed by Claire Lynn. A similar approach prior to the estimation of ‘apparent sample purities’ was done by myself.

(Figure 4.17M) and enhancer gain (Figure 4.17N).

Examples of recurrent somatic CAA Among the CAA present in $\geq 20\%$ of cases were multiple promoter and enhancers associated with known CRC driver genes. One such event was the loss of *CCDC6* promoter accessibility, observed in 11/24 cases, and the loss of accessibility of the GH10J059885 enhancer also associated with *CCDC6*. In general, mutations of CCDC6 are rare in CRC, but the loss of *CCDC6* expression has previously been suggested to play a role in the development of CRCs (Thanasopoulou et al. 2012). This previously identified loss of *CCDC6* in CRC is consistent with the loss of promoter CA at this gene, and the loss of an associated enhancer might play a functional role in this. An example of a *CCDC6* loss in one case from the cohort is shown in Figure 4.17I.

Other examples of CAAs affecting known CRC driver genes were the loss of CA around *SMAD3* (5/24) and *SMAD4* (6/24) promoters and a *SMAD3* associated enhancer (see Figure 4.17E&F). Both of these genes are involved in the regulation of the TGF-β signalling pathway, which is known to be deregulated in many tumour entities (Fleming et al. 2012). Consistent with this, the loss of genes from the SMAD family has previously been associated with tumour invasiveness and poor prognosis in CRCs (Fleming et al. 2012; Sodir et al. 2006; Xie et al. 2003; Isaksson-Mettävainio et al. 2006). Similar CAAs affecting a CRC driver gene were alterations of enhancers involved in the regulation of *ARID1A* (loss in 7/24 carcinoma and 1/8 adenoma, Figure 4.17J), *MAP3K1* (loss in 6/24 carcinoma) and *NCOR2* (gain in 6/24 carcinoma). Summarised, this suggests that a subset of these CRC drivers were also affected by CAA in addition to somatic mutations. Profiling of somatic mutations alone will miss such important non-genetic aberrations.

Additionally, several highly recurrent and potentially interesting CAAs were identified. Among these was the increase of *JAK3* promoter accessibility, a kinase thought to play a role in CRC oncogenesis (Lin et al. 2005), which occurred in 16/24 cases (Figure 4.17G&H). Other examples were the frequent opening of two *FOXQ1* promoters (21/24 and 7/21) an oncogene frequently overexpressed in CRC with angiogenic and antiapoptotic effects (Kaneda et al. 2010; Peng et al. 2015), the opening of a *LAMA5* promoter (12/24), a gene that appears to promote the growth of liver metastasis and angiogenesis (e.g., Pyke et al. 1994; Hlubek et al. 2001; Bartolini et al. 2016; Galatenko et al. 2018; Gordon-Weeks et al. 2019) or gain of *TBX20* a gene which could potentially play a role in angiogenesis (Meng et al. 2018). Among the CAA present in $\geq 20\%$ of cases, multiple promoter and enhancers associated with known CRC driver genes were also found to be present.

Presence of recurrent somatic CAA in adenomas To elucidate, when the identified CAAs arise during the development of carcinomas, the presence of 235 recurrent CAAs ($\geq 20\%$ of cases) in the 8 profiled adenomas was assessed. The hypothesis was that recurrent CAAs found to be frequently present in adenomas as a precursor lesion of CRCs would likely arise very early during tumorigenesis and rather be involved in the initiation of dysplastic growth. In contrast, recurrent CAAs found to be absent in all adenomas might instead be later arising events and more likely involved in the progression towards a malignant phenotype.

Of the 235 CAAs assessed (Figure 4.17E&F), only 32 (i.e., $\approx 14\%$) were found to be present in adenomas. This suggests that most of these epigenetic alterations arise relatively late during carcinogenesis and that some of these might play an important role in the malignant transformation of tumour cells. This has striking similarities to the observations by Cross et al. (2018) who found that somatic driver mutations explain the adenoma-carcinoma transition rather poorly, but that the — potentially punctuated — accumulation of CNAs appear to mark this transition. Similarly, most of the CAAs identified here appeared to be accrued during the transition from adenomas to carcinomas. This highlights the importance deregulation of the transcriptional machinery might have in this context. While certainly not all CAAs observed here are functionally important, and instead, the results of large-scale deregulation of the transcriptional program, detailed analysis and validation of some of these might be worthwhile.

Further, while most recurrent CAA were found to either be present in all or no region of the carcinomas, some of these did indeed show evidence of being confined to one region of the tumour (see Figure 4.17E-F). Still, the exact reason for why these arise and if they confer a selective advantage to the affected cells is unclear. To draw a comparison to somatic alterations, these could either be bona fide epigenetic driver alterations or, like most SNVs, be acquired as passenger mutations during the tumour expansion.

Role of epigenetic drift Epigenetic alterations, including DNA methylation and chromatin modifications, have a fundamental role in the regulation of cellular identity in complex multi-cellular organisms (Atlasi and Stunnenberg 2017). Originally used as an abstract concept to describe the link between the observable genotype and phenotype (Waddington 1942), the term epigenomics is now used to describe the entirety of concrete heritable non-genetic mechanisms that control the expression of genes. It is widely recognised that epigenetic alterations play an essential role in tumorigenesis (Jones and Baylin 2007), but how such alterations are related to genetic intra-tumour heterogeneity is not well understood (Black and McGranahan 2021). The presence of many clonal highly recurrent CAA identified here (Figure 4.17E&F) suggest that these are stable alterations passed on over many generations and maintained in distant tumour regions (i.e., several centimetres apart) with potentially very different microenvironments. Still, it is unclear whether this is caused by the absence of epigenomic drift, stabilising selection or cell-intrinsic regulatory programs.

Here I will use the concomitant epigenetic (ATAC-seq) and genetic (WGS) measurements on single colorectal glands from different tumour regions available in the EPICC study to characterise the relationship between epigenetic and genetic diversity. Initial exploratory analysis showed that the low number of reads, differences in purity, and TSSe might complicate the interpretation of the obtained measurements. The application of classic phylogenetic methods (Pagel 1999; Blomberg, Garland, and Ives 2003) used to explore the relationship between traits and genetic distances provided little insight into the 'phylogenetic signal' present in the per-loci CA measurements (e.g., Figure S.64, page 292). Given the limitation to obtain site-specific information due to the low number of reads per sample, this is not entirely surprising.

Still, the 'global distance'[18] between ATAC-seq measurements of samples obtained from the same region compared to those from different regions was generally smaller, suggesting the presence of general epigenetic intra-tumour heterogeneity. This signal was also present after accounting for differences in the TSSe and the total number of reads through regression (Figure S.56, page 287). In order to test whether these residual differences were significantly associated with the region labels, an ANOVA was conducted. As the pairwise comparisons between samples are not statistically independent, a Monte Carlo method was used to estimate the expected distribution of the F-statistic under the null hypothesis of no differences across groups. This permutation approach is equivalent to the PERMANOVA described by Anderson (2001). The analysis showed that a significant relationship between the type of region label pairs (i.e., between and within regions) and the 'global epigenetic distances' generally existed (Figure S.58, page 288). The observed coefficients implied that distances within regions were systematically smaller (Figure S.59, page 289). This result is compatible with both, epigenetic plasticity due to micro-environmental factors and

[18] Euclidian distance of coverage across all peaks.

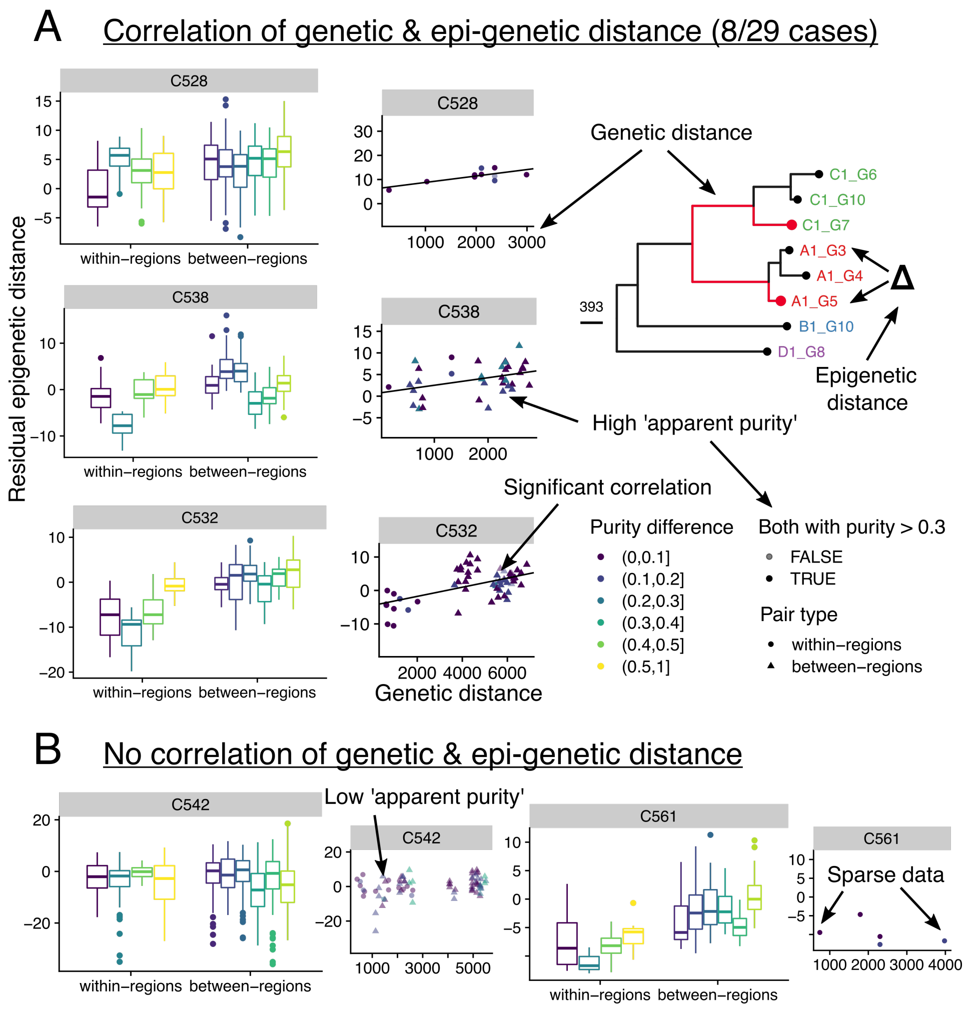


Figure 4.18: Relationship between global epigenetic and genomic distances. A) The distances between and within regions (left) and correlations with the genetic distance (right). B) Cases in which no correlation with the genetic distances existed data were often from low purity samples or sparse.

epigenetic drift in parallel with the genome during the clonal expansion, but demonstrates pervasive epigenetic heterogeneity within tumours.

To gain further insight into the relationship between genomic and epigenomic diversity, the correlation between the two was tested explicitly. For this test an extension of the

Mantel Test (Mantel 1967), similar to methods used by others (e.g., Legendre, Lapointe, and Casgrain 1994; Manly 1986), was used to elucidate whether the genomic distances encoded in the reconstructed phylogenies were significantly associated with the 'global epigenetic distances'. In brief, a linear regression of both measurements, the cophenetic distances between samples in the tree and their residual 'global epigenetic distances', was performed separately for each patient. The same regression was performed on data from randomly permutated trees. From these permutated datasets the significance of the observed coefficients was then determined. In a subset of cases (8/29) a significant positive association of the genetic and epigenetic distance of sample pairs was detected (Figure 4.18A and Figure S.61, page 290). In many of the remaining cases in which no relationship between the two measurements was observed, the available data were either sparse, or the 'apparent purity' of samples was low. In these cases, the lack of power did likely not allow to uncover such relationships in the first place (Figure 4.18B and S.65 on page 292).

While this analysis does not account for the relationship between genetic distances and region labels, as both of these are highly correlated. It still suggests that epigenetic drift might be a reasonable explanation for the observed 'global epigenetic distances' between samples. Still, it has to be noted that the data available for this analysis were imperfect, and future studies should be conducted to confirm these conclusions.

Conclusions Summarising, these reported findings show how information on somatic mutations alone might provide an incomplete picture of the alterations driving cancer evolution. A better understanding of this class of alterations is undoubtedly required to improve our understanding of why these alterations arise and what their specific effects are. The preliminary analysis of global differences of the CA across patients supports that these arise through epigenetic drift, but if this also applies to recurrent CAA or whether these arise 'punctuated' is unclear (e.g., from an altered epigenetic program) and should be explored in future.

4.3.5.2 Analysis of TF Binding Signatures

One mechanism that might cause such deregulation of focal chromatin accessibility are global changes in TF activities. TFs are proteins that bind to specific sequences of DNA and regulate the transcription rate of surrounding genes or the higher-order structure of the chromatin. Due to their ability to reduce the accessibility of the surrounding chromatin, the binding of TF to the DNA can leave a footprint in the signal observed from assays like

ATAC-seq. Likewise, caused alterations of the surrounding chromatin structure can lead to increased or decreased accessibility of the chromatin structure around a TF binding site.

In the following, I will present an analysis of these accessibility signals around predicted TF binding sites. In brief, I predicted binding sites for 870 known human TF binding motifs (Weirauch et al. 2014) and included additional experimentally determined binding sites from the ENCODE project (Dunham et al. 2012). I then split these predicted TF binding sites based on their distance to the closest TSS into those proximal to a TSS ('pTSS', $d \leq 2,000\,\mathrm{bp}$), close to a TSS ('cTSS', $2,000 < d < 10,000\,\mathrm{bp}$) and distal to a TSS ('dTSS', $d > 10,000\,\mathrm{bp}$). Each set was further divided into those overlapping a called peak ('oPeak') and those not overlapping a called peak ('nPeak').

For each of these six sets of TF binding sites, signals were calculated from the average number of insertions across an interval of ±1000bp around the centre in all tumour or normal samples. These were adjusted for differences in the insert-size distribution as well as the total number of reads, followed by the subtraction of the observed 'background signal' (see Methods section for details). Figure 4.19A shows the average number of insertions (CPM) in cancer samples from case C542 for different insert sizes (y-axis) and positions relative to the centre of the TF binding site (x-axis). The resulting distribution summarises the insertion distribution or accessibility around a given TF motif in the genome. Identical distributions were calculated for all collected normal samples (Figure 4.19B). In order to identify potential differences between tumour tissues and normal samples the differences of these normalised signal of each sample and the average signal in all normal samples was determined (Figure 4.19C).

To provide a summary of the activity of all TFs in the genome across samples, the integral over the central region of these TF fingerprints (see Methods) was calculated. On these summary statistics of the TF activity in individual samples, linear regression with sample purity estimates (per patient) and TSSe as potential confounding variables was performed. Differences in the TSSe relative to the normal reference explained a large amount of the variability in the observed signals across samples, with a smaller part explained by differences in sample purity. In Figure 4.19D, some of the coefficients of the purity variable in different patients (columns) and TFs (rows) are shown as a heatmap. This heatmap provides a high-level summary of the average purity associated differences of TF activity in individual tumour tissues compared to the normal tissue background. Here, positive coefficients

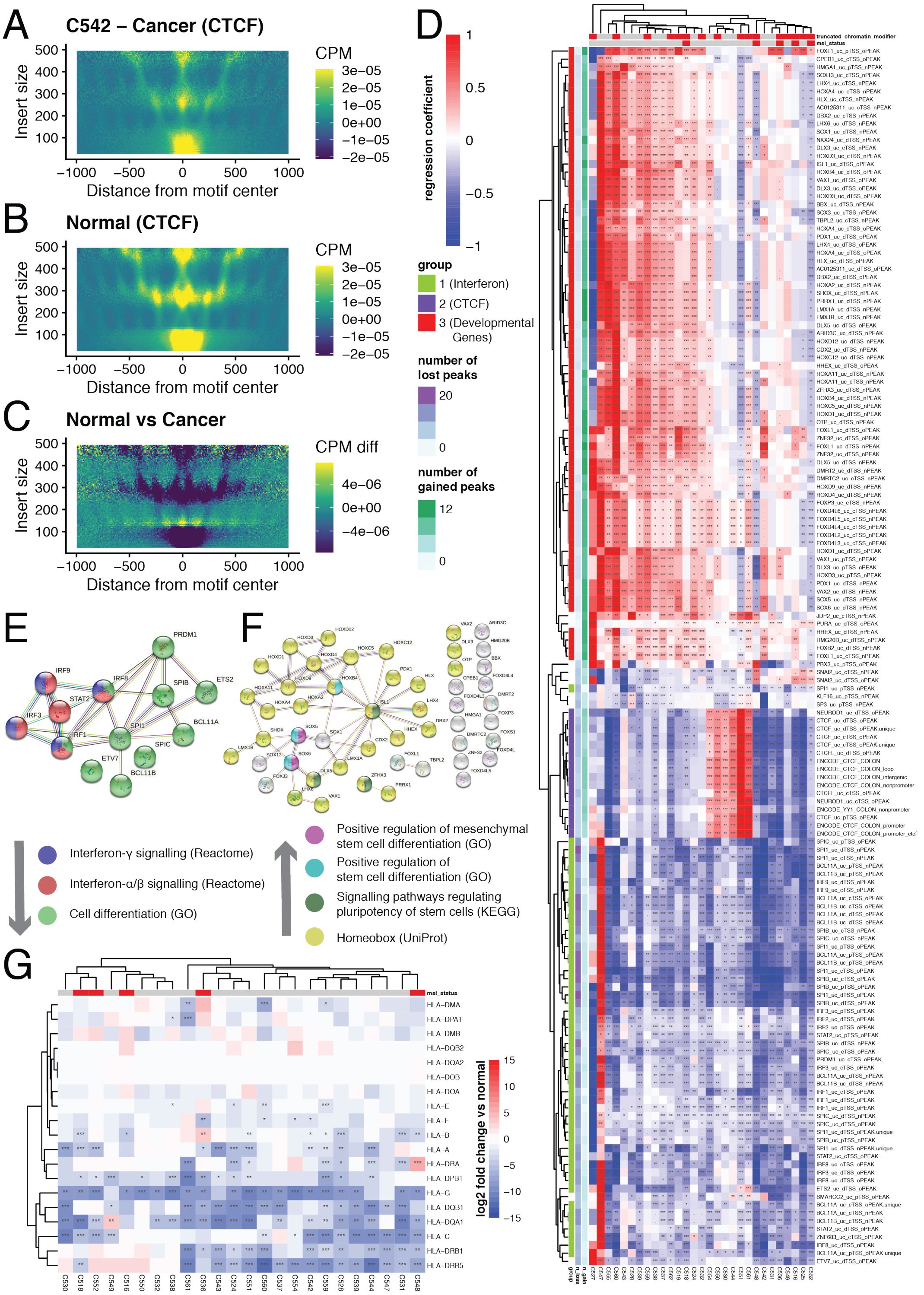


Figure 4.19: Recurrent epigenetic changes. A-B) The average distribution of insertions around the centre of the TF bindings sites in a tumour and normal samples respectively. C) The difference between the two. D) Regression was applied to data from C in each sample to identify changes associated with purity. Clear clusters were identified. E-F) Pathways overrepresented in the clusters. G) Shows loss of HLA expression in multiple cases.

mean that an increase in the number of cancer cells in the sample would be expected to cause an increased amount of signal around the TF binding sites and vice versa. In Figure 4.19D, the top 50 TFs that showed most frequently a positive and negative correlation of the signal with purity across patients are shown. In this context, only one TF of groups with a largely overlapping (i.e., $\geq 50\%$ within 100bp) set of predicted binding sites were retained (see Figure S.60, page 289). A separate test confirmed that similar values were observed for unique binding sites from these groups of TFs (Figure S.62, page 290). Still, given that binding sites might co-occur in the same overall regions, but at a larger distance from each other, this is not necessarily conclusive.

From the analysis of the accessibility signal across TF binding sites shown in Figure 4.19D, three clusters were apparent. These are shown by colours on the site of the heatmap. The first cluster (marked in green), showed an overall loss of signal in the majority of cases and appeared to contain several TF from the IRF family. These TFs have previously been identified to act as TSGs and loss of various IRF gene expressions have been identified in a variety of tumour entities (Tamura et al. 2008; Yanai, Negishi, and Taniguchi 2012). These observations were also confirmed by a pathway analysis (Figure 4.19E), which showed that TF involved in Interferon-γ signalling ($\text{FDR} = 4.2 \cdot 10^{-6}$), Interferon-$\alpha/\beta$ signalling ($\text{FDR} = 3 \cdot 10^{-8}$) and cell differentiation ($\text{FDR} = 5 \cdot 10^{-5}$) were overrepresented in this cluster. Notably, a particularly strong correlation with the purity of this cluster of TFs was observed in MSI cancers, with the cluster of cases on the right side of the heatmap being primarily composed of these ($p = 0.012$, Fisher's Exact Test). Consistent with the downregulation of anti-tumour immunity suggested by these observations, the analysis of gene expression data[19] from this set of patients showed a general pattern of HLA gene expression loss in many of the patients (Figure 4.19G).

A second cluster (marked blue) was primarily composed of CTCF, CCCTC and YY1 TFs. These TFs largely bind to similar genomic regions and are involved in the regulation of higher-order chromatin structures, insulation of enhancer-promoter interactions, and transcriptional regulation (Kim et al. 2007; Ghirlando and Felsenfeld 2016; Ong and Corces 2014; Gong et al. 2018; Wendt and Peters 2009). Loss of CTCF has been suggested to hamper the repair of double-strand-breaks (Lang et al. 2017) and cause alteration of gene expression due to atypical enhancer-promoter interactions (Lupiáñez et al. 2015; Hnisz et al.

[19] Kindly provided by Jacob Househam.

2016). A marked increase in CTCF signal was observed in a subset of cases, whereas loss of signal appeared to be generally more common in the remaining cancers. While the reasons for this CTCF loss is unclear, CTCF alterations, both losses and gains, were previously reported in bulk CRC samples (Fang et al. 2020). CTCF binding sites appear to be mutational hotspots and mutations of CTCF common in CRC (Katainen et al. 2015). Whether this might sufficiently explain the global loss of CTCF associated signal observed here is unclear, but this would be consistent with the findings that monoallelic CTCF loss predisposes to the development of cancer and acts as a putative haploinsufficient TSG (Filippova et al. 1998; Ohlsson, Renkawitz, and Lobanenkov 2001; Kemp et al. 2014; Marshall et al. 2017). Alternative explanations include global changes of the chromatin structure that lead to a general change of the ATAC-seq insertion distribution in the genome or inter-individual variation of TF binding (Phillips and Corces 2009).

A third cluster (marked red) mainly contained TFs from the HOX, FOX and SOX families that are involved in cell differentiation and developmental processes (Figure 4.19F): 'positive regulation of stem cell differentiation' (GO, $\mathrm{FDR} = 2.5 \cdot 10^{-4}$) 'mesenchymal stem cell differentiation' (GO, $\mathrm{FDR} = 9 \cdot 10^{-4}$), 'signalling pathways regulating pluripotency of stem cells' (KEGG, $\mathrm{FDR} = 0.047$), 'homeobox' (UniProt, $\mathrm{FDR} \leq 10^{-12}$), 'developmental protein' (UniProt, $\mathrm{FDR} \leq 10^{-12}$). In the majority of cases, higher signal levels were observed in the tumour cells, as indicated by the positive and significant coefficients of purity. This suggests that the reactivation of developmental genes in colorectal cancers might be an important step in tumorigenesis.

Fitting of the purity coefficients on samples from individual tumour regions generally revealed similar patterns in samples from different regions, but some region-specific effects did appear to exist. For example, a lower signal from CTCF binding sites at looping regions was observed in all samples from C543, but a higher signal was found to exist in one region from C543 (Figure S.63, page 291). While this might suggest a genuine difference in TF binding, region-specific biases might also explain these observations.

In general, this analysis of the average ATAC-seq signal around TF binding sites of a large number of human TFs suggests the presence of a general change in the global patterns, potentially driven by the deregulation of larger regulatory networks. Still, whether the observed signals are caused by large-scale changes of the chromatin surrounding individual TF binding sites, the differential binding of the corresponding TF themself or other mechanisms is not clear. This might even vary for the different TFs. Likewise, the nature of the analysis means that averages i) over different binding sites and ii) across cells were taken. For this reason, the observations could be caused by changes of a subset of binding sites or even temporary changes in a subset of cells.

4.4 Discussion

Here I have presented the results from a multi-region sequencing study of 30 CRCs in which concomitant profiling of single-glands with ATAC-seq, RNA-seq, and WGS was performed. In CRCs, this approach provides an alternative to single-cell sequencing, which remains challenging due to the large amount of noise resulting from DNA amplification and sequencing (Gawad, Koh, and Quake 2016). Concomitant multi-omics profiling of single-cells is also still in its infancy.

By sequencing single-glands, which are generally assumed to be formed by a small and closely related stem-cell population, some insight into the relationship of epigenetic and genetic heterogeneity in the CRCs was possible. While limited by the quality of the generated ATAC-seq libraries, epigenetic drift was found to provide a reasonable explanation for global differences in CA of different glands in the same tumour. In cases with a sufficient amount of data of reasonable quality, positive correlations of genetic and epigenetic distances were frequently observed.

Across patients, recurrent focal chromatin accessibility alterations were identified. Despite their relevance in cancer development, the role of these epigenetic events in tumour evolution remains relatively poorly understood (Black and McGranahan 2021). Some recent pan-cancer studies of chromatin accessibility across cancer types have primarily focused on how these are defined by the ‘cell of origin’ and the corresponding relationships with gene expression. A few studies that specifically focused on somatic alterations in cell lines (Akhtar-Zaidi et al. 2012) and large-scale chromatin structures (Johnstone et al. 2020) in CRC have been conducted. These highlighted the importance alterations of the chromatin structure have in this disease and the identified recurrent CAAs I reported here complement these previous findings. Indeed, profiling of somatic mutations alone might miss such important alterations in CRC and hence provide an incomplete picture of disease evolution.

This is especially important as mutations of common driver genes appear to only insufficiently explain the adenoma-to-carcinoma transition (Cross et al. 2018). Here, in addition to the selection of copy-number alterations, changes of the epigenome might provide a better explanation of CRC evolution towards increased malignancy. Likewise, some of the identified CAAs might, unlike genetic alterations, be able to predict the development of a metastatic ability in a subset of CRCs. The association of various CAAs with the clinical outcome of patients with be assessed as part of the follow-up of the study. Last but not least, the presence of CAA could provide an explanation of carcinoma in which no or few of the classic CRC driver genes (APC, K-Ras, and p53) were mutated. Indeed, cases that only harboured mutations of one of these three genes[20] showed many of the recurrent CAAs identified here. Larger studies might help in identifying these potential alternate pathways of CRC evolution.

I also conducted a comprehensive analysis of subclonal driver mutations in the cohort. In line with similar studies in smaller sets of patients (Sottoriva et al. 2015; Kim et al. 2015; Uchi et al. 2016; Cross et al. 2018) and previous analyses of bulk WGS datasets (Williams et al. 2016; Williams et al. 2018b) very few subclonal driver mutations were found. Other, rarely clonally occurring putative driver mutations were found to be present, but in most cases, it was unclear whether these caused a substantial fitness effect. The only exception were subclonal PIK3CA mutations, which an analysis across cohorts suggests to be present sub-clonally at a frequency identifiable by multi-region sequencing in $\approx 20\%$ of cases (Figure S.49A, page 283).

Overall, how to interpret the observed inter-tumour heterogeneity from the single-gland sequencing study presented here was not obvious. While a dN/dS analysis suggested the presence of some amount of subclonal selection in the cohort, only very few obviously elongated edges, which would provide evidence of subclonal selection, were found to be present in reconstructed phylogenetic trees. In those cases where these did exist, it was unclear whether these could likewise have arisen from genetic drift. While many of the issues previously identified for single-bulk sequencing data (Caravagna et al. 2020) do not exist with the single-gland sequencing data analysed here, interpretation was only straightforward in one case of a subclonal activating K-Ras p.G12V mutation. Nevertheless, patterns of spatial variegation were observed in a subset of cases (Table 4.3) and if a significantly different outcome exists for these will be tested in the follow up of the study.

Last but not least, the preliminary analysis of accidentality sequenced cancer-adjacent and additionally sequenced normal crypts have, similarly to Lee-Six et al. (2019), revealed

[20] C528, C532, C551, and C562.

the presence of *pks*$^+$ *E. Coli* associated signature SBS88 in normal crypts. The mutation rate in these normal crypts was found to be $\approx 25\,\mathrm{y}^{-1}$ for SNVs and $\approx 1\,\mathrm{y}^{-1}$ for InDels. Further, significantly more somatic variants were found in tumour-adjacent normal crypts compared to distant normal crypts obtained from the same patients, hence suggesting the presence of a mutagenic effect of the tumour micro-environment, the presence of a very early arising field-defect or some other influence of the tumour microenvironment on the behaviour of adjacent cells.

Chapter 5

Assignment of LP-WGS Samples to Trees

5.1 Motivation

While the costs of WGS drastically decreased over the last decades and even outpaced Moore's law, the costs of resequencing a whole-genome at a coverage appropriate for the analysis of somatic variants in a tumour are still substantial (Schwarze et al. 2020). LP-WGS has proven to be a cost-efficient alternative (Rohland and Reich 2012), which can be used for the reliable detection of CNAs (Carter et al. 2012; Oesper, Satas, and Raphael 2014; Muzny et al. 2012; Baker et al. 2019) and structural variants (Zhang et al. 2018).

Due to these low costs of LP-WGS, this method is sometimes used for the screening of cancer libraries to derive CNA based estimation of samples purities before deep sequencing (Lohr et al. 2014). Identical to this, several LP-WGS (coverage $0.5\times - 1\times$) datasets were generated for the multi-region single-gland sequencing study described above to identify glands with high tumour cell content for deep WGS sequencing at a higher coverage of $\approx 30\times$ (Table 5.2). Of these, many high-quality LP-WGS samples were never sequenced at a higher coverage and instead used for the analysis of copy-number alterations (Figure S.24, page 273 and Figure S.23, page 272).

While LP-WGS and deep WGS samples obtained from similar regions of the tumour tended to show nearly identical CNA profiles, exceptions did exist. Where samples showed subclonal CNAs similar to those from other regions, the data supported spatial variegation (Sottoriva et al. 2015). Still, CNA data are inherently sparse, tend to overlap (Beerenwinkel et al. 2015) and lack independence for frequently selected alterations (Zack et al. 2013; Cross et al. 2018). Consequently, the ability of CNAs to resolve phylogenetic relationships is rather poor (Zeira and Raphael 2020). Specialised methods for the inference of phylogenies from CNAs data have been developed (e.g., Chowdhury et al. 2014; Letouzé et al.

2010; Schwarz et al. 2014; Zaccaria and Raphael 2021), but some inherent limitations of CNA data are hard to overcome.

Due to the significant level of noise currently used NGS methods exhibit, (Gerstung, Papaemmanuil, and Campbell 2014; Gerstung et al. 2012) as well as the low number of divergent sites in individual tumours (Kandoth et al. 2013), reliable detection of SNVs from LP-WGS is impossible (Xu et al. 2014; Zaccaria and Raphael 2021).

In the context of the multi-gland sequencing study reliable SNV calls from deep WGS data were available and used to reconstruct MP sample phylogenies (see Figure S.51, 284 and Figure S.50, page 283). While the number of informative sites in LP-WGS samples is very low, known SNVs can provide information on the location of individual samples. In the following, a simple and fast ML method to estimate sample properties (i.e., background noise and purity) and the position of LP-WGS samples within the phylogeny will be described. Simulated LP-WGS samples and subsampled deep WGS sequenced samples will be used to assess the performance of the method. After the reconstruction of LP-WGS trees for the EPICC cohort, these will be compared to the results from the analysis of CNAs to demonstrate that these generally support the position of LP-WGS samples within the trees.

5.2 Method

5.2.1 Inference of Phylogenies and Ancestral Characters

A phylogenetic tree T is a directed graph that consists of a set of vertices or nodes V and a set of edges $E = \{(s,t) : s,t \in V \wedge s \neq t\}$ connecting nodes. Here I will use a simple MP method for the inference of phylogenetic trees from the observed data, but in principle, any other method that constructs a graph in which character changes (i.e., mutations) can be mapped to edges could be used. This also applies to ‘clone trees’ reconstructed through the use of clustering methods (e.g., Miller et al. 2014; Roth et al. 2014; Dentro, Wedge, and Van Loo 2017) from potentially heterogeneous, bulk WGS samples (e.g, Noorani et al. 2020).

As described previously, the MP trees were inferred with the Parsimony Ratchet method (Nixon 1999) implemented in the *phangorn* package for R (Schliep 2011) using a minimum of 100, a maximum of 10^6 iterations and termination after 100 rounds without improvement.

Various methods can be used to estimate ancestral character states for each internal node (i.e., those that are not the root or a leaf node). Here the accelerated transformation

(ACCTRAN) algorithm (Fitch 1971; Farris 1970; Swofford and Maddison 1987; Schliep 2011) was used. From the results of this method, a list of mutations that were acquired (i.e., state $0 \to 1$) or lost (i.e., state $0 \to 1$) on each edge of the phylogeny were obtained.

5.2.2 Likelihood

From these, the set of mutations M_e for each edge $e \in E$ of the tree that are uniquely mutated on it are kept. The number of variant reads $y_{s,i}$ observed from a mutated site i sequenced at coverage $n_{s,i}$ in a sample s are expected to follow a binomial distribution:

$$y_{s,i} \sim Bin(n_{s,i}, p_{s,i}).$$

The expected success probability $p_{s,i}$ is then a function of the sample's purity, ρ_s, the number of mutated alleles $m_{s,i}$ in tumour cells, the total copy-number $c_{s,i}$ of the site i in tumour cells and the copy-number in contaminating normal cells $c_n = 2$ given by

$$p_{s,i} = \frac{\rho_s m_{s,i}}{\rho_s c_{s,i} + (1-\rho_s)c_n} = \frac{\rho_s m_{s,i}}{\rho_s c_{s,i} + 2 - 2\rho_s}.$$

Due to a combination of different sources of noise, one might also expect to observe variant reads with a success probability $p_{0,s}$ at unmutated sites. Potential reasons for this would be a random misreading of bases during the sequencing process (Gerstung et al. 2012) or cross-contamination during the sample preparation. While the amount of noise might differ for mutated sites i, it is expected to be fairly low (i.e., $p_{0,s} \ll p_{s,i}$). Due to this and the generally low coverage $\bar{n}_s$, site-specific variations are ignored and only a sample-specific value $p_{0,s}$ is considered.

When a set of mutation M_e from a given edge e of T is considered all, none or a fraction π_m of these might be present in a given sample. The marginal likelihood of the observed data D_e of this set of mutations is then given by

$$p(D_e|\pi_m) = \prod_{i=0}^{|M_e|} \left(\pi_m \; p(y_{s,i}|n_{s,i}, p_{s,i}) + (1-\pi_m) \; p(y_{s,i}|n_{s,i}, p_{0,s})\right).$$

Assuming that mutated sites are not lost at any point in time, for a mutation from the edge $e = (s,t)$ to be mutated in a sample all variants on the path from the germline node r to the ancestral node s of this edge, i.e., $r \rightsquigarrow s$, also have to be mutated (i.e., $\pi_m = 1$). All remaining mutations, i.e., those that occur in the descendants of t or in different lineages of the tree, have to be absent (i.e., $\pi_m = 0$). For a position $x = (e, \pi_m)$ with $e = (s,t)$ the likelihood of the data D of all mutations that are part of the tree is hence:

$$\mathcal{L}(D, x, p_{0,s}, \rho_s) = \overbrace{p(D_e|\pi_{m,s})}^{\text{On edge e}} \overbrace{\prod_{e' \in r \rightsquigarrow anc(s)} p(D_{e'}|\pi_m = 1)}^{\text{Path from root to node s}} \overbrace{\prod_{e' \notin r \rightsquigarrow anc(t)} p(D_{e'}|\pi_m = 0)}^{\text{All others mutations}}.$$

ML estimates of the sample parameters $\hat{e} \in E$, $\pi_{m,s} \in [0,1]$, $\hat{p}_{0,s} \in [0,0.05]$ and $\hat{\rho}_s \in [0,1]$ were obtained, by minimising $-log(\mathcal{L})$.

5.2.3 CN and Mutation Multiplicity

To simplify the analysis, only mutations in sites for which all WGS samples had identical copy-number states (i.e., identical A&B alleles) were considered. It seems reasonable to assume that no CNAs accompanied the (subclonal) tumour evolution for these sites. It seems worthwhile to note that it would be possible that such CNAs still occurred (e.g., *ABB* → *AB* → *ABB*), but given that subclonal CNAs are comparatively infrequent in colorectal carcinomas (Cross et al. 2018) this appears relatively unlikely.

As it was assumed that no subclonal CNAs occurred throughout the evolution of the trees, $m_{s,i}$ was estimated across all samples s as

$$m_i = \underset{m_{s,i} \in 1,\ldots,c_{s,i}}{arg\,min} \sum_{s \in S} -log\left(\binom{n_{s,i}}{X_{s,i}} p_{s,i}^{y_{s,i}} (1-p_{s,i})^{n_{s,i}-y_{s,i}}\right) \mathbb{1}_{s,i},$$

with $p_{s,i}$ as defined above and where $\mathbb{1}_{s,i}$ indicates if the mutation i was detected in the sample s, $X_{s,i}$ is the number of mutated and $n_{s,i}$ the total reads in sample s. Due to potential issues with the accuracy of estimates of c_i and m_i, sites with a $c_i = 0$ or high copy-numbers $c_i > 4$ were excluded.

5.2.4 R Package

5.2.4.1 Code Availability

I included the code for the assignment of LP-WGS samples to a MP tree into a R package called *MLLPT* (`https://github.com/T-Heide/MLLPT`). In addition to functions available in base R (R Core Team 2020), methods from a number of additional R packages were used: *dplyr* (Wickham et al. 2020), *reshape2* (Wickham 2007) and *magrittr* (Bache and Wickham 2014) for general handling of data, *ggplot2* (Wickham 2016), *cowplot* (Wilke 2020), *ggtree* (Yu et al. 2017) for plotting of results, *ape* (Paradis and Schliep 2019), *phangorn* (Schliep 2011) and treeman (Bennett, Sutton, and Turvey 2017) for the manipulation of tree objects.

5.2.4.2 Usage

The main function of the R package is called *MLLPT::add_lowpass_sampled* and expects a number of different objects as input. The first one is an object of class *phylo* from the *ape* package for R (Paradis and Schliep 2019) containing the reconstructed phylogenetic tree. The second object contains the sequence data used for the reconstruction of the phylogenetic trees as object of class *phyDat* from the *phangorn* package (Schliep 2011). Mutation identifiers have to be added to the *phyDat* object as 'id' attribute (*attr(phy_data, 'id')*). The last argument required are the actual mutation data. These have to be passed as a named list of data frames containing the following columns: count of mutated alleles ('alt_count' or 'alt'), count of reads covering the site ('depth' or 'dp'), estimated copy-number of site ('cn_total' or 'cn') and mutation multiplicity of the mutated allele, defaults to 1 ('cn_mutated' or 'mm'). With the above objects available the main function can be called as follows:

```
tree_with_lp_added =
MLLPT::add_lowpass_sampled(
tree = tree,
phydata = phydata,
sample_data = samples
)
```

This function will print output similar to the one below and return a list containing the tree with the LP-WGS samples added to it (see Figure 5.1A) and a data frame containing the parameters estimates.

```
-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
Processing sample: EPICC_C518_C1_G1_L1 (1/2)
=> Optimizing estimates for purity, background.
New values:
- Background rate: 0.01 -> 0
- Purity: 1 -> 0.8314805
- MLL: -343.0637 -> -192.6926
-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
```

Several optional arguments can be passed to the function in order to adjust its behaviour. A description of these can be found in the package documentation (see *?MLLPT::add_lowpass_sampled*).

5.2.4.3 Plotting Methods

Multiple functions that allow plotting the results of the maximum-likelihood estimation (MLE) of the position of LP-WGS samples on the edge of the tree (see Figure 5.1A, 5.1B and 5.1D) as well as the ML estimates of the per-sample parameters (see Figure 5.1C) were added to the R package.

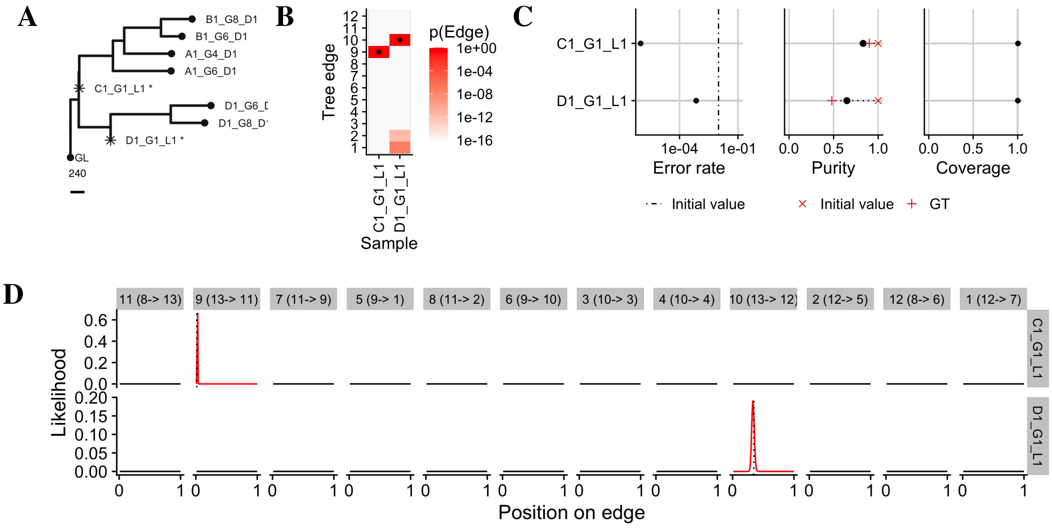


Figure 5.1: Plotting methods in the MLLPT package. A) Phylogenetic tree with LP-WGS samples (marked with stars) added to the position indicated by the ML estimates (*plot_tree*). B) Heatmap showing likelihood that the samples are associated with a given edge (*plot_lp_loglik*). C) Maximum-likelihood estimate of per sample parameters. D) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. (*plot_lp_loglik_edge*).

5.3 Results

5.3.1 Tests Using Simulated Data

I first applied the method to a set of simulated sequencing data generated under a wide range of sequencing parameters. In all cases, realisations of a spatial tumour simulation using the *CHESS* R package (Chkhaidze et al. 2019) were generated, and a single-gland tree containing a total of $N = 20$ samples was sampled. The ML method was then applied to these, using the true mutation tree and simulated LP-WGS samples to characterise the performance of the method. For the simulation of LP-WGS datasets, all possible combinations of the following parameters were used: sample purity values $\rho_s \in \{0.25, 0.5, 0.75, 1\}$, average sample coverage $\bar{n}_s \in \{0.1, 0.5, 0.75, 1\}$, rates of background noise at unmutated sites $p_{0,s} \in \{0, 0.01, 0.05\}$ and tree heights $height(T) \in \{1000\}$. For each of these $N = 48$ parameter combinations, 10 simulated LP-WGS datasets were generated, each consisting of 20 samples, resulting in a total of 9600 tests.

In Figure 5.2A an example of a true tree (on the left) and three corresponding LP-WGS trees are shown. From these the distance $\Delta x_s = |x_s - \hat{x}_s|$ between the true position x_s of the sample s and the ML estimate of the position $\hat{x}_s$ on the tree T was obtained and the relative error of the positions $\Delta x_s^{rel} = \Delta x_s / height(T)$ calculated. The distribution of this summary statistic is shown for all tested parameter combinations in Figure 5.2B.

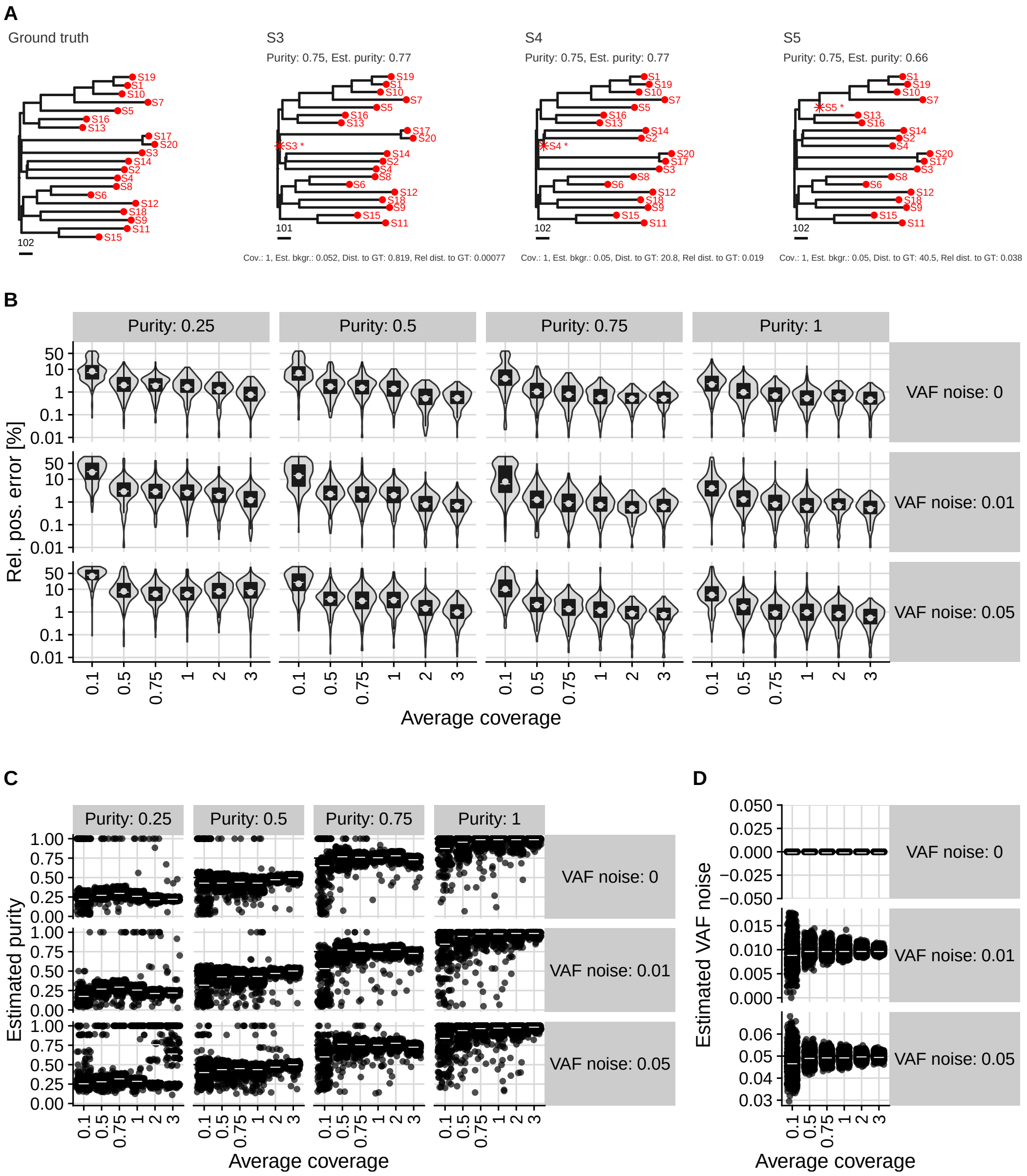


Figure 5.2: ML estimates of LP-WGS sample locations with simulated WGS data. A) Examples of simulated phylogenetic trees and three trees reconstructed using the ML method (marked by stars). B) Relative difference between estimated and true positions ($|x_s - \hat{x}_s|/height(T)$). C) Estimated vs real purity values. D) Estimated background error rates of sequencing. Abbreviations: Est. bkgr. - Estimated background rate, Dist. to GT - Distance to ground-truth positions, Rel. pos. error - Relative position error

Table 5.1: Error of sample purity estimates ($\Delta p_{0,s}$) with simulated sequencing data.

$p_{0,s}$	0%				1%				5%			
ρ_s	0.25	0.5	0.75	1.0	0.25	0.5	0.75	1.0	0.25	0.5	0.75	1.0
$\lvert\rho_s - \hat{\rho}_s\rvert \leq 0.1$	0.97	0.6	0.97	0.84	0.92	0.59	0.96	0.8	0.79	0.6	0.93	0.74
$\lvert\rho_s - \hat{\rho}_s\rvert \leq 0.2$	0.97	0.92	0.98	0.96	0.96	0.9	0.96	0.94	0.81	0.86	0.96	0.93

Issue of high background noise & low purity From the results shown in Figure 5.2B it is evident that all three samples parameters, $\bar{n}_s$, $p_{0,s}$ and ρ_s, affected the accuracy of the $\hat{x}_s$ estimates. Due to the low differences between the expected VAF of mutated (i.e., $0.5\rho_s$) and unmutated sites $p_{0,s}$, a combination of high background rate and low purity did especially affect the estimation of $\hat{x}_s$. For example, only 58% (116/200) and 81% (163/200) of samples had $\Delta x_s^{rel} \leq 0.1$ and $\Delta x_s^{rel} \leq 0.2$ respectively at $p_{0,s} = 0.05$, $\rho_s = 0.25$ and $\bar{n}_s \geq 0.5$. This effect was largely independent of $\bar{n}_s$ (range: $49.5\% - 68.5\%$, $99 - 137/200$ and $74\% - 90\%$, $148 - 181/200$). In the absence of a significant background rate, the estimates of $\hat{x}_s$ were more reasonable at low sample purity $\rho_s = 0.25$. Here 94.5% (189/20) and 99.5% (199/20) had $\Delta x_s^{rel} \leq 0.1$ for all but the lowest coverage $\bar{n}_s = 0.1$ tested.

Accurate estimations of position for higher purities For higher purity values (i.e., $\rho_s \in \{0.75, 1\}$) and coverage $\bar{n}_s = 1$ almost all samples had a relative error $\Delta x_s^{rel} \leq 0.1$ ($\geq 98.9\%$) and $\Delta x_s^{rel} \leq 0.2$ ($\approx 100\%$) respectively (Figure 5.2B). At a high background $p_{0,s} = 0.05$ for purity $p_{0,s} = 0.75$ and $p_{0,s} = 1$ Δx_s^{rel} was $\leq 4.3\%$ and $\leq 3.3\%$ for 90% of samples respectively. At a lower purity of $\rho_s = 0.5$ differences were larger, but $\Delta x_s^{rel} \leq 20\%$ in almost all cases. Here around 99% (198/200), 95.5% (191/200) and 85% (170/200) had $\Delta x_s^{rel} \leq 0.1$ at a background rate $p_{0,s}$ of 0, 0.01 and 0.05 respectively (Figure 5.2B).

Estimation of sample purity Despite the absence of any clonal variants and the relatively low number of mutations per lineage ($height(T) \approx 1000$), estimations of sample purities were reasonably accurate (Figure 5.2C). In most cases, the error on the purity estimates was below 20% across a relatively large range of sample parameters (see Table 5.1).

In this context, it has to be considered that the number of variant sites that contain information on ρ_s (i.e., those shared with another sample) can be much lower than $height(T)$. Overall, the estimations of sample purities obtained together with the ML estimation of LP-WGS sample positions were reasonably accurate when information on somatic variants from WGS samples was available. The estimated background VAF values $\hat{p}_{0,s}$ are shown in Figure 5.2D. As expected, estimated values of $p_{0,s}$ are centre around the underlying true simulated value, with the variance decreasing with increasing coverage.

5.3.2 Tests Using Subsampled WGS Samples

While the ML estimates of sample positions obtained from simulated data (see above) were generally reasonably accurate, it is of course possible that an essential aspect of the data was not simulated accurately. For this reason, the method was further tested on sub-sampled WGS samples, which in practice should result in data very similar to LP-WGS samples. The exact estimates of the copy-number $c_{s,i}$, mutation multiplicity $m_{i,s}$, as well as the reconstructed mutation tree T were used for this, and the actual assignment of LP-WGS samples are described later. Due to this, the results should also, at least to some extent, reflect the influence of inaccurate inputs (T, c and m), as well as the effects copy-number states different from AB, have on the estimates (i.e., one mutated and one unmutated allele).

A jackknife method was applied to a total of $N = 188$ deeply sequenced WGS glands from the EPICC cohort (excluding case C522). These samples span a range of purity values (see Figure 4.5, page 125,) and were grouped into bins of $(0,0.25],(0.25,0.5],(0.5,0.75],(0.75,1]$ containing 6, 29, 32 and 121 samples respectively. In short, one s from the tree of a given case was removed, the WGS data of s subsampled to a LP-WGS coverage equivalent $\bar{n}_s \in \{0.1,0.5,0.75,1,2,3\}$ and the estimates of c_i and m_s obtained from all samples were used as input for the ML method. This procedure was repeated one sample at a time for all samples.

Summary statistics In Figure 5.3A the results of the ML estimation applied to three samples from case C532 are shown alongside the original MP tree on the left. From these data shown in Figure 5.3A the absolute distance $\Delta x_s = |x_s - \hat{x}_s|$ between true x_s and estimate position $\hat{x}_s$ of sample in T were calculated, equivalent to the simulated data before. Since the main interest was the correctness of the subclonal structure of T, the relative error of $\Delta x_s^{rel} = \Delta x_s / height_{sc}(T)$ where $height_{sc}(T)$ is the height of T after keeping subclonal variants of cancer samples (i.e., those not present in all samples) was calculated.

Error of location estimates Figure 5.3B summarises these relative errors of the sample locations Δx_s^{rel} for different $\bar{n}_s$ in each of the four purity groups. The marginal distribution of the estimated background rate $p_{0,s}$ is shown in Figure 5.3C. From Figure 5.3C it is obvious that, apart from s with low purity $0 < \rho_s \leq 0.25$, the majority of $p_{0,s}$ lie between 1% and 5% with the median and the central 90% range being 1.5% (0.89% − 3.0%), 1.5% (0.87% − 3.6%) and 1.8% (0.98% − 4.8%) for the purity intervals $(0.25,0.5]$, $(0.5,0.75]$ and $(0.75,1]$

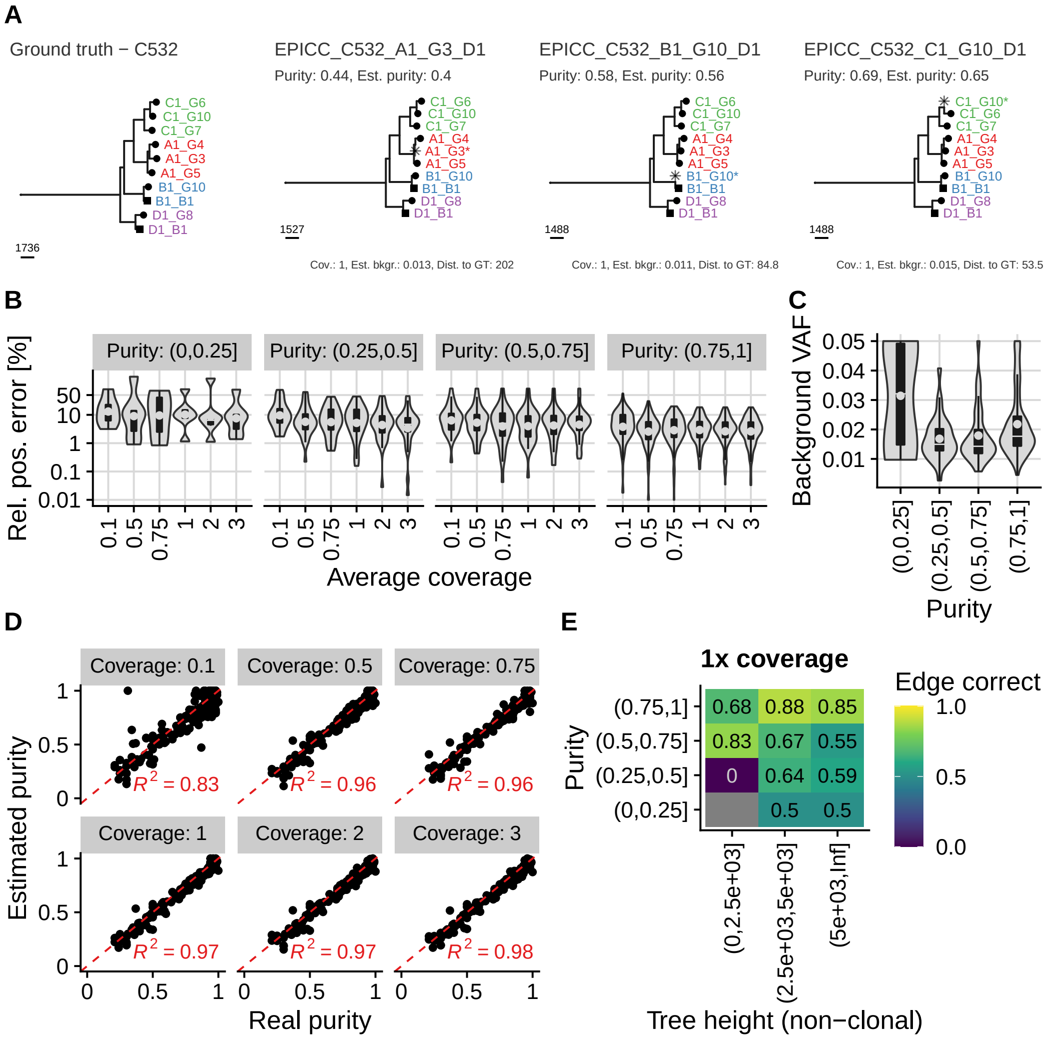


Figure 5.3: Jackknifed ML estimates of sample locations with subsampled deep WGS data. A) An example of a phylogenetic tree constructed from deep WGS data (left) and three trees reconstructed using the ML method after removal and subsampling of one sample (marked by stars). B) Relative difference between estimated and true positions ($\Delta x_s^{rel} = |\Delta x_s|/height_{sc}(T)$). C) Estimated background rate of unmutated variant sites. D) Correlation of estimated and real purity values (R^2 values in red). E) Fraction of samples assigned to the correct edge of the tree split by purity and tree height. Abbreviations: Est. bkgr. - Estimated background rate, Dist. to GT - Distance to ground-truth positions, Rel. pos. error - Relative position error

respectively. Possible reasons for and implications of these relatively high values of $p_{0,s}$ will be discussed in more detail below (see Section 5.3.3.3).

After taking into account the estimates of $p_{0,s}$ the data shown in Figure 5.3B are fairly consistent with results obtained from simulated LP-WGS data presented before (see Figure 5.2B). First, for samples with $\rho_s \leq 0.25$ reliable estimation of x_s was not possible. Instead, the data were explained by high values of $p_{0,s}$ with s being put close to the root of T for all tested values of $\bar{n}_s$. Secondly, for samples with $\rho > 0.25$, the accuracy of $\hat{x}_s$ generally increased for higher values of ρ_s. At $\bar{n}_s = 1$ for around 66% (19/29), 75% (24/32) and 90% (109/121) of samples $\Delta x_s^{rel} \leq 0.1$ and for around 86% (25/29), 94% (30/32) and 100% (121/121) of samples $\Delta x_s^{rel} \leq 0.2$ in each of the purity intervals $(0.25, 0.5]$, $(0.5, 0.75]$ and $(0.75, 1]$ respectively. These results are consistent with the fraction of samples for which the edge the sample was assigned to were correct at $\bar{n}_s = 1$ as shown in Figure 5.3E.

Summarised the estimates of $\hat{x}_s$ were quite accurate despite the fairly high background error rate of $\approx 1.7\%$ for values of $\rho > 0.25$ across a wide range of $\bar{n}_s \geq 0.5$ and T observed in the cohort (e.g., $height(T)$).

Purity estimates Accurate purity estimates can sometimes be hard to obtain for LP-WGS samples with no or very few CNAs or where the majority of CNAs are LOH events. Indeed, estimation of purity and copy-number values failed for 38/347 (11%) LP-WGS samples sequenced as part of the LP-WGS dataset described below. Accordingly, the accuracy of the ML estimates of ρ_s during the sample assignment was assessed. Figure 5.3D shows a scatter plot of $\hat{\rho}_s$ against the independently estimated $\hat{\rho}_s'$ for different coverage values. In general, a large part of the variance of ρ_s was explained by the ML estimates with $R^2 \geq 0.96$ for coverage values $\bar{n}_s \geq 0.5$. Only for the lowest tested value of $\bar{n}_s = 0.5$ a substantially lower $R^2 = 0.83$ was observed.

In summary, this indicates that, at least with the number of clonal variants present in the WGS colorectal cancer samples of this cohort, reasonable estimates of ρ_s can be obtained within the context of the ML estimation of x_s.

5.3.3 Application to LP-WGS Samples From the EPICC Cohort

After evaluating the performance of the ML LP-WGS assignment method with both, simulated and subsampled deeply sequenced WGS samples, it was decided that additional LP-WGS sequencing of single-glands at $\bar{n}_s \geq 0.5$ would be performed. For this target coverage, previous results indicated that robust estimation of x_s, ρ_s and $p_{0,s}$ should be possible. The

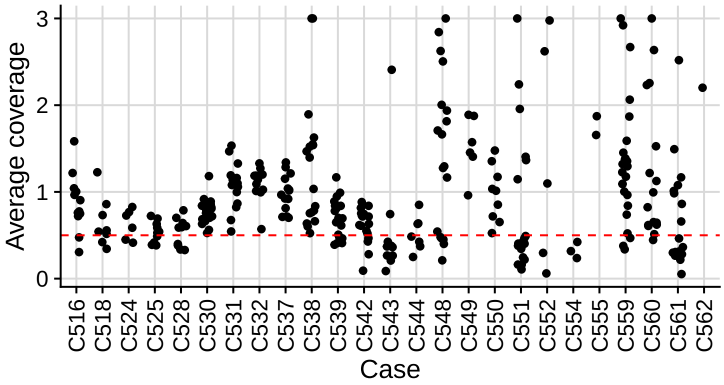


Figure 5.4: Average coverage values of EPICC single-gland LP-WGS . The dashed horizontal line indicates the target coverage of $\bar{n}_s = 0.5$.

average coverage across samples was 1.0 (median: 0.78, 90% upper range: $0.36 - 7.5$) and a plot summarising the per-sample coverage of single-gland samples in each case can be found in Figure 5.4.

The method was applied to all of these LP-WGS samples initially. A subset of these was later sequenced at a higher depth (i.e., $\bar{n}_s \approx 30$), after which the analysis was repeated with all samples. After comparing the relative position of the LP-WGS samples to the equivalent deep WGS sample, these 'duplicated' LP-WGS samples were removed from the final tree. Table 5.2 shows the total number of single-gland LP-WGS, deep WGS samples and those for which both data types were available.

5.3.3.1 Reconstructed ML LP-WGS Trees

After applying the method to all LP-WGS samples sequenced as described in the Methods section, a number of plots summarising the per-sample parameter estimates were created. Results for one representative case (C532) are shown as an example in Figure 5.5. Identical figures for the remaining 25 cases can be found in the Figures S.67-S.91 (page 294-306).

5.3.3.2 Removed Samples

As expected from the initial tests with simulated data (Section 5.3.1) and the subsampled deep WGS data (Section 5.3.2), the reconstruction of the LP-WGS position on trees did not succeed for some low-purity samples. Similar to previous observations, the ML estimates indicated the absence of any somatic variants (i.e., sample location close to the root), a high background rate and low purity (see for example case C538, Figure S.77A-B, page 299). While the analysis of these might have been potentially possible by constraining the parameter range, these problematic samples (see Table 5.1) were instead removed from all

Table 5.2: Number of EPICC glands per case used for ML LP-WGS sample assignment. The column 'WGS & LP-WGS Glands' contains the number of glands for which both deep WGS and LP-WGS data were generated.

Case	WGS	LP-WGS	LP-WGS (filtered)	WGS & LP-WGS
C516	7	13	8	4
C518	6	8	8	0
C524	10	6	3	3
C525	10	10	8	2
C528	6	10	10	0
C530	13	22	18	4
C531	11	15	11	4
C532	10	13	13	0
C537	7	14	8	1
C538	10	18	16	0
C539	14	19	11	5
C542	12	17	15	1
C543	6	10	7	1
C544	6	7	3	0
C548	8	18	18	0
C549	9	6	5	0
C550	7	9	5	3
C551	12	21	20	1
C552	5	5	5	0
C554	8	3	3	0
C555	5	2	2	0
C559	11	24	22	2
C560	7	16	16	0
C561	14	18	18	0
C562	4	1	1	0
Σ	240	305	254	31

Table 5.3: LP-WGS samples excluded from ML LP-WGS trees.

	Case	Sample	Purity (SNVs)	Purity (CNAs)	Ploidy (CNAs)
1	C516	A1_G3	0.24	0.32	2.00
2	C537	B1_G10	0.13	0.12	3.00
3	C537	B1_G2	0.19	0.19	3.00
4	C537	B1_G6	0.24	0.26	3.00
5	C537	B1_G8	0.25	0.27	3.00
6	C537	D1_G9	0.16	0.18	3.00
7	C538	A1_G1	0.10	0.11	2.00
8	C538	A1_G6	0.19	0.19	2.00
9	C539	B1_G4	0.17	0.17	3.00
10	C539	C1_G6	0.14	0.16	3.00
11	C539	D1_G2	0.20	0.21	3.00
12	C542	C1_G3	0.19	0.24	3.00
13	C543	A1_G4	0.21	0.19	2.00
14	C543	D1_G10	0.16	0.21	2.00
15	C544	C1_G1	0.21	0.23	3.00
16	C544	C1_G5	0.25	0.16	3.00
17	C544	C1_G6	0.16	0.19	3.00
18	C549	A1_G9	0.07	0.11	3.00
19	C550	B1_G8	0.05	0.12	2.00

further analyses to reduce the potential effects from errors. Of the excluded samples, all but one (C516_A1_G3_L1) had an estimated purity well below 25%, severely limiting the usability of these for most commonly performed analyses.

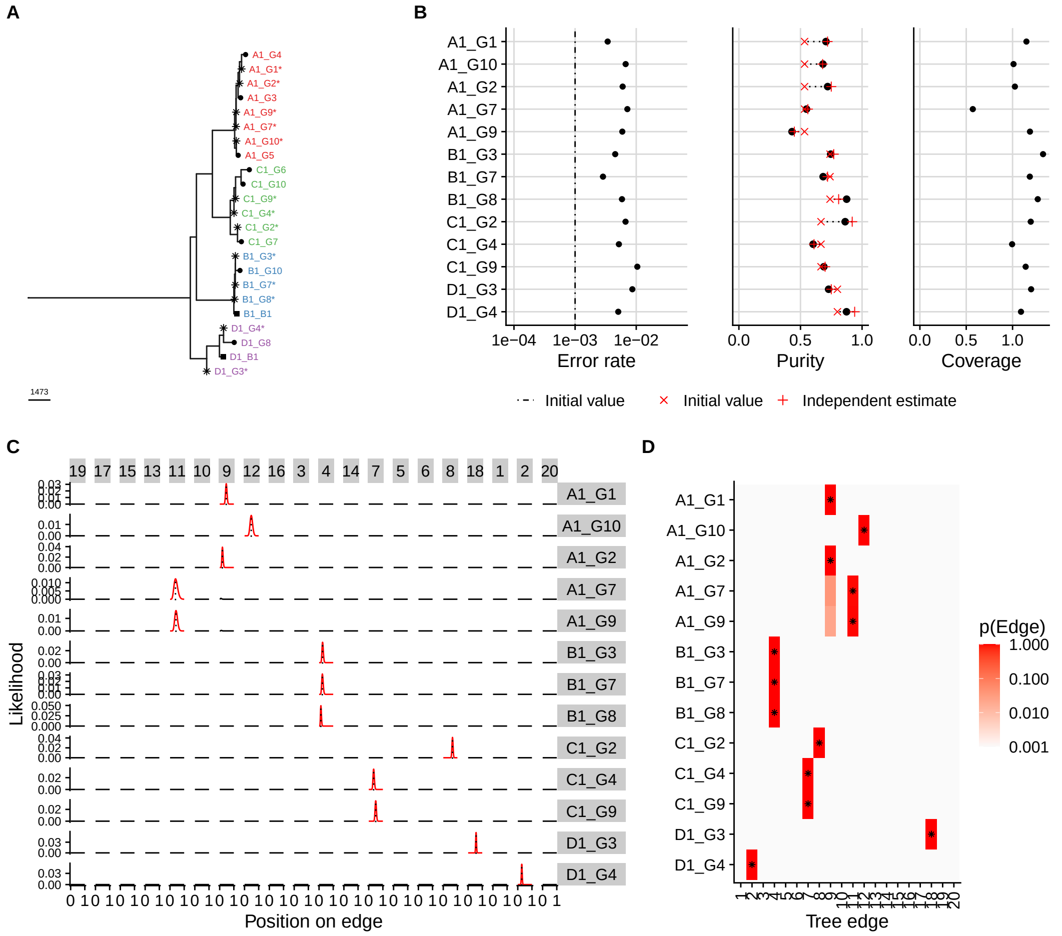


Figure 5.5: ML LP-WGS assignment results for case C532. A) ML tree reconstructed from WGS data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.

5.3.3.3 Purity & Background Rates

As shown in Figure 5.6B the majority of $\hat{p}_{0,s}$ were estimated to be between 0.3% and 3% with the average value being 1.1% (median: 0.8%, central 90% range: 0.31% – 3.1%). Notably, there exists a substantial difference in $\hat{p}_{0,s}$ estimates across cases (Figure 5.6B). Further, while the average values of $\hat{p}_{0,s}$ of each case were significantly correlated with those obtained from jackknifed deep WGS per case ($r = 0.65, p = 0.00032$), these were consistently lower for LP-WGS samples compared to the deeply sequenced WGS samples shown in Figure 5.3 (mean ratio: 2.3). The exact reason for this is elusive. The estimated values for $\hat{p}_{0,s}$ were several orders of magnitude larger than the expected error rates of NGS, which are typically expected to be around 10^{-5} – 10^{-4}.

Since $\hat{p}_{0,s}$ was only estimated on sites that were mutated in other samples of the same

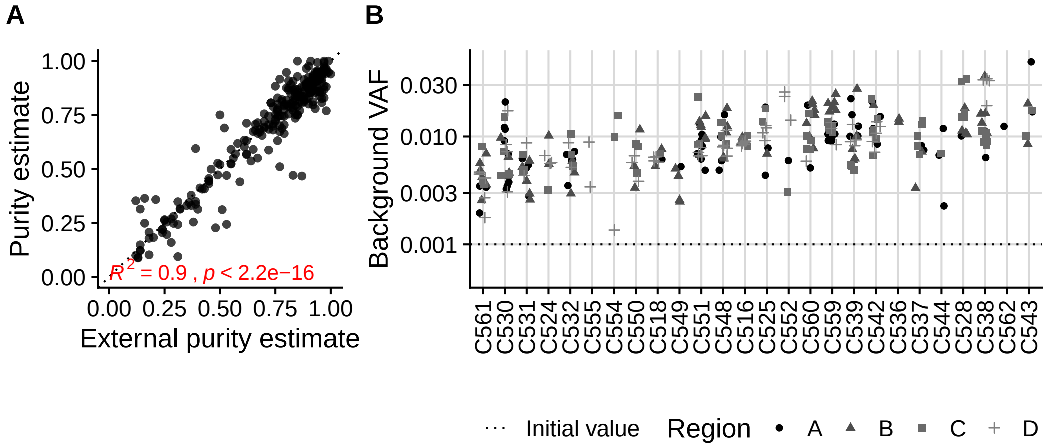


Figure 5.6: ML estimates of purity and background rate. A) Scatter plot of ML estimates of the purity (y-axis) and estimates obtained from an orthogonal analysis of copy-number alterations. B) Estimated background rates at unmutated sites.

tumour, one explanation for the high values and differences of $\hat{p}_{0,s}$ would be that the cells surrounding individual glands contaminated these to varying degrees. This contamination could have happened either as part of the gland structure itself or in the liquid medium transferred together with the glands. An alternative explanation could also be provided by some biological phenomenon that causes a heterogeneous cell population within a gland (e.g., a stem cell population). A minor subpopulation of cells could lead to a measurable presence of low-frequency mutations from a lineage fixed in another gland (i.e., have high VAF), but which are present only in a subset of stem cells in the observed gland (i.e., having low VAF).

Consistent with both of these hypotheses the background rates on sites mutated in other samples from one tumour region (i.e., intra-region, mean: 0.025%) were significantly higher than those from sites only mutated in samples from different regions (i.e., inter-region, mean: 0.73%) as shown in Figure 5.7A. It was tested whether patterns in samples that appeared intermixed within the tree[1] differed and if these had a higher background at sites mutated within the glands from the same clade (intra-clade). This could potentially support the hypothesis of a minor subclone present due to a stem-cell structure within the glands. In contrast, a higher frequency of mutation found in samples from the same region (intra-region), would rather support contamination of samples by surrounding tumour cells.

[1] C524_C1_G5: Region D, C531_A1_G8: Region C, C538_B1_G4: Region D, C559_D1_G5, C559_D1_G9: Region C, C560_C1_G8: Region B, C551_A1_G6, C551_B1_G7, C551_B1_G2, C551_A1_G9: C551_B1_G3) shown in Figures S.69 (page 295), S.72 (page 296), S.77 (page 299), S.89 (page 305), S.86 (page 303) and S.87 (page 304).

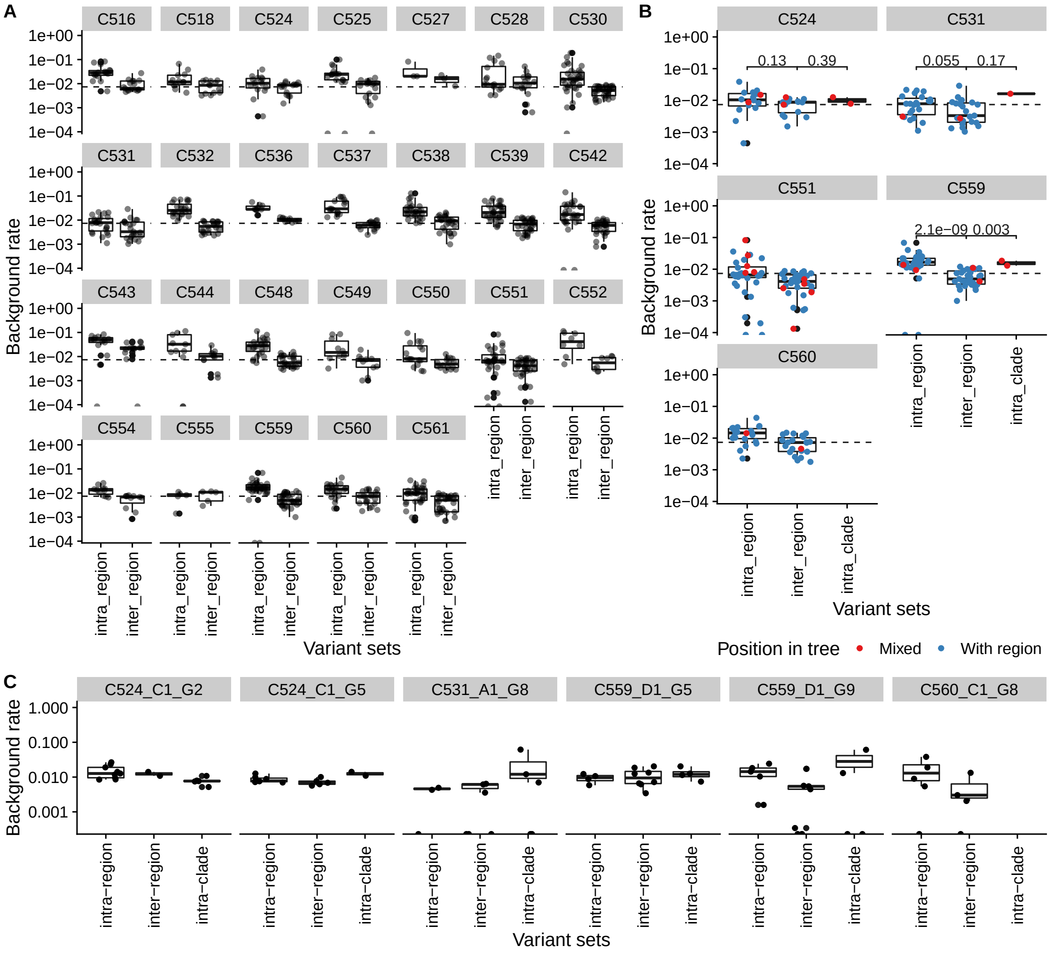


Figure 5.7: Observed background rates in data used for the ML LP-WGS sample assignment. A) Background estimates per edge were generally larger for those present in other samples from the same region. Due to the general structure of the tree, this does not allow to distinguish effects related to ancestral relationships and regional properties. B) For a subset of cases, intermixed samples (red dots) allow distinguishing these two effects, showing elevated rates for both mutations present in samples from the same region and those in samples from the same clade. C) The per edge estimates of the background rate for individual intermixed samples.

The results from this analysis are summarised in Figure 5.7B and 5.7C. While conclusions from this small set of samples are certainly limited, the background rate of mutations present in other samples from the same clade, but not region, where elevated in at least some samples (e.g., C524_C1_G5, C531_A1_G8, and C559_D1_G9), suggesting that incomplete fixation of cells in a stem-cell structure might, at least partially, contribute to the signal observed. Still, mutations from samples of the same region also had an elevated background rate in some samples (e.g., C524_C1_G2, C559_D1_G9, and C560_C1_G8), indicating that cross-gland contaminations are also an important contributing factor.

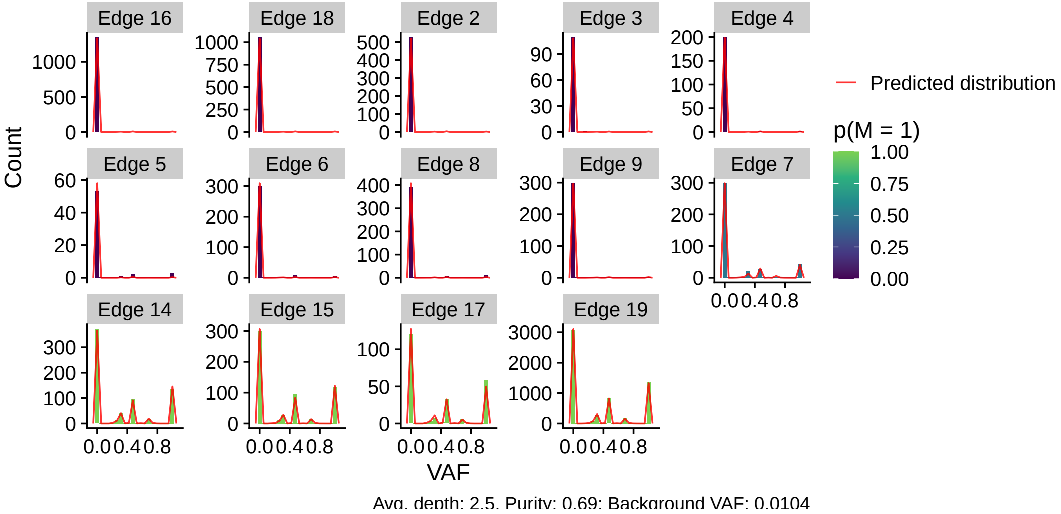


Figure 5.8: VAF distribution of a representative LP-WGS sample EPICC_C532_C1_G9_L1 for each edge of the tree.

5.3.3.4 Model Fits

For the ML estimation of LP-WGS sample position in the tree, the trees, copy-number and mutation multiplicity estimated from the WGS were assumed to be accurate. It was further assumed that each sample was a monoclonal population of cells. While this appears reasonable for the single-gland LP-WGS sequencing data analysed here, this is not guaranteed to be the case. Therefore, the observed VAF data were compared to the distribution one would expect to see under the ML estimates of the model parameter. For this purpose, plots equivalent to the one shown in Figure 5.8 were generated.

Generally, the observed VAF data were concordant with the expected ones. Each fit was manually reviewed and a total of 5 samples for which some deviations existed were identified. For sample C516_A1_G3_L1 (Figure S.92, page 307) a relatively large number of variants with a low frequency along the edge of the tree were observed and this sample was independently identified as problematic and excluded (Table 5.3). For the remaining four samples less than expected variants were observed for one edge of the phylogeny: C524_B1_G3_L1 (Edge 13, Figure S.93, page 307), C538_B1_G1_L1 (Edge 11, Figure S.94, page 307), C543_B1_G9_L1 (Edge 7, Figure S.95, page 308), and C548_C1_G1_L1 (Edge 14, Figure S.96, page 308). In the latter more than expect variant reads were observed from on a different edge (Edge 13) as well. For these four cases, some degree of polyclonality might exist, explaining the observed mismatch. Nevertheless, overall, the model fitted the single-gland LP-WGS data extremely well, supporting the initial assumptions on the monoclonal nature of individual glands.

5.3.3.5 Sample Intermixing

In the majority of cases, the structure of reconstructed trees indeed recapitulated the overall structure of the original WGS trees. This meant that for the majority of cases clear segregation of samples according to their respective regions occurred in trees. Such structures can clearly be seen in trees reconstructed for cases C516 (Figure S.67, page 294), C518 (Figure S.68A, page 294), C532 (Figure 5.5A, page 184), C537 (Figure S.75A, page 298), C538 (Figure S.77A, page 299), C539 (Figure S.79A, page 300), C552 (Figure S.80A, page 300), C554 (Figure S.84A, page 302), C555 (Figure S.88A, page 304), C561 (Figure S.90A, page 306) or C562 (Figure S.91A, page 306). In these cases, LP-WGS samples from regions without deeply sequenced WGS samples were usually assigned close to the MRCA of all samples. This can, for example, be seen in data from C518 (Figure S.68A, page 294), C538 (Figure S.77A, page 299) or C552 (Figure S.80A, page 300). This matches the assumption that divergence in space occurred at a time point early during the tumour growth with data obtained from a spatially sampled star-shaped phylogeny.

Trees reconstructed from a subset of cases showed some interspersed samples in clades primarily formed by samples from different regions of the tumour. This is equivalent to the pattern of spatial variegation of clones within space described in Sottoriva et al. (2015). One example of a potentially variegated LP-WGS sample (i.e., A1_G8) can be seen in the tree of C531 (Figure S.72A, page 296). In this case, no supporting CNAs were present (S.103, page 310), but the ML estimate clearly supported the position within the tree (Figure S.72C–D, page 296).

Another example of spatial intermixing was found in case C538 (Figure S.77, page 299), here one sample from region B (i.e., B1_G4) was located outside of the clade formed by all remaining samples of region B&C. In this case, the analysis of CNAs, specifically the absence of a loss on chr5 in B1_G4, provided independent support for the relative position in the tree (S.106, page 311). Similar patterns were observed for three LP-WGS samples in C551 (Figure S.87, page 304, and S.114, page 313), one LP-WGS sample in C559 (Figure S.89, page 305) and likewise for one sample in C560 (Figure S.86, page 303). These were again supported by the presence of region-specific CNAs (Figure S.119, page 314).

The total number of these intermixed samples are summarised in Table 5.4. A Fisher's Exact Test conducted on the tabulated counts of the cases suggested variability of proportions across patients ($p = 0.007$). Due to the relatively low number of glands per case and the generally low proportion of intermixed glands (i.e., 2.4% across the entire cohort), little more than that some cases appeared to show some intermixing between regions whereas others did not, can be said. For this reason, the most appropriate hypothesis to put forward for testing in follow up appears to be that cases that exhibited some evidence of spatial variegation (i.e., those in Table 5.4) might differ in their outcome compared to the rest.

Table 5.4: Number of glands in clades formed by samples from a different region.

Case	Mixed WGS	Mixed LP-WGS	Σ Mixed	N glands
C524	2	0	2	16
C531	0	1	1	26
C538	0	1	1	25
C551	0	4	4	31
C559	1	1	2	35
C560	0	1	1	22
Other	0	0	0	296

5.3.3.6 Validation Using CNAs

LP-WGS data are typically used to analyse CNAs in a large number of samples. Similar to this, integer copy-number values were estimated from these. After the assignment of the LP-WGS samples onto the trees — notably only using sites at which no copy-number alterations occurred — the general pattern of CNAs was compared to the reconstructed tree topologies. Generally, data from both were consistent.

An example showing copy-number alterations that support the overall structure of the reconstructed trees and specifically the placement of LP-WGS samples within it is shown in Figure 5.9. In this case, the added LP-WGS samples allowed to time some subclonal relative to the position of the LP-WGS samples. For example, the added sample A1_G9 exhibited the chr10 gain, but not the chr3p loss observed in the other samples from the tumour region A. This provides evidence for the relative order of these variants and could potentially be used to improve the timing of CNAs from the number of SNVs as done for example by Cross et al. (2018). I will refrain from a detailed analysis of these CNAs in the individual cases, but plots equivalent to that in Figure 5.9 are shown in Figure S.97-S.121 (page 308-314) for the remaining cases.

5.3.3.7 Validation Using Matched LP-WGS & WGS

To assess the consistency of the ML estimates one last time, distances between LP-WGS and matched WGS samples for which both data types were available were used (see Table 5.2). One example of such a set of LP-WGS samples is shown in Figure 5.10A. While

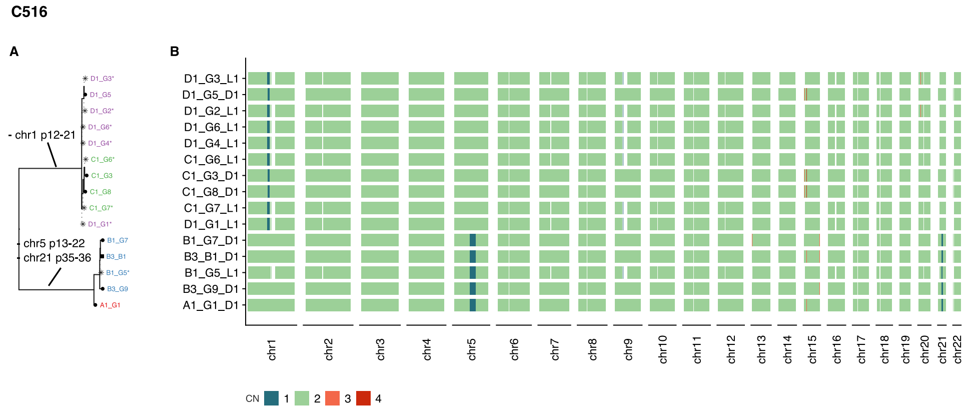


Figure 5.9: Subclonal copy-number alterations and LP-WGS tree C537. A) Subclonal structure of the LP-WGS SNV tree. Clonal mutations are not shown. B) Copy-number states of the samples.

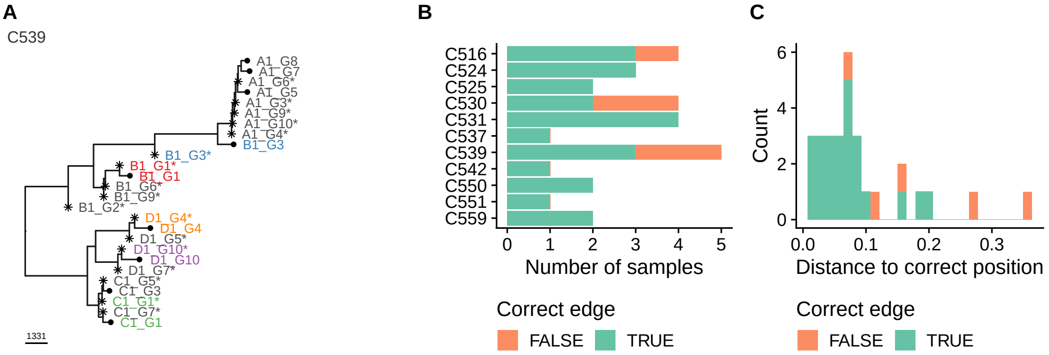


Figure 5.10: Distances between matched LP-WGS and WGS samples in ML LP-WGS trees. A) Example of a case with three LP-WGS samples with matched deep WGS data (highlighted in red). B) Fraction of samples per case in which the LP-WGS sample was added to the correct edge (i.e., that of the tip of the matched sample). C) Relative distance between LP-WGS samples and the match WGS sample.

the majority of LP-WGS samples, apart from sample B1_G3, were located close to the matched WGS tip, 2/5 samples (i.e., B1_G3 and C1_G1) were not directly assigned on the corresponding edge.

Similar observations were made in the cases in which matched samples existed. A total of 24/29 (83%) samples were assigned to the correct edge (see Figure 5.10B) and close to the tip node (see Figure 5.10C). This observation is consistent with observations from the downsampled deep WGS samples (Figure 5.3B&E) and corroborates the previous conclusion that the method is reasonably accurate.

5.3.4 Application to ATAC-seq Samples From the EPICC Cohort

While ATAC-seq is in principle a different assay, reads obtained from a cancer sample should contain mutant alleles more or less identical to that of WGS. For this reason, the

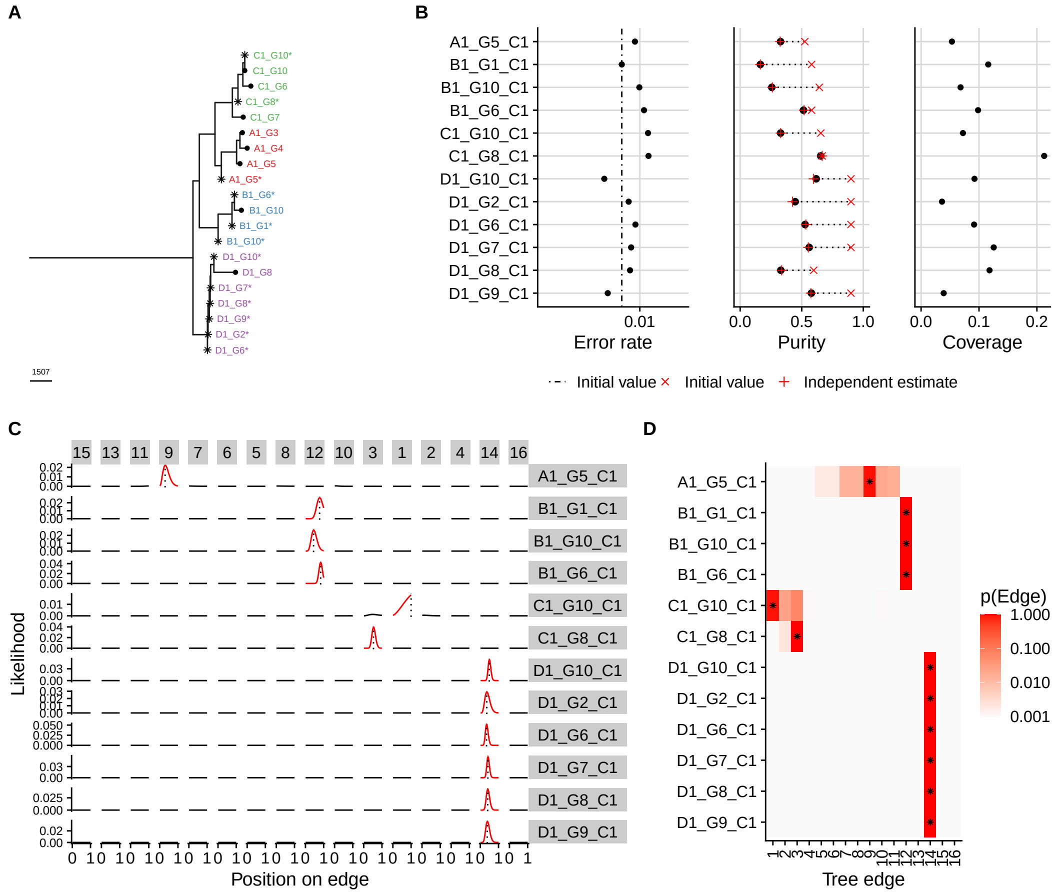


Figure 5.11: ML ATAC-seq assignment results for case C532. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing the distribution of likelihood that the samples are associated with a given edge.

same ML method I used to assign LP-WGS samples to an edge of the tree was applied to all 1060 ATAC-seq samples (177 bulks and 883 glands). The results of one representative example (case C532) are shown in Figure 5.11.

Similar to this example, the computed position of most samples was consistent with the structure of the tree resolved by shallow and deep WGS samples of the same case. Due to the relatively low average coverage of somatic sites (Figure 5.11B), little information on the relative position was available. Consistent with this, for ATAC-seq samples the position estimates of samples were relatively unresolved, as indicated by the wide distribution of the position likelihoods across edges (example C531 in Figure S.124C–D, page 317). Due to this, all samples in which the likelihood of the best edge was less than 90% of the sum of the likelihood of all edges and samples that were assigned close to the root of the trees were removed.

These two filtering criteria were passed by $342/883$ (39%) of glands and $79/177$ (45%) of bulks, which are $421/1060$ (40%) of samples overall. Equivalent to the LP-WGS samples for this subset of samples, the distance to any matched WGS samples (i.e., those obtained from the identical piece of tissue) and whether the ATAC-seq samples were assigned somewhere on the edge to the tip node of this matched WGS sample was determined. Of the 61 glands for which a matched deeply sequenced WGS sample existed, the majority, i.e., $33/61$ (54%), were assigned to the correct edge and for $28/61$ (46%) the relative distance (i.e., $\delta x/height_{sc}(T)$) was ≤ 0.2 (Figure S.125, page 318).

While these numbers are substantially lower than those obtained with the LP-WGS samples, given the limited coverage and the generally low purity of the samples, these data reveal a strong sample-specific signal of somatic variants and show that the observed structures were indeed that of the matched samples. Still, even for these high-quality samples, the apparent purity of ATAC-seq compared to matched WGS samples, while significantly correlated, was generally substantially lower (S.126, page 318). Since the reason for this bias was unclear, these trees were not used as the basis for any other analysis (i.e., ABC inference and Expression quantitative trait locis (EQTLs) analysis).

5.4 Discussion

In general, the simple ML method described above enabled the identification of a reasonable estimate for the position of individual LP-WGS sample (0.5–2 coverage) within the MP trees inferred from deeply sequenced WGS samples. The method's limits and overall accuracy were assessed on simulated and subsampled WGS data for different sample purities, background rates, and sequencing coverages.

This analysis showed that low purity, especially in combination with a high background rate at unmutated sites, severely limits the ability to estimate sample positions within the tree. Despite this, the analysis of the majority of LP-WGS samples generated as part of the EPICC study was possible. These ML LP-WGS trees, with their large number of added samples, allowed me to find additional examples of intra-region intermixing (i.e., spatial variegation), which might potentially be indicative of general phenotypic properties of the associated tumours. The separate analysis of CNAs supported the general structure of the reconstructed trees. In this context, the use of the method might potentially also guide the identification of convergent CNAs in different parts of a phylogenetic tree, the relative timing of CNAs or the general prioritisation of samples for deep WGS.

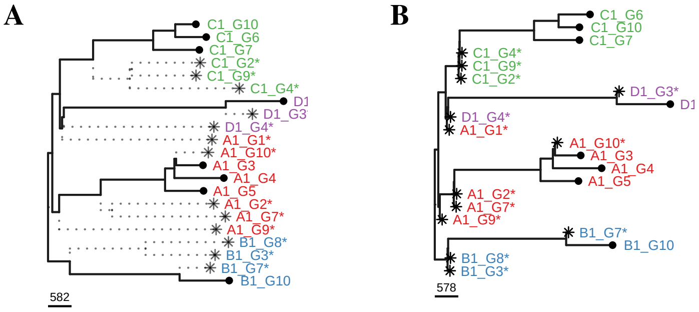


Figure 5.12: Loss of subclonal structures in the ML LP-WGS trees. A) The true latent structure of trees obtained from spatial simulations. Lineages that are expected to be loosed in the LP-WGS trees are shown as dashed lines. B) The corresponding ML LP-WGS tree. While the overall structure of the inferred tree is, apart from small error, correct, much of the complexity of the true tree is lost.

Nevertheless, the loss of the subclonal structure of added samples and the lack of information on the number of private mutations present in these limits the general applicability of the method. This problem is best exemplified by the comparison of a simulated latent tree shown in Figure 5.12A and the corresponding ML LP-WGS tree inferred from simulated WGS data shown in Figure 5.12B. While the position samples are assigned to are, apart from some predictable error, correct, the overall information content of the shown tree is severely reduced. It is, for example, unclear whether the assigned samples are part of a lineage with shared ancestry, like those in region C or not, like the samples D1_G4 and A1_G1. Still, the added LP-WGS samples can provide useful insight into the relative age of the MRCA of a set of samples (i.e., the relative genetic diversity within a region) or the presence of intra-regional mixing (i.e., spatial variegation). The latter was beneficial in cases where CNA that could otherwise be used for the same purpose to quantify the frequency of these events were absent.

While it is obviously impossible to reliably detect individual somatic point mutations in LP-WGS samples, one might in principle be able to estimate the unseen somatic mutation burden of a sample. This mutation burden could then provide information on the length of the tip edge of added samples. In the same way, it might be possible to analyse the number of mutations shared between individual LP-WGS samples added to the tree, giving some insight into the overall structure of these. In light of the substantial background rates observed in the dataset of the EPICC cohort (i.e., VAF $\approx 1\%$), these two ideas were not explored further.

In general, the ML LP-WGS assignment method described here enabled the identification of additional examples of 'spatial sample variegation' in some tumours, a pattern that was previously proposed to be of diagnostic value (Sottoriva et al. 2015; Ryser et al. 2018). It is expected that these improved estimates of the prevalence of 'spatial sample variegation' in the cases analysed will help to assess its diagnostic value as part of the prospective follow up of the study.

The increased number of samples with information on the position in the reconstructed phylogenies has also allowed the inclusion of a much larger number of samples into a separately conducted analysis of EQTLs[2], which was based on samples for which matched RNA-seq, CNA and mutation data were available. Similarly, the reconstructed ML LP-WGS trees will be used as the basis for an ABC inference of spatial dynamics that will be described in the following. The ABC method allowed to take the unique properties of added LP-WGS samples into account, and generally, the inclusion of additional LP-WGS samples lead to the improvement of the results obtained.

[2] Done by Jacob Househam.

Chapter 6

ABC-SMC Inference

In the previous two chapters, I have described the mutational landscape of a total of 30 multi-region single-gland sequenced colorectal cancers from the EPICC cohort. In this context, I described the rare intermixing of glands from different regions and hypothesised that this might be a phenotypic property of some cancers. Further, I described the status of mutations in putative driver genes previously identified in other studies (e.g., Muzny et al. 2012; Martínez-Jiménez et al. 2020; Martincorena et al. 2017). The vast majority of these mutations were found to be clonal mutations present in all glands of a tumour. Such clonal mutations were likely accumulated before the initiation of the corresponding tumour or have been part of a subclone that effectively swept through the population. A few examples of potential subclonal driver mutations (e.g., one KRAS p.G12C and seven PIK3CA mutations) were also identified. In one of these cases, C539, the mutation was accompanied by a clear elongation of the associated branches of the phylogenetic tree. This branch elongation indicates that the associated glands share a most recent common ancestor that went through many cell divisions to reach a higher frequency, the ‘hallmark’ of subclonal selection.

While the presence of a selected subclone, or more broadly speaking, changes of the evolutionary dynamics, seemed rather obvious in this case, it was unclear how to interpret the information contained in the trees in other cases. The same problem exists for many previous studies of the subclonal diversification at primary sites (Yates et al. 2015) or during metastasis (Gundem et al. 2015; Yates et al. 2017; Noorani et al. 2020). Despite being very impressive, these studies have provided little functional insights into the evolutionary dynamics driving these processes. Specifically, it remains unclear to what degree selection of adaptive phenotypic properties plays a role in the later stages of cancer evolution and what effect occasionally observed subclonal driver mutations have.

The majority of cancer driver genes were identified based on their recurrence across patients using statistical models like mutSigCV to determine if mutations are overrepresented across patients (e.g., Lawrence et al. 2014; Martínez-Jiménez et al. 2020). Still, generally recurrent mutations might have little or no effect in some genetic or environmental backgrounds, and often complex analyses and tedious experiments are required to uncover these relationships. An excellent example of such context-dependent selection can be found for PTEN mutations in prostate cancers and leukaemia (Berger, Knudson, and Pandolfi 2011). In these, the incomplete loss of PTEN in a TP53 wildtype background is tumorigenic, whereas total loss of PTEN would lead to the induction of senescence and hence no tumour formation. If the complete PTEN loss instead occurs after the prior loss of TP53 (i.e., in a TP53 mutant context), more aggressive tumour growth is instead observed. In order to study the effect of such driver alterations *in vivo* mouse models are often used, but these are costly, time-consuming and require specific hypotheses to test.

For self-evident reasons, longitudinal observation of solid tumours in their primary site cannot be conducted in humans.[1] For this reason, the strength of the selective advantage provided by driver mutations in actual tumours is not well studied. Here, I will apply a spatial computational inference framework to single-gland multi-region WGS data. Doing so, I will demonstrate how this approach can indirectly gain insight into the fitness effect of naturally arising somatic mutations in primary CRCs. This approach allows identifying relevant alterations occurring in primary tumours, which could be validated subsequently in controlled *in vivo* experiments. As such, this allows for prioritisation of relevant alterations observed in primary tumours for validation. Unlike other studies based on bulk sequencing data (e.g., Dentro et al. 2021), this approach has sufficient power to infer subclonal selection and also allows to characterise the specific genetic background that putative driver mutations occurred in (i.e., their respective lineage).

6.1 Bayesian Statistics

In order to understand the evolutionary dynamics observed in cancer genomic data, we cannot resort to our intuition or descriptive statistics. Instead, one optimally wants a statistical model that captures the process underlying the measured data well enough to allow to gain significant insight into it (Box, Launer, and Wilkinson 1979). Aspects of the model that influence its behaviour, the model parameters, can then provide a more interpretable sum-

[1] Some cases of non-resectable tumours or refusal of treatment can give the rare opportunity to do this.

mary of the data. Many methods to fit such a model to actual observations (i.e., statistical inference) exist. In order to apply most of these, one needs to be able to calculate the likelihood function $p(D|\theta)$, defining the probability of observing the data D under a given set of parameters θ from the parameter space $\theta \in \Theta$.

With a likelihood function available classic statistical methods can be used to identify parameters under which it would be most likely to observe the data, the MLE $\hat{\theta} = \arg\max_{\theta \in \Theta} p(D|\theta)$. This MLE is possible even if a closed-form solution of the ML is not available or hard to obtain.

An alternative approach, so-called Bayesian inference, is to use Bayes' theorem to instead calculate a probability distribution over the parameter space $p(\theta|D)$ the posterior distribution or short posterior. From Bayes' theorem, it follows that

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta')p(\theta')\,d\theta'},$$

where $p(D|\theta)$ is the likelihood and $p(\theta)$ is a probability distribution over the parameters space, the prior likelihood or short prior, that encodes the prior belief on θ. The marginal likelihood of the data $p(D)$ can be interpreted as a normalisation constant. Dropping this results in a distribution that is proportional to the actual posterior likelihood but does not sum to one (i.e., $p(\theta|D) \propto p(D|\theta)p(\theta)$).

6.1.1 Approximate Bayesian Computation

Unfortunately, a likelihood function for the spatial distribution of mutations in a growing tumour, potentially with several differently fast-growing subpopulations (i.e., neutrality vs selection), cell death and various modes of growth (i.e., exponential growth[2] vs boundary driven growth[3]) is not readily available and probably intractable.

Nevertheless, since it is possible to simulate the underlying process, a class of algorithms that allow performing Bayesian inference without a likelihood function can be used. These ABC methods use a generative process to approximate the likelihood conditional on a set of parameters θ (Karabatsos and Leisen 2018).

To do this ABC methods require a model $f(\cdot|\theta)$ from which random realisations $D^* \in \mathscr{D}$ can be drawn, a prior distribution $p(\theta)$ on the set of the inferred parameters $\theta \in \Theta$,

[2] By exponential growth a situation in which all cells grow at a rate proportional to their fitness throughout the development of the tumour is meant.

[3] By 'boundary driven growth' any growth that is dominated by the expansion of cells on the outer edge is meant. Here it will be assumed that this results from the spatial constraints and the inability of cells to push outwards within the tumour. In principle similar consequences could also arise from limited nutritional supply or oxygenation within the centre of the tumour.

multiple summary statistics $\eta : \mathscr{D} \to S$, a distance function $\rho : S \times S \to \mathbb{R}^+$, and a critical distance $\varepsilon \in \mathbb{R}^+$ below which random observations are assumed to match the observed data D. With these we then seek to sample from the marginal posterior distribution

$$p(\theta, D^*|D, \varepsilon) = \frac{p(\theta)f(D^*|\theta)\mathbb{I}_{A_{\varepsilon,D}}}{\int \pi(\theta)f(D^*|\theta)\, dD^* d\theta},$$

where $\mathbb{I}\mathbb{A}_{\varepsilon,\mathbb{D}}(x)$ indicates whether x is an element of the set $A_{\varepsilon,D}$ of observations with

$$A_{\varepsilon,D} = \{z \in \mathscr{D} : \rho(\eta(D), \eta(D^*)) \leq \varepsilon\}.$$

ABC methods were pioneered in the field of population genetics by Tavaré et al. (1997) to infer coalescence times from DNA sequence data and by Pritchard et al. (1999) to study the evolution of the Y chromosome. Since then, such methods have been used extensively (Underhill et al. 2000; Kaessmann et al. 2001; Glover et al. 2013). Examples of the application of ABC methods in other fields include molecular biology (Woods and Barnes 2016), pharmacology (Picchini 2014), epidemiology (McKinley, Cook, and Deardon 2009; Tanaka et al. 2006) or indeed cancer evolution (Sottoriva et al. 2015; Williams et al. 2018b).

Various extensions of the brute-force accept-reject method used by Tavaré et al. (1997) and Pritchard et al. (1999) exist, these seek to combine ABC with other algorithms to increase the efficiency of sampling in the parameter space. Examples include MCMC (Marjoram et al. 2003; Wegmann, Leuenberger, and Excoffier 2009), Sequential Monte Carlo (SMC) (Sisson, Fan, and Tanaka 2007; Del Moral, Doucet, and Jasra 2012; Filippi et al. 2013) or Population Monte Carlo (PMC) (Beaumont et al. 2008; Baragatti, Grimaud, and Pommeret 2012; Murakami 2014). Other modifications seek to replace the rejection based approximation of the likelihood with alternative estimators (see Karabatsos and Leisen 2018, for details). One example of this, which will later be used for the calculation of the expectation of the posterior predictive likelihood, are synthetic likelihoods (SLs) similar to those proposed by (Wood 2010). In the context of ABC, this method makes the assumption that the distribution of summary statistics follows a specific distribution. With these assumptions, likelihoods for a critical distance $p(D \leq \varepsilon) \ll 1/N$, that are impermissible to be used with rejection based methods, can be approximated. More detail on SLs will be provided below.

Here, two different algorithm will be used for the ABC inference of parameters describing the growth dynamics in individual tumours i) the rejection sampling described first by Pritchard et al. (1999) (see Section 6.2.4.1, page 205 for details), and ii) the ABC-SMC algorithm proposed by Del Moral, Doucet, and Jasra (2012) (see Section 6.2.4.2, page 206 for details). In brief, the number of subclones $max(i)$, their relative growth rate λ_i compared to the ancestral clone $i = 0$ with $\lambda_0 = 1$, the respective coalescent population sizes t_i, and a global parameter d_{push} describing the distance from the outer rim of the tumour at which glands can grow and 'push' outwards will be inferred. I will also explore if a global increase in death rates μ would explain the observed data better. An overview of all the inferred and constant parameters of the model can be found in Table 6.1 (page 205). A more detailed explanation of the simulation setup and how a simulated sampling scheme equivalent to the one used in the actual experiments was generated will be provided in Section 6.2.1 and Section 6.2.2 (pages 199, pages 201) respectively. A summary of the statistics and distance metrics used to compare the observed data to simulated datasets will be provided in Section 6.2.3 (pages 203). A general overview of the implemented inference framework is shown in Figure 6.1.

6.2 Methods

6.2.1 Spatial Simulations

As outlined above, a model $f(\cdot|\theta)$ from which one can sample simulated observations D^* given a set of parameters θ is required to apply ABC based inference to the whole-genome sequencing data described in the previous chapter. For this, a slightly modified version of the spatial tumour simulator (Chkhaidze et al. 2019) described before (Chapter 3) will be used and generate synthetic sequencing data according to a spatial sampling scheme that is equivalent to the one used to generate the actual data (Chapter 4). For the inference, simulations of a two-dimensional tumour were used. This choice was made, based on the observation that colorectal cancers grow, at least during the initial stages, primarily in a two-dimensional plane through crypt fission (Greaves et al. 2006; Chen et al. 2005; Shen et al. 2005; Bernstein et al. 2008). While this might be a simplification, the simulation of a tumour in two dimensions should still give some insight into a three-dimensional tumour's general growth dynamics.

In well and moderately differentiated colorectal carcinomas the majority of the tumour consists of glandular structures (Fleming et al. 2012; Nagtegaal et al. 2020). These structures are reminiscent of the crypts, small finger-like invaginations into the underlying tissue that normally forms the colorectal epithelium (Humphries and Wright 2008). Similar to normal crypts, these glands are assumed to expand spatially through a process of gland fission

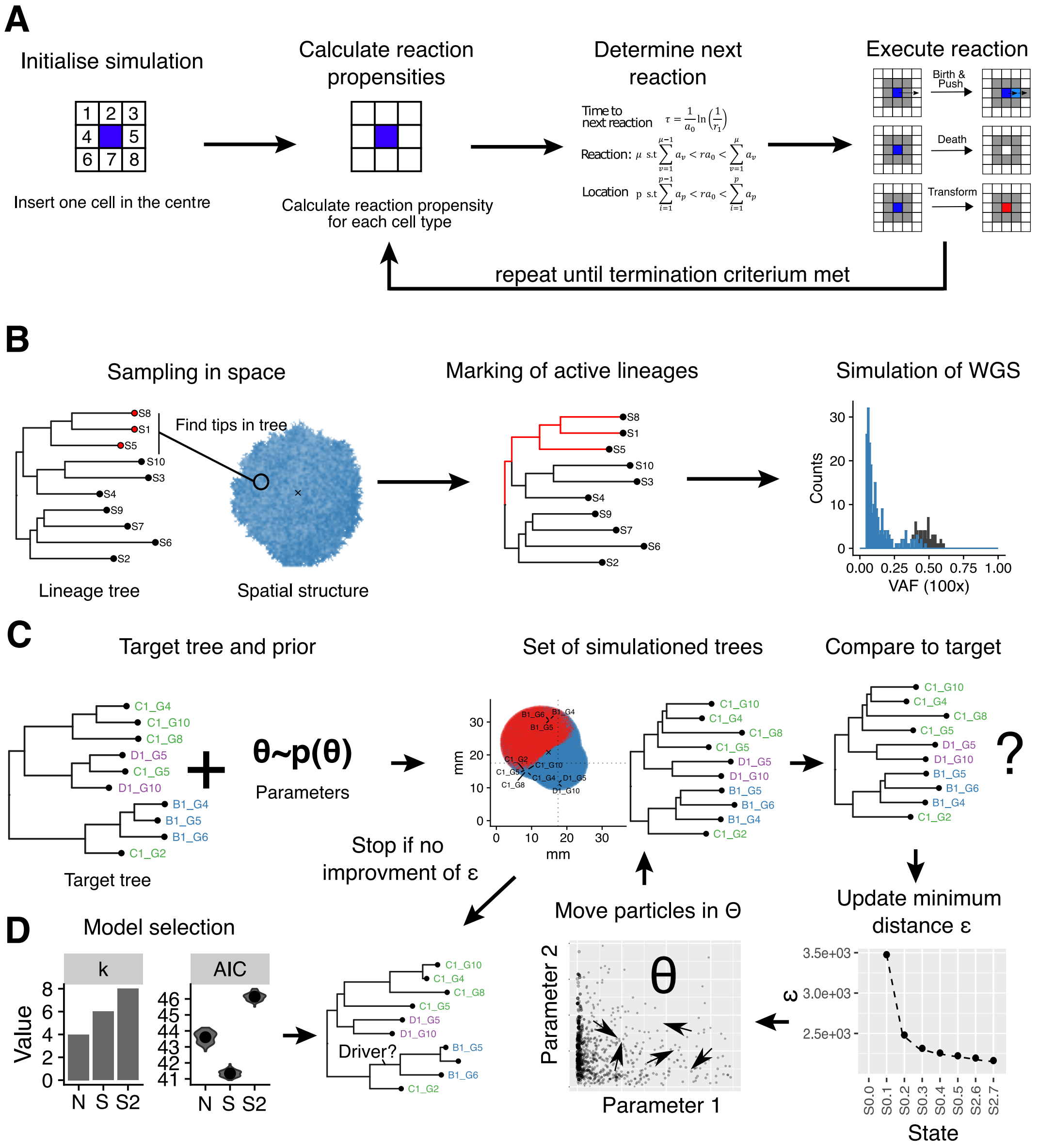


Figure 6.1: The ABC-SMC inference framework. A) The spatial simulation is obtained from a tumour simulator using the Gillespie algorithm. B) For the simulation of WGS data a subset of cells is selected in space, then active lineages in the tree are marked, and finally, the active part of the tree is traversed to simulate WGS data. C) Using an ABC-SMC algorithm (Del Moral, Doucet, and Jasra 2012) parameters of each model are inferred. D) Lastly, a model selection procedure is applied to the fitted models obtained from C to select the most appropriate one.

(Graham et al. 2011; Garcia et al. 1999; Bruens et al. 2017), which also drives neoplastic growth of colon tumours (Wong et al. 2002; Preston et al. 2003). In summary, individual glands are the ‘clonal units’ of colorectal cancers (Baker et al. 2014), and for this reason, each cell of the spatial simulation will be assumed to represent a single gland. Due to the fast replacement of stem cell lineages compared to the rate of crypt fission, the complexity of the population structure within each gland will also be disregarded.

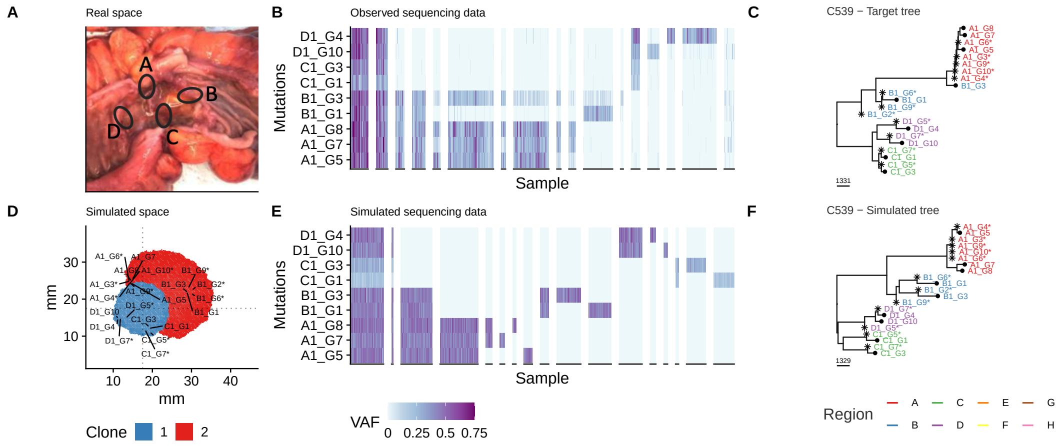


Figure 6.2: An illustrative example of simulated and observed WGS sequencing data. **A** and **D** show the spatial layout of the tumour in space and the sampling locations for the actual and simulated tumours respectively. **B** and **E** summarise the VAF of actual and simulated WGS sequencing data for deep WGS samples respectively. **C** and **F** show the actual and simulated trees reconstructed from the data respectively.
VAF: variant allele frequency.

The diameter of colonic crypts in normal colon tissue is about 60 microns, and the colorectal epithelium contains roughly 100 crypts per square millimetre of colon (Nguyen et al. 2010). Assuming a similar number of glands per square millimetre of tumour tissue, a colon tumour with a diameter of $\approx 3.5\,cm$ can be represented by a grid size of $350x350$ (Figure 6.2D). Such a tumour would contain $\approx 96,000$ glands, of which each is consists of $\approx 2,000-10,000$ cells. The number of $\approx 10^9$ cells simulated by these $350x350$ simulations are roughly similar to that present in human malignancies (Del Monte 2009) and are still fast enough $\leq 3\,s$ to allow the generation of a sufficient number of simulations for ABC inference.

6.2.2 Equivalent Sampling Scheme

After the generation of a simulated tumour, a set of samples that reflected the sampling schema used (Figure 6.2A, & Figure 4.1E, page 109) to generate the actual single-gland WGS data (Figure 6.2B) was generated as shown in Figure 6.3. First, a random angle ϕ from the centre of the tumour $O(x_O, y_O)$ along which the first sample region (i.e., the centre of the region 'A') should be placed is generated first sampling from

$$\phi_A \sim U(0, 2\pi).$$

For each region A-D an offset ϕ_i' was added to ϕ_A. Here two different methods were used i) a constant offset where $\phi' = (0, 0.5\pi, \pi, 1.5\pi)$ for the regions A-D respectively and ii) a randomly varied sampling schema with relative angles between adjacent regions

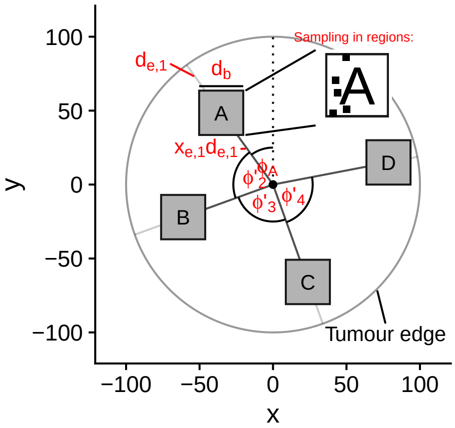


Figure 6.3: Diagram illustrating the simulated sampling schema of the EPICC cohort. For the simulation of the sampling first, the angular position of the region 'A' $\phi_A \sim U(0, 2\pi)$ relative to the centre (black dot) of the tumour (outline as a grey circle) was generated. The angular position of the remaining regions were defined relative to this region using the offsets ϕ' (i.e., ϕ'_2, ϕ'_3 and ϕ'_4). The centre of the sampling squares (grey filed boxes, with width d_b) was then placed at a relative position to the edge $x_e d_e$, where d_e is the distance to the tumour edge along ϕ. The sampled tumour cells (black squares in the inset) were randomly drawn from the sample squares without replacement.

sampled from a Dirichlet distribution $\mathbf{x}_\phi \sim Dir(K = 4, \alpha)$ with $\alpha = (0.25m_\phi, ..., 0.25m_\phi)$. Here m_ϕ denotes the prior strength of the prior and the angle offsets ϕ'_i are given by $\phi'_i = 2\pi \sum_{j=0}^{i} x_{\phi,j}$.

Along each of these vectors, the distances to the most distant occupied grid point $\mathbb{I}_o(x, y)$ (i.e., the edge of the tumour) were searched using a half-interval search between O and the edge of the simulated space, to identify

$$d_{e,i} = \underset{r \in [0, d_max]}{\arg\max}\ r\, \mathbb{I}(r\, cos(\phi_A + \phi_i) + x_O, r\, sin(\phi_A + \phi_i) + y_O)).$$

In cases with a high death rate, a high number of grid points within the centre of the tumour are empty. Hence, positions were only assumed to be empty when grid points along the vector defined by ϕ up to a distance of 10 from the evaluated position were also unoccupied. At the most extreme values of the parameter range considered ($\mu \leq 0.5$ and $d_{push} = 1$) up to $\approx 8.5\%$ of grid points in the centre of the tumour can be empty, but even at these values, the observation of 7 or more consecutive empty grid points is very unlikely ($\ll 10^{-}7$).

After the identification of the distance to the edge $d_{e,a}$ the centre of sampling regions were placed at a relative position $x_e \in [0, 1]$ along the vector with the coordinates being given by:

$$x = [x_e\ d_{e,a}\ cos(\phi_A + \phi_i) + x_O],$$
$$y = [x_e\ d_{e,a}\ sin(\phi_A + \phi_i) + y_O].$$

Similar to the angle offsets ϕ_i', two different methods were used to define x_e i) a constant fixed value of $x_e = 0.75$ and ii) a random value sampled from a Beta distribution with a given prior strength m_d and mean μ_d with $x_e \sim B(\mu_d m_d, (1-\mu_d) m_d)$.

After the definition of the centre of the four regions, random grid points within a rectangular area of edge lengths d_b around these were sampled randomly without replacement until the required number of samples from the region were obtained. Sampled, but unoccupied grid points were rejected. Figure 6.2A and 6.2D show an illustrative example of the equivalent sampling scheme described above applied to a simulated tumour and the actual macroscopic sampling locations in the real tumour respectively.

6.2.3 Distance Function

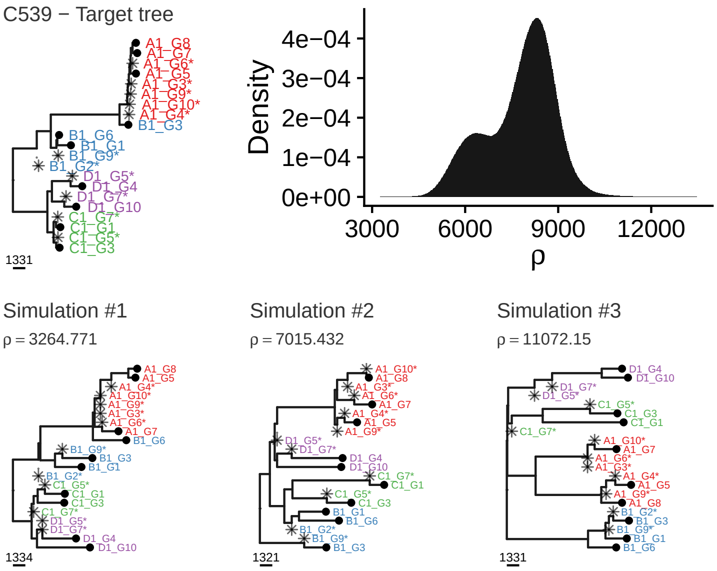


Figure 6.4: Simulated trees with different distance ρ to the target tree (top left). On the top right, the distribution of distances between the target and simulated trees is shown.

Due to spatial information obtained for the real data, a distance metric that is dependent on the sample labels was used. Labels of samples with the same characteristics were swapped (i.e., sequencing type and sample region) to minimise the distance metric. Since clonal mutations do not inform on the subclonal dynamics, these were removed from the trees T and T' prior to the calculation of the distance between them. Next, the patristic distances $d(i,j)$, that is the sum of the lengths of the edges that link two nodes i and j in the tree, for all pairs of tip nodes, were determined and scaled by the height of the tree (i.e., the maximum distance from the root 0 to a tip).

For a maximum parsimony tree without any homoplasy, this is equivalent to the distance calculated from the mutation data itself as

$$d_{j,k} = (|M_j \cup M_k| - |M_j \cap M_k|)/max\{|M_i| : i \in 1,...,n\},$$

where M_i denotes the set of mutations found to be present in the sample associated with the tip i and n is the total number of tips present in the tree.

The distance between two trees T and T' is then calculated from the differences of the scaled patristic distances using the L2-norm

$$\rho(T,T') = \left(\sum_{i=0}^{|V^1|} \sum_{j=i}^{|V^1|} \left(\frac{d_T(i,j)}{h(T)} - \frac{d_{T'}(i,j)}{h(T')} \right)^2 \right)^{\frac{1}{2}}, \text{where } h(T) = \max_{i \in V^1} d_T(0,x).$$

This distance takes into account the tip labels of the tree, but samples of the same type (i.e., WGS or LP-WGS) and region (i.e., A-D) can be considered to be equivalent. Accordingly, such equivalent tip labels in the tree T' were swapped until the distance between both trees could not be reduced by swapping any additional labels as outlined in Algorithm 1.

Algorithm 1: Label swapping in trees

```
Data: Trees $T$ and $T^*$
Result: Minimised distance between $T$ and $T^*$
Initialise list $L$ of all label pairs in $T^*$ of same type and region.;
$T^{*'} \leftarrow T^*$;
do
    $T^* \leftarrow T^{*'}$;
    $d \leftarrow \Delta(T,T^*)$;
    $\Delta d \leftarrow 0$;
    foreach $l \in L$ do
        $T_l^{*'} \leftarrow T^*$ with labels $l$ swapped;
        $d_l \leftarrow \Delta(T,T^{*'})$;
        $\Delta d_l \leftarrow d_l - d$;
        if $\Delta d_l < \Delta d$ then
            $T^{*'} \leftarrow T_l^{*'}$;
            $\Delta d \leftarrow \Delta d_l$;
while $\Delta d < 0$;
return (d)
```

It is worth noting that this gradient descent does not necessarily result in the tree with the smallest distance possible, which could only be found by exploring all swaps. As the number of possible ways to label a tree is $\prod_{i=0}^{|N|} = N_i!$ where N are the labels in the label group i, this would be infeasible for all but the smallest trees. For the tree shown in Figure 6.4 for example there are $N = (2,2,2,2,2,3,3,5)$ labels per group resulting in $138,240$ trees for which the distance ρ would have to be calculated. This is computationally infeasible and instead, only the closest local optimum is searched.

In Figure 6.4 a couple of simulated trees with variable distances to a given target are shown to illustrate how changes of the tree topology and branch length are reflected in the distance metric.

6.2.4 ABC Algorithms

As mentioned before, two ABC inference algorithms were applied to the datasets to conduct the statistical inference. In the following, the simple ABC rejection sampling algorithm (Pritchard et al. 1999) will be described first. As this method severely suffers from the 'curse of dimensionality', a more complex ABC-SMC algorithm (Del Moral, Doucet, and Jasra 2012) that is less affected by this problem will be described following this.

The parameters inferred using the ABC algorithms are summarised in Table 6.1. The death rate μ, mutation rate m and 'push distance' d_{push} were assumed to be global properties of the tumour, whereas the number of subclones $max(i)$ and the associated birthrate λ_i and clone start time t_i are assumed to be clone specific parameters.

Table 6.1: Overview of model parameters for the spatial tumour model. Variables with a subclone index *i* are set individually for each subclone. All other variables are assumed the be constant for the whole tumour.

Symbol	Name	Description	Limits
$max(i)$	Subclone number	Number of subclones	[0,2]
λ_i	Birthrates	Rate of cells division	[1,20]
t_i	Clone start times	Population size at introduction	$[1,\lfloor N_{max}/2 \rfloor]$
a_i	Fathers	Index of Ancestor	$i-1$
μ	Deathrates	Likelihood of death during division	$0 \vee [0,0.5]$
d_{push}	Push distance	Distance from the edge cells grow	$[0,x/2]$
m	Mutationrates	Number of mutations during division	$\sim h(T^*)$
d_b	Sample box size	Diameter of the sampling region	[15,25]

6.2.4.1 ABC Rejection Sampling

The simplest ABC inference algorithm is that of rejection sampling. While this brute force method can be used the obtain an approximation of the posterior, it is computationally very costly as a time proportional to the density of the prior on θ is spent on the generation of samples, even in regions with very low probability. Due to this shortcoming, many more efficient alternative algorithms do exist. Still, due to its simplicity, this algorithm will be used to test the output of the ABC-SMC algorithm described later and provide a short introduction to ABC in general.

The ABC rejection algorithm was first used by Tavaré et al. (1997) and a more generalised version, introducing the explicit definition of a distance function $\rho(\eta(D^*), \eta(D)) \leq \varepsilon)$ by Pritchard et al. (1999). The general procedure used in both papers is principle identical and described by the Algorithm 2.

Algorithm 2: ABC rejection sampler

```
Data: Target $y$, distance function $\Delta$, prior distribution $p(\theta)$, simulator $f(\cdot|\theta)$.
Result: A set of N particles $P$ approximating the posterior distribution.
for $i \leftarrow 0$ to $N$ do
    repeat
        $\theta \sim \pi(\cdot)$; // sample parameter from prior
        $z \sim f(\cdot|\theta)$; // generate simulation
    until $\Delta(y,z) \leq \varepsilon$;
    $P \leftarrow P \cup \{\theta\}$; // append $\theta$ to set of particles
return $(P)$
```

This procedure results in the generation of a set number N of particles for which the distance between the summary statistics of simulated and observed data fall below a predetermined distance threshold ε. These particles can be used to approximate the posterior distribution of the parameters θ.

6.2.4.2 ABC-SMC Algorithm

Due to the general shortcomings of the simple rejection ABC algorithm, especially in high dimensional problems with many parameters, a different ABC inference algorithm was used for most cases. This algorithm is an adaptation of sequential Monte Carlo methods for the ABC context and was described by Del Moral, Doucet, and Jasra (2012). It has a number of advantages over previous ABC-SMC methods described by others (e.g., Sisson, Fan, and Tanaka 2007; Toni et al. 2009; Beaumont et al. 2009). First, it has linear complexity $O(N)$ in the number of particles N instead of $O(N^2)$ for these previously proposed methods. Secondly, the ABC-SMC algorithm by Del Moral, Doucet, and Jasra automatically adjusts the distance threshold ε_n, whereas other methods require the explicit definition of a distance schedule. The careful adjustment of this schedule on ε is often critical, as a too fast reduction can reduce the performance of the inference or even lead to its collapse (Del Moral, Doucet, and Jasra 2012). The ABC-SMC algorithm conceived by Del Moral, Doucet, and Jasra only requires the definition of one parameter α that controls how fast the critical distance ε is reduced.

The SMC algorithm consists of a number of steps, in which a set of N particles is updated repeatedly. Each $i = 1, ..., N$ of these particles consists of a set of parameters $\theta^{(i)}$,

M simulated observations $X^{(i)}_{j=1...M,n=0}$ and a weight $W^{(i)}_{n=0}$. The weight of each particle is proportional to the number of simulated observations $j \in \{1,...,M\}$ that are part of the set $A_{\varepsilon_n,y} = \{z \in D : \rho(y,z) < \varepsilon_n\}$ and can be calculated as

$$W^{(i)}_n = \frac{\sum_{j=0}^{M} \mathbb{I}_{A_{\varepsilon_n,y}}(X^{i}_{j,n=0})}{\sum_{i'=0}^{N}\sum_{j=0}^{M} \mathbb{I}_{A_{\varepsilon_n,y}}(X^{i'}_{j,n=0})},$$

where $\mathbb{I}_{A_{\varepsilon,y}}(x)$ indicates whether the observation x is an element of the set $A_{\varepsilon_n,y}$.

The effective sample size (ESS) of a set of particles with the set of weights $\{W^{(i)}_n\}$ is

$$ESS(\{W^{(i)}_n\}) = \frac{1}{\sum_{i=0}^{N} {W^{(i)}_n}^2}.$$

This ESS is a measure of the complexity of the particle set (Liu 2008) and is used to update the distance threshold ε and trigger a resampling of particles in different steps of the SMC. The individual steps of the ABC-SMC algorithm are described in detail below:

0. **Initialisation** At the beginning of the SMC, i.e., at step $n = 0$, an initial set of N particles is generated. For each of these parameters and random, simulated observations are generated with

   $$X^{(i)}_{j,n=0} \sim f(\cdot,\theta^{(i)}),\ \theta^{(i)}_{n=0} \sim p(\theta),\ \text{for}\, j = 1,...,M \text{ and } i = 1,...,N.$$

   As the distance threshold is initialised as $\varepsilon_{n=0} = \infty$, with $W^{(i)}_{n=0} = 1/N$ and $ESS = N$ at this point.

1. **Update of weights and ESS** At the beginning of each SMC step $n \leftarrow n+1$. Then the distance threshold ε_n gets updated so that $ESS(W^{(i)}_n) \leq \alpha ESS(W^{(i)}_{n-1})$ for the new value. If the relative reduction of the distance threshold $\Delta\varepsilon_{rel} = \frac{\varepsilon_{n-1}-\varepsilon_n}{\varepsilon_{n-1}}$ falls below a critical value the inference is terminated at this point.

2. **Resampling step** If $ESS(W^{(i)}_n) \leq N_R$ a resampling step is triggered. For this, N particles are sampled from the set of all particles with resampling at a rate proportional to the particle weight. After the resampling, all particle weights are set to $W^{(i)}_n = \frac{1}{N}$.

3. **MCMC step** For each particle i with $W^{(i)}_n > 0$ a proposed new parameter set θ^*_i is sampled from the transition kernel $\theta^*_i \sim K(\theta_i)$. Here, K is a truncated multivariate normal distribution $\theta^*_i \sim MVN(\theta_i, 2\hat{\Sigma}, \mathbf{a}, \mathbf{b})$, where $\hat{\Sigma}$ is the empirical estimate of the covariance, and a and b the lower bound of the parameters respectively.

For each of these proposed new parameter values θ^*, a random set of observations $X^*_{1:M}$ are sampled: $X^{*(i)}_j \sim f(\cdot|\theta^*_i)$, where $j = 1,...,M$. Each proposed particle is then accepted with the likelihood given by the Metropolis-Hastings ratio

$$A(X^{*(i)}_j, X^{(i)}_{1:M}) = min\left(1, \frac{\sum_{j=1}^{M} \mathbb{I}_{A_{\varepsilon,y}}(X^{*(i)}_j)}{\sum_{j=1}^{M} \mathbb{I}_{A_{\varepsilon,y}}(X^{(i)}_j)} \frac{q(\theta,\theta^*)}{q(\theta^*,\theta)}\right).$$

At which point the algorithm continues at step 1.

Parametrisation of the ABC-SMC For the purpose of the ABC-SMC inference conducted here a value of $\alpha = 0.95$ was used in all cases. Initially conducted tests on a subset of cases using a value of $\alpha = 0.99$ obtained essentially identical results and multiple independent runs of the ABC-SMC on a subset of cases using $\alpha = 0.95$ also converged to essentially identical solutions. This supported the stability of the results for the value of $\alpha = 0.95$ and while a larger value might in principle still have been preferable, this would have also increased the associated computational costs significantly. For the transition kernel, a truncated multivariate normal distribution from the *tmvtnorm* package (Wilhelm and G 2015) was used, giving a proposal of $\theta^*_i \sim MVN(\theta_i, 2\hat{\Sigma}, \mathbf{a}, \mathbf{b})$, where $\hat{\Sigma}$ is the empirical estimate of the covariance obtained from the particle parameters $\{\theta^{(i)}_n\}$ weighted by $\{W^{(i)}_n\}$ existing at the time point n. The upper and lower limits a and b for each parameter set are listed in Table 6.1.

As many different spatial samples from a single simulated tumour can be generated, a second layer of sampling doing this was introduced. The number of multiple observations drawn from a simulation is denoted as M' below. The number of independent realisations of simulations as M instead. In general, two setups were used: i) $N = 500$, $M = 25$ and $M' = 100$ for the first set of models with fixed sampling position parameters $x_e = 0.75$, $\phi_i = (0, 0.25, 0.5, 0.75)$ and $d_b = 25$ and ii) $N = 5000$, $M = 1$ and $M' = 100$ for a second set of models with variable sampling positions with prior strength $m_\phi = 50$ and $m_d = 20$ (see Chapter 6.2.2 for details). In all cases, the threshold at which resampling was triggered was set to $N_R = 0.75N$.

Termination criteria The ABC-SMC algorithm was terminated when the relative distance reduction $\Delta\varepsilon_{rel}$ decreased below 1% for $t_{\Delta\varepsilon} > 3$ steps (assumed convergence) or until a total of $n = 40$ steps were run. In a couple of cases, a larger value for $t_{\Delta\varepsilon}$ was tested to evaluate if, after longer mixing of the chain, a further decrease in ε could be archived. An example of a distance schedule derived by the algorithm and the associated statistics of the ABC-SMC chain is shown in Figure 6.5. In this example, the algorithm's parameters were $\alpha = 0.95$, $N = 5000$ and $N_R = 3750$ and the inference was terminated after 14 steps.

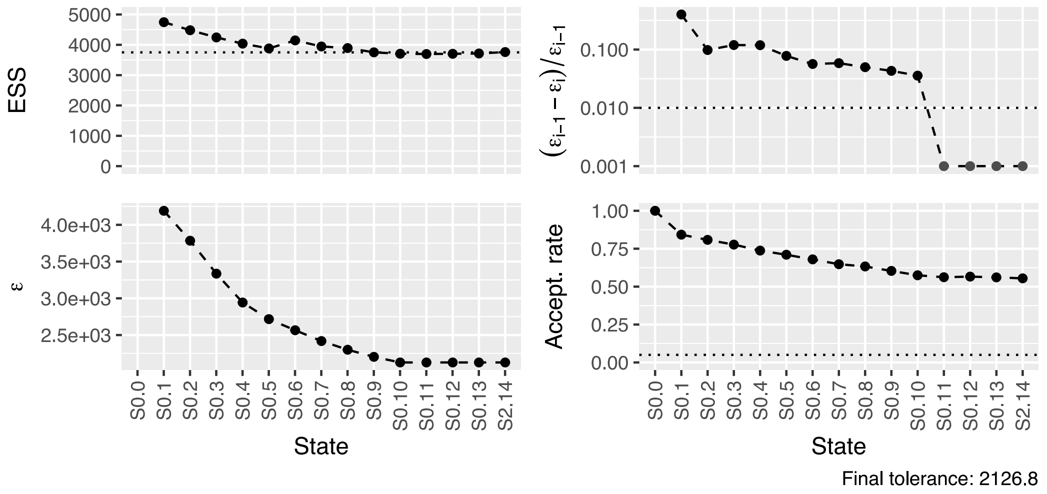


Figure 6.5: Example of an ABC-SMC chain. The plot summaries i) the change of the ESS for the chain with $\alpha = 0.95$, $N = 5000$ and $N_T = 3750$, ii) the resulting schedule of the distance threshold ε, iii) the relative change of the distance $\Delta\varepsilon_{rel}$ and iv) the acceptance rate in the MCMC step for different states of the SMC chain.

6.2.5 Model Selection

6.2.5.1 Synthetic Likelihoods

Generally, the distances under point estimates for neutral simulations were distributed approximately normally, as indicated by quantile-quantile-plots (see Figure 6.6). Further, the likelihood to observe a single datum with a distance Δ_{D^*} below the critical distances ε was often very low ($\ll 0.1\%$). This makes the rejection based approximation of the likelihood $L\tilde{(\theta)} = p(\rho(\eta(D), \eta(D^*)) \leq \varepsilon)$ computationally very expensive and a SL approach (Wood 2010) was used instead in cases where the observed fraction of random realisations D^* with distance below ε was less than 1%. In these cases, the likelihood was approximated as

$$\hat{L_s}^N(\theta) = \Phi\left(\frac{\varepsilon - \mu}{\sigma}\right) = \frac{1}{2}\left[1 + erf\left(\frac{\varepsilon - \mu}{\sigma\sqrt{2}}\right)\right],$$

where

$$\hat{\mu}_\theta = \frac{1}{N}\sum_{i=1}^{N} \rho(\eta(D), \eta(D_i^*)),\ \hat{\sigma}_\theta^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\rho(\eta(D), \eta(D_i^*)) - \hat{\mu}_\theta).$$

are empirical estimates of the variance σ^2 and mean μ calculated from a minimum of $N = 1000$ realisation for each θ.

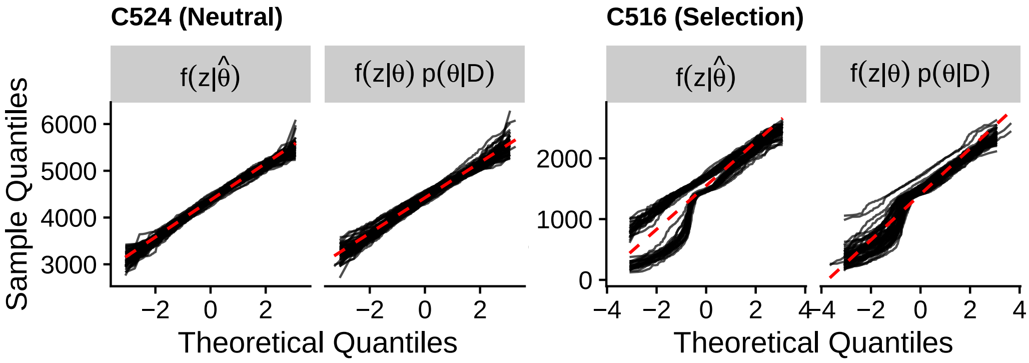


Figure 6.6: Quantile–quantile plot demonstrating that the distance between observed and simulated data (z) generated under the maximum-a-posteriori ($\hat{\theta}$) as well as the posterior distribution ($p(\theta|D)$) are approximately normal for neutral simulations (left), but not for models with selection (right) .

While it has been shown that the SLs can be fairly robust to violations of normality assumptions (Everitt 2018; Price et al. 2018; Grazian and Fan 2019), an improved semi-parametric version has been proposed (An, Nott, and Drovandi 2019). In the context of the ABC-SMC conducted here, distances obtained from neutral models were approximately normally distributed, whereas non-neutral models typically exhibited multiple modes across the observed distance distribution. To these univariate Gaussian mixture models with a variable variance and $K = \arg\min_{K \in 1,\ldots,9} BIC(M_K)$ using the *mclust* R package were fitted instead and estimates of the likelihood were obtained from these (Scrucca et al. 2016; Fraley and Raftery 2002; Fraley et al. 2012).

For each step that involved the estimation of SLs, a simple bootstrap method (Efron 1992) was applied to the observed distances and the Gaussian mixture models fitted to these permutated datasets to estimate the variability of the estimates and quantify errors arising from the limited Monte-Carlo integration.

6.2.5.2 AIC

The AIC was used to penalise the different models for the number of free parameters. The marginal likelihood $p(D|m,\varepsilon) = \int_\theta p(D|\theta,m,\varepsilon)p(\theta|m)\mathrm{d}\theta$ was estimated for each model m using Monte-Carlo integration across a minimum of 200 parameter sets θ obtained from the posterior particles weighted by $\{W_n^{(i)}\}$. For $p(D|\theta,m,\varepsilon)$ the SL approximation $\hat{L}_{s,\varepsilon}^{N}(\theta)$ obtained from the distribution of distances $\Delta(D,D^*)$ of a minimum of 100 simulated datasets $D^* \sim f(\cdot|\theta)$ was used. For the critical distances $\varepsilon = \min_{m \in M} \varepsilon_{end,m}$, where M is the set of models on which model selection was performed, was tested. A large range of ε across possible distances were also evaluated to characterise how the results of model selection were affected by ε.

I then used the obtained estimate of the marginal likelihood $p(D|m,\varepsilon)$ to calculate $\hat{\text{AIC}}_{\varepsilon} = 2k - 2ln\left(p(D|m,\varepsilon)\right)$ where k is the number of free model parameters. An example of such estimated marginal likelihoods and AIC values for given values of ε are shown in Figure 6.9 D&I. Here the negative marginal log-likelihood (NMLL) of the 'Neutral' and 'Selection' models are very similar, leading to the 'Neutral' model being preferred over the alternative models due to its lower AIC.

6.2.6 Code Availability

The code used to perform the ABC-SMC inference on the EPICC cohort and to create other figures shown here can be found on GitHub: `https://github.com/T-Heide/EPICC_inference`.

6.3 Results

For the ABC-SMC inference, two sets of trees were generally used as input i) the MP trees reconstructed from the mutation data (Figure S.51, page S.51), and ii) the MP trees with assigned LP-WGS samples (Figure 4.11, page 143). The simple accept-reject ABC algorithm was initially applied to a couple of test cases. Two of these — one for which the ABC inferred boundary driven growth and another with non-boundary driven growth — will be shown as a simple example first. Following this, results using the ABC-SMC algorithm from the entire cohort will be summarised.

6.3.1 Rejection ABC

6.3.1.1 Inference of non-boundary driven growth in C561

ABC rejection sampling was, together with a couple of other examples, initially applied to the tree of C561 shown in Figure 6.7B and a total of $1,000,000$ simulated trees from $10,000$ particles were generated. These were filtered to retain 0.05% of simulations with a distance $\varepsilon < 3987.2$. One of the trees from the posterior tree set is shown in Figure 6.7C. This tree and most other trees in the set reflect the topology and general structure of the target tree, suggesting that the distance method and ε chosen were able to select trees with a good fit to the target.

Only a single (neutral) model without cell death (i.e., $\mu = 0$) and fixed sampling positions relative to the centre (i.e., $d_b = 25$ and $d_{e,a} = 0.75$) was considered. For this reason,

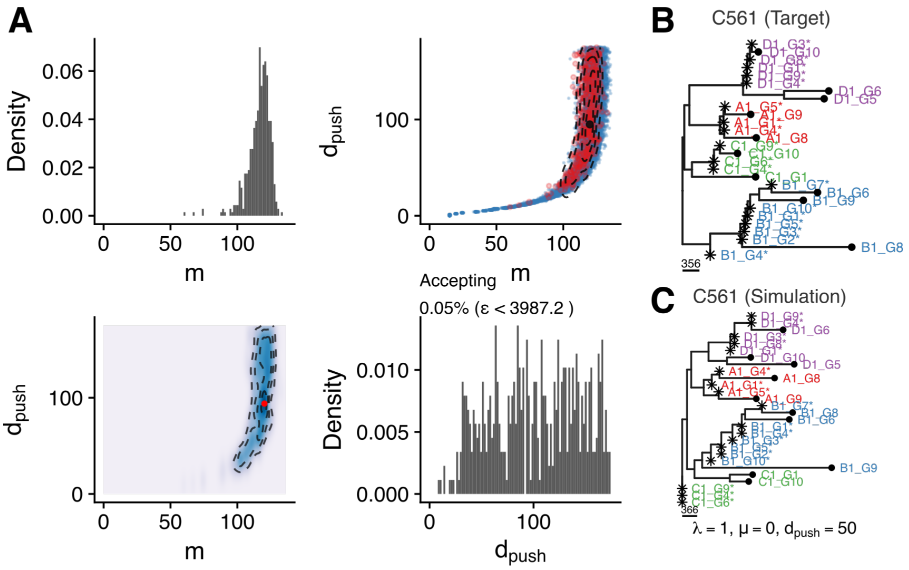


Figure 6.7: Results of the ABC rejection algorithm applied to the tree of case C561. A) The estimated marginal and joined posterior density distributions of the d_{push} parameter in the interval $[0, 175]$ and corresponding mutations rates, m. The marginal densities of the parameters, shown across the diagonal of the plot suggest that $m \approx 110$ and $d_{push} > 20$. A weak correlation of m and d_{push} can be seen in the joined posterior density (bottom left grid). The posterior mode is shown as red (bottom left grid) and black (top right grid) dot in these plots. The estimated posterior densities are shown by black lines. The top right grid shows the rejected (blue points) and accepted particles (red points). B) The target tree y and C) a simulated tree z from the posterior set of trees with $\varepsilon < 39$.

only two parameters, the mutation rate m and pushing distance d_{push} were inferred. The posterior distribution of both is shown in Figure 6.7A. These posterior distribution suggest that simulations similar to the target tree are more likely to be observed under conditions with weak spatial constraints (i.e., $d_{push} > 20$) at a mutation rate of $m \approx 110$ per gland division. The bottom left and top right grid in Figure 6.7A show the joined posterior probability distribution of m and d_{push}. This plot only shows a weak correlation of the two parameters across the posterior distribution.

6.3.1.2 Inference of boundary driven growth in case C356

The same ABC rejection sampling was also applied to the tree of case C536 shown in Figure 6.8D. For the distance threshold, the value of $\varepsilon = 760$ — which identical to the final ε of the ABC-SMC algorithm described below — was used. This allows a direct comparison of the results both algorithms. At this ε, all accepted trees were extremely similar to the observed tree. A representative example of a simulated tree from the posterior particle set is shown in Figure 6.8E.

Again, only a neutral model with fixed values for $\mu = 0$, $d_b = 25$ and $d_{e,a} = 0.75$ was

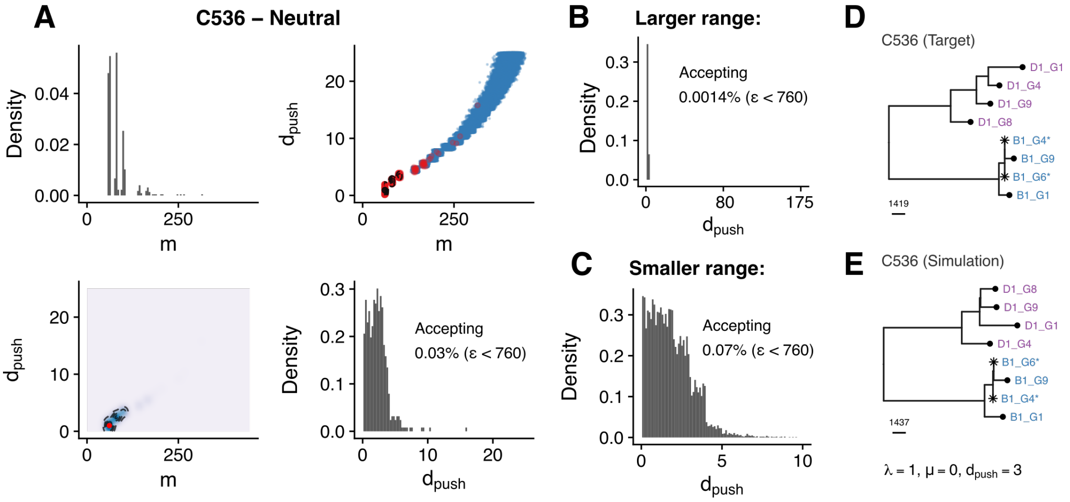


Figure 6.8: Results of the ABC rejection algorithm applied to the tree of case C536 (see D). A) The estimated posterior density distribution of the d_{push} parameter in the interval $[0, 25]$ and corresponding mutations rates, m. The posterior mode is shown as red (bottom left grid) and black (top right grid) dot in these plots. The top right grid shows the rejected (blue points) and accepted particles (red points) as well as the estimated posterior density (black lines). B) As A but for the full range of d_{push} ($[0, 175]$). C) As A but for a smaller range of d_{push} ($[0, 10]$). D) The target tree y. E) One simulated tree z from the posterior set.

considered, meaning that only two parameters, m and d_{push} were inferred. The posterior distribution of these two parameters is shown in Figure 6.8A. The posterior distribution of d_{push} indicates clear boundary driven growth with $0 \leq d_{push} \leq 5$ and a mutation rate of $m \approx 75$ per gland division.

As seen in the lower right grid of Figure 6.8A the prior distribution for d_{push} was restricted to the interval $[0, 25]$. In this interval $\approx 0.03\%$ of the proposed trees were accepted. If proposals were instead drawn from the full range of the prior (i.e., [0,175]), the acceptance rate was with $\approx 0.0014\%$ substantially lower (Figure 6.8B). Not a single tree with a distance of $\varepsilon < 760$ simulated from particles in the range of $20 < d_{push} < 175$ was observed, hence reducing the total fraction of accepted trees substantially. Likewise, a narrower range across the parameters are shown in Figure 6.8C. From the data shown, one can estimate that over the whole distribution of the d_{push} parameter, approximately $1.6 \cdot 10^{-5}$ of proposed simulations are accepted, and to generate 500 accepted trees, one would have to simulate $> 3.2 \cdot 10^{7}$ trees.

In multivariate setups with additional parameters, the fraction of accepted particles across the entire parameter space can become even lower. This can make inference computationally infeasible. Even in the above case, the acceptance of simulations with $d_{push} > 20$

is rare, and the resulting low number of accepted simulations would lead to a poor approximation of the posterior. Alternative ABC algorithms — like the ABC-SMC algorithm by Del Moral, Doucet, and Jasra (2012) used below — were developed for this exact reason.

6.3.2 Fixed Sampling Schema

In the two examples using the ABC rejection sampling algorithm shown before, only simulations with one subclone were considered. In the following, the number of subclones $\max(i)$ and the corresponding clone parameters ($\lambda_i \geq 1$, $t_i \geq 0$) will also be inferred from the data.

6.3.2.1 General Classification Framework

ABC-SMC inference In order to select the most appropriate model, in this example, the tree of the case C536 (Figure 6.9A), first the parameters of all considered models were inferred using ABC-SMC (Figure 6.9B). In this case a fully neutral model ('Neutral') with variable strength of the boundary driven growth d_{push}, a neutral model with additional stochastic death 'Neutral+Death' and models with one or two selected subclones 'Selection' and 'Selection x 2' were considered.

The inference for each of these was run till convergence. At this point, a set of particles from which one can estimate the posterior distribution was obtained. In Figure 6.9C the prior and posterior distribution of d_{push}, the only inferred parameter of the fully neutral model, are shown. These can be compared to the posterior obtained from the rejection ABC described earlier (see Figure 6.8). This comparison shows that both methods agree and that they suggest the presence of strong boundary driven growth.

Model selection A model selection procedure was then used to decide if any of the alternative models were more likely to generate the observed data. In brief, a random set of parameters θ was sampled from the posterior predictive distribution of the model and used to estimate the marginal likelihood using Monte Carlo (MC) integration in combination with SLs (see Section 6.2.5 for details). Given the estimates of the marginal likelihood, the AIC was used to penalise each model for the number of free parameters.

In Figure 6.9D the results of the model selection procedure for C536 are shown. Here, number of parameters of the model k and NMLL are shown. NMLL was estimated as $-ln\ \int p(D|\theta,M)p(\theta|M)\mathrm{d}\theta$ through MC integration across $1,000$ simulations with 500 trees each. To ensure that this resulted in a reasonably accurate estimation of the NMLL, a bootstrap of the data was applied during the estimation of the SL. The different NMLL

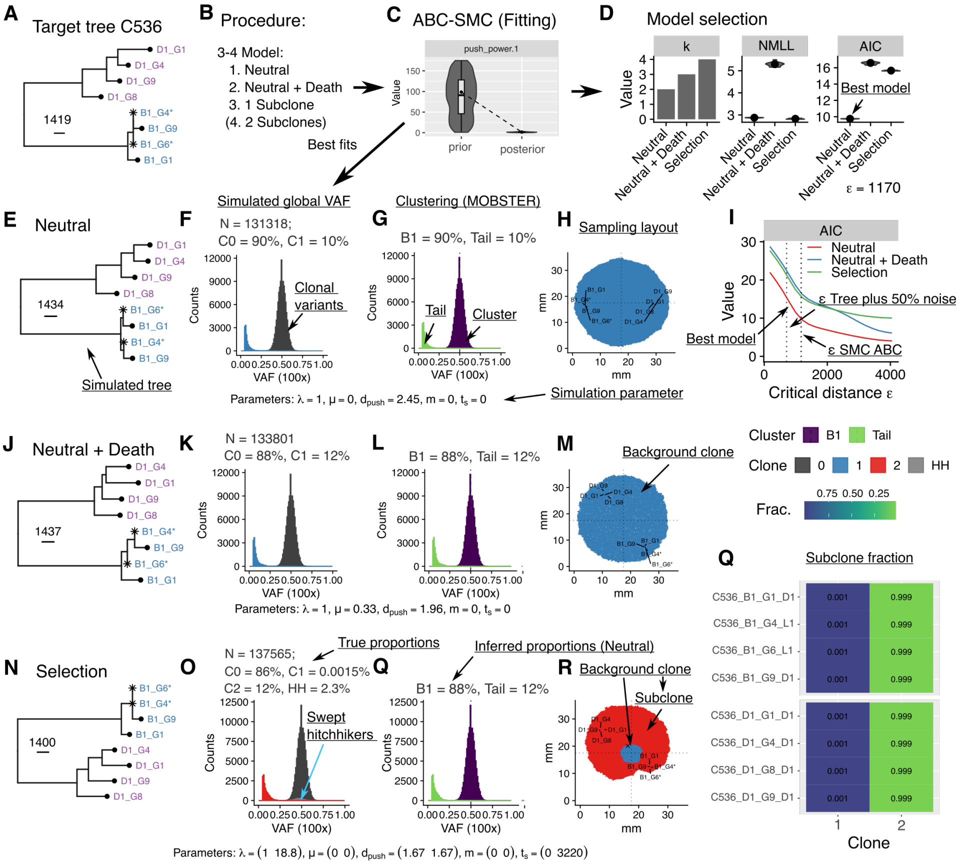


Figure 6.9: ABC-SMC inference framework applied to the ML LP-WGS tree of C536. A) The target tree of C536. B) For various models, parameters are inferred using the ABC-SMC inference. C) The resulting posterior distributions of each model are then used to estimate the marginal likelihood of the data at a given critical distance threshold ε. D) A model selection procedure using the AIC was then used to select the best model (i.e., 'Neutral'). E-H), J-M) and N-R) The best simulated tree, the associated simulated VAF spectrum of the entire tumour, the result of clustering this with MOBSTER and the spatial distribution of clones in space for the 'Neutral', 'Neutral+Death' and 'Selection' model respectively. I) The distribution of the AIC for various distances for all three fitted models obtained from C). Q) The fraction of accepted simulations in which samples were located in the ancestral (i.e., clone #1) and the selected clone (i.e., clone #2) respectively.

estimates that were obtained from the bootstrapped data are shown as violin plots around the point estimates in Figure 6.9D. In this case, no overlap of the bootstrapped distributions existed, demonstrating that inaccuracies of the MC integration were negligible in this case. Further, it can be seen that virtually no difference between the 'Neutral' and 'Selection' model with regard to the likelihood to observe matching trees under the two models existed. For the 'Neutral + Death' model, the marginal likelihood to generate matching data was even lower than for the less complex 'Neutral' model, a effect that is sometimes referred to as Bayesian Occam's razor (Murray and Ghahramani 2005). After penalisation for the number of free parameters of each model, the simplest 'Neutral' model was selected. To test whether this pattern was consistent for different distance thresholds ε, identical values were also calculated for a range of values. The results of this procedure are shown in Figure 6.9I, indicating that the Neutral model was consistently preferred over all alternative models for a wide range of ε.

Best model fits and additional statistics In Figure 6.9E, 6.9J and 6.9N simulated trees with the smallest distance to the target tree (Figure 6.9A) for the 'Neutral', 'Neutral + Death' and 'Selection' model are shown respectively. All of these were essentially indistinguishable from the target tree. This was, as shown in Figure S.127 (page 319), also true for the accepted trees in general. For each of the simulations from which the simulated trees were generated, a 'synthetic tumour bulk sample', with several clonal variants equivalent to the ones present in the sample tree, was simulated. These are shown in Figure 6.9F, 6.9K and 6.9O for each of the three trees respectively. In each of these Figures, variants are coloured by the actual clone or combination of these in which they were present. As seen in Figure 6.9F, around 90% of the simulated variants detected at a coverage of 100x are from the clonal cluster, with the remaining mass being located in the subclonal power-law tail. A similar pattern can be seen for the simulation with stochastic death (i.e., 'Neutral+Death') as shown in Figure 6.9K. Due to the different scaling of the tail in the presence of death, a slightly larger number of variants are present within the tail. From the VAF spectrum of the 'Selection' model, it is evident that the subclone effectively swept through the population, with a small number of 'hitchhiker mutations' present at a VAF > 0.45. This subclonal sweep can makes the obtained observations 'effectively neutral'.

Consistent, with these observations, the MOBSTER clustering method (Caravagna et al. 2020) correctly identified the presence of a neutral tail and the corresponding fraction of mutations for the two neutral simulations (Figure 6.9G, 6.9L) as well as for the essentially neutral simulation of the 'Selection' model (Figure 6.9Q). The spatial extend of simulated clones, and the locations of the individual samples are shown in Figure 6.9H, 6.9M and 6.9R for the three simulations respectively. From Figure 6.9R, it can be seen that none of the samples were taken from the ancestral clone (blue), which is consistent with the general pattern across all accepted simulations (Figure 6.9Q).

6.3.2.2 An Example of a Selected Subclone - C539

In contrast to this neutral case, the inference on the case C539, which was previously identified to contain an activating KRAS mutation (p.G12C) and associated elongated internal edge (see Figure 6.10A), suggested the presence of a selected subclone (Figure 6.10B). In addition to the three models considered in the example of C536 and in all other cases, various additional models were explored in this case (i.e., one subclone plus death, two subclones and two subclones plus death). While the model with two subclones and death had the lowest NMLL, after penalisation for the number of parameters using the AIC, the model with one subclone was selected consistently over the alternative model (Figure 6.10C). The best-fitting simulations of three selected models — neutral, one subclone and two subclones — are shown in Figures 6.10D,H&L respectively.

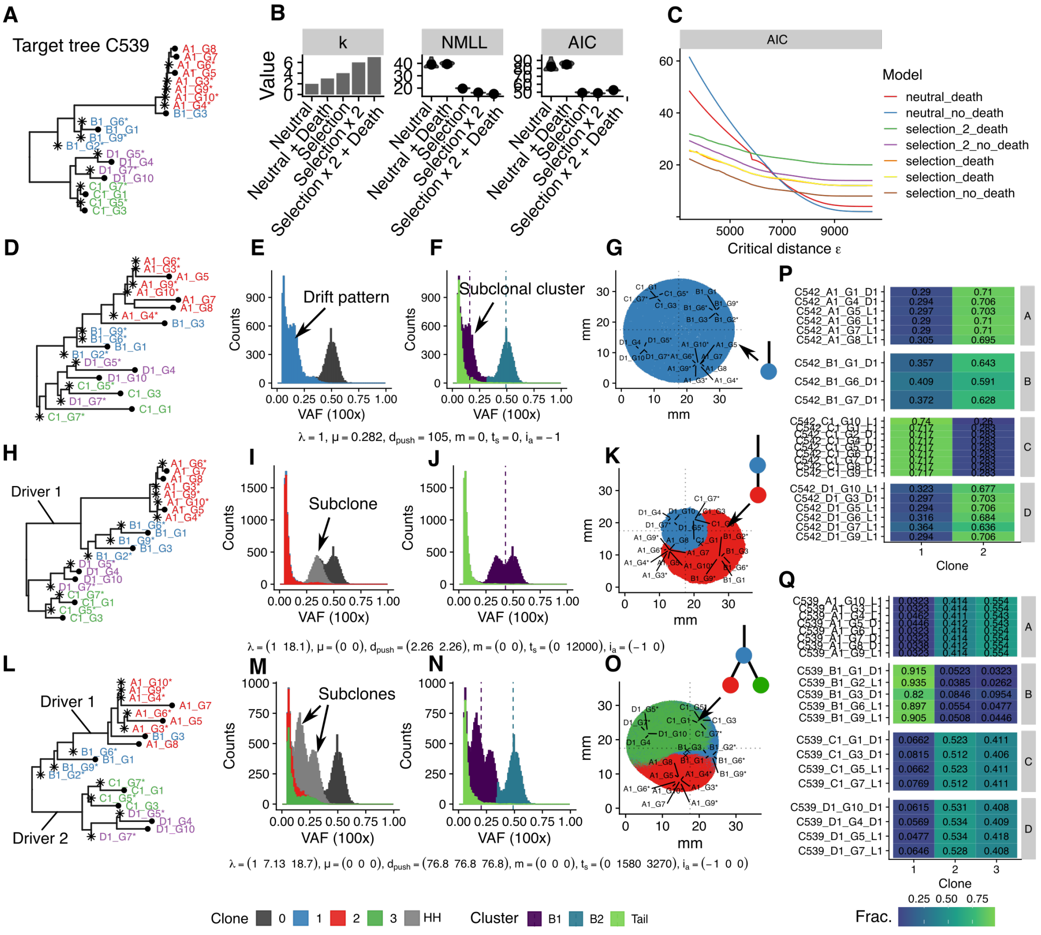


Figure 6.10: ABC-SMC inference framework applied to ML LP-WGS tree of C539. A) The target tree (WGS&LP) of C539. B) Model selection indicates that the preferred model is ‘Selection’ (i.e., one subclone). C) This is consistently the case over a wide range of ε. D-G) The simulated neutral tree closest to the target tree. E) The best simulation with one subclone. F) The best simulation with two subclones. G-K) Proportion of times individual samples fall into a given clone for simulations with one subclone. L-O) Like G-K for the model with two subclones. P&Q) The fraction of simulations in which each sample was associated with a given clone for the model with one and two subclones respectively.

Neutral model As seen in 6.10D the neutral model, indeed provided a poor fit to the target. While a substantially elongated internal edge, similar to that of the target, was present, the overall structure of the resulting trees was inconsistent with that of the target. From the simulated global VAF (Figure 6.10E) a shift of non-selected variants to a higher frequency resulting from stochastic drift can be seen. This pattern does not fit the expected power-law distribution and consequently gets detected by Mobster as a subclonal cluster (Figure 6.10F). Such extreme drift is very unlikely to occur, which is reflected by the very low marginal-likelihood (Figure 6.10B).

Model with one subclone The model with one selected subclone (Figure 6.10G) instead resulted in simulated trees reasonably similar to the target. From the sample locations drawn into the simulated tumour (Figure 6.10K), it can be seen that the samples from regions A&B were both located within the subclone and the corresponding edge of the simulated tree shown in Figure 6.10H is highlighted by the label 'Driver 1'. Figure 6.10P shows the fraction of accepted simulations in which each sample was located within the two clones (i.e., the ancestral clone and the subclone). From this summary of the accepted particles, it can be seen that all samples from regions A&B were consistently assumed to be located within the subclone and samples from C&D within the background clone. The simulated global VAF spectrum also showed clear evidence for the presence of a selected subclone (Figure 6.10I&J), which would likely be detected in a representative sample of the entire tumour.

Model with two subclones Finally, Figure 6.10L shows the best fit obtained from the model with two selected subclones. Here the edge of the clade formed by samples from regions C&D is explained through the introduction of a second selected subclone. This simulated driver is as highlighted by the 'Driver 2' annotation in the tree. In the instance shown here, samples from region B, except for sample B1_G3, were located in the ancestral clone (blue), samples from A within the first subclone (red) and samples from regions C&D within the second subclone (green) (Figure 6.10O).

Again, Figure 6.10Q shows that this is a pattern that was generally supported by the posterior set of simulated trees, with the majority of representative bulk samples obtained from the tumour showing the presence of selected subclones similar to those in Figure 6.10 M&N.

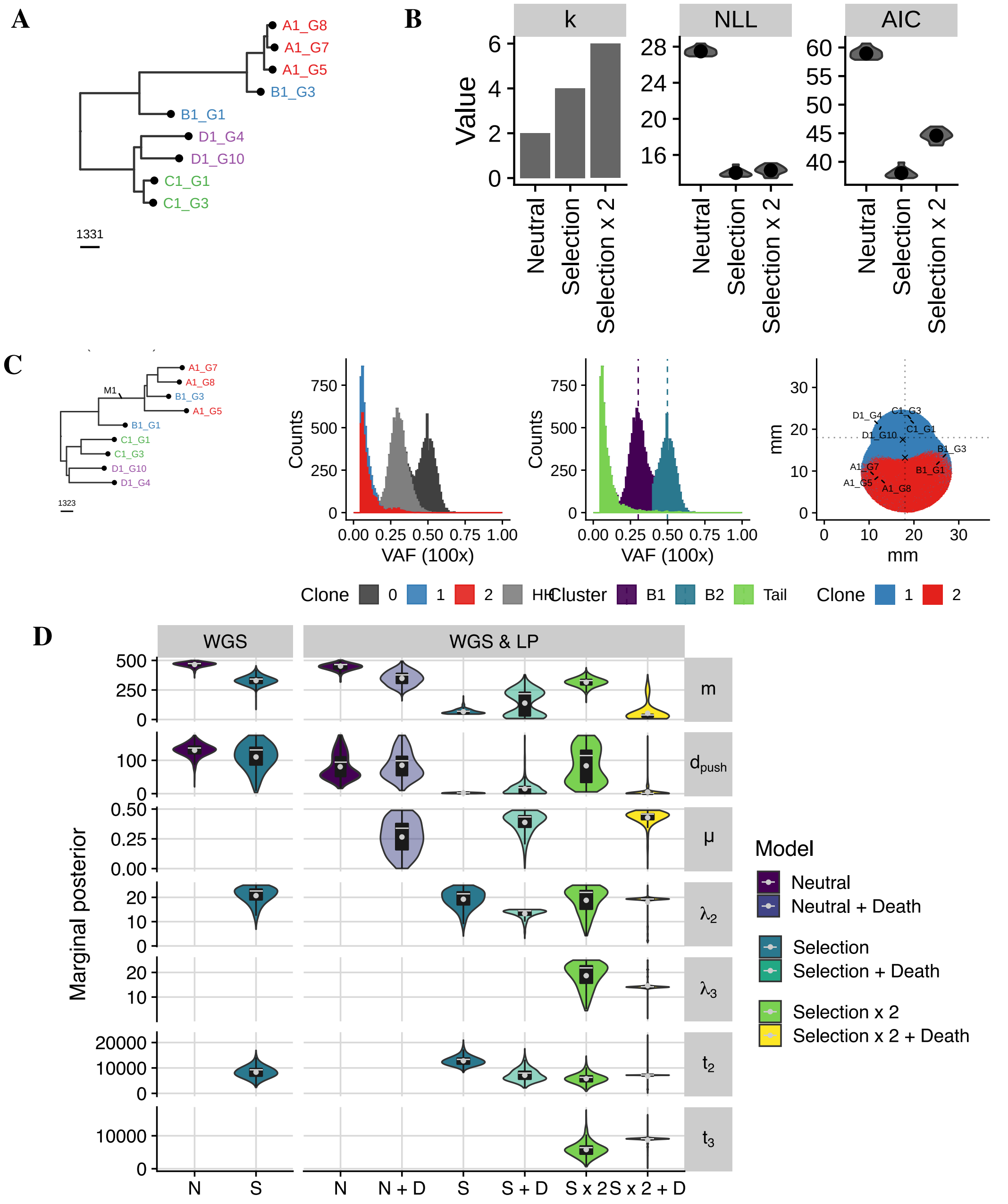


Figure 6.11: Result of ABC-SMC inference for the WGS tree of C539. A) Target tree of C539 containing only deeply sequenced WGS samples. B) The result of the model selection indicates that the model with one subclone ('Selection') and two subclones ('Selection x 2') provide a reasonable explanation of the data. After penalisation of the models for their respective complexity with AIC the 'Selection' model is picked. C) The best fit of the 'Selection' model. D) The posterior distribution of the parameters for both the models fitted to the WGS tree and the LP-WGS tree also including the LP-WGS samples (see Figure 6.10).

Posterior distributions To evaluate whether the assignment of the LP-WGS sequenced samples could have potentially altered the results of the inference and subsequent model selection; the same procedure was also applied to the trees containing only the deeply sequenced WGS samples (Figure 6.11A).

Reassuringly, the inference obtained very similar results from this tree in terms of both i) the selected model (Figure 6.11B) and the inferred posterior distribution of model parameters (Figure 6.11D). Figure 6.11C shows a representative example of a simulated tree from the posterior tree set of the 'Selection' model, which again has a similar structure to those inferred from the trees that included LP-WGS samples (see Figure 6.10H).

6.3.2.3 Model Selection Results Across the EPICC Cohort

In order to establish how many of the analysed cases exhibited evidence of subclonal selection, the ABC-SMC inference — described for two representative cases above — was applied to the entire cohort. Many different criteria can be used for the selection of models, and here, three different criteria were used, the marginal likelihood, the AIC and the AIC with a minimum discriminatory power of $\Delta\text{AIC} > 4$. Each of these penalises the model complexity differently. The result of the number of times each of the tested models was selected based on these criteria is summarized in Figure 6.12.

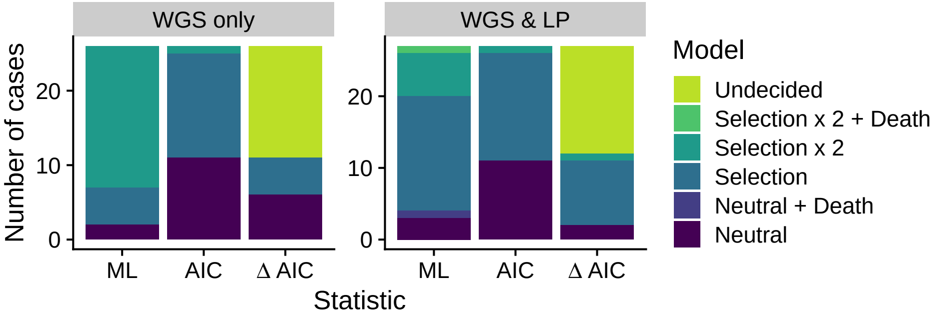


Figure 6.12: Summary of model selection results from ABC-SMC with the fixed sampling setup. The results shown are based on the two tree sets (WGS and LP+WGS), using different model selection criteria. In order of increasing penalty, these are the marginal likelihood (ML), the AIC and lastly ΔAIC where only models with a minimum difference $\Delta\text{AIC} > 4$ were classified.

Marginal likelihood As expected, given the increased flexibility of the non-neutral models, the marginal-likelihood often selected more complex models with multiple subclones over the neutral ones. This was especially the case in the absence of additional LP-WGS samples (1/27 neutral, 9/27 one subclone, 17/27 two subclones), but also for the trees in which the LP-WGS samples were included (1/27 neutral, 16/27 one subclone, 4/27 two subclones).

Consistent with the previously described results, the addition of stochastic death[4] never provided a better description of the data compared to the alternative models. The two cases, C544 and C555, in which the marginal likelihoods of the 'Neutral + Death' model were lower, differences were within the error of the MC integration (see Figure S.137, page 325 and Figure S.138, page 326). This observation makes sense in light of the tissue architecture of colorectal cancers, which are composed of individual crypts with multiple, closely related stem cells.

AIC After penalising the different model for the model complexity using AIC $= 2k - 2ln(L)$, in $\approx 40\%$ $(11/27)$ of cases, the neutral model was preferred over the alternative models, suggesting that in the remaining $\approx 60\%$ some structures of the trees required subclonal selection to be explained. In only one case, C518 (see Figure S.136, page 324), a model with two subclones was selected. I next assessed whether trees were consistently classified as neutral or non-neutral for both available tree sets. As shown in Figure 6.13 this appeared to be the case for the majority of cases $(18/26, \approx 70\%)$. Still, four cases (C559, C562, C543, C560) were classified as neutral when using the WGS trees and non-neutral when using the WGS & LP-WGS samples. The opposite was the cases for four other cases (C544, C528, C530 and C554).

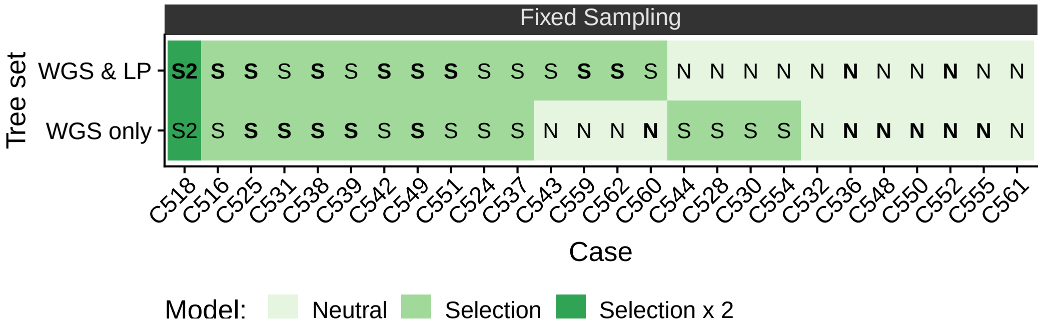


Figure 6.13: Agreement of the model selection based on LP+WGS sample and WGS sample trees for the fixed sampling layout. Model sets with ΔAIC > 4 are highlighted by bold text within the tiles.

A review of the corresponding model selection data (see Figure S.128-S.135, page 320-323) showed that for all but C559 and C562 the differences in the AIC between models was very small (i.e., < 4). This indicates that the best model was only marginally preferable over the alternative one, and as such, the less complex neutral model might be preferred over the more complex model in C543, C560 (ΔAIC < 4) as well as C544, C528, C530

[4] Tested on the ML LP-WGS trees only.

and C554 (i.e., all cases where the addition of LP-WGS lead to classification as neutral). In this case, $16/26$ or $\approx 50\%$ of cases would be classified as neutral. For, C559, one of the two cases in which the errors of the MC integration did not explain the differences in model selection, the more complex 'Selection' model indeed provide a much better explanation of the observed data (Figure S.139, page 327). Why the addition of a single LP-WGSsample to the case of C562 had such a drastic effect on the selected model (Figure S.140, page 328) is unclear. Most like this resulted from overfitting despite the penalisation using AIC. In general, results of model selection obtained for the two different tree sets were either identical or explainable by small insignificant differences in the AIC.

I also tested whether cases inferred to be neutral had a lower number of samples on average, as this could indicate the lack of power to detect the presence of selection in these cases. Reassuringly, in line with the relative consistency of the results of the ABC-SMC classification of cases across both datasets, no such differences were observed (Figure 6.14),

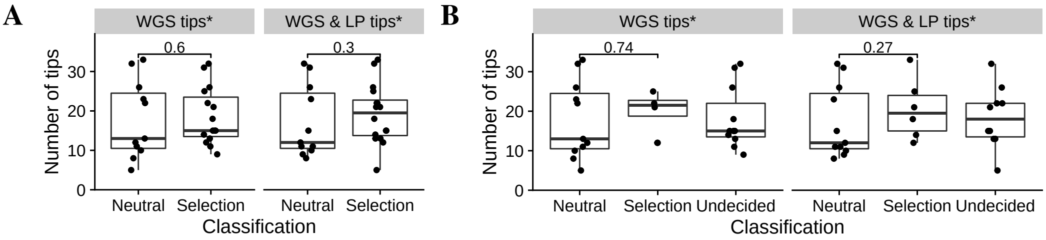


Figure 6.14: Relationship of the number of samples on the selected model for the fixed sampling setup. A) Shows that for neither of the two datasets a significant difference in the number of available samples existed between the cases classified as neutral ('Neutral') and non-neutral ('Selection') for the inference using the fixed sampling scheme. This suggests that insufficient power did not cause the classification of a large number of cases as 'Neutral'. B) The same, but with all samples where discriminatory power was insufficient to choose between two or more alternative model (i.e., $\Delta\text{AIC} \leq 4$) shown as 'Undecided'. Again, no difference in the number of samples between cases classified as neutral and non-neutral was evident.
* Cancer samples only.

6.3.2.4 Posterior Model Parameters

While the differences in the selected model were reasonably small for both datasets (i.e., WGS and ML LP-WGS trees), the inclusion of the LP-WGS samples into the inference led to some differences in the inferred posterior distribution of the parameters. These distributions are summarized for all cases in Figure 6.15. As shown in this figure, the ABC-SMC inference suggested that the data supported a model of boundary driven growth in a subset of cases[5] (i.e., $11/27 \approx 41\%$) and non-boundary driven growth in other cases[6] (i.e., $7/27$

[5] C536, C532, C544, C552, C554, C548, C543, C559, C538, C549, and C539

[6] C561, C530, C528, C522, C550, C551, and C518

≈ 26%).

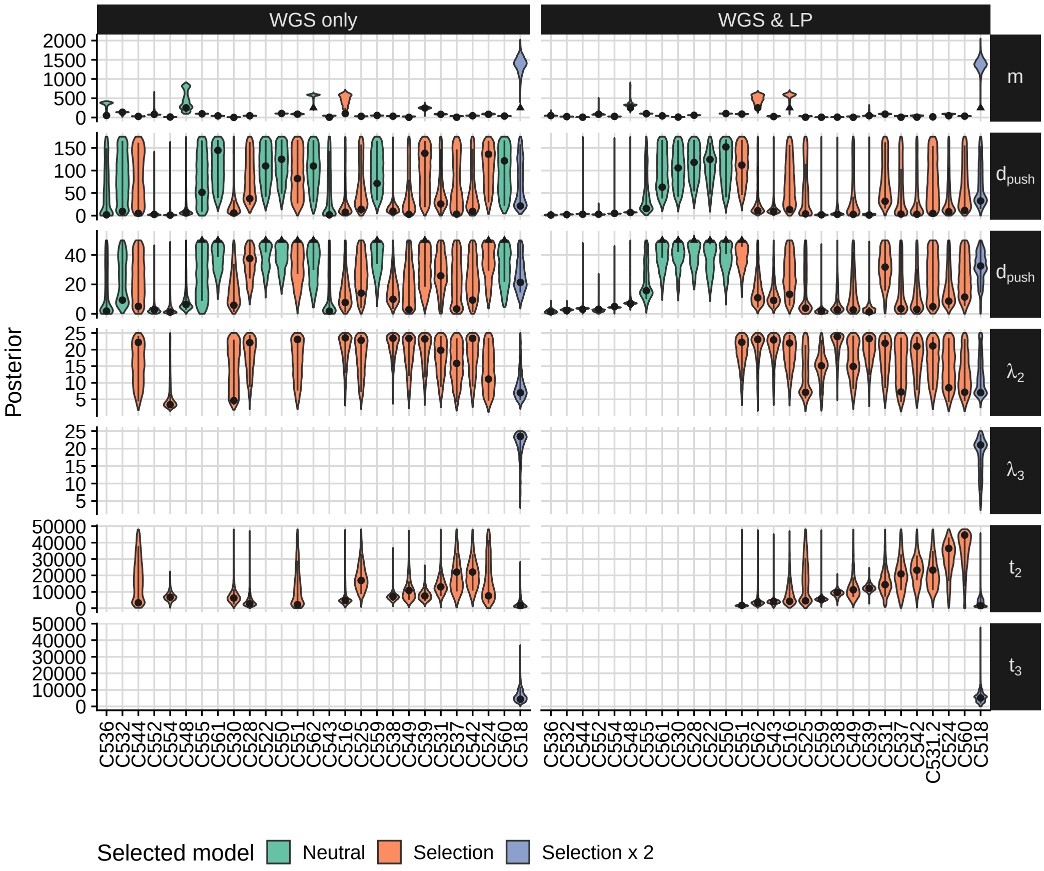


Figure 6.15: Marginal posterior distribution of inferred parameters obtained from the ABC-SMC using the fixed sampling setup. For the parameter estimates of the width of the growing outer rim d_{push} a smaller window $(0,50)$ of the posterior is shown (fourth row). Mutation rates m with a value over 250 (i.e., cases with MSI) are truncated (indicated by a small triangle). Abbreviations: m mutation rate, d_{push} width of growing edge, λ_i birthrate of subclone i and t_i start time of subclone i. The lines indicate the central 90% intervals of the marginal posterior distribution, dots the multivariate estimates of the maximum a posteriori probability.

Secondly, the posterior on the birthrates of subclones was in general extremely wide. In light of the sampling scheme used, this makes sense as the information on the spatial extent of an expanded subclone was severely limited. Essentially, only the number of regions over which a subclone spreads can be identified, leaving a significant uncertainty in the actual clone size. As this size of a clone, in combination with t_i, provides information on the value of λ_i, the large prediction interval observed are expected.

A high level overview of how the inferred parameter and subclones are related to overall structure of the trees, is shown in Figure S.154 (page 334, WGS trees) and Figure S.153 (page 333, LP-WGS trees). From these two figures, it is obvious that the introduction of selected subclones explained different aspects of the trees incompatible with the neutral model. This is foremost the presence of elongated internal edges, like in C539 or C538, but also the presence of clade formed by samples from different regions with a more recent MRCA as seen, for example, in C518.

6.3.3 Variable Sampling & Overdispersion

Further evaluation of the posterior trees derived using the fixed sampling layout showed that, while the structure of the accepted trees was generally similar to the ones observed, specific aspects of the tree topology and shape were not. For example, compared to internal edges, the length of terminal edges was often consistently too long (see for example Figure 6.10H or 6.10L). I suspected that such differences would be reduced by locating the sampling regions closer to the edge of the simulated tumour, which some simple tests confirmed. Further, questions regarding the effect of the relative sampling position along the edge of the tumour arose, especially whether moving two sampling regions closer to each other would emulate features otherwise attributed to subclonal selection (e.g., clades formed by two adjacent regions).

Variable sampling model For this reason, I derived an alternative simulated sampling schema that took the uncertainty of the conducted sampling layout into consideration. The details of this are outlined in Section 6.2.2 above, but in brief, a Dirichlet prior on the angle ϕ between sampling regions and a Beta prior on their distance from the edge d_e were added to the model. The average position relative to the edge was assumed to be around 90%, which upon reconsideration of the actual sample collection performed appeared to be more realistic. The average angular position was again assumed to be at 90, 180, 270 and 360 degrees.

In addition to this variation of the sampling positions, I also allowed for increased variability of individual edge lengths within the tree. This was motivated by the assumption that spatio-temporal variations of the mutation rate or drift within the stem-cell compartment of glands could cause such an overdispersion to arise. For this I replaced the Poisson distributed number of mutations accumulated per generation with a Negative-Binomial distribution. The Negative-Binomial distribution was parametrised so that an additional parameter $\psi \geq 0$ would control the amount of overdispersion. This additional parameter ψ was then also inferred as part of the ABC-SMC. For $\psi = 0$, the distribution of edge length is equivalent to a Poisson distribution and as ψ increases, the length of individual edges in the tree become more variable.

6.3.3.1 Results of Model Selection

Again, the classification and inference framework — this time using the modified version of the simulated sampling — was applied to both WGS and ML LP-WGS trees from the entire cohort. The results of the model selection for the alternative model are summarised in Figures 6.16 and 6.17. Summary plots of the results in each case are shown in the Figures S.164-S.192 (page 340-368).

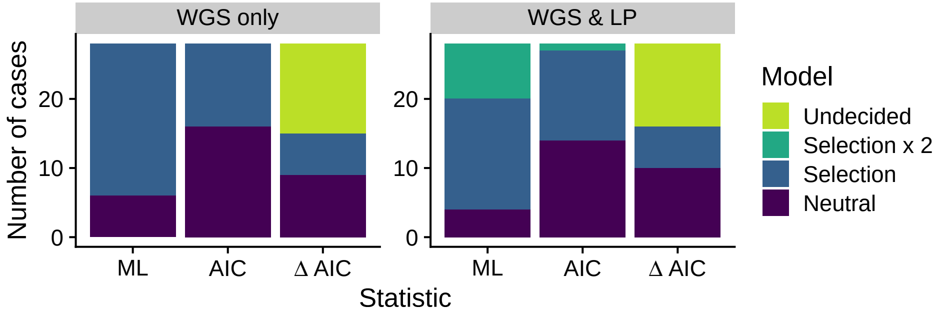


Figure 6.16: Summary of model selection results from ABC-SMC with the variable sampling setup. The results shown are based on the two tree sets (WGS and LP+WGS), using different model selection criteria. In order of increasing penalty, these are the marginal likelihood (ML), the AIC and lastly ΔAIC where only models with a minimum difference $\Delta\text{AIC} > 4$ were classified.

When using this alternative setup, a substantially larger fraction of cases were considered to be compatible with the neutral model. Specifically, $16/27$ ($\approx 59\%$) and $15/27$ ($\approx 55\%$) of cases were classified as neutral based on the WGS and ML LP-WGS trees respectively. An even larger fraction of cases — $10/16$ ($\approx 62\%$) and $13/17$ ($\approx 76\%$) of cases respectively — of those in which the $\Delta\text{AIC} > 4$ suggested sufficient discriminatory power, were classified as neutral.

Comparison of the different classifications of individual cases showed that these were generally consistent with results obtained previously (Figure 6.17). Notably, the six cases[7], which were specifically discussed before (i.e., undecidable with regard to the preferred model), were classified as neutral in this setup. As such, no unexpected changes of the general classifications did occur, but the overall structure of the simulated trees was more similar to the observed ones.

Finally, I assessed, as before, whether the number of tips differed significantly between cases classified as neutral and non-neutral to elucidate whether the number of tips might have limited the ability to classify cases. As shown in Figure 6.18, this was again not the

[7] C543, C560, C544, C528, C530, and C554

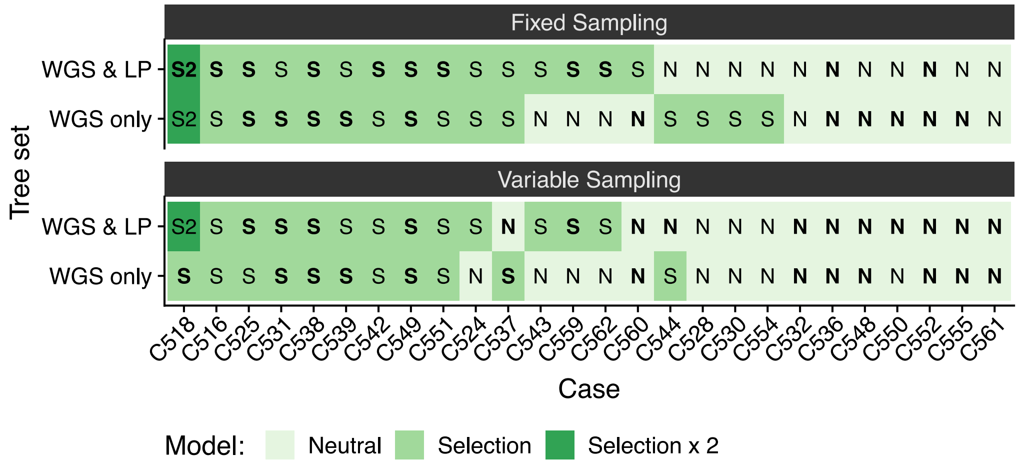


Figure 6.17: Agreement of the model selection based on LP+WGS sample and WGS sample trees for both sampling setups. Model sets with ΔAIC > 4 are highlighted by bold text within the tiles.

case, indicating that the number of assessed samples did not generally limit the ability to detect selection.

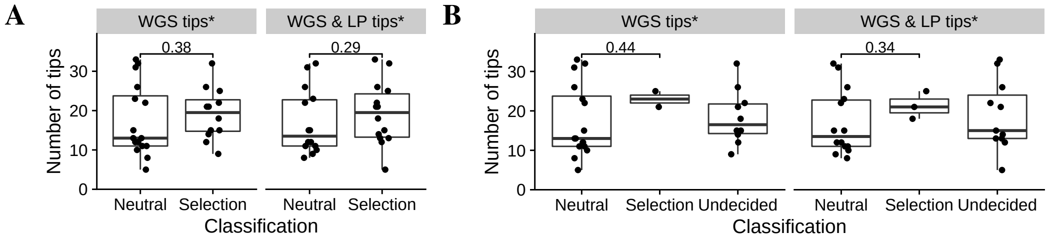


Figure 6.18: Relationship of the number of samples on the selected model. A) Shows that for neither of the two datasets a significant difference in the number of available samples existed between the cases classified as neutral ('Neutral') and non-neutral ('Selection') for the inference using the variable sampling scheme. This suggests that insufficient power did not cause the classification of a large number of cases as 'Neutral'. B) The same, but with all samples where discriminatory power was insufficient to choose between two or more alternative models (i.e., ΔAIC ≤ 4) shown as 'Undecided'. Again, no difference in the number of samples between cases classified as neutral and non-neutral was evident.
* Cancer samples only.

Summarising, the results from the classification obtained using the ABC-SMC inference suggest that the trees observed in $\approx 40\%$ of cases indicated the presence of a selected subclone in at least one of the regions. This is between the estimates of Williams et al. (2016) of $\approx 60\%$ of analysed colon cancers with deviations from the expected VAF spectrum and the $\approx 20\%$ of analysed cases reported by Williams et al. (2018b). I applied the $1/f$ test statistic to simulations from the posterior of the ABC-SMC in each case and observed that both, cases with boundary driven growth (i.e., $d_{push} < 10$) and subclonal selection, frequently had values of $R^2 < 0.98$ (see Figure S.157A, page 336 and Figure S.158A-B, page

337). This indicated that the $1/f$ statistic used by Williams et al. (2016) might also have identified a subset of neutral colorectal cancers with non-exponential growth and provides an explanation for the slightly higher fraction of non-neutral tumour identified by them. This has indeed been suggested by Wang et al. (2018a) before. Given the improved power to detect selection events in the single-gland sequencing data of the EPICC cohort, the slightly larger fraction of cases with evidence for the presence of subclonal selection compared to the study by Williams et al. (2018b) appear reasonable.

6.3.3.2 Overdispersion and Sampling Locations

I next assessed the posterior distribution of the dispersion parameter ψ and the distribution of the sample locations of the accepted particles to identify cases in which the added flexibility of the model was required to explain the data sufficiently. The marginal posterior distribution of ψ is shown, together with those of the other parameters, in Figure 6.21. As seen here, the posterior distributions of ψ generally suggested that only a relatively small amount of overdispersion was present in the data. Notable exceptions from this observation were the cases C516, C549, C538, and C562 (see Figure 6.21, compare Figure S.155, page 335).

In two of these cases, a very small number of samples were obtained (i.e., C516 and C562), and here the larger value of ψ only improved the fit of the edge to the sample closest to the root of the tree. The remaining two cases with large posterior values of ψ were still inferred to contain a selected subclone (i.e., not 'Neutral'), suggesting that some general variability within the tree was insufficiently explained by the introduction of a selected subclone into the model and absorbed by increased variability of the edge lengths. While the specific reason for this is elusive, it is important to note that the larger degree of dispersion allowed by the model did not cause a different classification of either of these two cases compared to the model that did not include such a parameter (see Figure 6.17).

A similar observation was made for the relative positioning of samples within the tumour. Here the variability of ϕ and d_e meant that samples could have theoretically been placed much closer to the edge of the tumour or closer to each other. If such changes in the positioning of samples would have consistently led to improved fits of the model, these patterns should be observable in the marginal distribution of the sample locations within the tumour. Still, in general little deviation of the sample locations from the prior locations did occur. A representative example of this can be seen in Figure 6.19 (C561).

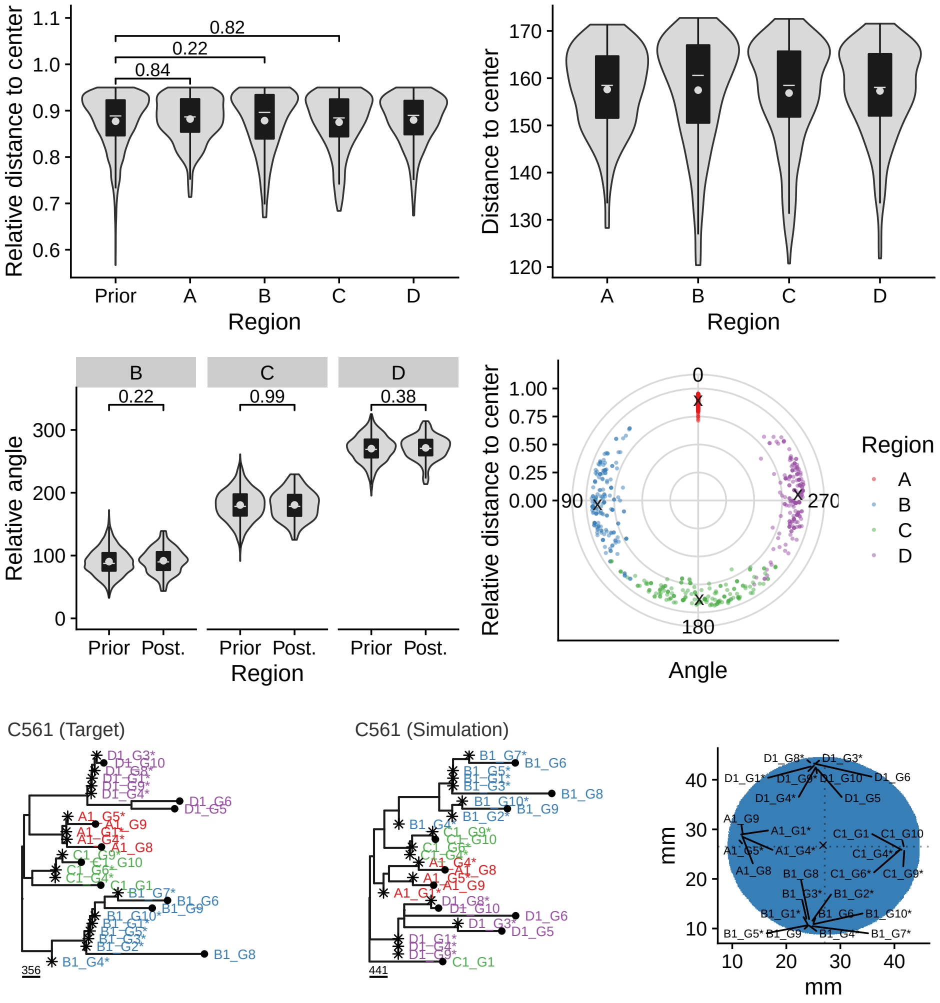


Figure 6.19: Regional sample positioning in case C561.

As for the majority of other cases, only small deviations from the prior did occur in this case, with the mean of the region position being at the average position of the prior.

A counter-example showing a clear deviation from this pattern can be seen in the fit of the, still inadequate, neutral model fitted to tree of C518 shown in Figure S.156 (page 336). Here, the regions A&B were consistently moved close to each other in space, thus sampling from a clone patch with a more recent MRCA. Given the priors defined on the positioning of samples in space, the non-neutral models were still favoured over the neutral model, leading to the selection of the model with two subclones in this specific case.

In summary, these results confirm that while the relative sampling positions were relatively unimportant in many cases, a couple of exceptions did exist. In these, non-neutral models were sometimes preferred to explain the observed patterns (e.g., C518), thus leaving the question of whether alternative growth models (e.g., 3D or growth along existing spatial structures) would explain the observed trees better.

6.3.3.3 Subclonal dN/dS After ABC-SMC Classification

While the general consistency of the ABC-SMC across sampling setups and different datasets was on its own reassuring, I next sought to test if an excess of non-synonymous mutations suggested the presence of subclones in the cases classified as ‘Neutral’. For this, an approach similar to the one used by Tarabichi et al. (2018) in their criticism of the $1/f$ method was followed. In brief, subclonal variants identified in the entire cohort were split based on the microsatellite stability of cases and dN/dS ratios for various sets of genes estimated with the *dndscv* model (Martincorena et al. 2017). Based on the results of the model selection of the ABC-SMC inference, all cases were divided into a ‘Neutral’ and ‘Non-Neutral’ group. dN/dS ratios were then calculated on the subclonal variants of these two groups. A summary of this analysis can be seen in Figure 6.20A and the obtained dN/dS estimates for the subclonal variants in the different sets of patients are summarised in Figure 6.20B.

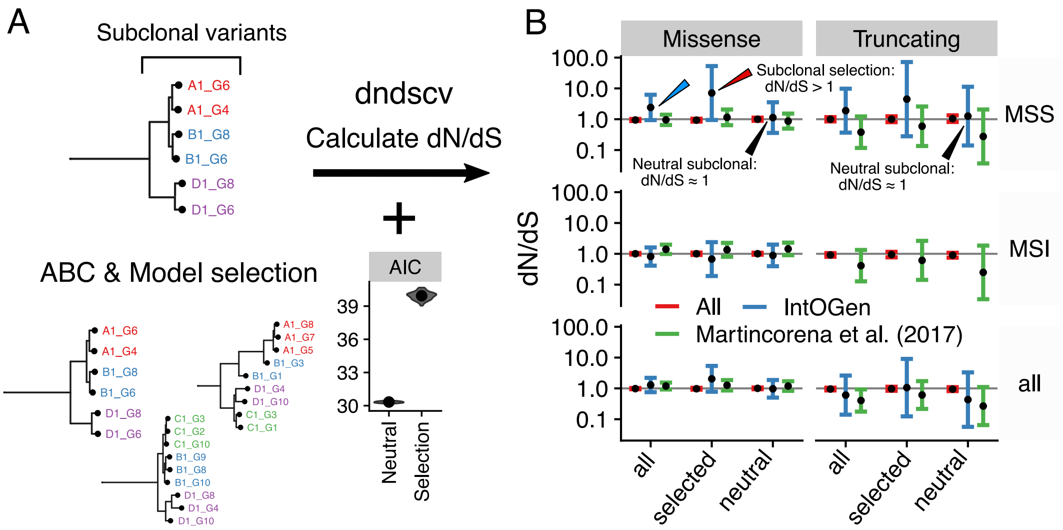


Figure 6.20: dN/dS estimates from subclonal variants based on ABC-SMC classification. A) The set of all subclonal variants was split based on the model selection performed by the by the ABC-SMC inference method.B) dN/dS estimates by *dndscv* (Martincorena et al. 2017) from subclonal variants of all, MSI and MSS colorectal carcinomas (y-axis grids) for missense and truncating variants (x-axis grids) in the entire genome (‘All’), colorectal driver genes defined by IntOGen (Martínez-Jiménez et al. 2020) and a set of pan-cancer drivers from Martincorena et al. (2017). The results show a general excess of subclonal non-synonymous variants in driver genes for MSS cases (blue arrow). After classification of cases with the ABC-SMC inference method, evidence of positive selection was found in those classified ‘selected’ (red arrow), but not in those classified as ‘neutral’ (black arrows).

As seen here, subclonal dN/dS estimates of known driver genes were markedly elevated above one when all cases were analysed (blue arrow in Figure 6.20), thus indicating the presence of subclonal selection in a subset of cases. Reassuringly, the point estimates of the dN/dS values for which the presence of a subclone was inferred as part of the ABC-SMC inference were even higher, consistent with the presence of a clone with a selective advantage due to a somatic mutation (red arrow in Figure 6.20). In contrast, the dN/dS point estimates for subclonal variants of the remaining cases (i.e., those classified as neutral), was almost exactly dN/dS $= 1$ (black arrow in Figure 6.20). This is consistent with the absence of selected subclones in the observed part of the tumours.

Altogether, this analysis suggests that the ABC-SMC classification was indeed able to identify trees for which the structure suggested the presence of subclonal selection in some parts of the tumour. While potentially underpowered due to the relatively small number of analysed cases (27), the majority of cases (i.e., $\approx 60\%$) appeared to be better explained by a very simple spatial model without subclonal selection. Consistent with this, the orthogonal dN/dS analysis also did not reveal an excess of non-synonymous mutations. While it is certainly possible that non-genetic drivers could have cause widespread subclonal selection, a analysis of chromatin accessibility changes and differential expression[8] using ATAC-seq and RNA-seq conducted as part of the EPICC study did not reveal any evidence for this. However, as mentioned before, alternative model including more complex spatial dynamics, like immunoselection, necrosis or structures of the surrounding tissue, should certainly be considered to assess whether these would provide a better fit to the data.

6.3.3.4 Posterior Distributions

To evaluated the posterior distribution of the individual parameters, the marginal posterior distribution of all parameters were determined. These are summarised for both datasets and all cases in Figure 6.21. In the following, I will discuss specific aspects of these marginal posterior distributions.

Boundary vs non-boundary driven growth Consistent with the previous results, some cases were found to only be compatible with boundary driven growth (see panel d_{push} in Figure 6.21). Among these were $5/6$ of the neutral cases[9] for which the inference using the fixed sampling setup suggested the presence of boundary driven growth (i.e., $d_{push} \leq 20$). Similarly, the 5 neutral cases[10] for which the inference previously suggested non-boundary

[8] This work was done by Jacob Househam.

[9] C536, C532, C544, C552, and C554. Not C548.

[10] C522, C528, C530, C550, and C561.

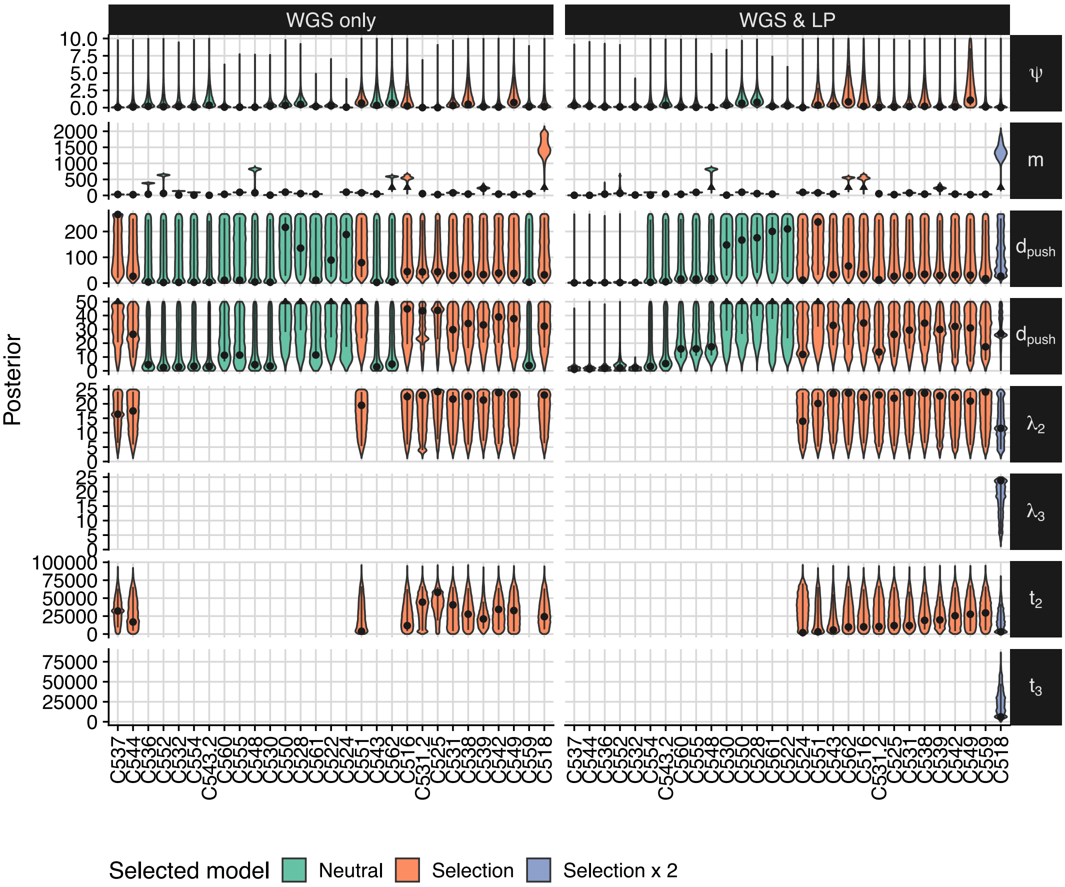


Figure 6.21: Marginal posterior distribution of inferred parameters obtained from the ABC-SMC using the variable sampling setup. For the parameter estimates of the width of the growing outer rim d_{push} a smaller window $(0, 50)$ of the posterior is shown (fourth row). Mutation rates m with a value over 250 (i.e., MSI cases) are truncated (indicated by a small triangle). Abbreviations: ψ dispersion parameter, m mutation rate, d_{push} width of growing edge, λ_i birthrate of subclone i and t_i start time of subclone i. The lines indicate the central 90% intervals of the marginal posterior distribution, dots the multivariate estimates of the maximum a posteriori probability.

driven growth, were generally also so in the new variable sampling setup. One additional case[11], which previously classified as non-neutral, was also found to be consistent with neutral boundary driven growth. In the remaining cases, wide posterior intervals suggested that little information on the strength of boundary driven growth was contained in the data.

Mutation rates Mutations are expected to accumulate proportional to the number of alleles (i.e. copy-numbers). For this reason, mutation rates obtained from the posterior distribution of the m parameter were adjusted to account for the differences in the ploidy of individual cases. After this correction, the median of the maximum a posteriori probabilitys (MAPs) of the mutation rates in the tumours was 78.6 and 540 per gland division per diploid genome for MSS and MSI cases respectively. Assuming an effective genome size of $2.9 \cdot 10^9$ base

[11] C537.

pairs, this corresponds to a mutation rate of $1.35 \cdot 10^{-8}$ and $9.31 \cdot 10^{-8}$ per base and gland division respectively. The relative difference of a ≈ 6.8 fold difference between MSI and MSS cases is generally consistent with those reported previously. Still, the rates themselves are lower than those estimated by Williams et al. (2016) and Williams et al. (2018b) from bulk WGS data.

Previous studies have suggested that boundary driven growth, which is not part of the model used in either of the two studies, would lead to the overestimation of mutation rates (Fusco et al. 2016; Schreck et al. 2019). To confirm this, I tested the performance of the $1/f$ test statistic on these cases and found that in those with boundary driven growth, estimates of the mutation rates were indeed larger than the true rates (Figure S.158C, page 337C and S.157B, 336). Still, as the majority of cases with strong boundary driven growth had a $R^2 < 0.98$ the majority of these would have been expected to be excluded from the analysis performed in Williams et al. (2016) (see Figure S.158A&B, page 337 and S.157A, page 336).

I also assessed whether a high false negative rate during the calling of mutations in single-gland samples with Mutect2 might explain the observed discrepancy (see Figure S.163, page 339), but found that the power to identify clonal mutations in individual samples was generally above $> 90\%$ (i.e., FNR < 0.1). It might be possible that a significant amount of false positive low-frequency variant calls caused the estimates in previous studies to be too high. As this problem would not exist for the single-gland WGS data analysed here and for this reason the obtained estimates might indeed be more accurate (Salcedo et al. 2020).

Subclone parameters Consistent with the previous observations, the posteriors of the clone specific parameters λ_i and t_i were very wide. Especially, the birthrate varied over a large range of values. The likely reasons for this were described in detail above, but in brief, very little information on the size of individual clones is contained in the generated data. As this is the main measure that allows estimating λ and t_i, large posterior intervals are expected. For more precise estimates, a different sampling layout would have had to been used.

Conclusion Assuming that the inferred estimates of d_{push} are informative, one would assume that the absence of spatial limitations on the tumour growth might be a property that is associated with a generally more invasive phenotype. For this reason, cases with a low estimate for d_{push} might be expected to have a worse prognosis. Likewise, the presence of a faster-growing subclone might be indicative of a worse outcome. As the entire study was planned as prospective study data on the outcome will be available in the future and as these come available, evaluation of these two hypotheses should be possible.

6.3.3.5 Tree Topologies

In Figure 6.22 critical parameters of the ABC-SMC inference, as well as the classification of cases along the trees used for the inference, are shown alongside each other. Edges of the tree that were consistently associated with a simulated 'Driver' mutation in the 200 best fitting trees are highlighted in colour for the cases classified as non-neutral. In cases where somatic subclonal drivers were identified as part of the previous analysis, these were also added as labels to the tree.

For a number of these highlighted (i.e., selected) edges a subclonal driver mutation was identified. These observations suggest that the subclonal mutations in Table 6.2 had detectable fitness effects in the respective genetic and environmental context in which they occurred. In the following paragraphs I will discuss each of these putative driver mutations in detail and in context of existing literature. In the majority of cases, this assessment showed that the identified subclonal driver mutations did indeed provide a reasonable explanation for the observed selection.

Table 6.2: Likely subclonal driver mutations identified by the ABC-SMC inference on the trees of the EPICC cohort.

Case	Gene	Mutation	Type	Figure
C518	PTEN	p.C136R	Second hit of TSG	S.141, page 329
C524	PIK3CA	p.C378R	Oncogene	S.142, page 329
C525	PIK3CA	p.Q546P	Oncogene	S.144, page 330
C531	SMAD4	p.A118V	TSG	S.145, page 330
C538	RNF43	p.D153*	TSG	S.146, page 330
C539	K-Ras	p.G12C	Oncogene	S.147, page 331
C542	chr1p	Loss	CNA	S.108, page 311

Activating K-Ras mutation Activating mutations in classic colorectal oncogenes, like K-Ras and PIK3CA, lead to the over-activation of signalling pathways[12] and are expected to have a dominant effect. For this reason the observation of a fitness altering effect for these mutations is not surprising.

Especially for the K-Ras p.G12C mutation observed in C539, these would be expected.

[12] The RAS and the PI3K/AKT pathway respectively.

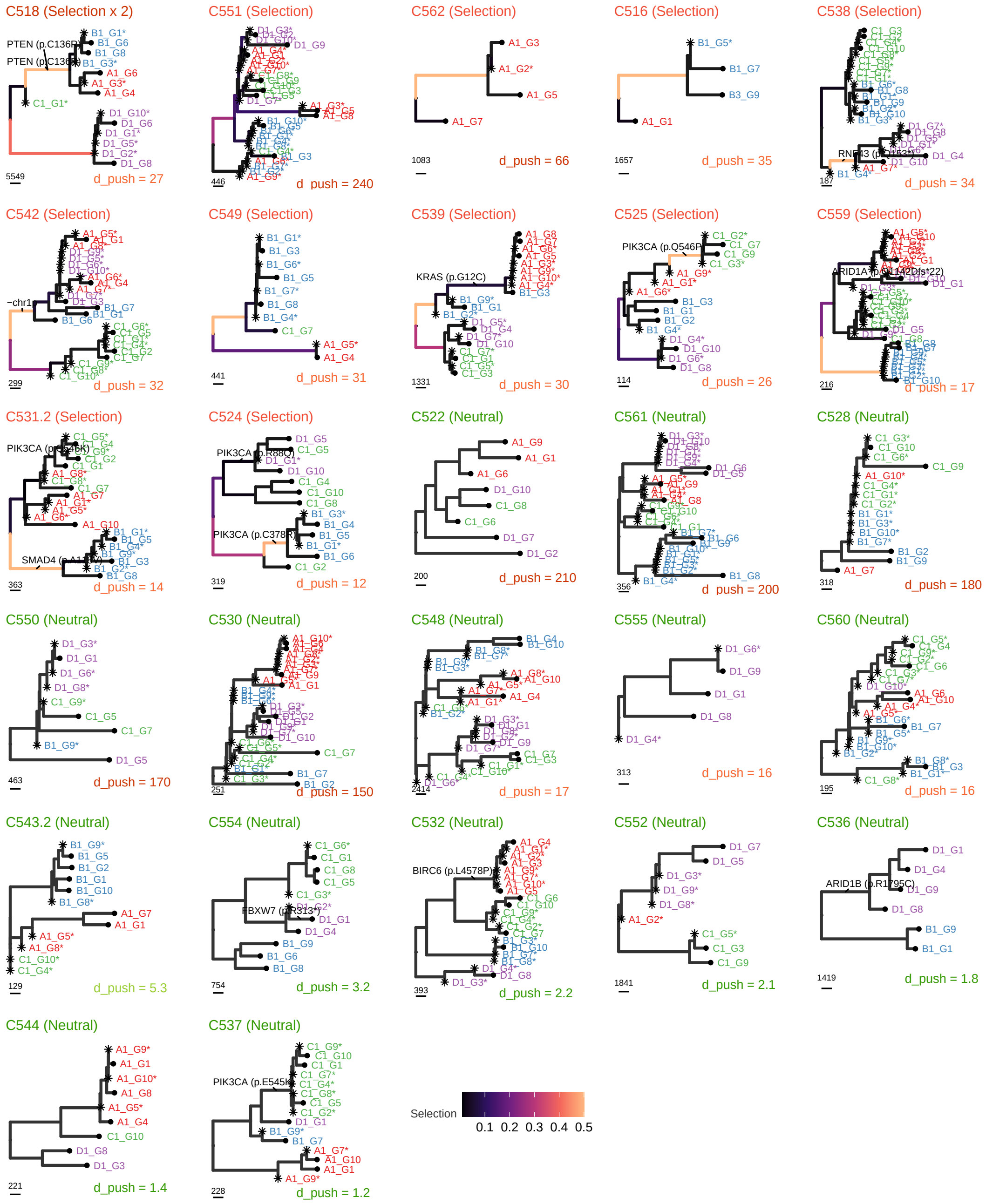


Figure 6.22: ML LP-WGS trees and associated results of ABC-SMC inference. The classification of cases based on the AIC (i.e., 'Neutral', 'Selection' and 'Selection x2') are shown above trees. The MAP estimates of the strength of boundary driven growth for the selected model are shown below the trees. The frequency colours edges of the trees that these were associated with the introduced subclonal 'Driver' mutation in the 200 best fitting trees from the posterior distribution (legend in the bottom left).

Alterations of RAS genes are the single most common oncogenic mutation acquired in various human malignancies (Bos 1989). In colorectal cancers p.G12D, p.G12V, and p.G13D mutations are typically found to be present. A smaller subset of $\approx 3\%$ of cases also harbour the p.G12C mutation (Prior, Lewis, and Mattos 2012) observed in C539. The most likely explanation for these tissue-specific differences is that the specific substitutions — a C[C>A]A mutation for the p.G12C K-Ras mutation — are more likely to occur under specific mutagenic processes active in these tumour entities (Prior, Lewis, and Mattos 2012; Temko et al. 2018). In the case of C539, the analysis of mutational signatures indicated the presence of a relatively large number of C>A mutations (Figure S.159, page 337), thus providing a reasonable explanation for the presence of the relatively unusual K-Ras p.G12C mutation.

Since, the K-Ras mutation arose in the context of a full loss of APC and p53 activity this also provided a nice counter-example to the classic adenoma-carcinoma sequence. This model of CRC evolution suggests that oncogenic K-Ras mutations are important for the initiation of adenomas and that they typically occur before p53 mutations (Vogelstein et al. 2013). Summarised the analysis showed that activating subclonal K-Ras mutations also lead to a substantial fitness effects in an already established colorectal carcinoma.

PIK3CA mutations Interestingly, neither of the two PIK3CA mutation for which a significant fitness effect was inferred (i.e., p.C378R and p.Q546P) was one of the most common 'hotspot mutations' (i.e., p.E542K, p.E545K or H1047R). Nevertheless, previous screening of the phenotypic effects of such rarer PIK3CA mutations showed, these can also have growth-promoting activity (Dogruluk et al. 2015). The data analysed here suggested that, in the genetic background of a activating K-Ras mutation, these two variants cause significant phenotypic effects *in vivo*.

Still, while the lack of phenotypic effects might explain the absence of selection for some of the rare PIK3CA variants (i.e., p.R88Q and p.G118D), the p.Q546K mutation observed in C531 was previously found to have growth-promoting effects. The phenotypic effect of this p.Q546K PIK3CA variant was indeed predicted to be similar to the p.Q546P PIK3CA variant for which clear evidence of selection was identified in C525. While it is certainly possible that the environmental or genetic background of this specific case played a role, it seems to be more likely that the variant arose very recently within the carcinoma and that selection did not have sufficient time to act on this variant. In this case, it would be expected that only a small increase of the edge length would occur. This in turn might cause problems to identify deviations from the neutral expectation.

A similar lack of power to detect selection might explain why the p.H1047Q and p.545K PIK3CA mutations found in C544 and C537 respectively, were not identified. Alternatively, the absence of a co-occurring K-Ras mutations might have reduced the phenotypic effect of these PI3KCA mutations. This hypothesis is also supported by observations made by others (e.g., Wang et al. 2013a; Stintzing and Lenz 2013; Phipps, Makar, and Newcomb 2013; Green, Trejo, and McMahon 2015; Oda et al. 2008; Wang et al. 2018b). This observation is also supported by a recent paper that found a co-mutation of K-Ras and PIK3CA, but not PIK3CA alone to be associated with poor overall survival (Luo et al. 2020)

PTEN mutation in C518 In case of C518, the inference suggested that the presence of a PTEN p.C136R mutation caused selection of the mutated subclone. Importantly, this subclonal PTEN mutation occurred in the background of a clonal truncating PTEN mutation (p.K267Rfs*9) that should have caused the loss of the alternate allele. Consistent with this a separate analysis of the RNA-seq data[13] from the EPICC cohort confirmed that the expression of the second allele was completely lost in samples from regions A and B of the tumour (see Figure S.160, page 338).

The identified PTEN p.C136R mutation causes a substitution close to the NG_2-terminal phosphatase domain of the PTEN protein (Han et al. 2000). However, the mutation is not within one of the frequently mutated hotspots of the protein (Bonneau and Longy 2000; Dillon and Miller 2014). Still, similar PTEN mutations (i.e., p.C136Y) have previously been found to disrupt the phosphatase activity of the PTEN protein itself (Han et al. 2000). Further evidence for the phenotypic effect of the PTEN p.C136R mutation comes from the observation that germline PTEN p.C136R mutations are associated with Cowden syndrome — a genetic disease caused by germline PTEN mutations — as well as the demonstration of a reduced stability and loss of phosphatase activity of the encoded protein in vitro (He et al. 2013). Taken together, it seems likely that the second (subclonal) PTEN mutation led to the complete loss of PTEN activity.

PTEN itself is considered a CRC tumour suppressor gene. It inhibits the conversion of phosphatidylinositol-4,5-bisphosphate to phosphatidylinositol-3,4,5-triphosphate (PIP3) (Molinari and Frattini 2014). Loss of PTEN activity, thus causes the accumulation of

[13] This work was done by Jacob Househam.

PIP3 within the cells, leading to the over-activation of downstream PI3K pathway (Song, Salmena, and Pandolfi 2012). This activation of the PI3K pathway, mediated through its various targets, is associated with increased proliferation, inhibition of cell death, and stimulation of angiogenesis (Song, Salmena, and Pandolfi 2012; Molinari and Frattini 2014). Summarised, the subclonal PTEN mutation identified in C518 provides another example of subclonal alterations of the PI3K pathway in the analysed CRCs.

SMAD4 mutation in C531 The ABC-SMC inference applied to the tree of C531 suggested the presence of a putatively selected subclone in region B of the tumour. The assessment of potential somatic driver alterations in genes reported in the IntOGen database revealed the presence of a single SMAD4 mutation.

Mutations in SMAD4 occur in about 8.5% of CRCs (Fleming et al. 2013). In the EPICC cohort three clonal SMAD4 variants in CRCs with MSI were identified in addition to the SMAD4 p.A118V mutation in C531. SMAD4 mutations have previously been found to be associated with poor outcome in a metastatic setting (Alazzouzi et al. 2005; Mizuno et al. 2018; Kawaguchi et al. 2019), chemo-resistance to 5-FU (Zhang et al. 2014a), and resistance to Cetuximab (Lin et al. 2019). These studies highlight the general importance of SMAD4 loss in the late-stage evolution of CRCs. Genes of the SMAD family of genes are transcription factors that control the expression of genes as part of the TGF-β signalling pathway. Activation of TFG-β signalling has tumour suppressive and metastasis promoting effects (Bierie and Moses 2006).

The formation of SMAD2-4 complexes is necessary for the transduction of growth-inhibiting effects of TGF-β signalling in the nucleus and their loss appears to cause a shift towards negative effects of TGF-β signalling in CRC (Zhang et al. 2010). The specific subclonal driver mutation observed in C531 (i.e., SMAD4 p.A118V), occurred in a frequently mutated position of the gene and is likely pathogenic (Iacobuzio-Donahue et al. 2004; Jones et al. 2008; Fleming et al. 2013).

Still, the mutation only affected one of the two SMAD4 alleles (see Figure S.145, page 330). Some previous studies indicate that mutations of genes of the SMAD family can act dominantly-negative (Hoodless et al. 1999; Xu et al. 2000; Alberici et al. 2006), but bi-allelic mutation of SMAD4 are frequently observed in cancer genomic data (Fleming et al. 2012). This suggests that the mutation or loss of the second allele would likely be required. For this reason, I assessed the expression of SMAD4 in general and the SMAD4 p.A118V variant specifically in all the samples of C531. Surprisingly, this analysis of the RNA-seq data of the EPICC project showed that virtually no expression of SMAD4 occurred in samples from the mutated region B (see Figure S.161, page 338). This complete loss of SMAD4 expression — potentially due to the down-regulation of expression by a second independent event — in combination with the pathogenic SMAD4 mutation, suggests that a complete loss of SMAD4 activity occurred in the corresponding lineage of the tumour. This provides a reasonable explanation for the observed subclonal selection in region B.

RNF43 mutation in C538 Another example of a subclonal mutation of a known CRC driver gene on an edge for which the ABC-SMC suggested the presence of a positively selected alteration was identified in C538. Here all samples found in region B and one sample from region A (A1_G7) were part of a putatively selected subclone and an associated truncating mutation of RNF43 (p.Q153*) was identified. Truncating mutations of RNF43 are relatively frequently observed in colorectal, endometrial, ovarian and pancreatic carcinoma (Jiang et al. 2013; Zou et al. 2013; Ryland et al. 2013; Giannakis et al. 2014).

RNF43, together with ZNRF3, typically plays a vital role in the inhibition of Wnt signalling through the degradation of Wnt receptors of the Frizzled family (Koo et al. 2012; Jiang et al. 2015; Tsukiyama et al. 2015). Since the expression of RNF43 and ZNRF3 is induced by Wnt/β-catenin signalling itself, these proteins provide a negative feedback loop required for this signalling cascade. The deletion of both RNF43 and ZNRF3 was shown experimentally to induce tumour formation through activation of the Wnt pathway (Koo et al. 2012). This activation of the Wnt pathway is also thought to be the reason APC loss, the most common driver mutation of CRCs, is selectively advantageous. Consistent with this APC and RNF43 mutations were previously found to be mutually exclusive in CRCs (Giannakis et al. 2014). The degradation of FZD receptors by RNF43/ZNRF3 is thought to be caused by the activation of endocytosis through ubiquitination by the RING domain of these proteins (Hao et al. 2012; Koo et al. 2012) and the recognition of Frizzled family receptors is mediated by DVL binding to the DIR domain of RNF43 and ZNRF3 (Jiang et al. 2015).

The variant observed in C531 is predicted to cause the loss of both of these domains (i.e., DIR and RING). While such a mutation certainly leads to the loss of function of the mutated allele, only one of the two alleles was found to be mutated in C531 (Figure S.146, page 330), meaning that the second allele could potentially compensate for the loss of RNF43 function. Similarly, it has previously been suggested previously that ZNRF3 can potentially compensate for the loss of RNF43 and that only loss of both proteins leads to tumour formation and altered Wnt/β-catenin signalling (Koo et al. 2012; Lannagan et al. 2019). Assessment of the mutation and expression status of ZNRF43 revealed no evidence for somatic mutations or altered expression of ZNFRF43 in the corresponding samples. The expression of the mutant RNF43 variant showed that the gene was dominantly expressed in one of the three samples from tumour region B but not in the others.

Interestingly, some previous studies found that missense mutations observed in CRC (Koo et al. 2012; Tsukiyama et al. 2015; Yu et al. 2020) and truncated RNF43 variants missing the RING domain (Hao et al. 2012; Tsukiyama et al. 2015) might have a dominant-negative effect. Still, the overwhelming majority of truncating variants tested in a large screening of RNF43 variants by Yu et al. (2020) and specifically a p.Q152* mutation (i.e., one amino-acid differences to the one observed here) were found to only cause to loss of function and not to be dominant-negative. A single truncating frameshift variant of APC was observed (p.E1309Dfs*4) in this patient. Assuming that a second APC variant affecting the alternate allele was present but not detected, a mutation like that of RNF43, which also activates the Wnt/β-catenin pathway, would be expected to not lead to any or very little fitness increase.

Evidently, the results suggest that the mono-allelic loss of function mutation of RNF43 p.Q153* in C538 might have a fitness altering effect, but this appears to be at odds with existing literature and the assumed mode of function of this mutation. Given the APC mutated background and the inconclusive analysis of the gene expression data, it might be possible that a different, possibly undetected, somatic mutation or non-genetic alteration was responsible for the observed effect. Alternatively, while certainly speculative, mono-allelic RNF43 mutations in the APC depleted background present in C538 might be haploinsufficient and hence lead to further activation of the Wnt/β-catenin pathway with corresponding fitness effects.

Other cases Similarly, non-genetic or unidentified somatic mutations of driver genes might provide an explanation for four cases in which no corresponding somatic mutation in a driver gene could be identified. These were C542 (region B&C, Figure S.148, page 331), C542 (region A, B & D, Figure S.148, page 331), C549 (region B&C, Figure S.150, page 332), C551 (Figure S.151, page 332) and C559 (region B, Figure S.152, page 332). In C542 for example, the subclonal loss of chr1p — an alteration that was found to be recurrent in Stage III CRCs (Xia et al. 2020) and more frequently mutated in metastatic CRC (Ghadimi et al. 2006; González-González et al. 2014) — could explain the subclonal selection of one tumour region. In general, a larger cohort of cases would be needed to assess whether cases in which no subclonal driver mutations could be identified are simply false positives or if rare or non-genetic drivers are indeed present in these.

I found no conclusive evidence of subclonal alterations of gene expression using the available RNA-seq data or changes of chromatin accessibility that would explain the observed selection, but future studies powered to perform a comprehensive analysis of such alterations in putatively selected subclones should certainly be considered.

6.3.3.6 Prediction of Clone-Size Distributions

The marginal posterior of the ABC-SMC inference can also be used to estimate the distribution of subclones within the tumour. To demonstrate this, the 100 simulations that produced trees with the closest distance to the observed data were obtained and used to estimate the marginal distribution of each subclone in space. The locations of simulated cells were first rotated by ϕ_A, assigned to the closed grid point and then used to calculate the fraction of times specific subclones were observed at these positions (see Figure 6.23A). The results of this procedure are shown for one example (C539) in Figure 6.23B. Here the inference suggested that the subclone (Clone 2) extended over a large area of space.

Various methods to detect and map somatic variants *in situ* have been developed (Bagasra 2007; Larsson et al. 2004) and these could in principle be used to improve or verify predictions of the algorithm.

6.3.4 Limitations of the Conducted Analysis

While the performed inference appeared to generally provide sensible results and able to recover subclonal selection due to somatic mutations of bona fide CRC drivers, a couple of limitations of the analysis, which will be discussed in the following, have to be considered.

Limitations of the used sampling schema The applied sampling schema provides on its own a minimal amount of information on the spatial extent of selected subclones within the tumours. If, for example, all samples from one region of the tumour would be sampled from a subclone — assuming that this could be identified by an elongated edge in the reconstructed tree in this case — only very little would be known about the size of the clone itself. From the observed data, it would be reasonable to assume that the subclone covers at

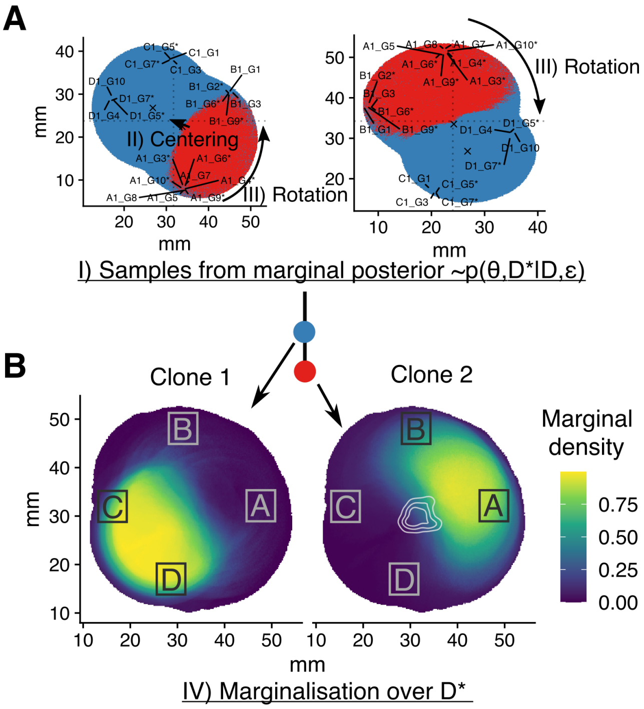


Figure 6.23: Marginal clone size estimate for C539. A) Two realisations drawn from the set of particles approximating $p(\theta, D^{*}|D, \varepsilon)$. These are centred and rotated to calculate B) The marginal fraction each grid point contains a specific clone.

least the sampled region (i.e., $d \approx 5mm$) and less than a quarter of the tumour (i.e., from the left to the right neighbouring region).

Since the size of a subclone and the time it arose provide information on the relative strength of selection a subclone is experiencing, large credibility intervals (CIs) on λ would be expected. These large CIs are precisely what was observed from the inference, with potential values of λ ranging from little more than the background clone to a, certainly unrealistic, 25-fold increase of the division rate of selected subclones (see Figure 6.21).

Performing systematic spatial sampling while keeping a record of the relative sampling positions in space or deep bulk sequencing of a huge tissue sample would provide a much better information on the spatial extend of selected subclones. Other, more practical methods might also allow a better idea of the spatial extent of these. For example, mixing the spatial sampling performed here with several random small punch biopsies along the outer diameter and subjecting these to pooled deep sequencing might also allow a more accurate estimation of relative clone sizes.

Similarly, large parts of each tumour were not observed, potentially reducing the ability to detect expanding subclones in general. The sampling locations themselves only cover a tiny area of the tumour as a whole (i.e., $< 5\%$), and while selected subclones by definition grow to cover a large area, they could, in principle, still be present between sampled regions. This assumption is encoded in the model itself and hence the inference framework, but a detailed analysis of the likelihood of this under different model parameters might be worth conducting.

Three-dimensional structures Furthermore, the three-dimensional structure CRCs exhibit was simplified as a two-dimensional model (see Figure 6.2A). This might be reasonable if cells primarily grew along the horizontal plane, but adenoma and carcinoma frequently exhibit complex three-dimensional structures growing into the colorectal lumen. Whether subclonal dynamics within these are sufficiently recapitulated in the simpler two-dimensional model used here could undoubtedly be questioned. Especially as some conclusions made here are on the strength of boundary driven growth, questions remain if and how these would be recapitulated in three dimensions.

Alternative growth models should also be considered. The spatial model used here is equivalent to the ‘constant crust’ model of tumour growth (Mayneord 1932; Conger and Ziskin 1983). Such simple models were suggested to be insufficient to even explain growth dynamics of tumour spheroids in culture (Marušić et al. 1994) or *in vivo* (Marušic et al. 1994). Despite a long-held interest in the growth laws of human malignancies (Steel and Lamerton 1966; Gerlee 2013), the exact nature of these in tumour entities in general (Chignola and Foroni 2005; Talkington and Durrett 2015) and CRCs in specific (Burke et al. 2020) are surprisingly not fully understood. It would certainly be of interest if genomic measurements could give insight into the growth law (i.e., exponential vs boundary driven growth) on a patient-by-patient basis, especially this might even predict the growth rate at related metastatic sites. Still, other models should be considered as well. For example, inclusion of desquamation at the surface the tumour — a proposes that was proposed by Spratt (1961) to explain the slower growth rates of CRCs at the primary site compared to the metastatic site — or the inclusion of central growth inhibition (Mayneord 1932; Steel and Lamerton 1966) might significantly alter the predictions made by inference on the genomic data.

Stem cell dynamics Another factor that should probably be considered are the stem cell dynamics within single-glands as well as the genetic heterogeneity of stem cells existing in these. These additional complexities were ignored here, which is only reasonable if the replacement rate of stem cells within glands/crypts is swift compared to the division rate of these. Previous studies have shown that this is to be accurate, but this might not generally be the case.

Siegmund et al. (2009b) for example, have suggested that 'palm shaped' phylogenies, which were found to be explained reasonably well by boundary driven growth alone, might arise due to the presence of a few long-lived cancer stem cells (CSC) and that star-shaped phylogenies, here explained by non-boundary driven growth, instead arise in the presence of many CSC. Not having included these dynamics into the spatial model appears to be one of the major shortcomings of the analysis. It is unclear whether large stem cell pools and growth dynamics could be distinguished from each other and how data generated under a combination of both of these would behave.

Changes of mutation rates Last but not least, spatio-temporal changes of mutation rates might occur during the evolution of individual tumours. The previously described analysis of mutational signatures, as a surrogate of such mutation rates, did not identify any particularly prevalent subclonal changes. While the analysis of mutational signatures did suggest that the relative contribution of the individual process was not altered over time, it could certainly be that their absolute contribution changes over time. Given that one of the signatures of subclonal selection is the elongation of individual edges compared to the remaining tumour, it might be hard to distinguish changes of mutation rates from subclonal selection.

6.4 Discussion

Here I presented results from statistical inference, which used an extended version of a spatial tumour model previously described by us (Chkhaidze et al. 2019) to perform computational inference on single-gland multi-region sequencing data from a total of 26 CRCs. Using this framework, it was possible to quantify tumour-specific properties that describe their evolutionary dynamics. Specifically, cases were distinguishable based on i) the presence of subclonal selection, ii) their growth law (i.e., slow, boundary driven vs fast exponential growth), and iii) their mutation rates. Importantly, these are — unlike a mere collection of features provided by the measurement of gene expression (Uhlen et al. 2017) or somatic mutations (Campbell et al. 2020) — interpretable descriptions of the growth dynamics occurring in individual tumours.

While speculative, these properties might also predict the speed with which already exiting, but still, invisible metastasis grow or be associated with a malignant phenotype.

As such, the corresponding parameter estimates appear to be a reasonable proposal for an ‘evolutionary biomarker’. The quality of these will be assessed as part of the prospective follow-up of the EPICC study in the near future.

A specific aspect of the inference that appears to be worthwhile to validate further is the prediction that some CRC evolved under boundary driven growth, whereas others did not. This would imply that individual CRCs are growing at a substantially different speed. Measuring such ‘growth laws’ of CRCs in general and especially in individual patients appears to have been challenging (Burke et al. 2020), but of general importance (Friberg and Mattson 1997; Sachs, Hlatky, and Hahnfeldt 2001; Comen, Morris, and Norton 2012; Rodriguez-Brenes, Komarova, and Wodarz 2013).

The presented results also corroborate previous studies on the prevalence of subclonal selection in CRCs. In many of the characterised tumour lineages, I was unable to identify any putative driver alterations. These cases were sufficiently explained by a spatial model without selected subclones, suggesting that the observed parts of the tumours evolved effectively neutral. This is consistent with previous observations made in bulk sequencing data (e.g., Williams et al. 2016; Williams et al. 2018b) and supports the reply to previous criticism (e.g., Williams et al. 2017; Williams et al. 2018a; Heide et al. 2018) of these studies by others (e.g., Balaparya and De 2018; Tarabichi et al. 2018; McDonald, Chakrabarti, and Michor 2018). In a subset of cases, subclonal driver mutations were identified. Here the ABC-SMC inference framework frequently suggested the presence of positively selected subclones. This demonstrates that the method was sufficiently powered to detect deviations from neutrality in general. An orthogonal analysis of dN/dS values, inspired by the criticism of Tarabichi et al. (2018), support these conclusions.

The presented analysis demonstrates that single-gland multi-region sequencing can provide a general framework to measure the effect of subclonal driver alterations *in vivo*. Due to the ability to fully reconstruct the lineage of individual subclones, this approach allows analysing the effect of driver alterations in the genetic and environmental context they occurred in. Such context-specific effects are thought to be of importance (Berger, Knudson, and Pandolfi 2011) and a large-scale application of the approach used here could be used for the discovery of refined models of *in vivo* driver gene activity.

Due to the scarcity of subclonal selection and the relatively small cohort analysed, a comprehensive analysis of novel driver mutations could not be performed. Still, I am optimistic that in the future, more extensive studies using a similar approach might allow to gain significant insight into the role of genetic and epigenetic alterations in subclonal tumour evolution. Here, similar approaches could distinguish the large amount of 'neutral signal' present in measurements of intra-tumour heterogeneity from that associated with meaningful subclonal selection.

The identification of signals from selected subclones could be especially beneficial for the analysis of non-genetic drivers (Black and McGranahan 2021) that lack appropriate models to describe their dynamics. Indeed, a small subset of cases in which inferred somatic driver mutations did not explain subclonal selection was identified as part of this work. In these cases, non-genetic events could have had a critical role.

6.5 Conclusion

The spatial computational inference on the single-gland sequencing data of the EPICC cohort presented here has allowed to i) predict the presence of selected subclones and ii) estimate the growth laws in individual tumours. Furthermore, the conducted analysis demonstrated that the developed method allows to identify relevant subclonal selection events and could thus guide the discovery of novel or rare genetic driver alteration in the future.

Chapter 7

Summary and Outlook

7.1 Summary

The aim of this thesis was to investigate intratumor heterogeneity of CRCs concomitantly on the genetic and epigenetic levels. For this experimental data from a multi-omics sequencing method able to perform WGS, ATAC-seq and RNA-seq on single-gland colorectal glands were utilised. This was done in particular to i) identify, so far understudied, epigenetic alterations that might contribute to carcinogenesis in CRCs, ii) demonstrate that the information encoded in the genomic data accessible by single-glands WGS can be used to infer the presence of subclonal selection in individual patients, and iii) derive 'evolutionary biomarkers' with a potential predictive value that could be assessed during the follow up of patients. The analysis was driven by the goal to integrate such measurements into a rigorous statistical inference framework able to make predictions on the evolutionary dynamics and properties of subclones within individual tumours, especially with regard to subclonal selection, to then identify responsible epigenetic and genetic alterations.

A combination of theoretical models and genomic data (i.e., sequence information) have allowed gaining profound insight into the evolution and population dynamics of species. Population genetics, which is concerned with the study of alleles as the source of genetic variation of individuals in populations, is indeed one of the few areas of biology that is based on a well defined theoretical foundation. The existence of such a theoretical framework allows performing quantitative experiments that are able to determine fundamental properties of the studied system (e.g., in the form of model parameters) and to make generalisable predictions from these. This is in stark contrast to many other areas of biology that lack a clear theoretical basis and hence rely much more on experimental validation of specific hypotheses.

Interestingly, in the field of cancer evolution, which studies the effects and dynamics of somatic alleles in tumour tissues, similar models have been applied comparatively little so far. Instead, much of our knowledge of which genes contribute to the development of cancer and their effects they have been derived through the use of biological experiments that introduced mutant alleles into biological systems to observe their effects. For this cell lines and mouse models have been used extensively. Still, these model systems often do not allow to draw direct conclusions about the effect of mutations in human tumours and they are often costly and time-consuming. The use of animals for cancer research has also been criticised due to ethical concerns.

Since the advent of NGS, which allows for routine high throughput screening of somatic variants detectable in individual tumours, statistical methods have been used to systematically identify mutations with a role in the development of various tumour types. These are often based on the relative frequency with which genes or specific sites of a protein are mutated in a large cohort of cases. While these studies have been invaluable for the systematic identification of genes that contribute to tumourigenesis, they are unable to predict the exact phenotypic effect such mutations have in vivo.[1]

Here the utilization of models similar to those used in population genetics, promises to allow a direct inference of the relevant phenotypic properties of cells in terms of their relative fitness compared to other cells in a tumour. In tumour evolution, the pervasive selection of adaptive mutations is often assumed by default. This is in contrast to the interpretation of sequence data in species evolution where rigorous statistical tests — since these make assumptions that are violated in tumours these can unfortunately not directly be used in the context of tumours — are applied to come to such conclusions. Many studies of tumour heterogeneity in the field of tumour evolution focus on the accurate reconstruction of the ancestral relationships of cells or subclones. The observation of a large number of subclonal mutations in many multi-region sequencing studies, some of which are also putative driver mutations, can give the illusion of a functional interpretation of these variants. However, it is important to note that these data can only provide the data basis — a detailed summary of the ancestral relationships encoded in the genomic information of tumour cells — for a subsequent interpretation.

The debate of if and how the detection of selected subclones could be done with se-

[1] The co-occurrence of mutations can provide some insight into dependencies and in combination with previous knowledge of the function of genes some deductions of their effects can be made.

quencing data from tumours is ultimately what contributed to the controversy surrounding a study by Williams et al. (2016). Here, Williams et al. used a simple summary statistic — the $1/f$ test, which compares the VAF spectrum to that expected under neutrality (i.e., the LD mode) as a null model — to identify tumours with evidence for subclonal selection. Despite the fact that this null model was rejected in the majority ($\approx 65\%$) of cases, the criticism primarily focused on the fact that the null model was not rejected in all cases. This debate provided much of the motivation for the work described herein.

In Chapter 2 of this thesis I presented a detailed analysis of one of these criticisms (i.e., Tarabichi et al. 2018) that I conducted at the beginning of the project. Here, I showed that Tarabichi et al. severely mischaracterised the weaknesses of the $1/f$ statistic and that their assertion that it might constitute a worse than random classifier was simply false. Instead, much of the parameter space analysed by Tarabichi et al. (2018) was found to give rise to 'effectively neutral' VAF distributions as a consequence of subclones with an increased mutation rate. This caused the large 'neutral tail' of the hypermutant subclone to mask the deviations the clonal variants of the subclone caused in the VAF distribution. Still, even under perfect conditions, the power of the $1/f$ statistic was generally found to be insufficient. In a subsequent study, which instead used a Bayesian classifier, Williams, Sottoriva, and Graham also showed that the detection of subclonal selection from the VAF distribution alone is generally hard (Williams, Sottoriva, and Graham 2019).

In Chapter 2 I have also described how the VAF spectrum of mutations generated by a branching process model of tumour evolution might be uninterpretable by frequently used clustering methods at the currently used sequencing depth. In Chapter 3 I expanded on some of these observations. Here, I explored in more detail whether multi-region bulk sequencing data, which can be used to study the evolutionary relationships of cells in tumours (e.g., Gerlinger et al. 2012), would allow the detection of subclonal selection more easily. For this, I developed a simple spatial tumour simulator together with Ketevan Chkhaidze. I describe this model in detail in Chapter 3, but summarised it allows to model the effects of spatial crowding and boundary driven growth, subclonal selection, changes of mutation rates, and cell death. I used this simulator, to identify some general problems associated with spatially sampled bulk WGS data. These issues, arising from artefacts of spatial sampling, mean that the shape of reconstructed trees can be distorted in various ways. For this reason, the presence of these caveats has to be taken into account for the interpretation of such data in light of selection. In this context of this analysis, I have also shown that the application of clustering methods to multi-region WGS will likely cause the identification of many subclonal clusters and that these should not generally be interpreted to suggest the presence of selected subclones. Again, due to the aforementioned spatial artefacts, the interpretation of the results of such clustering methods is not always intuitive. For example, the mutation burden identified in individual samples or clusters can vary widely and should not be used to make direct predictions of the ‘age’ of a subclone. However, a similar analysis of simulated single-cell sequencing data showed that these would be ideal methods for the detection of subclonal selection and boundary driven growth as a consequence of spatial crowding. The absence of distortions introduced due to the spatial sampling of cells means that such data are generally more well behaved. Subclonal selection should be revealed by a clear elongation of internal branches and the presence of boundary driven growth, by a characteristic alteration of the internal branching structure. Still, the reliable detection of mutations from single-cell sequencing data is currently still challenging.

The limitations of single-cell sequencing were the reason why the multi-region sequencing study ‘EPICC’ that I analysed in Chapter 4 instead conducted sequencing of single colorectal tumour glands. Since CRC glands are thought to be the clonal units of CRCs this approach should have a similar resolution to single-cell sequencing in other tumour entities and therefore allow the detection of subclonal selection from the genomic measurements obtained from these glands. In Chapter 4, I have presented a detailed analysis of the conducted concomitant WGS, ATAC-seq and RNA-seq of individual colorectal tumour glands. Based on the WGS data obtained from the glands, I showed that relatively many spurious subclonal mutations in putative driver genes can be identified. A dN/dS analysis, similar to the one I described in Chapter 2, also indicated the presence of subclonal selection. Still, since the majority of the observed mutations occurred in genes that rarely show clonal mutations in CRCs and since most of these genes were not recurrent across patients either it was not clear how to interpret these data. Neither a careful visual examination of reconstructed phylogenetic relationships nor a summary statistic motivated by the $1/f$ test of (Williams et al. 2016) were able to give a conclusive answer on which driver mutations were under active selection in the majority of cases. The only exception from this were a single subclonal K-Ras mutation and a small number of subclonal PIK3CA mutations. In general, a pervasive and widespread subclonal selection of driver mutations, as suggested by other studies (e.g., Dentro et al. 2021), was certainly not apparent on the genetic level.

Non-genetic alterations, such as changes of the chromatin accessibility, are also thought to contribute to tumorigenesis and they can, just like genetic alterations, also be subject to selection. Here, the concomitant profiling of glands using ATAC-seq allowed me to provide some insight into the prevalence of epigenetic alterations in the CRCs. For this, I used the data of multiple glands obtained from each tumour and the matched normal crypts to identify genuine somatic chromatin accessibility alterations in the carcinomas. Across cases, this lead to a list of several hundred genomic regions that showed a highly recurrent alteration of chromatin accessibility compared to healthy colorectal glands. It is not unreasonable to assume that some of these might constitute bona fide epigenetic drivers of CRCs and that a careful evaluation of these might eventually lead to the identification of novel therapeutic targets. Consistent with this hypothesis, the majority of the identified alterations were not found to be sub-clonally altered in different regions of the tumour, suggesting that the identified alterations were, just like known genetic drivers, primarily early clonal events.

It has to be mentioned though, that the obtained single-gland ATAC-seq profiles were of comparatively poor quality and that additional biases further reduced the ability to identify subclonal changes of chromatin accessibility. For this reason, it was also not possible to conduct a detailed genome-wide analysis of chromatin alterations on a single-gland level. Instead, a preliminary analysis of the data suggested that the chromatin accessibility of the majority of functional sites of the genome might be under drift and thus evolve in parallel with the genome. Still, specific alterations, potentially different from those recurrently altered in CRCs, might arise due to selection or phenotypic plasticity. Further studies are clearly required to elucidate these aspects further.

In Chapter 6, I have presented the results of a computational ABC based inference that I applied to the genomic data of the EPICC cohort. For this, I integrated the spatial tumour simulator from Chapter 3 into an ABC-SMC based inference framework, that allows the identification of the number of selected subclones and the degree of boundary driven growth based on information contained in phylogenetic trees reconstructed from single cells or glands. The approach explicitly simulates the performed spatial sampling and is hence able to take into account potential biases arising from spatial sampling. To further increase the size of the reconstructed dataset and hopefully the ability to identify examples of selection, I also assigned additional LP-WGS glands onto the reconstructed phylogenies using a simple maximum-likelihood method that I described in Chapter 5 and modified the inference framework to take their specific properties into account.

Summarised, I was able to conduct the inference on 26 CRCs with a sufficient number of whole-genome sequenced single-glands. In a subset of 8 CRCs showed strong evidence ($\Delta AIC > 4$) for subclonal selection. In $5/8$ of these cases, putative subclonal driver mutations were found to be present on the exact edges of the trees that the inference suggested to be under selection. Specifically, I found that the subclonal selection in these tumours was likely driven by a PIK3CA p.Q546P, a RNF43 p.Q153*, a KRAS p.G12C, a SMAD4 p.A118V, and a PTEN p.C136R mutation. Overall, a total of $12/26$ cases showed weak evidence for subclonal selection ($AIC_S < AIC_N$), with the remaining $14/26$ trees being consistent with neutrality. The orthogonal dN/dS based analysis of IntOGen driver genes found dN/dS ratios of ≈ 1 in 'neutral' and $dN/dS > 1$ in 'non-neutral' cases, thus confirming that the computational inference framework was indeed capable to identify subclonal selection.

While certainly speculative, the results from a ABC based inference also suggested that it was possible to use the single-gland WGS and LP-WGS data to identify the 'growth-law' individual tumours adhered to. For $7/14$ tumours the inference suggested the presence of strong boundary driven growth and nearly exponential growth for the remaining tumours. If this is indeed true it seems not unreasonable to assume that the former cases might have a somewhat better prognosis. However, given the limited evidence, further investigation of this is certainly warranted. The same is certainly true for many other aspects of spatial tumour growth, that were not considered in this analysis. Specifically, the effect of three-dimensional growth, changes in mutation rates or the presence of tumour stem-cell populations should be considered.

In summary, this thesis aimed to provide an assessment of the genetic and epigenetic heterogeneity of individual CRCs on a functional level. For this, I have used statistical and computational approaches to interpret genomic measurements obtained from tumours. This functional interpretation of the information encoded in the spatial distribution of mutations was the goal of the computational inference that I presented in the last chapter. This approach, which is rooted in populations genetics, allowed the identification of a small number of bona fide driver mutations that were likely under active selection in these tumours. The statistical analysis of the ATAC-seq dataset has provided the so far largest collection of somatic epigenetic driver mutations in CRCs, but these were found to be exclusively clonal.

7.2 Outlook

I hope that similar approaches will ultimately provide us a much better understanding of the fitness landscape of tumours. This knowledge would be crucial to allow predictions of the future evolutionary trajectory of cells in a tumour. This might eventually allow to routinely exploit population dynamics between different tumour subpopulations, as for example suggested by Zhang et al. (2017), in the clinical practice. Other ideas like the evolutionary steering of tumour cells (Nichol et al. 2015; Acar et al. 2020) towards a more manageable state might also be possible, but likewise, they also hinge on a better understanding of the fitness landscape of tumours. For this further technological advancements of sequencing methods will likely be essential. Excitingly, much progress is already being made in this regard and many limitations of currently conducted WGS might indeed be resolved soon. These limitations mainly stem from two factors: i) the relatively high error rates of the necessary PCR based amplification steps and ii) the loss of phasing information. Due to this high error rate, it is currently only possible to detect mutations at a relatively high frequency. Here new methods now make the reliable detection of mutations occurring in single DNA molecules possible (Abascal et al. 2021). Ultra-deep sequencing using such methods might soon be able to routinely detect mutations occurring only in a very small fraction of cells. Indeed the comparison of different variants might itself enable to gain fundamental insight into the fitness landscape existing within individual tumours. Similarly, recent advancements of single-cell sequencing now permit the reliable detection of single-nucleotide variants in individual cells (Xing et al. 2021). Such an approach preserves the phasing information of mutations and would thus make the perfect reconstruction of the ancestral relationships of a large number of cells possible. Applied to tumours, these or similar methods will undoubtedly provide an excellent data basis to research how individual tumours evolve.

Similar progress is being made on single-cell multi-omics methods. A relatively large number of different approaches combining WGS sequencing with single-cell ATAC-seq and RNA-seq are currently being researched. In the not too far future many of these will likely allow to conduct single-cell multi-omics and provide a much clearer insight into how individual somatic mutations and epigenetic alterations affect the phenotypic properties of tumour cells.

Much progress will certainly be made on the various statistical methods required to interpret these new sequencing data. It seems obvious that many of the currently used approaches, like dN/dS or phylogenetic reconstruction, will be extended to single-cell or low-frequency mutation detection. Likewise, approaches that look at the relative frequency of subclonal 'driver mutations' for individual sites will likely be used. In terms of the interpretation of single-cell phylogenies, computational approaches, like the one I used, could of course be used. However, such methods have serious drawbacks and statistical problems. For this reason, it seems likely that new mathematical models or even some closed-form solutions of specific aspects of spatial tumour evolution will be identified and used to answer questions regarding the interpretation of ancestral relationships between cells.

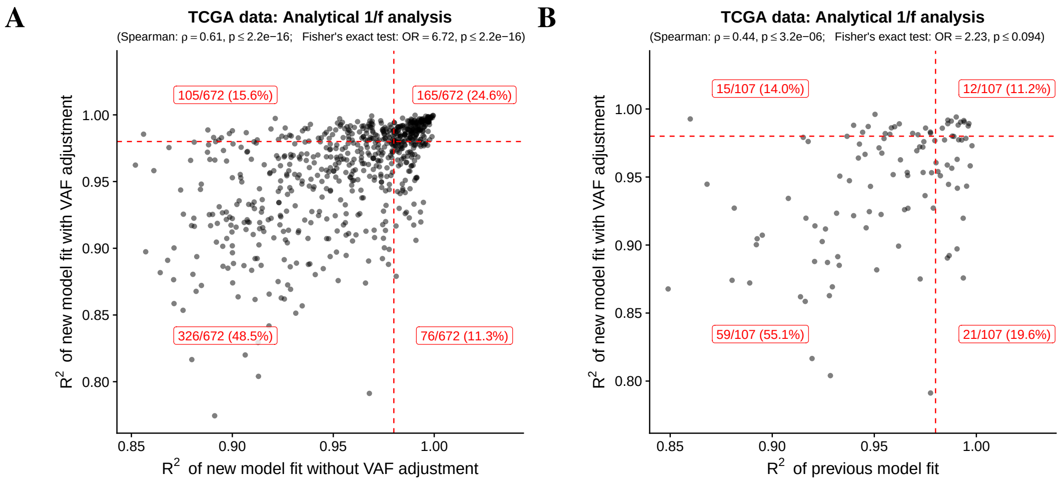
Together these approaches might permit us to obtain such a detailed understanding of the evolutionary dynamics within tumours, that prediction of their behaviour under various conditions becomes possible. This might in turn enable us to approach the treatment of tumours as a disease in a more informed way. Instead of the 'brute-force' application of targeted drugs, chemotherapies or radiation, which eventually fail due to evolution of resistance, an 'evolutionarily informed' treatment of tumours might allow for the much more efficient use of already existing drugs. The detailed knowledge of evolutionary properties of tumour cells might even allow the identification of new drug targets or alternative therapeutic approaches.

Supplementary Data

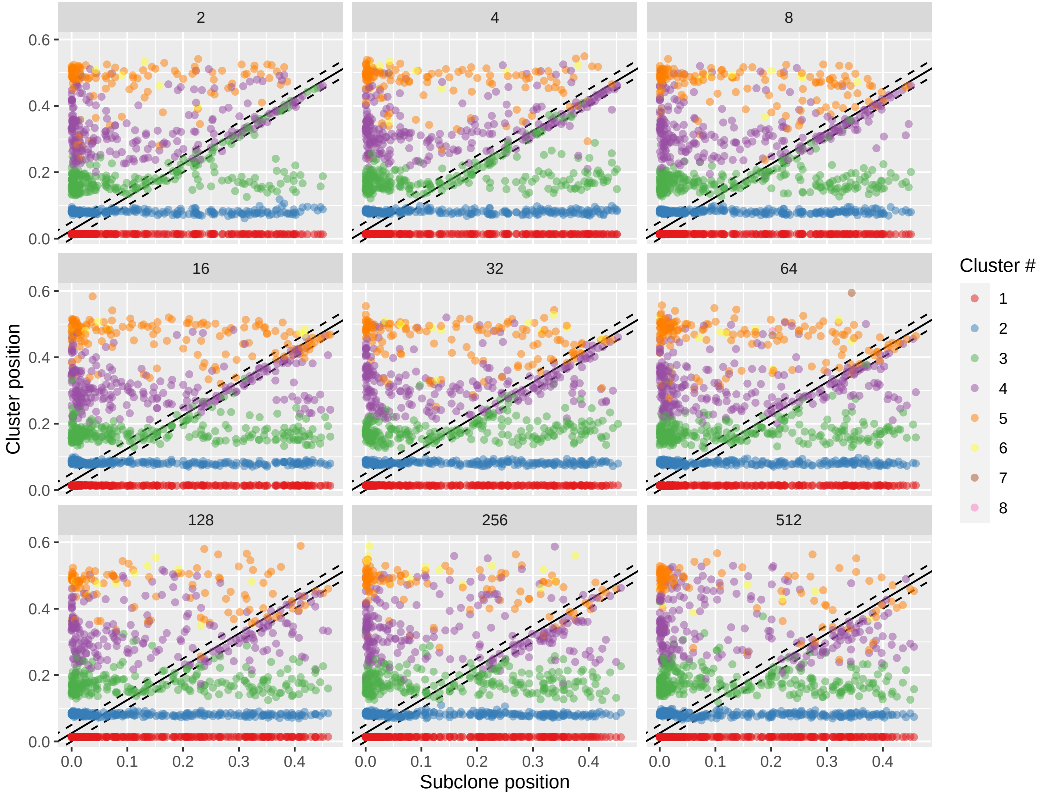
List of Supplementary Figures

List of Supplementary Tables

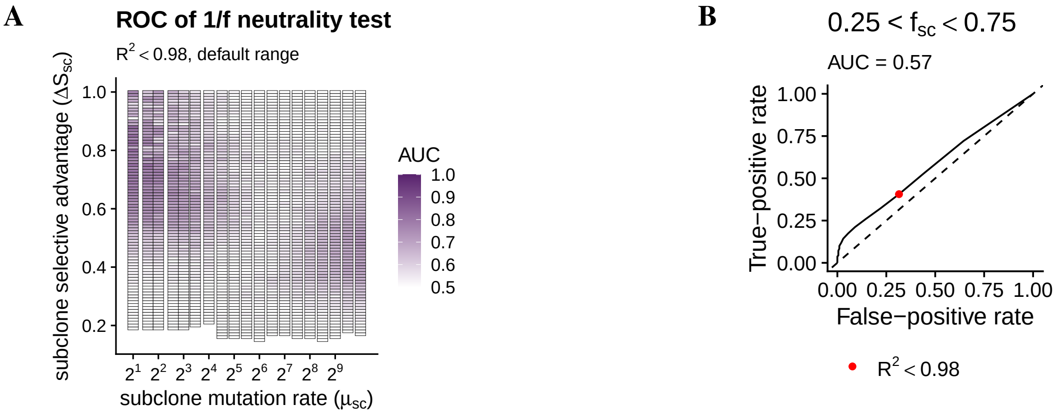
S.1 Neutral Tumour Evolution



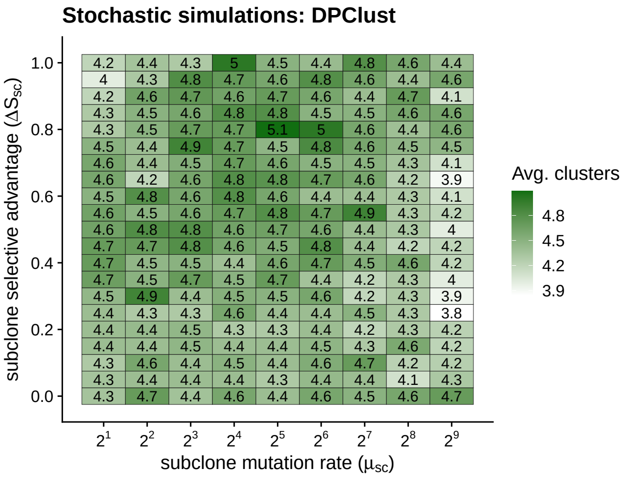
Supplementary Figure S.1: Effect of the inclusion of CN correction on the $1/f$ classification in A) the entire analysable TCGA cohort and B) the subset assessed by Williams et al. (2016).



Supplementary Figure S.2: Positions of clusters inferred by DPClust in simulated WGS. Data in this figure should be compared to the results shown in Figure 2.4.

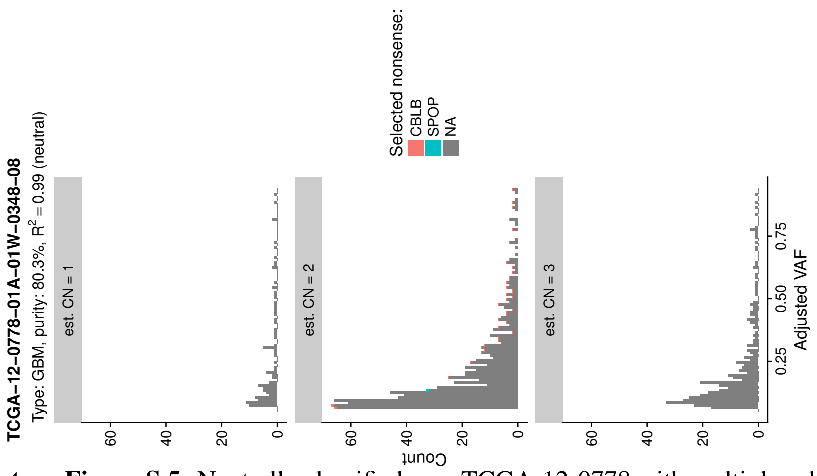


Supplementary Figure S.3: ROC of the $1/f$ test on stochastic simulations. A) Shows the AUC for the $1/f$ test applied to non-neutral simulations with a subclone fraction $0.25 < f_{sc} < 0.75$ at various parameter combinations. B) Shows the ROC curve for the $1/f$ test applied to all simulations with a subclone fraction $0.25 < f_{sc} < 0.75$ across all of these parameters.

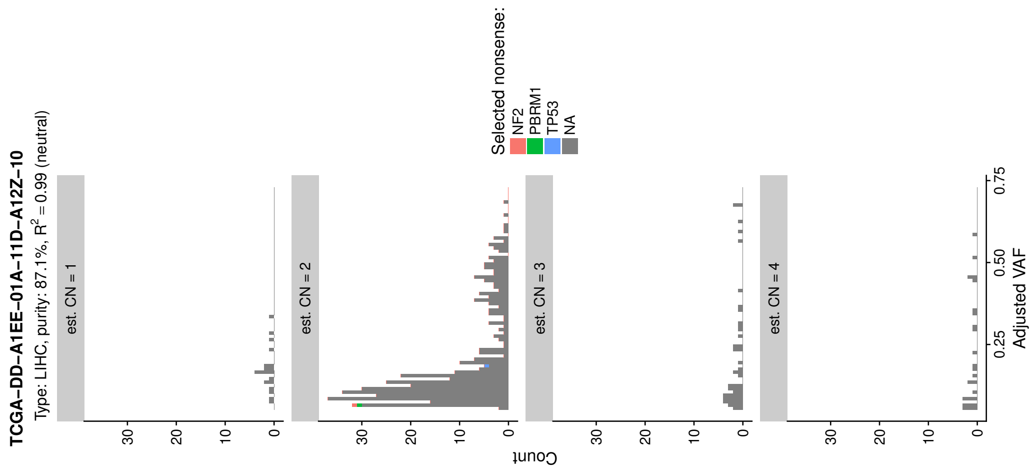


	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9
1.0	4.2	4.4	4.3	5	4.5	4.4	4.8	4.6	4.4
	4	4.3	4.8	4.7	4.6	4.8	4.6	4.4	4.6
	4.2	4.6	4.7	4.6	4.7	4.6	4.4	4.7	4.1
	4.3	4.5	4.6	4.8	4.8	4.5	4.5	4.6	4.6
0.8	4.3	4.5	4.7	4.7	5.1	5	4.6	4.4	4.6
	4.5	4.4	4.9	4.7	4.5	4.8	4.6	4.5	4.5
	4.6	4.4	4.5	4.7	4.6	4.5	4.5	4.3	4.1
	4.6	4.2	4.6	4.8	4.8	4.7	4.6	4.2	3.9
0.6	4.5	4.8	4.6	4.8	4.6	4.4	4.4	4.3	4.1
	4.6	4.5	4.6	4.7	4.8	4.7	4.9	4.3	4.2
	4.6	4.8	4.8	4.6	4.7	4.6	4.4	4.3	4
	4.7	4.7	4.8	4.6	4.5	4.8	4.4	4.2	4.2
0.4	4.7	4.5	4.5	4.4	4.6	4.7	4.5	4.6	4.2
	4.7	4.5	4.7	4.5	4.7	4.4	4.2	4.3	4
	4.5	4.9	4.4	4.5	4.5	4.6	4.2	4.3	3.9
	4.4	4.3	4.3	4.6	4.4	4.4	4.5	4.3	3.8
0.2	4.4	4.4	4.5	4.3	4.3	4.4	4.2	4.3	4.2
	4.4	4.4	4.5	4.4	4.4	4.5	4.3	4.6	4.2
	4.3	4.6	4.4	4.5	4.4	4.6	4.7	4.2	4.2
	4.3	4.4	4.4	4.4	4.3	4.4	4.4	4.1	4.3
0.0	4.3	4.7	4.4	4.6	4.4	4.6	4.5	4.6	4.7

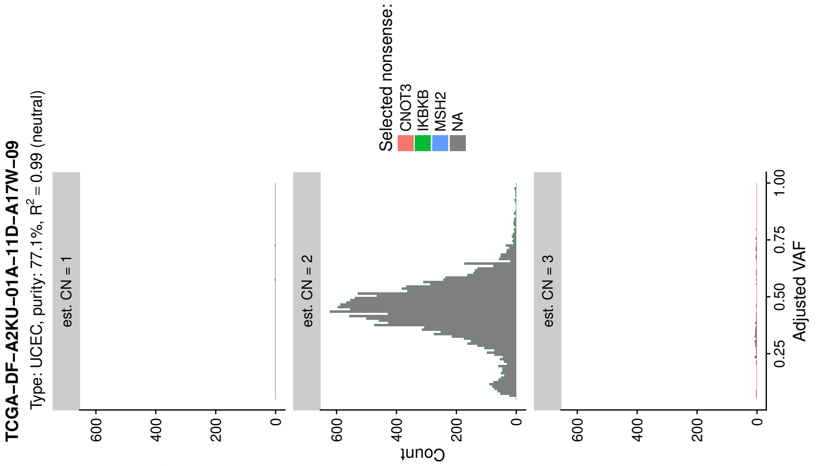
Supplementary Figure S.4: Average number of clusters inferred by DPClust in simulated WGS. Data in this figure should be compared to the sensitivity shown in Figure 2.4G.



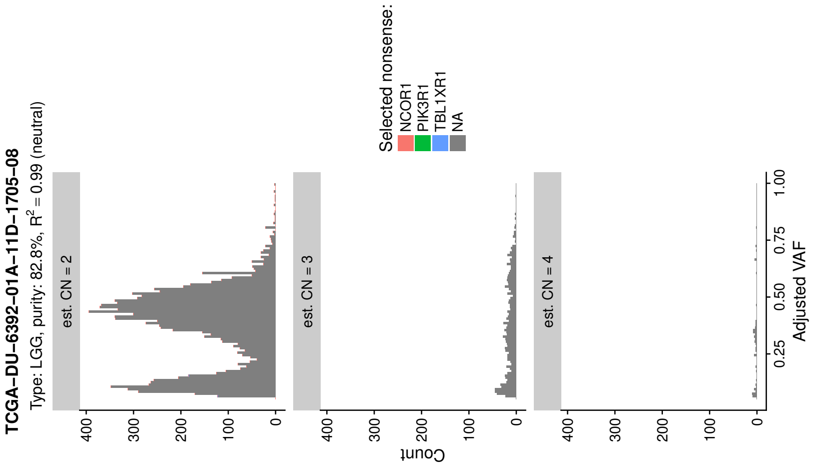
Supplementary Figure S.5: Neutrally classified case TCGA-12-0778 with multiple subclonal nonsense mutations. This is one of the 11/290 neutrally classified tumours with three or more subclonal nonsense mutations in cancer driver genes.



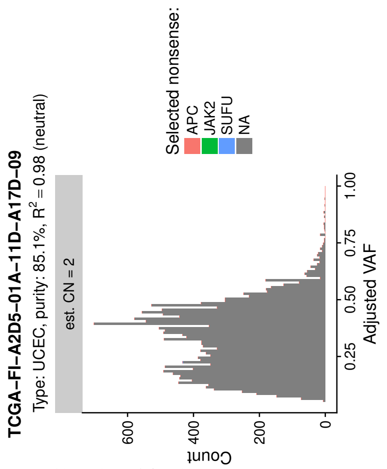
Supplementary Figure S.6: Neutrally classified case TCGA-DD-A1EE with multiple subclonal nonsense mutations. This is one of the 11/290 neutrally classified tumours with three or more subclonal nonsense mutations in cancer driver genes.



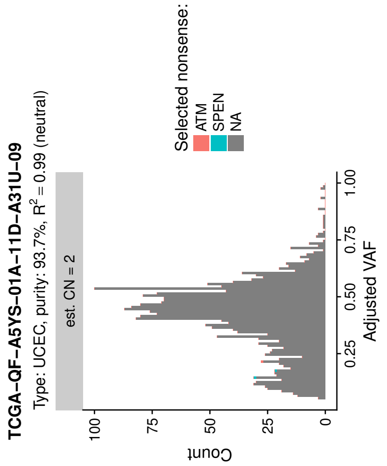
Supplementary Figure S.7: Neutrally classified case TCGA-DF-A2KU with multiple subclonal nonsense mutations. This is one of the 11/290 neutrally classified tumours with three or more subclonal nonsense mutations in cancer driver genes.



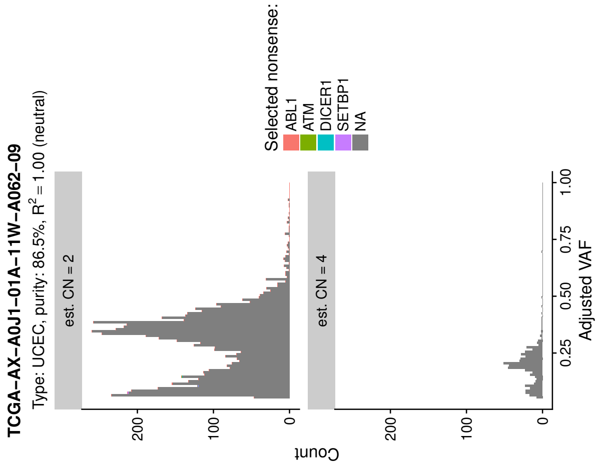
Supplementary Figure S.8: Neutrally classified case TCGA-DU-6392 with multiple subclonal nonsense mutations. This is one of the 11/290 neutrally classified tumours with three or more subclonal nonsense mutations in cancer driver genes.



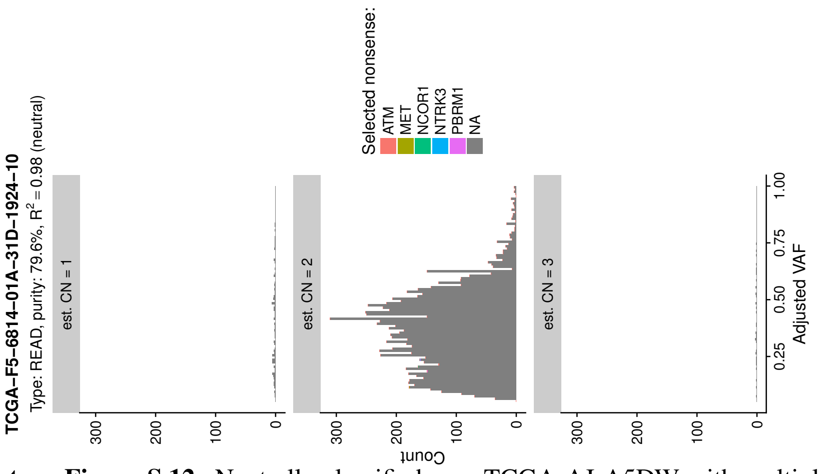
Supplementary Figure S.9: Neutrally classified case TCGA-FI-A2D5 with multiple subclonal nonsense mutations. This is one of the 11/290 neutrally classified tumours with three or more subclonal nonsense mutations in cancer driver genes.



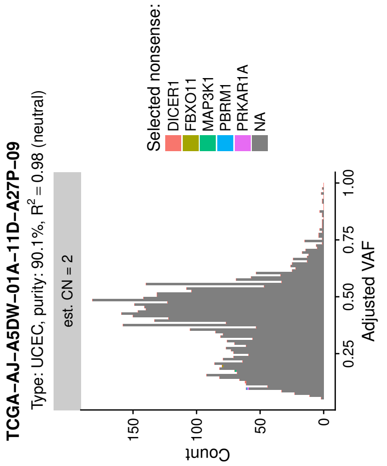
Supplementary Figure S.10: Neutrally classified case TCGA-AX-A0J1 with multiple subclonal nonsense mutations. This is one of the 11/290 neutrally classified tumours with three or more subclonal nonsense mutations in cancer driver genes.



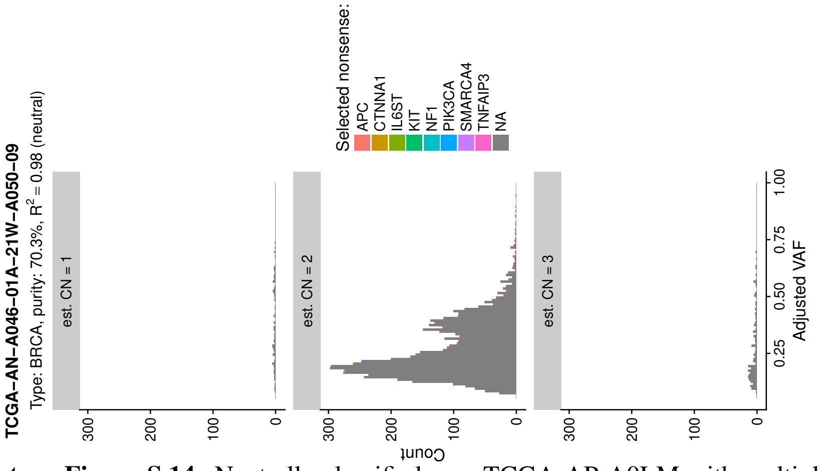
Supplementary Figure S.11: Neutrally classified case TCGA-F5-6814 with multiple subclonal nonsense mutations. This is one of the 11/290 neutrally classified tumours with three or more subclonal nonsense mutations in cancer driver genes.



Supplementary Figure S.12: Neutrally classified case TCGA-AJ-A5DW with multiple subclonal nonsense mutations. This is one of the 11/290 neutrally classified tumours with three or more subclonal nonsense mutations in cancer driver genes.



Supplementary Figure S.13: Neutrally classified case TCGA-AN-A046 with multiple subclonal nonsense mutations. This is one of the 11/290 neutrally classified tumours with three or more subclonal nonsense mutations in cancer driver genes.



Supplementary Figure S.14: Neutrally classified case TCGA-AP-A0LM with multiple subclonal nonsense mutations. This is one of the 11/290 neutrally classified tumours with three or more subclonal nonsense mutations in cancer driver genes.

S.2 Spatial Simulations

S.2.1 Details of the R Package CHESS

S.2.1.1 Implementation

The spatial simulator described in Section 3.1.1 (page 77) was implemented in the C++ general-purpose programming language and integrated into a package for the R statistical programming language (R Core Team 2020). For this, methods from the Rcpp package (Eddelbuettel and Francois 2011; Eddelbuettel 2013), which allows seamless integration of R and C++, were used. Further, methods from the following R packages were used: *ape* (Paradis and Schliep 2019), *phangorn* (Schliep 2011), *reshape* (Wickham 2007), *magrittr* (Bache and Wickham 2014), *dplyr* (Wickham et al. 2020), *tmvtnorm* (Wilhelm and G 2015), *ggplot2* (Wickham 2016), *ggtree* (Yu et al. 2017), *ggrepel* (Slowikowski 2020), *ggtern* (Hamilton and Ferry 2018), *ggrastr* (Petukhov, Brand, and Biederstedt 2020).

S.2.1.2 Pushing of Cells

During divisions in which non neighbouring grid cells were found to be empty dividing cells were assumed to make room by pushing cells along a vector **v**. This vector v was either chosen to be a random vector in space or a vector that points to the closest edge of the tumour.

Pushing of cells Cells along the identified vector **v** were generally moved until an empty grid point was found or the distance d_{push} was reached. In the first case, creating room for the daughter cell was considered successful, and the cell would undergo division. In the latter case, the daughter cell's creation was unsuccessful, and the division was aborted. For the moving of cells along the grid, a 3D version of the Bresham line drawing algorithm (Bresenham 1965; Heckbert 1990), which allows to quickly traverse each grid point along **v**, was used.

Vector for random pushing The vector for random pushing in space was generated by sampling

$$\phi = U(-\pi,\pi),\; u \sim U(-1,1)$$

followed by the determination of new target coordinates along the vector as

$$x' = x + sqrt(1-u^2)\cdot cos(\phi)d_{push},$$

$$y' = y + sqrt(1-u^2)\cdot sin(\phi)d_{push},$$

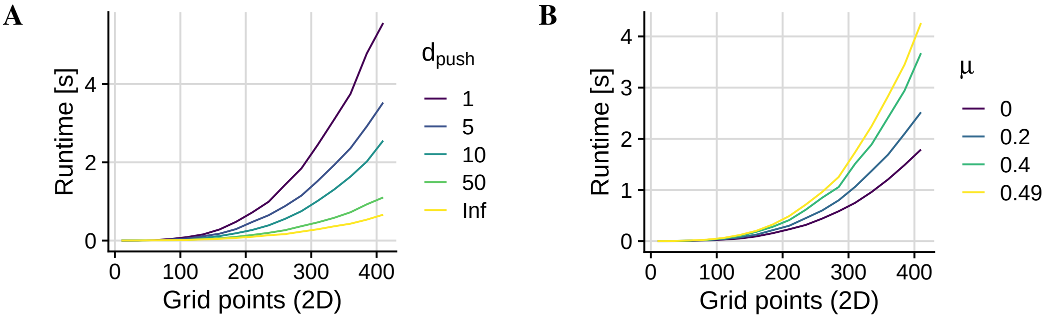
$$z' = z + ud_{push},$$

where x, y and z are the current coordinates and d_{push} the maximum distance to traverse along **v**. Pushing along **v** was terminated if a empty grid point was encountered.

Vector for pushing to the closest edge To speed up the identification of the closest edge of the tumour a simple heuristic was used: The search for the closest edge from the position $p_k = (x_k, y_k, z_k)$ of the cell k was started along the vector defined by $\phi = atan(y_k - y_c, x_k - x_c) + \phi_{os}$ with $\phi_{os} = U(-0.05\pi, 0.05\pi)$. Given the maximum distance to the edge d_{push} a quick 'sweep' around the position of the cell was conducted to identify an offset vector $\phi + \phi_{os}$ along which the position at distance 1.5 $d_{push} + 5$ was empty. For this, the values ϕ_{os} from $(0.2, -0.2, 0.4, ..., -0.8)$ were tried in turn until such a position was found. If none of these positions contained an empty position, the search was aborted. From the identified vector a binary search was conducted to find vectors with a closer edge. For this, an offset of $\phi_{os} \in \pi$ was tested, then an offset of $\phi_{os} \in \{-0.5\pi,\ 0.5\phi\}$ and so on, until a difference of $5\pi/360$ was reached. The search for the distance to the edge along the vector $\phi + \phi_{os}$ was done using a binary search between the maximum distance possible and the location of the cell. The code for this was only created for two-dimensional simulations.

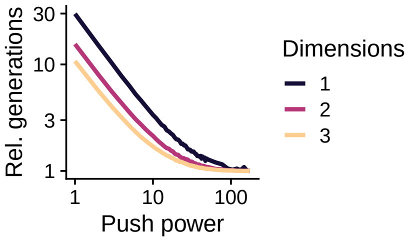
S.2.1.3 Runtime

The runtimes of the simulator for different 2D tumours with various diameter and different values of d_{push} and the death rate μ are shown in Figure S.15. The shown run times were obtained on a 1.4 GHz Intel Core i5 and are the average of 20 replicates. For cases with boundary driven growth (i.e., $d_{push} < r_{tumour}$) a much larger number of generations are required to reach a specific tumour size (compare Figure 3.3) and as such it is expected that this parameter significantly increases the runtime of the simulator as shown in Figure S.15A. Generally, the runtimes of the simulator were sufficiently short to simulate several thousand realisations — for example as part of a ABC inference — of a tumour containing up to $5 \cdot 10^5$ cell/glands in a feasible time frame (i.e., less than a week).

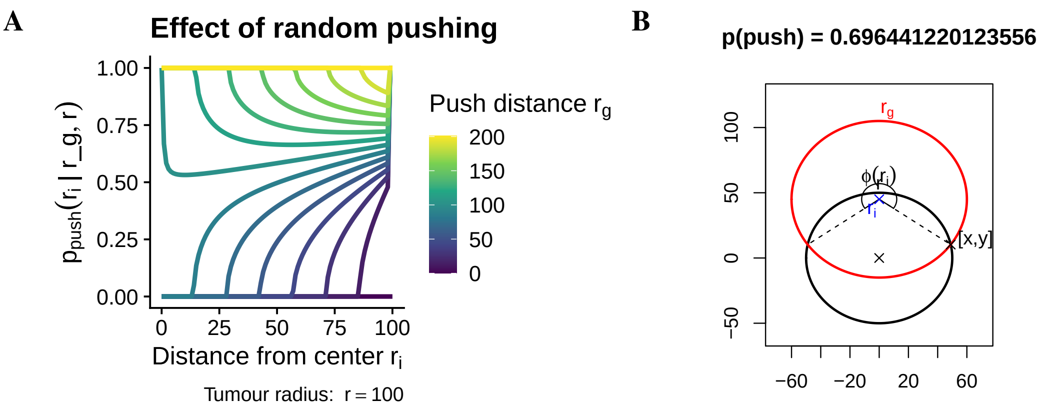


Supplementary Figure S.15: Average runtime for the generation of a spatial 2D tumour simulation as a function of the simulated tumour size. A) The left figure shows the effect of the push distance parameter (d_{push}) on the simulation runtime in the absence of any cell death ($\mu = 0$). B) Similarly, the effect of different death rates (μ) under non-exponential growth ($d_{push} = 20$) is shown on the right.

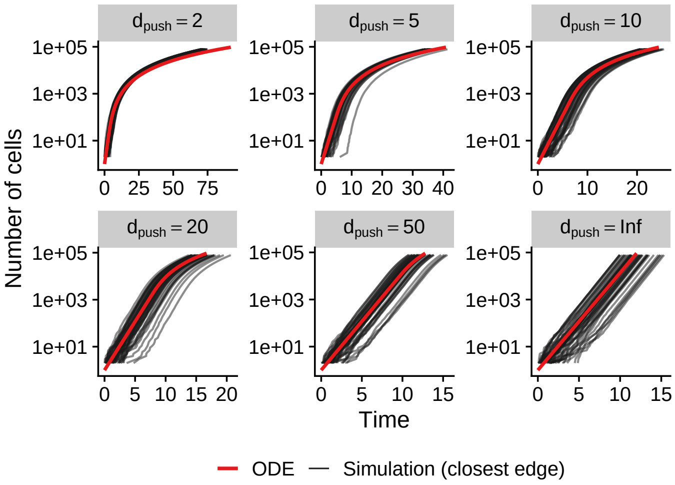
S.2.2 Supplementary Figures



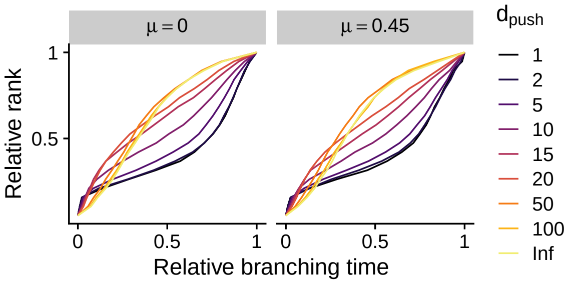
Supplementary Figure S.16: Expected time for a tumour to reach a radius of $r_{end} = 175$ under various degrees of boundary-driven growth and in different dimensions.



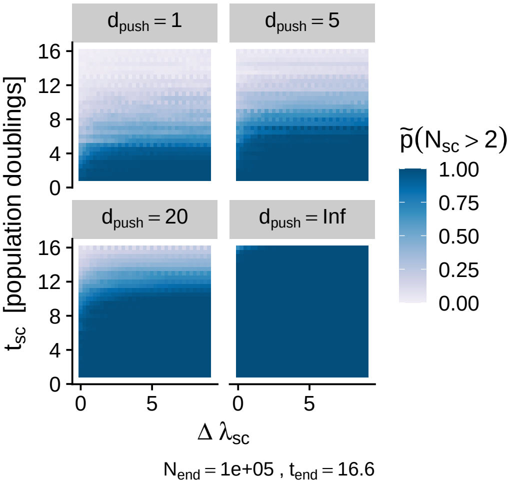
Supplementary Figure S.17: Likelihoods of random pushing.



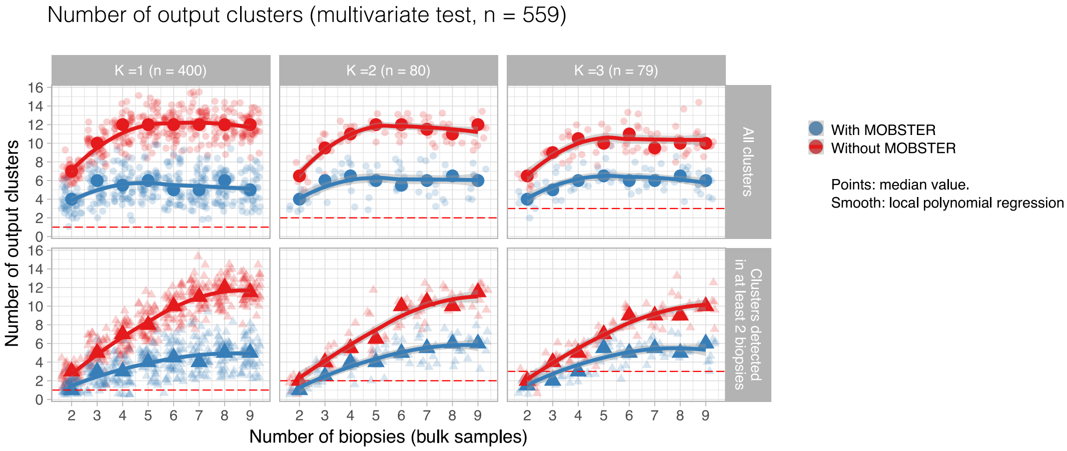
Supplementary Figure S.18: Comparison of growth curves expected for different values of the d_{push} parameter. Shown are the growth curves of a two dimensional neutrally growing tumour with a final radius of $r_{end} = 175$. The red line shows the expected growth dynamic from an ODE, the black lines show 50 random realisations obtained from the stochastic simulator.



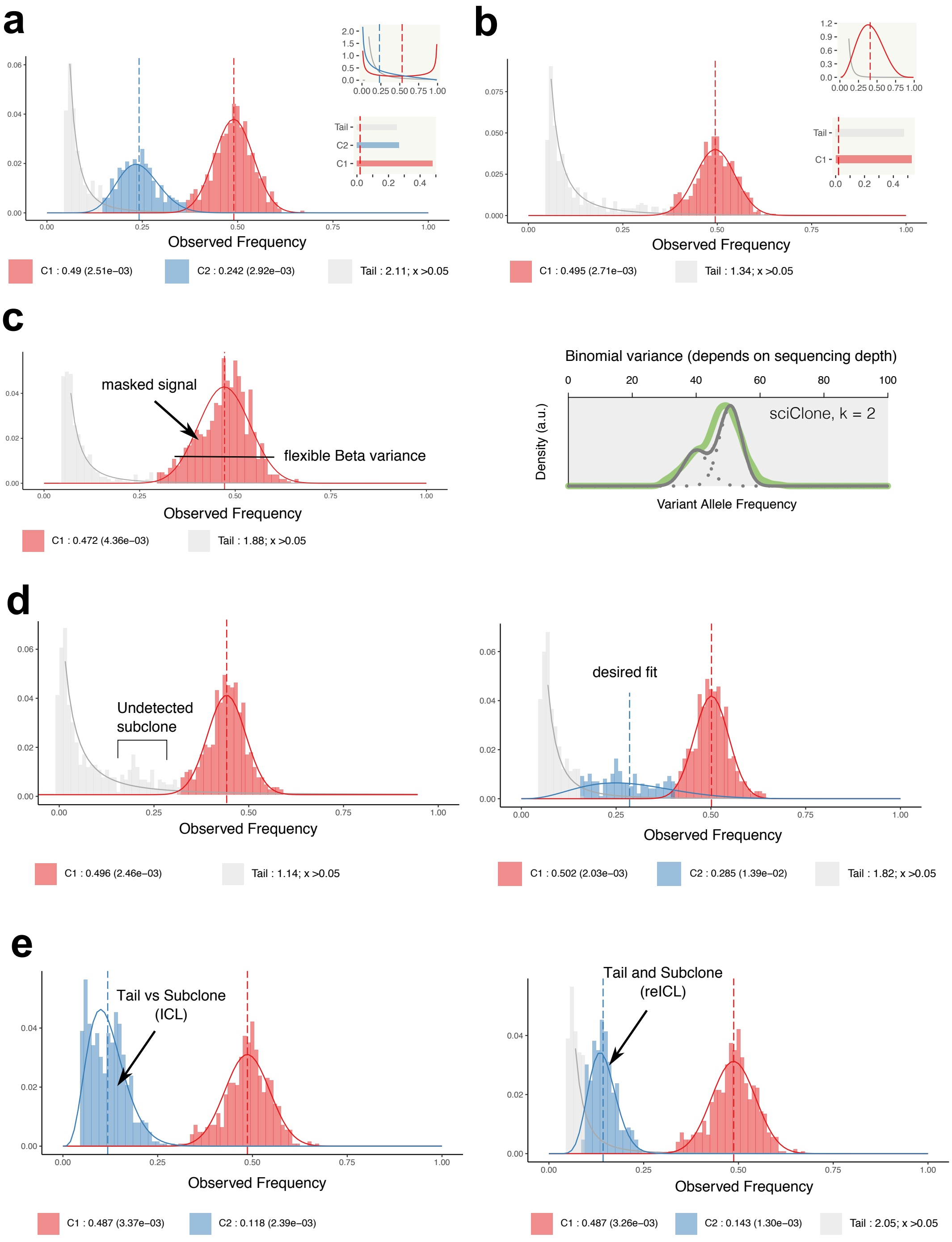
Supplementary Figure S.19: Average branching times under different degrees of boundary driven growth.



Supplementary Figure S.20: Fraction of expanding clones under different degrees of boundary driven growth.

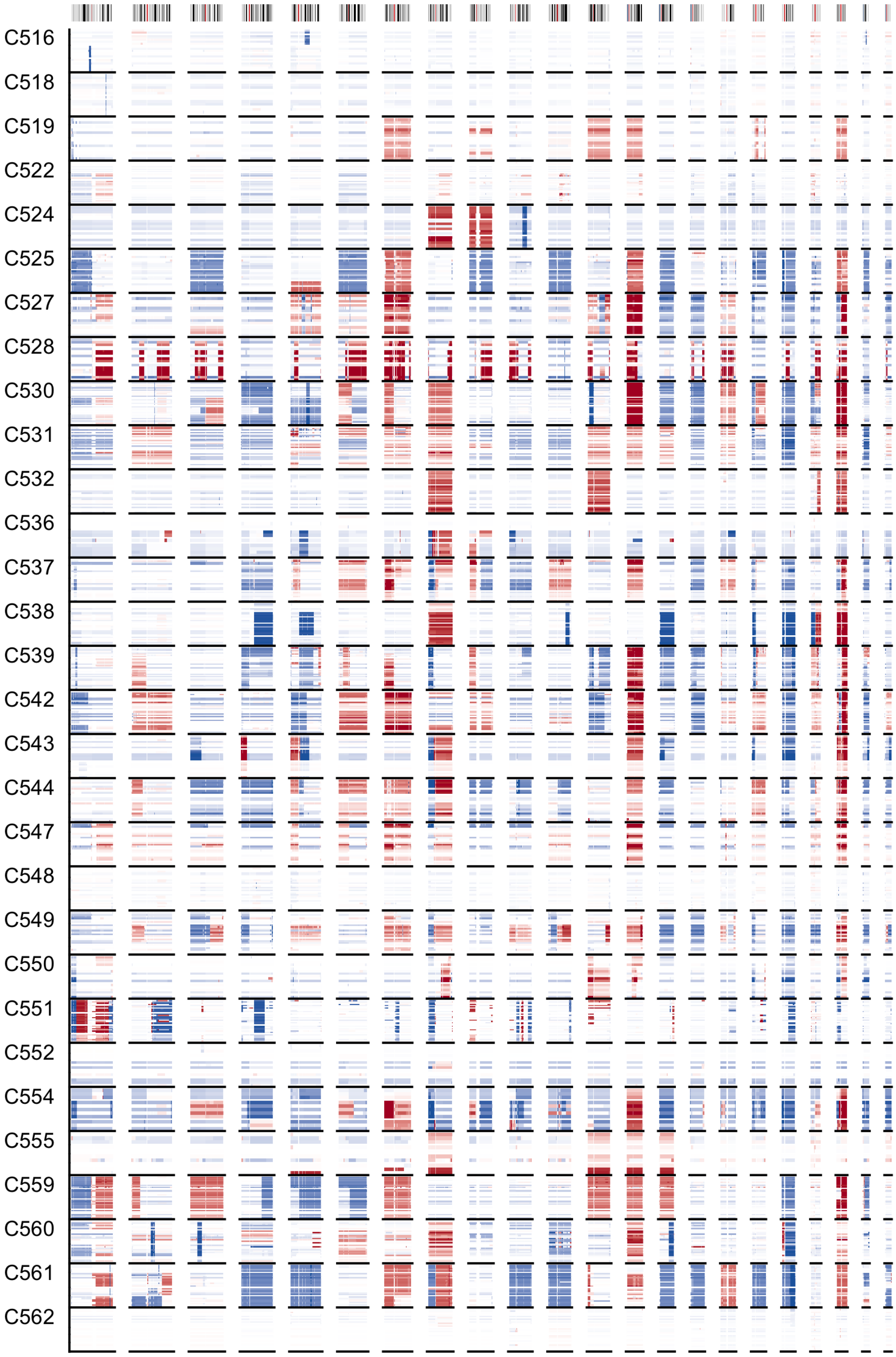


Supplementary Figure S.21: Performance of clustering methods in multivariate datasets.

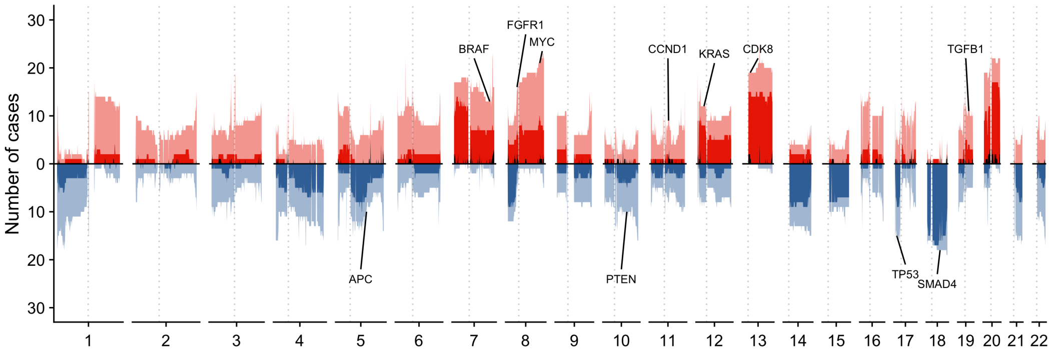


Supplementary Figure S.22: Examples fits of MOBSTER, SciClone and PyClone on univariate tumour samples.

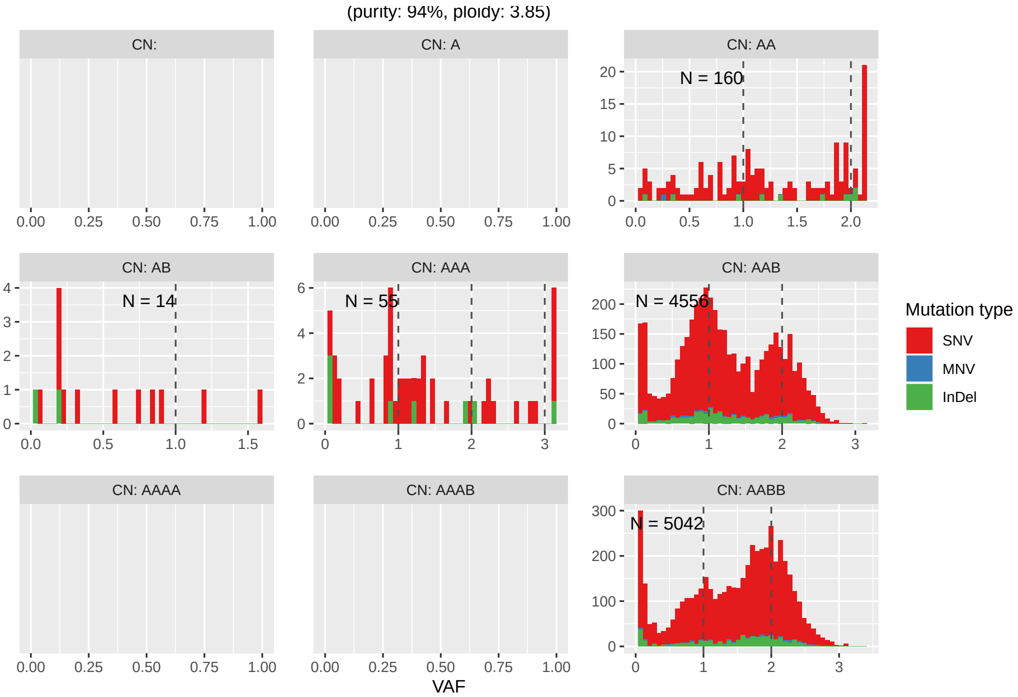
S.3 EPICC Data



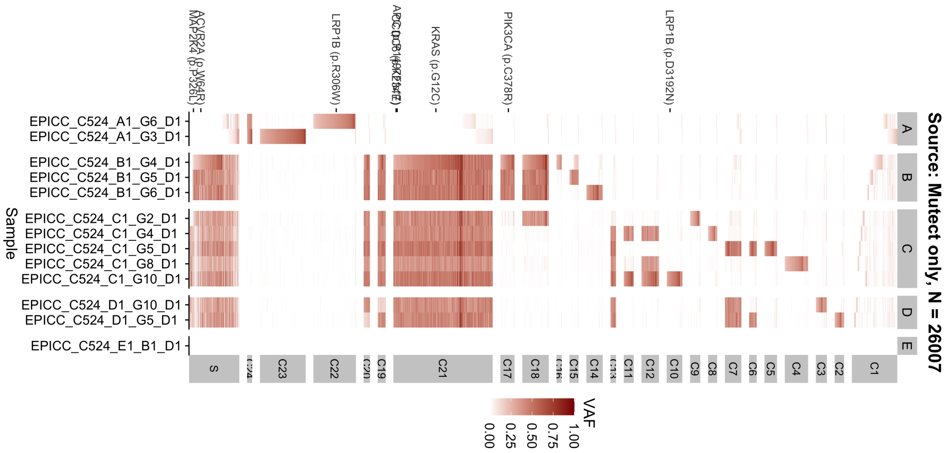
Supplementary Figure S.23: EPICC: CNA alterations



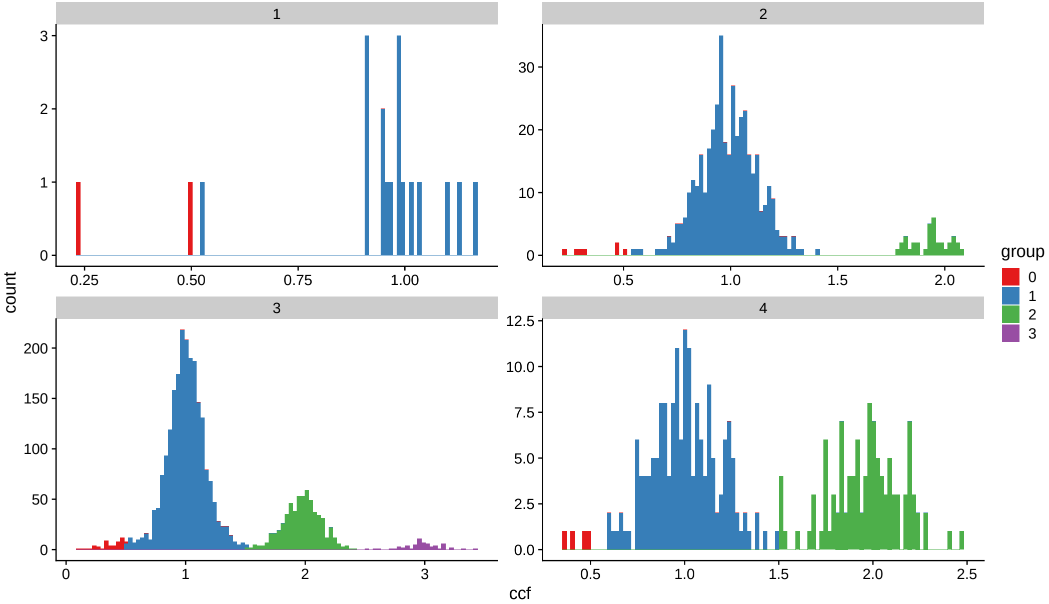
Supplementary Figure S.24: EPICC: Frequency of clonal and subclonal CNAs.



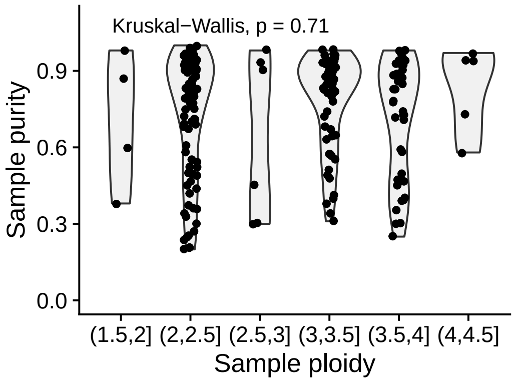
Supplementary Figure S.25: Example VAF distribution in a tetraploid tumour (C531).



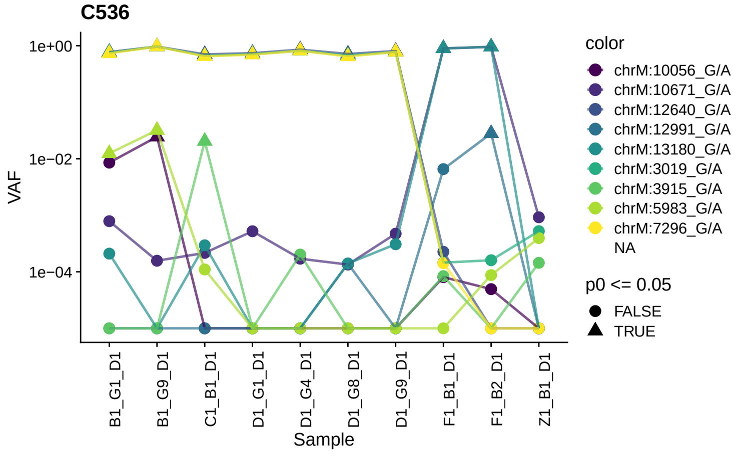
Supplementary Figure S.26: Example of normal adjacent glands.



Supplementary Figure S.27: Example of marginal mutation multiplicity estimates (C539).



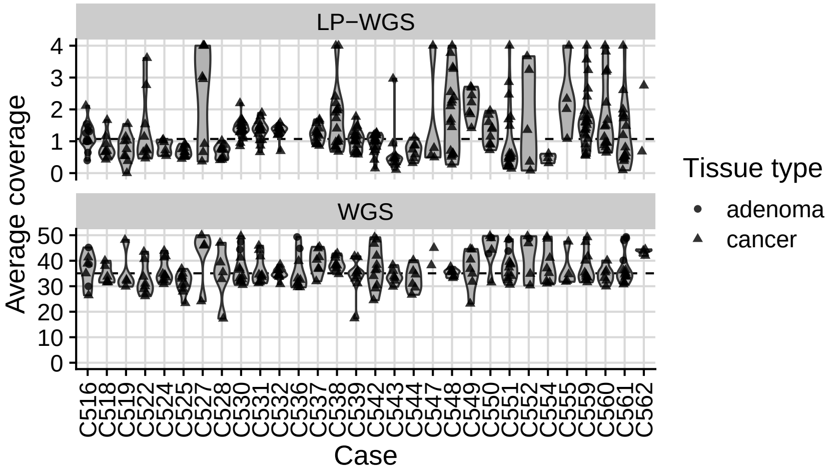
Supplementary Figure S.28: Correlation of estimated ploidy and purity in cancer WGS samples.



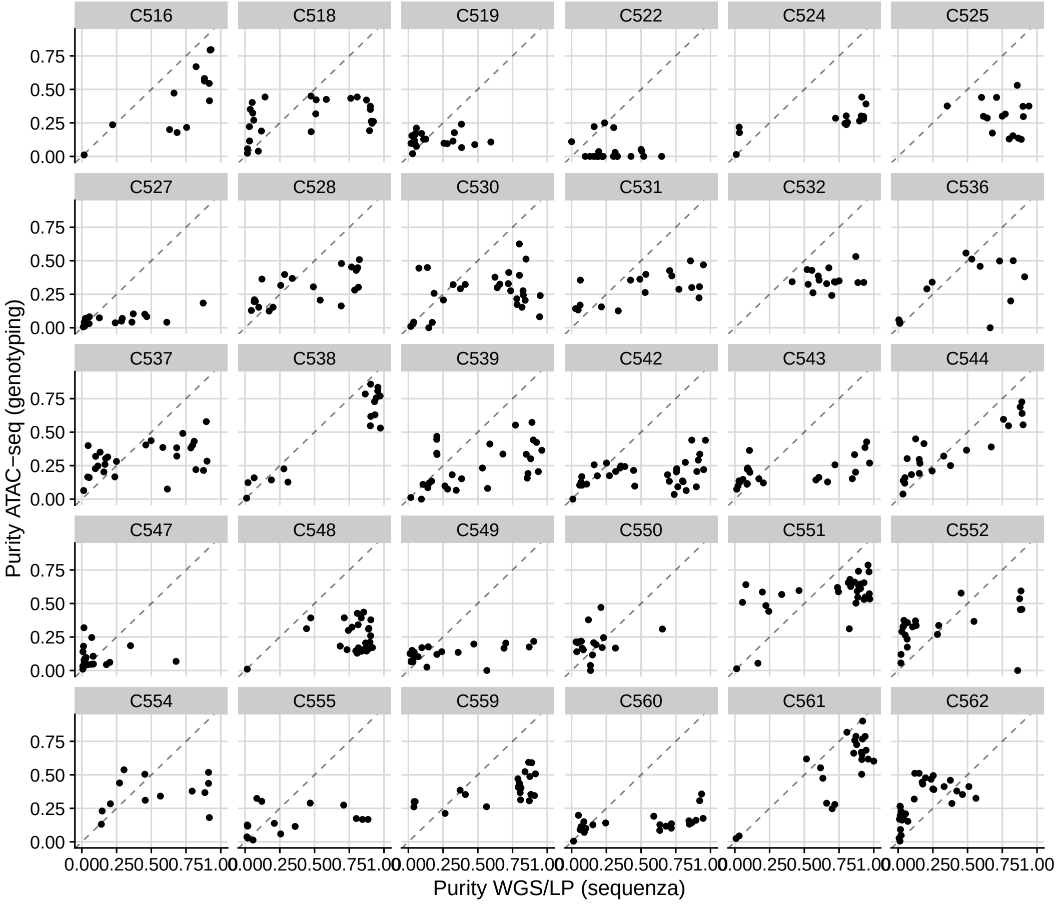
Supplementary Figure S.29: Example of mitochondrial variant calls in C536.

Supplementary Table S.1: Cases of the EPICC cohort. EM: Extramural, IM: Intramural, PN: Perineural.

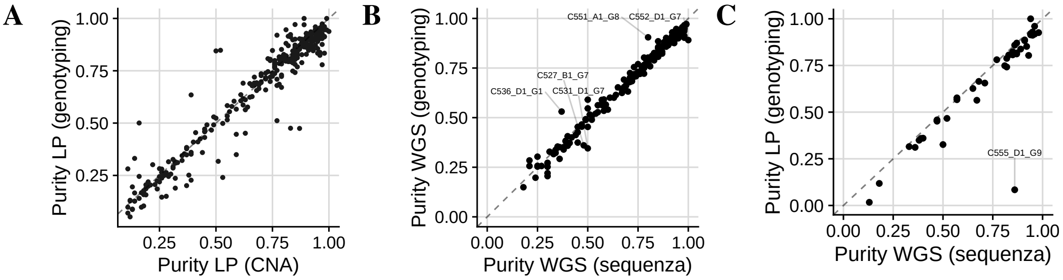
Case	Grade	TNM	TNM Stage	Gender	Venous Invasion
C516	2	pT3 N1 M0	IIIB	F	IM
C518	2	pT3 N0 M0	IIA	M	None
C519	2	pT3 N0 M0	IIA	M	IM
C522	3	pT4b N2 M0	IIIC	M	IM, EM, PN, lymphovascular
C524	2	pT3 N2 M?	IIIC	M	submucosal venous invasion
C525	2	pT3 N1 M0	IIIB	M	EM
C527	1	PT2 N1 M0	IIIA	M	None
C528	3	pT3 N2b M0	IIIC	F	EM
C530	?	pT3 N1b M0	IIIB	M	possible EM
C531	2	pT4a N2a M0	IIIC	F	IM, EM, PM
C532	1	pT4a N0 M0	IIC	M	None
C536	3	pT4a N1a M0	IIIB	M	IM
C537	2	pT4a N2a M0	IIIC	F	EM
C538	1	pT2 N0 M0	I	F	EM
C539	2	pT3 N0 M0	IIA	F	None
C542	2	pT3 N0 M0	IIA	M	EM
C543	2	pT3 N0 M0	IIA	M	EM
C544	2	pT2 N0 M0	I	M	IM
C547	2	pT3 N0 M0	IIA	M	EM, PN
C548	2	pT3 N0 M0	IIA	M	submucosal venus invasion
C549	2	pT2 N0 M0	I	F	None
C550	2	pT3 N1b M0	IIIB	M	IM
C551	2	pT3 N0 M0	IIA	M	None
C552	2	pT4a N1c M0	IIIB	M	EM lymphatic
C554	2	pT3 N0 M0	IIA	M	IM
C555	2-3	pT4a N1c M1	IIIB	F	EM
C559	2	pT4a N0 M0	IIB	M	EM
C560	2	pT3 N0 M0	IIA	M	None
C561	2	pT3 N2b M0	IIIC	M	EM
C562	2	pT3 N1b M0	IIIB	M	IM venous & lymphatic



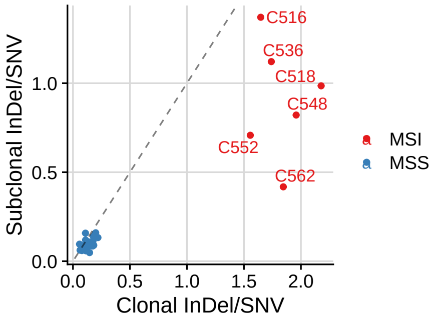
Supplementary Figure S.30: Median coverage of WGS samples.



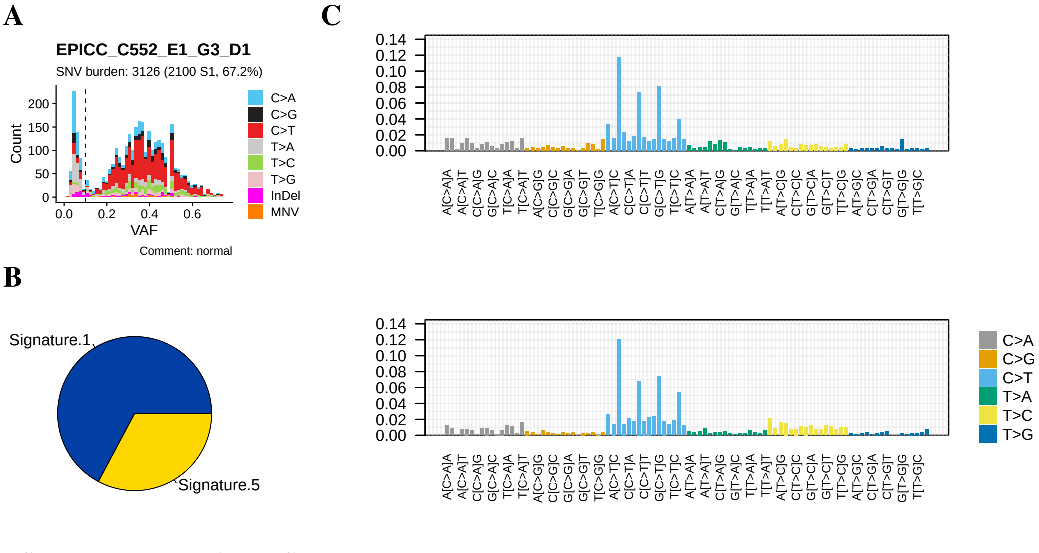
Supplementary Figure S.31: Correlation of ATAC-seq and WGS purity.



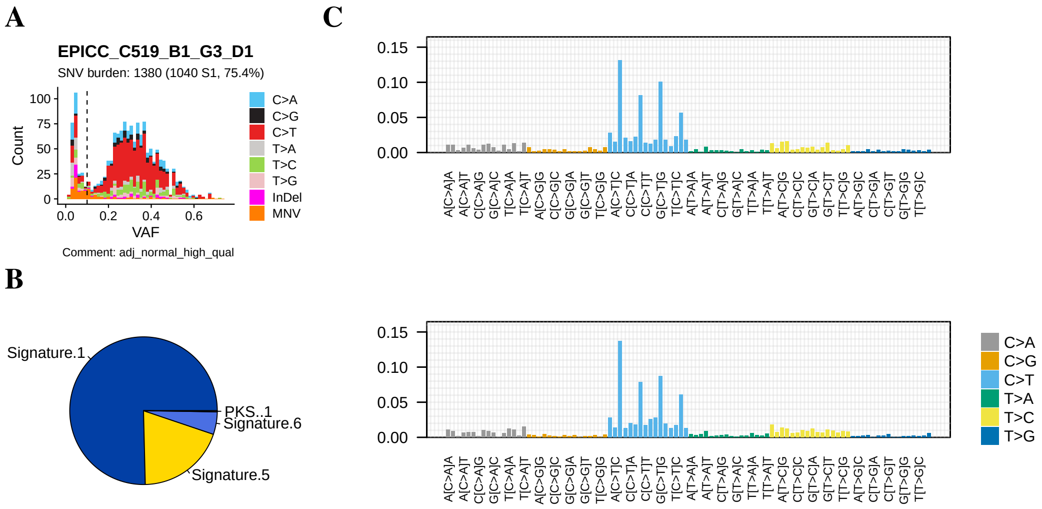
Supplementary Figure S.32: ML purity estimates. A) Correlation of ML and CNA LP-WGS purity estimates. B) ML estimates of WGS sample purities based on somatic variants. C) ML estimates of LP-WGS sample purities based on somatic variants.



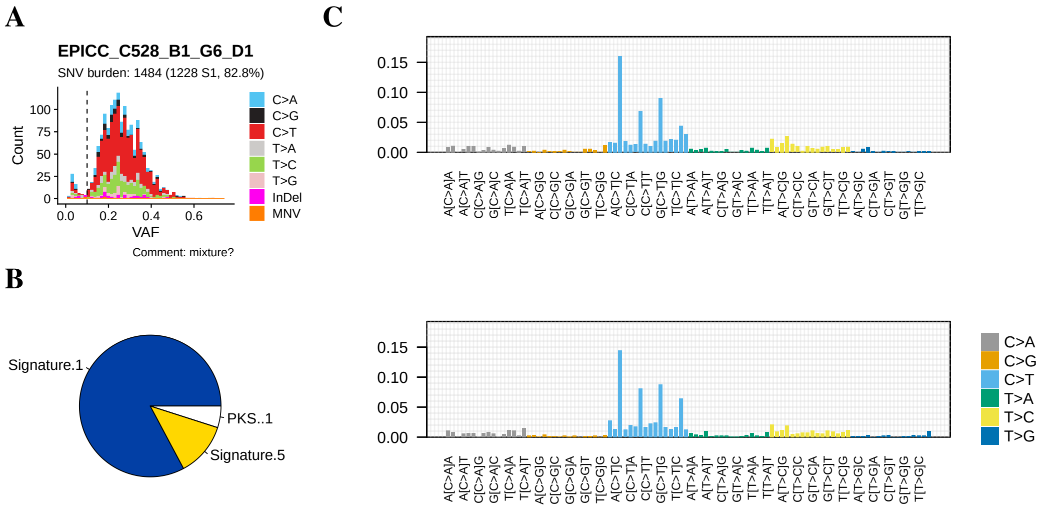
Supplementary Figure S.33: EPICC: Clonal and subclonal InDel/SNV ratio in MSI and MSS cases.



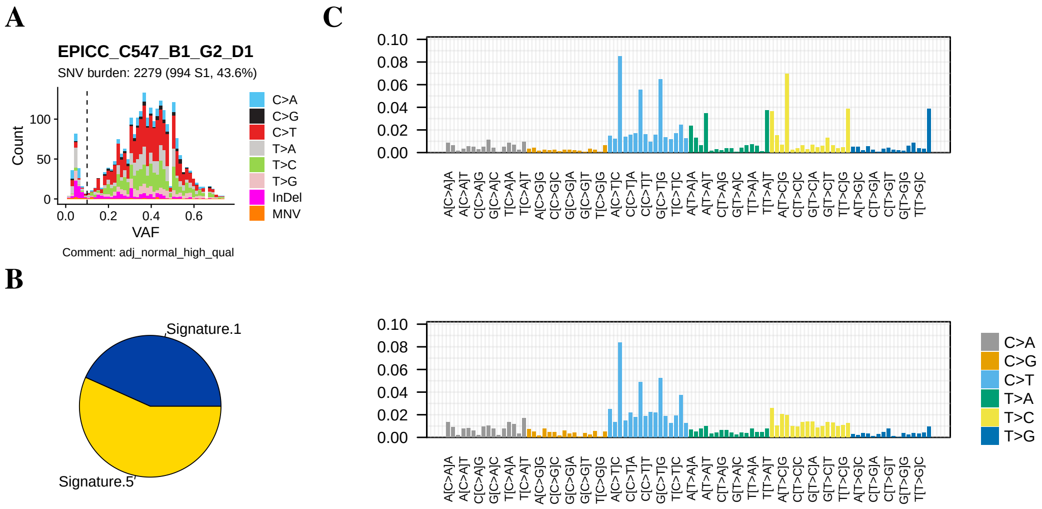
Supplementary Figure S.34: Mutation spectrum of a representative normal crypt C552_E1_G3.



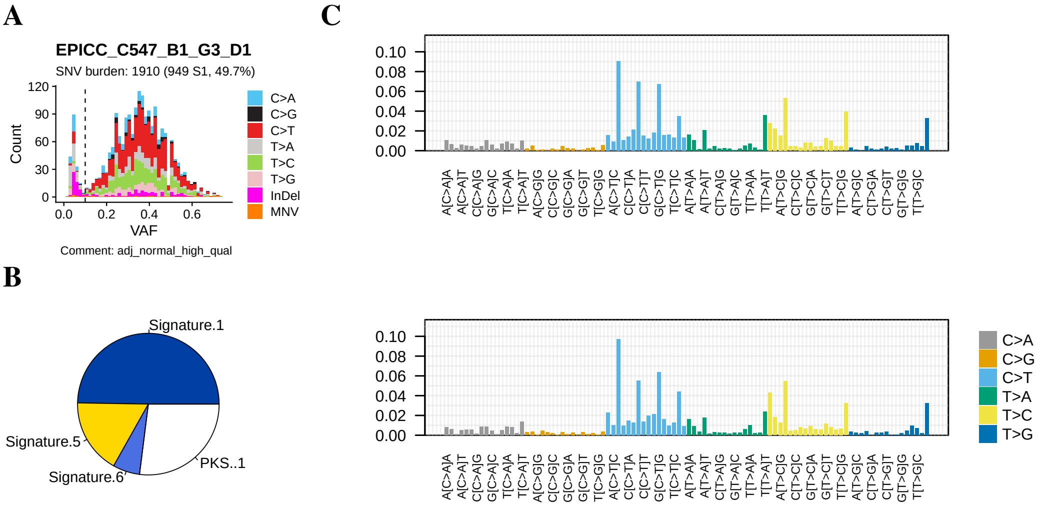
Supplementary Figure S.35: Mutation spectrum of a representative cancer adjacent normal crypt C519_B1_G3.



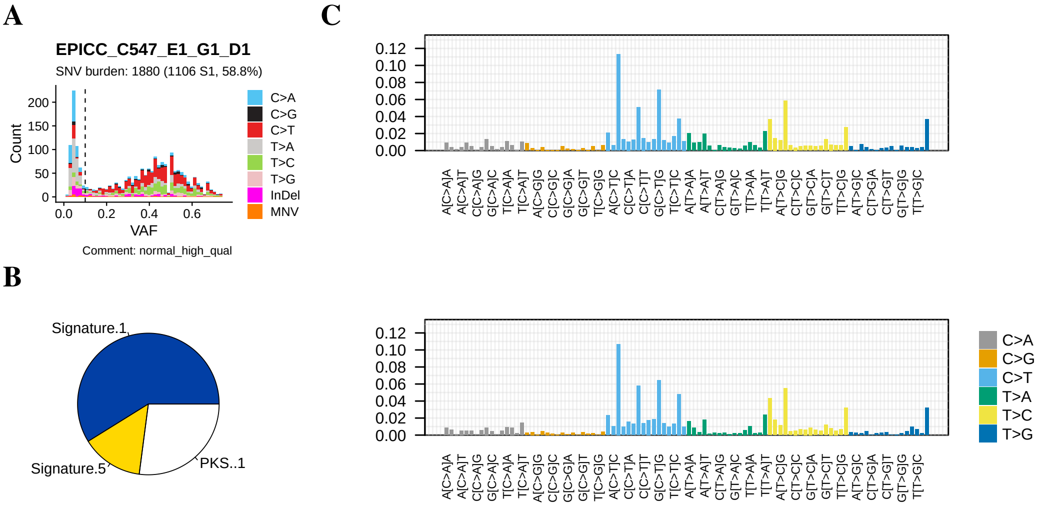
Supplementary Figure S.36: Mutation spectrum of a representative intermixed normal crypt C528_B1_G6.



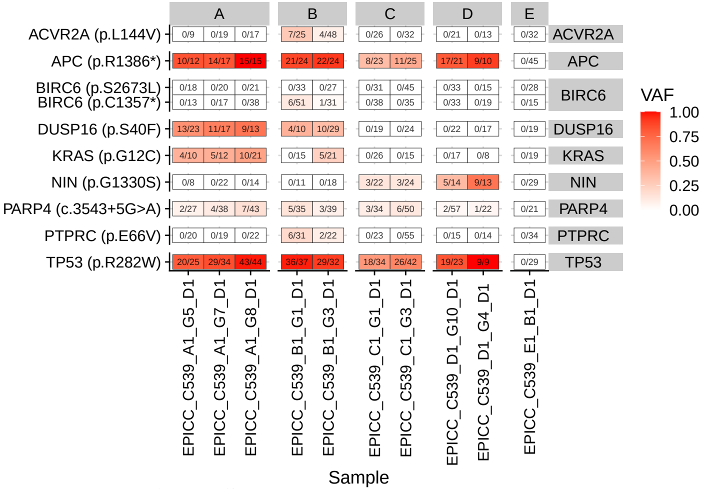
Supplementary Figure S.37: Mutation spectrum of a normal adjacent crypt with the *pks*$^+$ mutational signature C547_B1_G2.



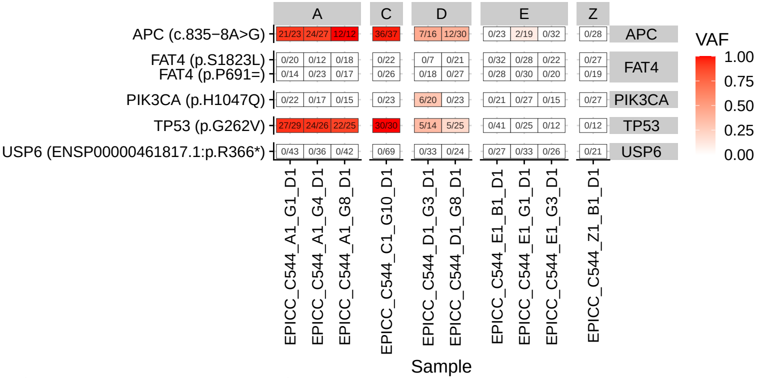
Supplementary Figure S.38: Mutation spectrum of a normal adjacent crypt with the *pks*$^+$ mutational signature C547_B1_G3.



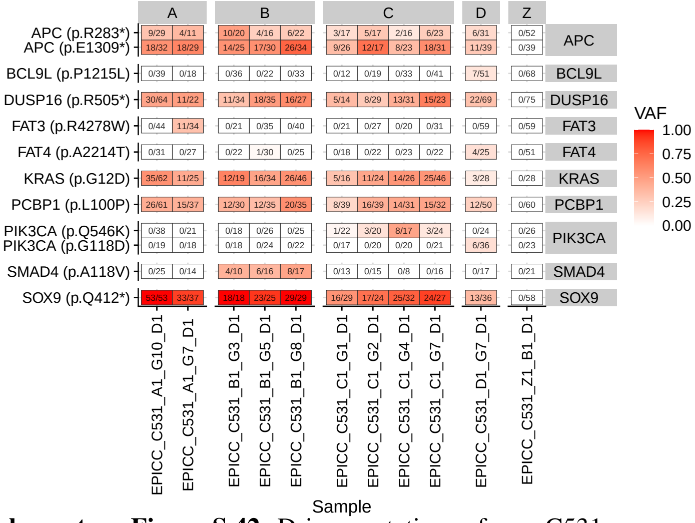
Supplementary Figure S.39: Mutation spectrum of a normal adjacent crypt with the *pks*$^+$ mutational signature in sample C547_E1_G1.



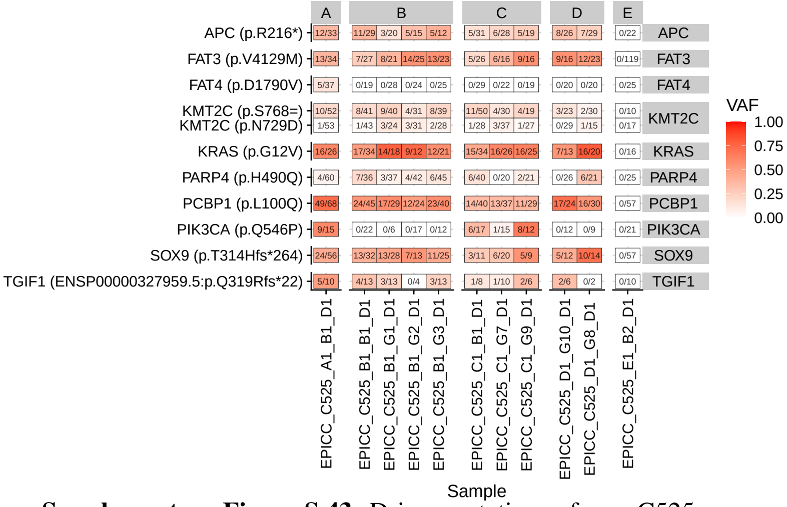
Supplementary Figure S.40: Driver mutations of case C539.



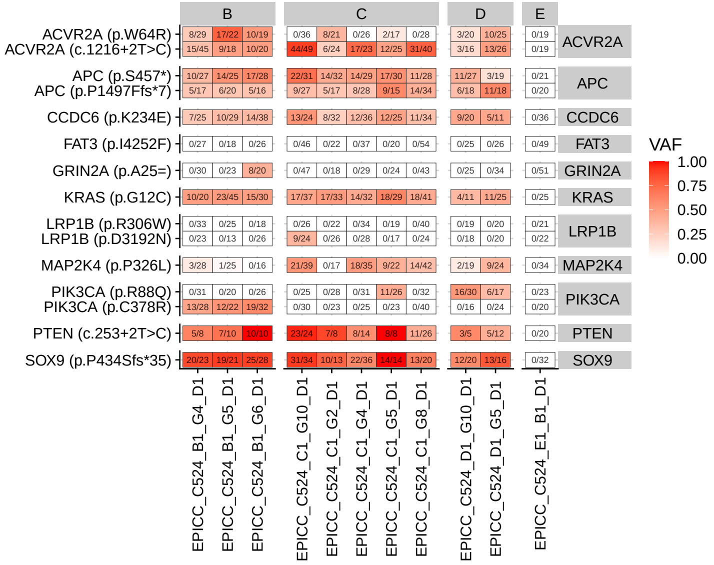
Supplementary Figure S.41: Driver mutations of case C544.



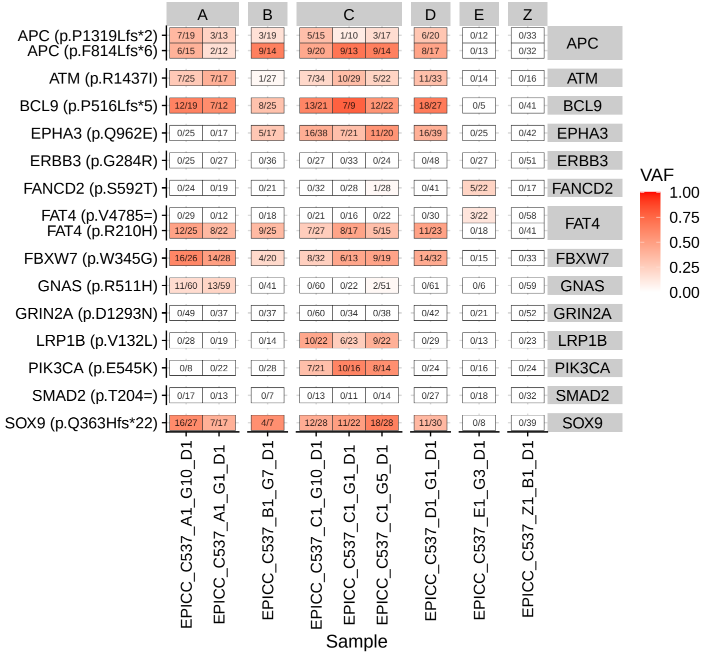
Supplementary Figure S.42: Driver mutations of case C531.



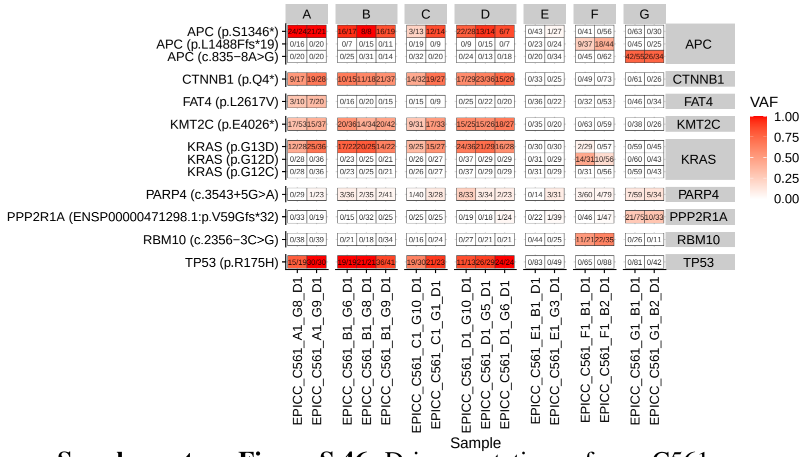
Supplementary Figure S.43: Driver mutations of case C525.



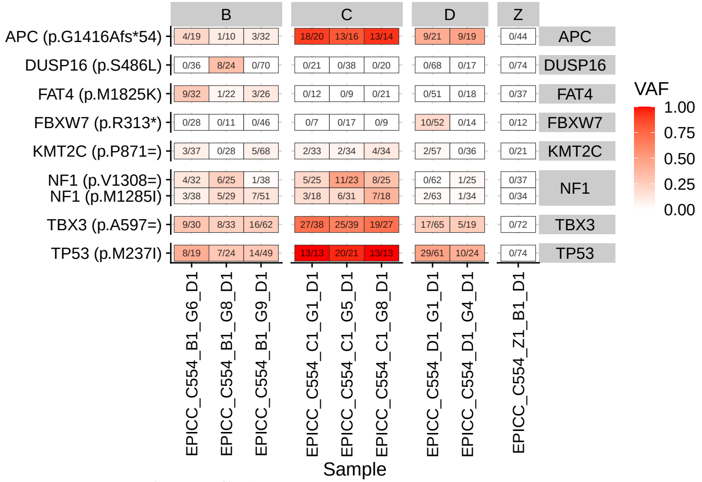
Supplementary Figure S.44: Driver mutations of case C524.



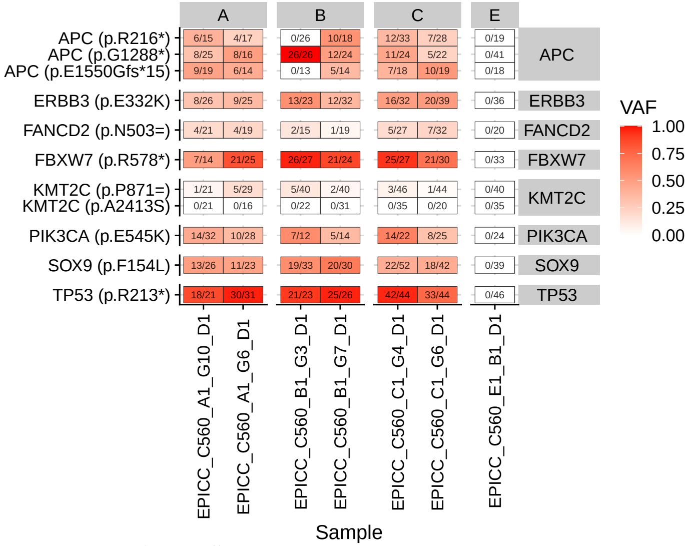
Supplementary Figure S.45: Driver mutations of case C537.



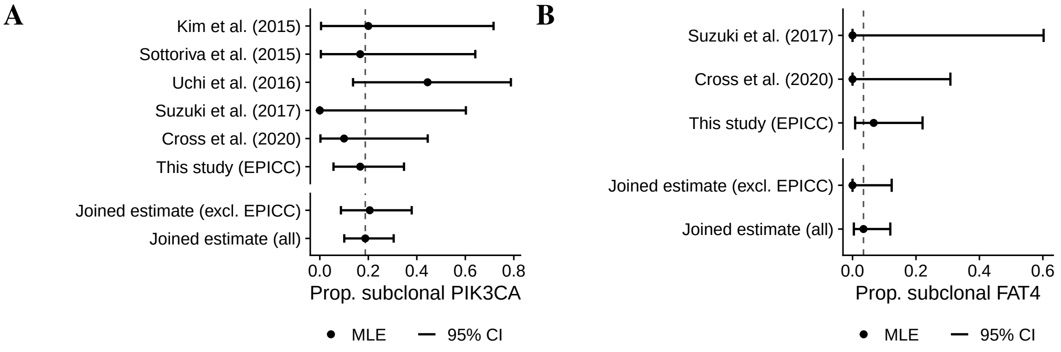
Supplementary Figure S.46: Driver mutations of case C561.



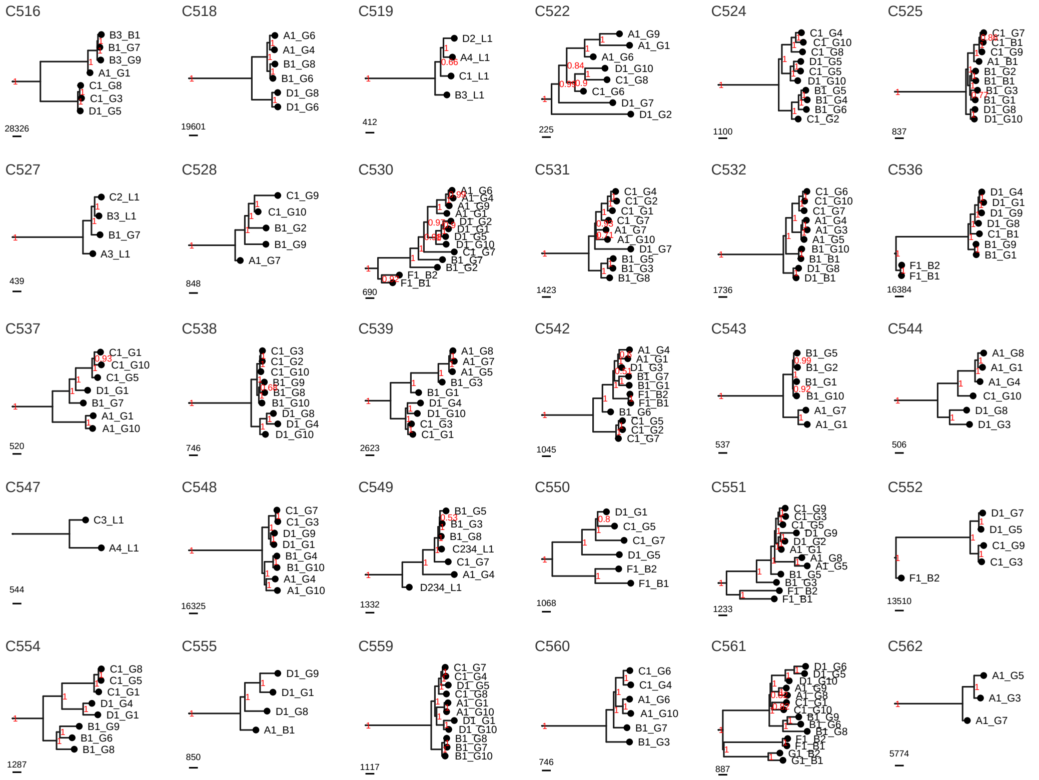
Supplementary Figure S.47: Driver mutations of case C554.



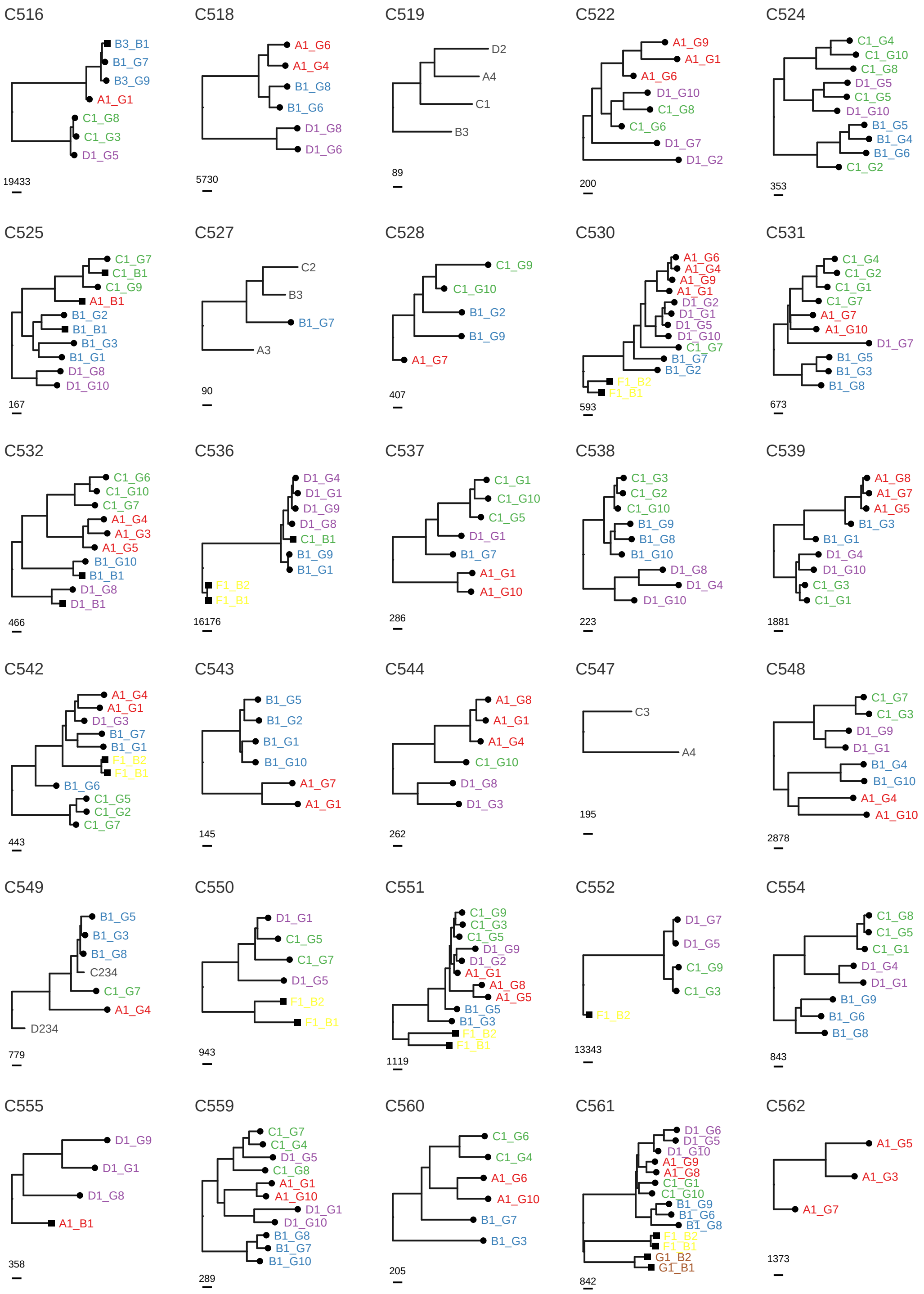
Supplementary Figure S.48: Driver mutations of case C560.



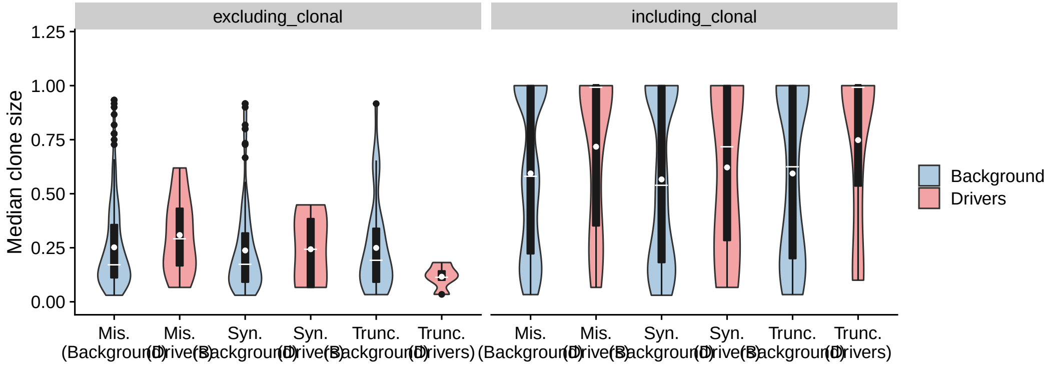
Supplementary Figure S.49: Meta-analysis of subclonal PIK3CA (A) and FAT4 (B) mutations in CRC. For each study and the joined analysis 95% confidence intervals (CI) and maximum likelihood estimates (MLE) of the fraction of CRC with subclonal mutations are shown.



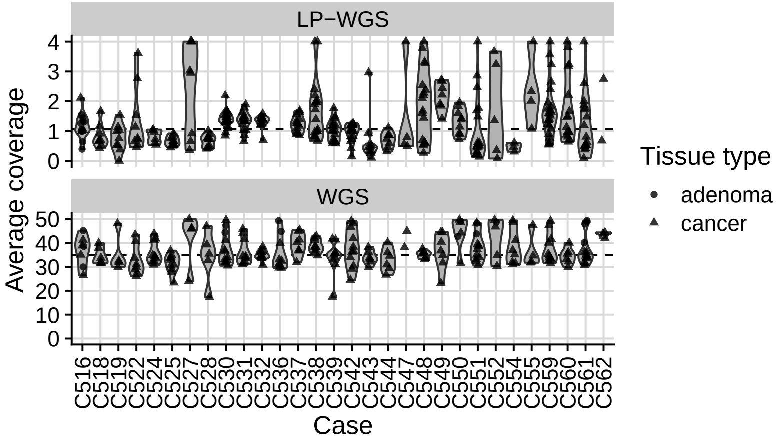
Supplementary Figure S.50: Bootstrap support of maximum-parsimony trees from the EPICC cohort.



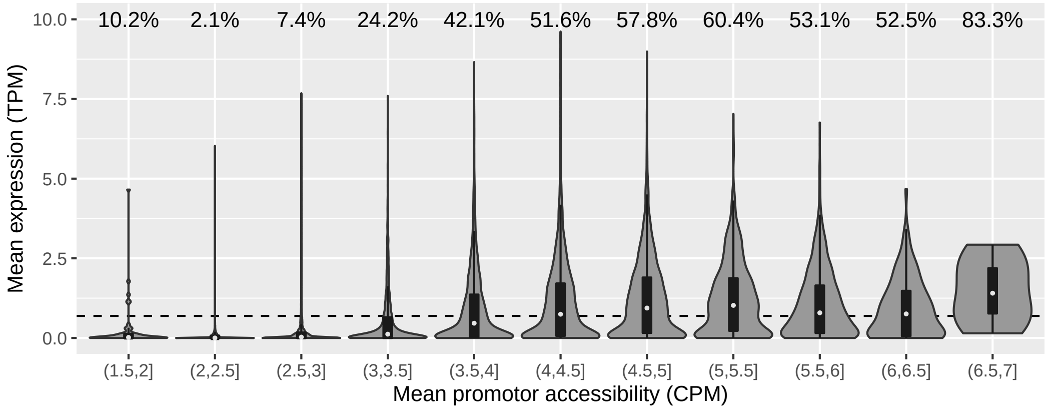
Supplementary Figure S.51: MP trees reconstructed from all mutations.



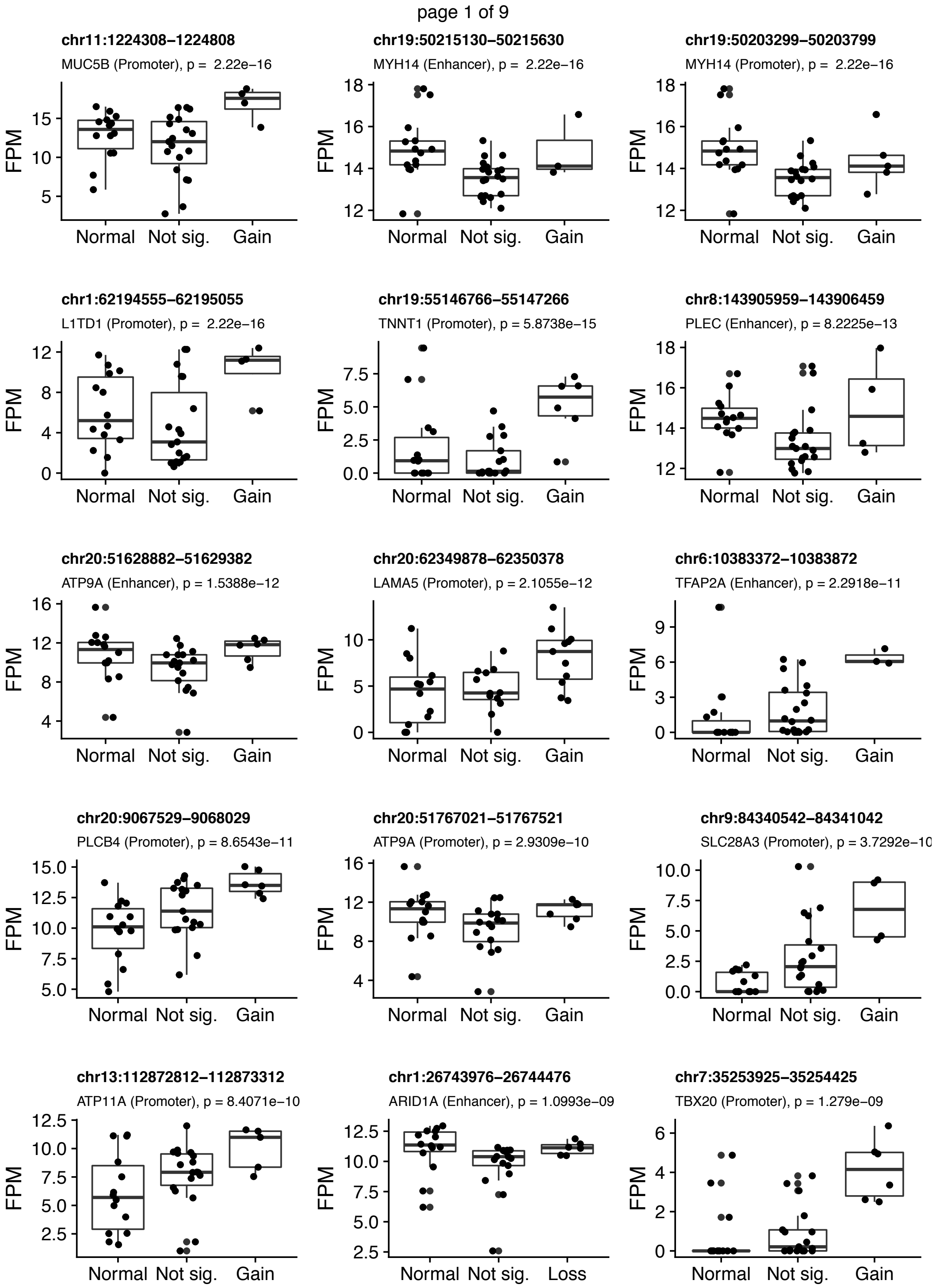
Supplementary Figure S.52: Median clone size for driver and background genes.



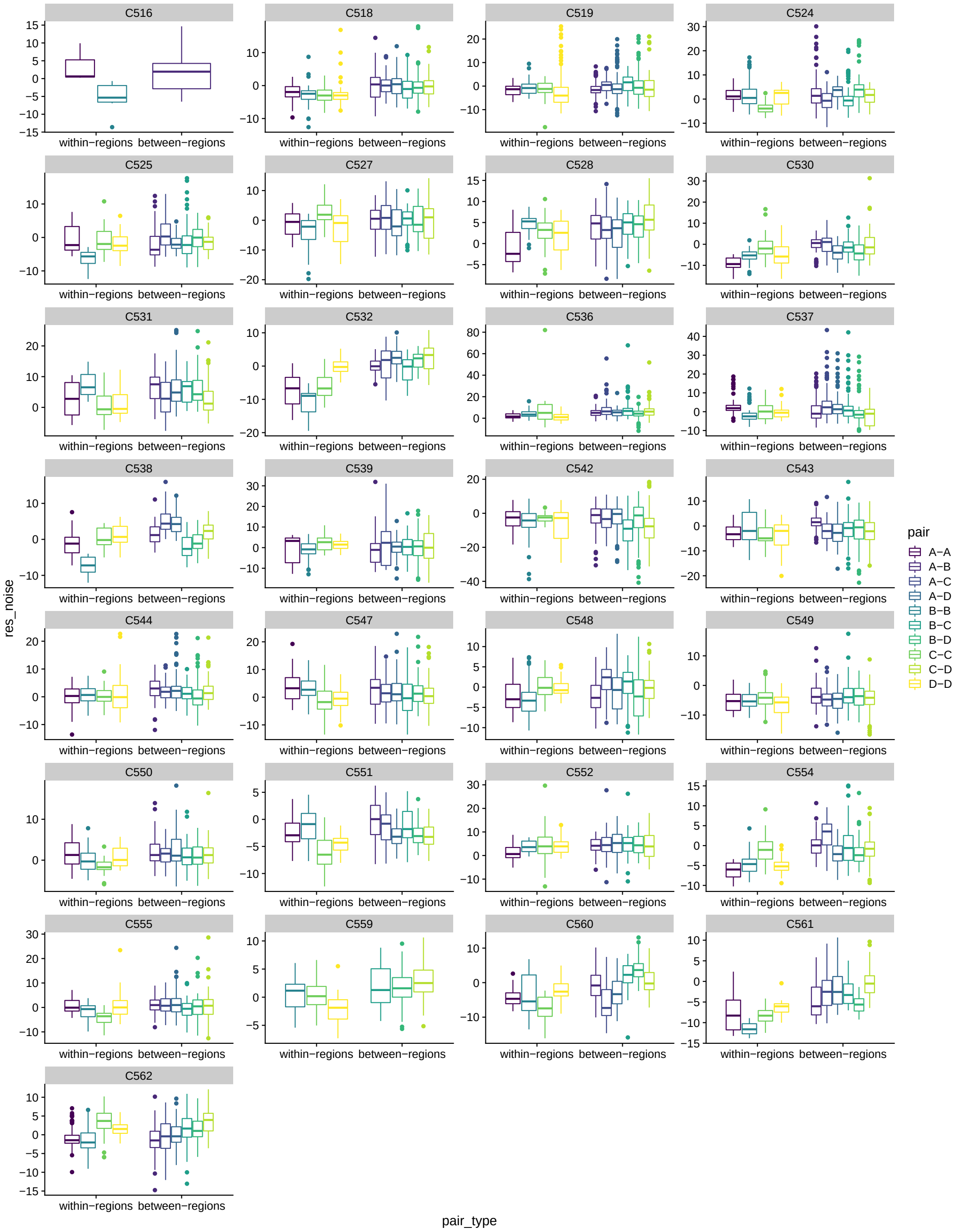
Supplementary Figure S.53: EPICC: WGS and LP-WGS per patient coverage data.



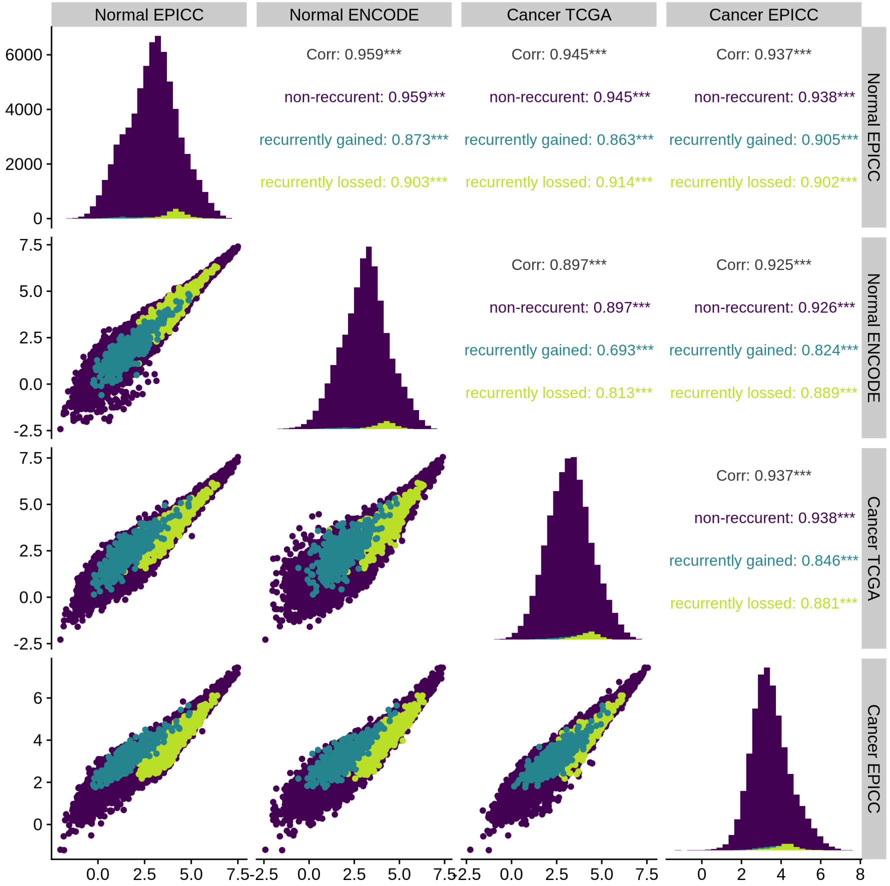
Supplementary Figure S.54: Relationship of promoter accessibility measured with ATAC-seq and expression of matched genes measured by RNA-seq. Shown values are the averages over all obtained normal colon glands. Genes with an average promoter accessibility below 3.5 CPM are generally not expressed. A subset of $\approx 60\%$ of those with a promoter accessibility above 3.5 CPM are expressed (i.e., TPM > 1).



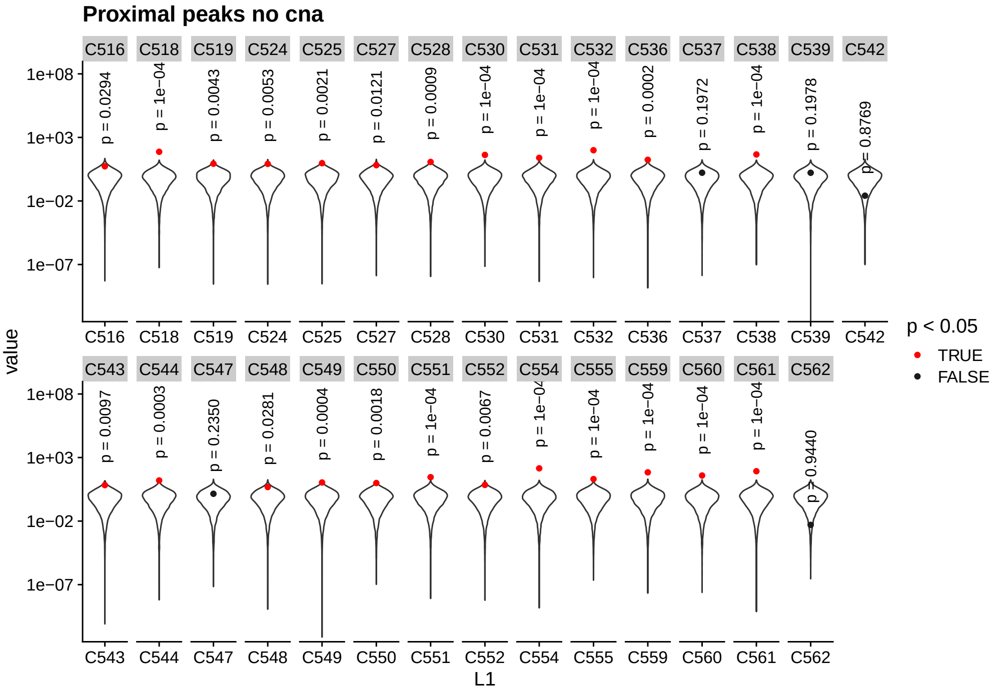
Supplementary Figure S.55: Examples of recurrently altered focal chromatin accessibility alterations with matched changes of gene expression.



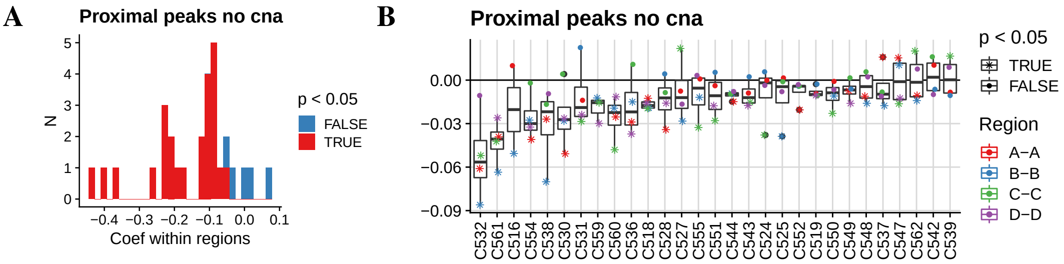
Supplementary Figure S.56: Global ATAC-seq differences between and within regions.



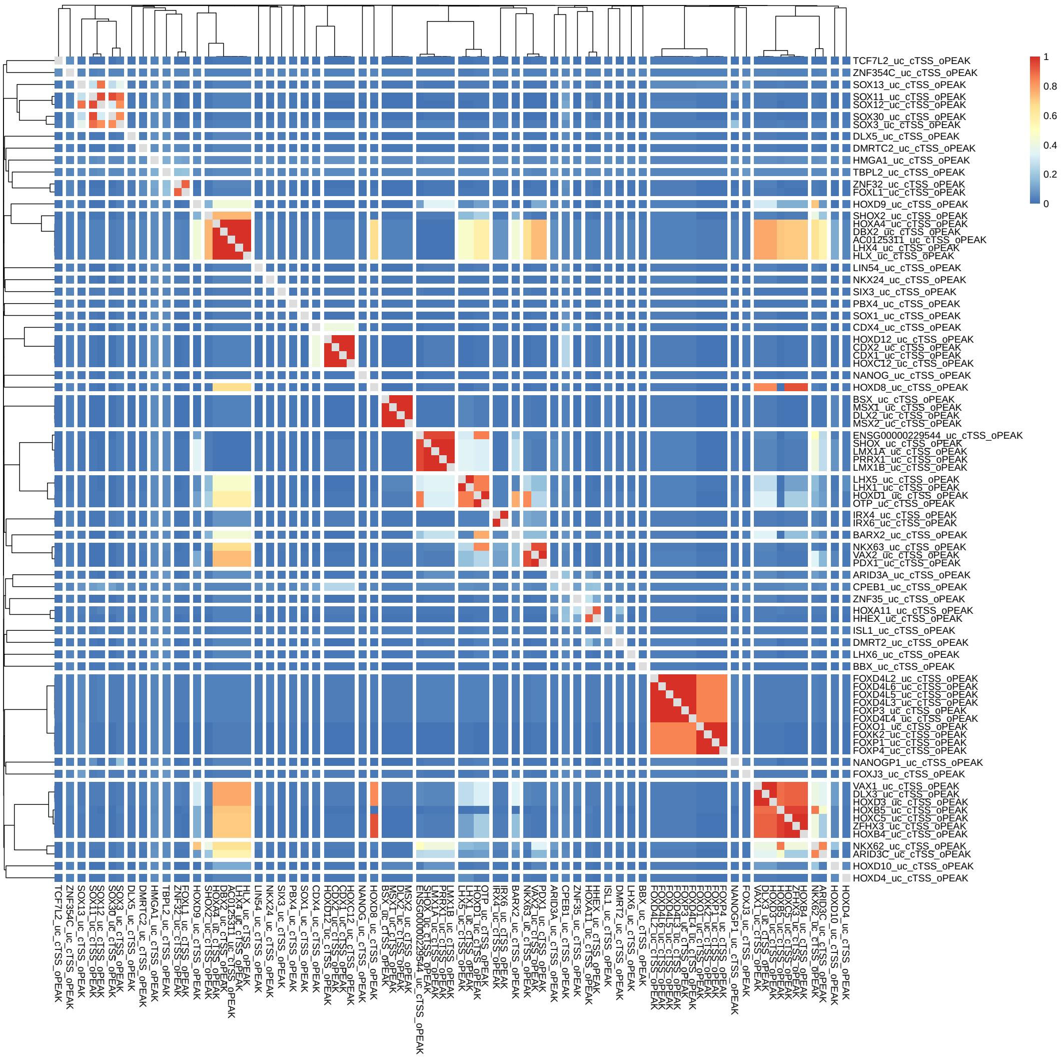
Supplementary Figure S.57: Correlation of CPM values of reads in regions of open chromatin in EPICC (normal & tumour), ENCODE (normal) and TCGA (tumour) datasets.



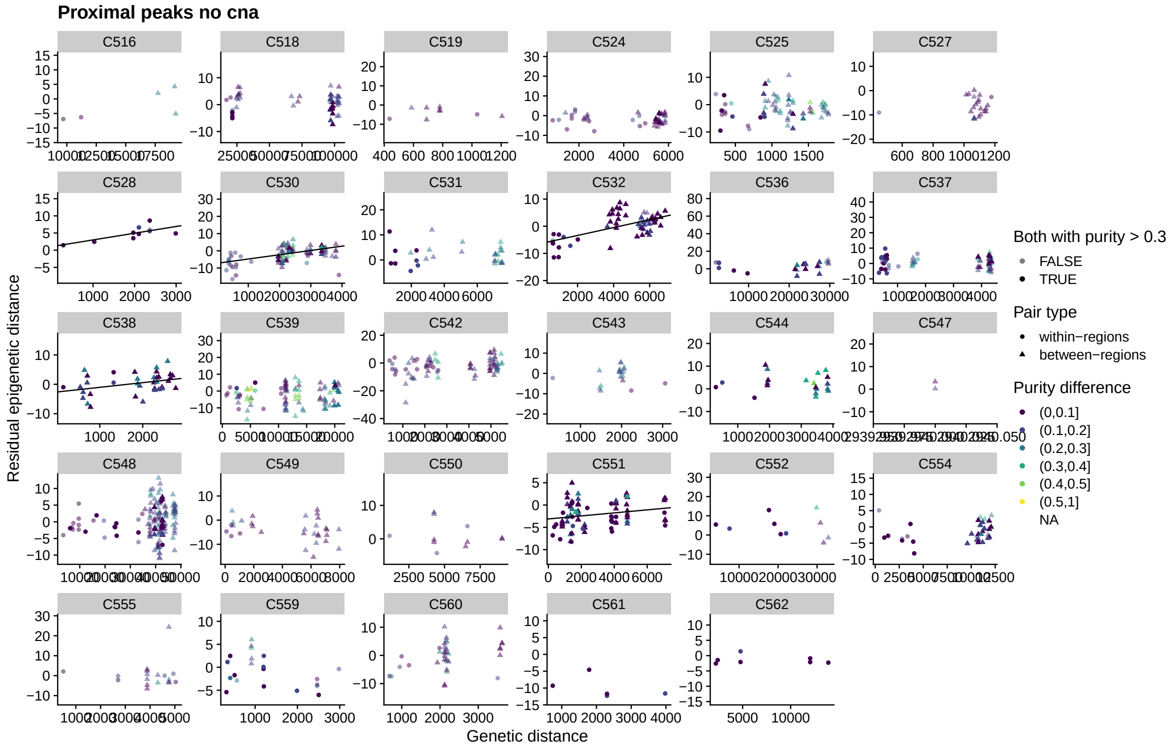
Supplementary Figure S.58: ANOVA of ATAC-seq differences between and within regions.



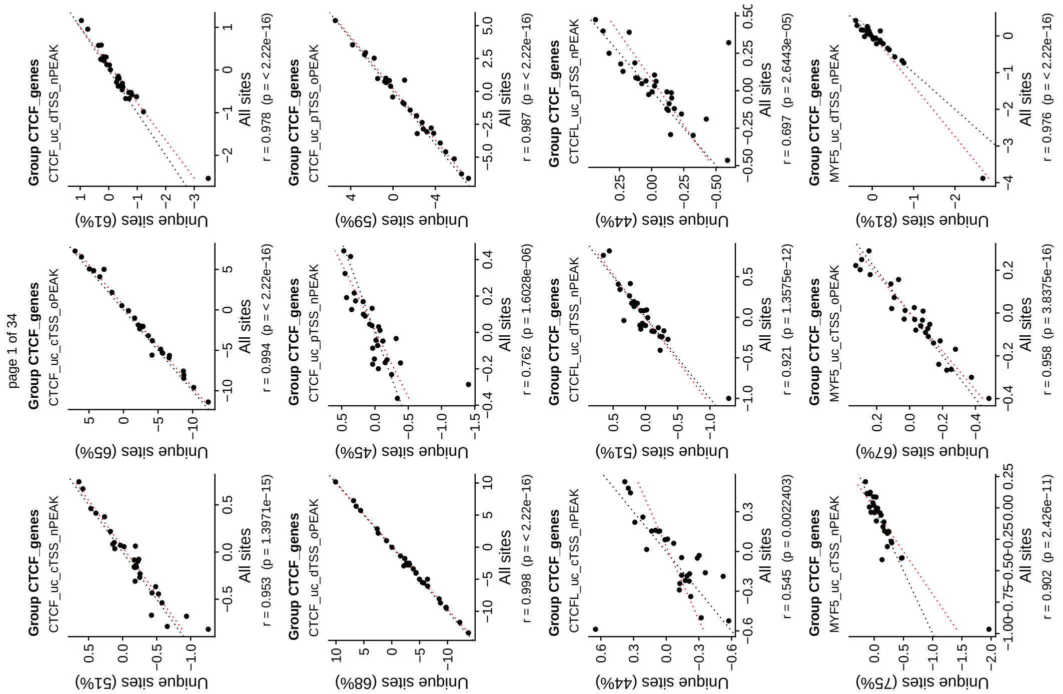
Supplementary Figure S.59: Coefficients from ANOVA of ATAC-seq differences between and within regions.



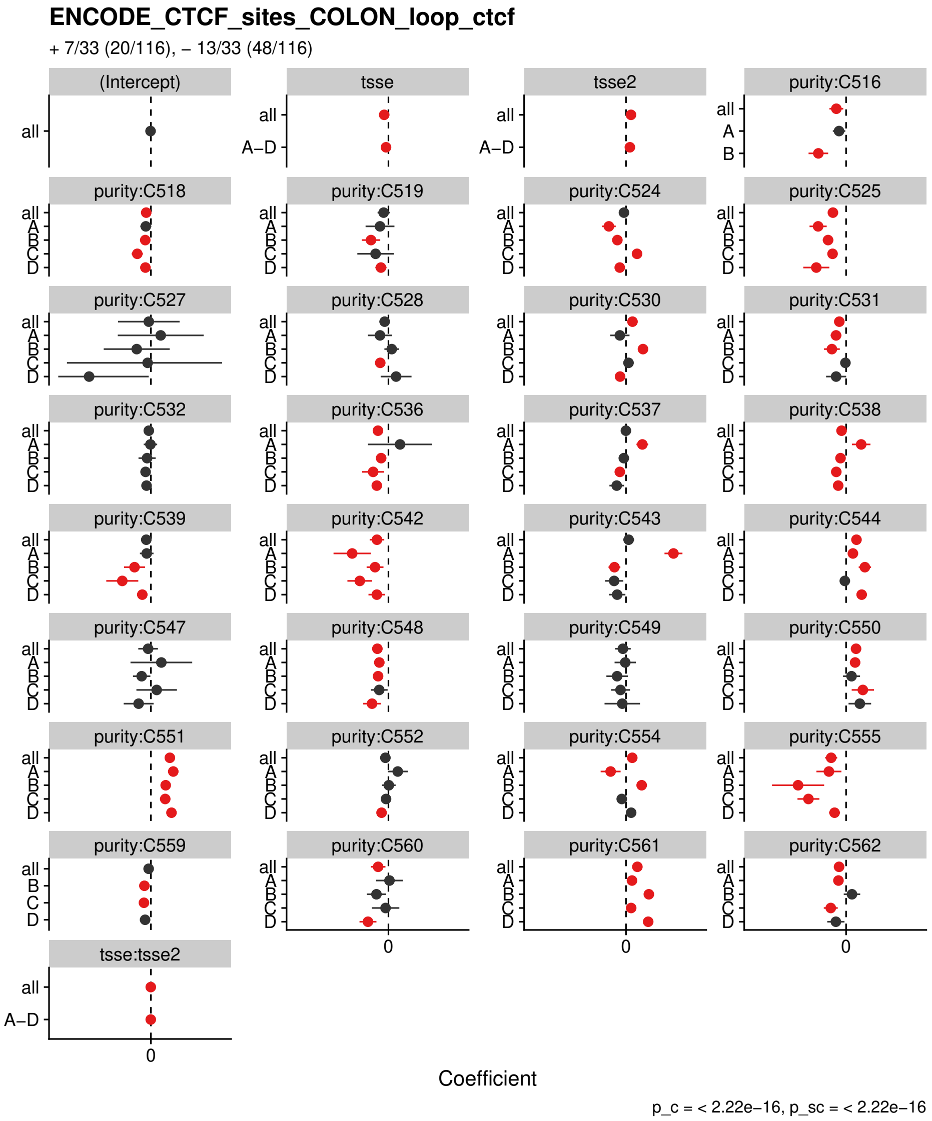
Supplementary Figure S.60: Example of TF binding site overlaps.



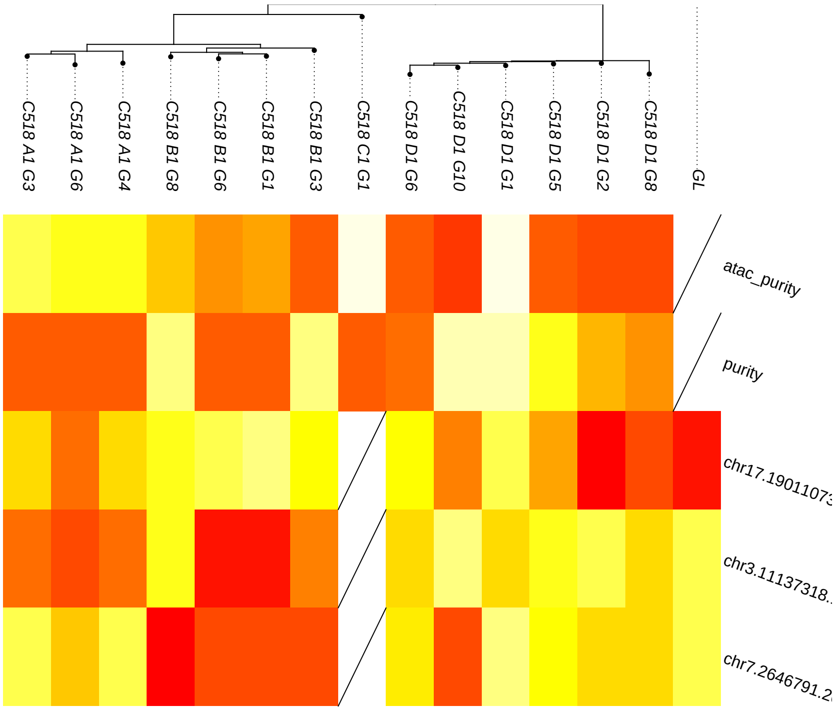
Supplementary Figure S.61: Correlation of genetic and epigenetic distances between samples pairs.



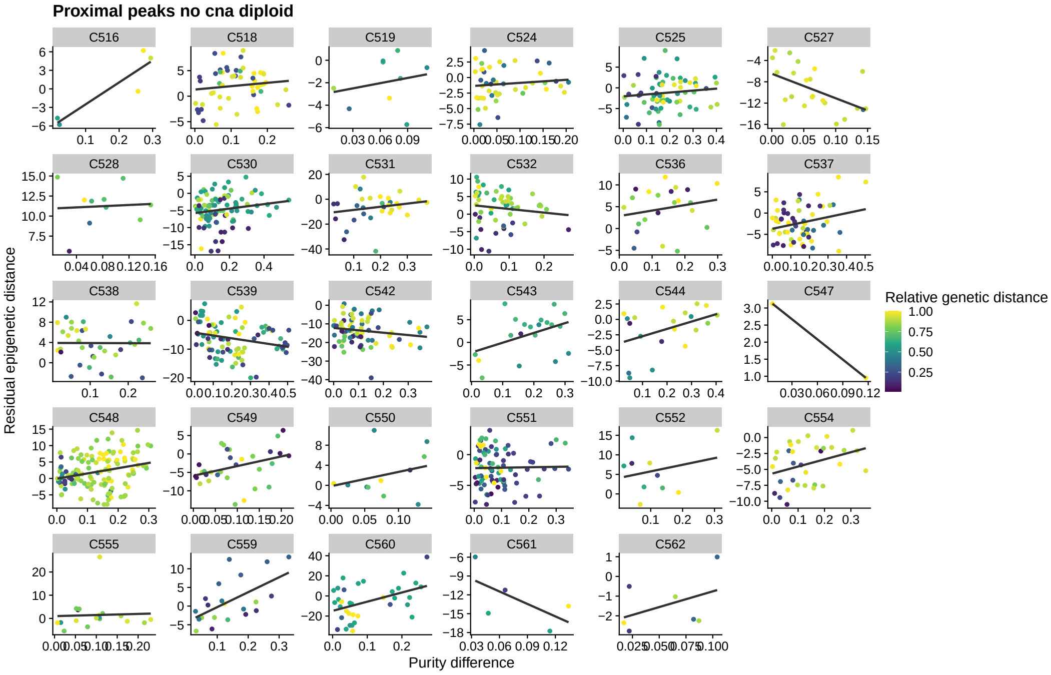
Supplementary Figure S.62: Signal of unique TF binding sites compared to all binding sites for groups with large overlap (see Figure S.60).



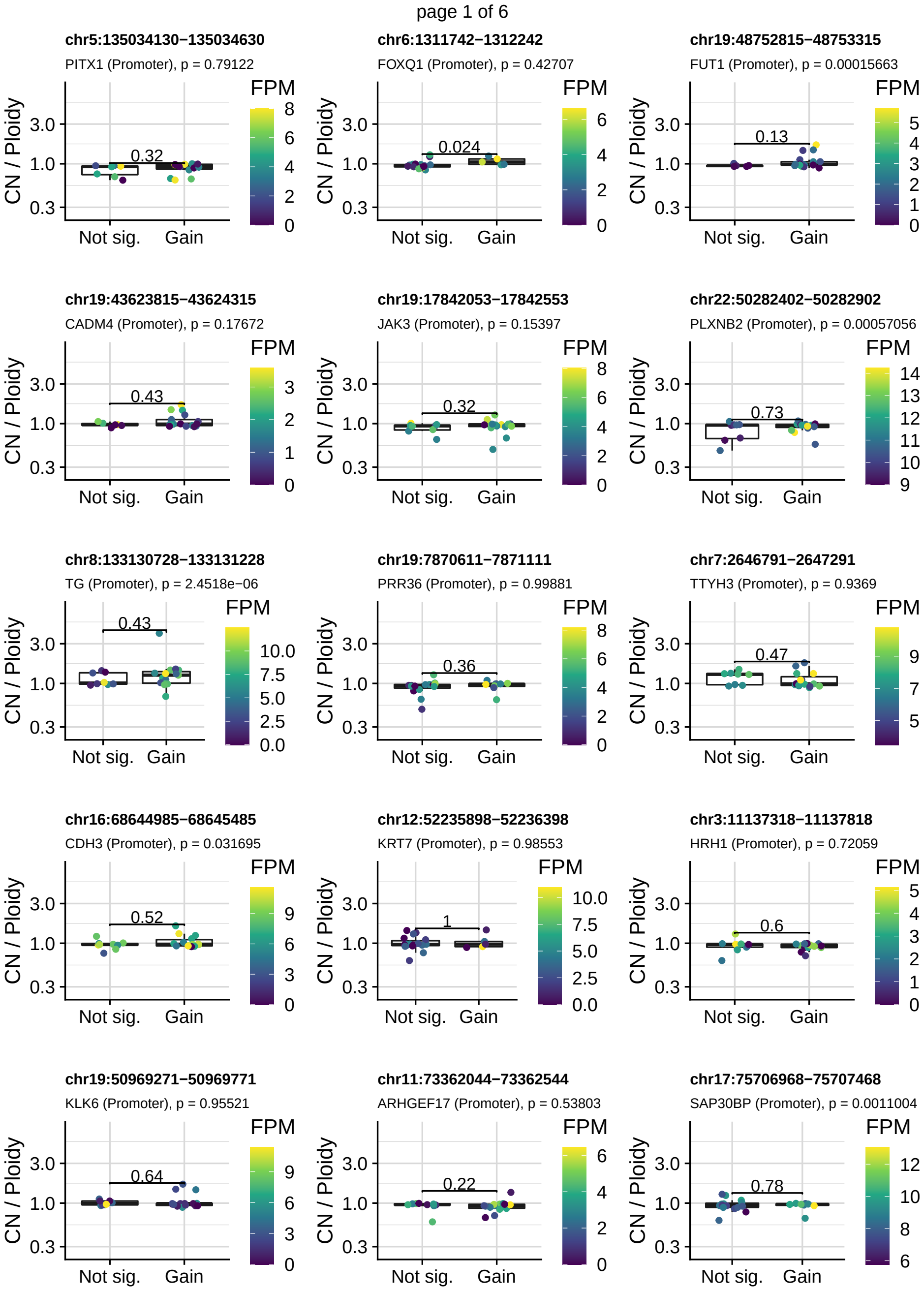
Supplementary Figure S.63: Coefficients of linear regression on TF binding site signals across samples.



Supplementary Figure S.64: Results of phylogenetic regression on individual ATAC-seq sites in C518. Shown are sites of ATAC-seq peaks with significant 'phylogenetic signal' using Pagel (1999) λ including estimates of (Ives, Midford, and Garland 2007). Tests were conducted with methods from *phytools* package for R (Revell 2012). This case is representative of 5/28 cases in which a small number of loci with phylogenetic signal were identified.



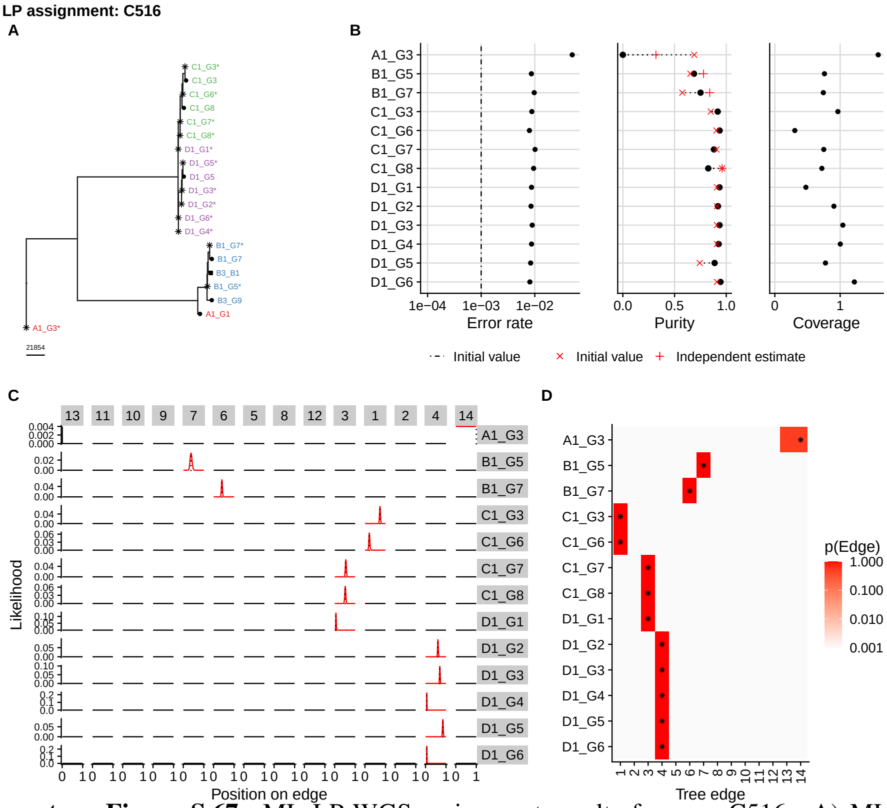
Supplementary Figure S.65: Correlation of residual TF signals with purities differences of samples.



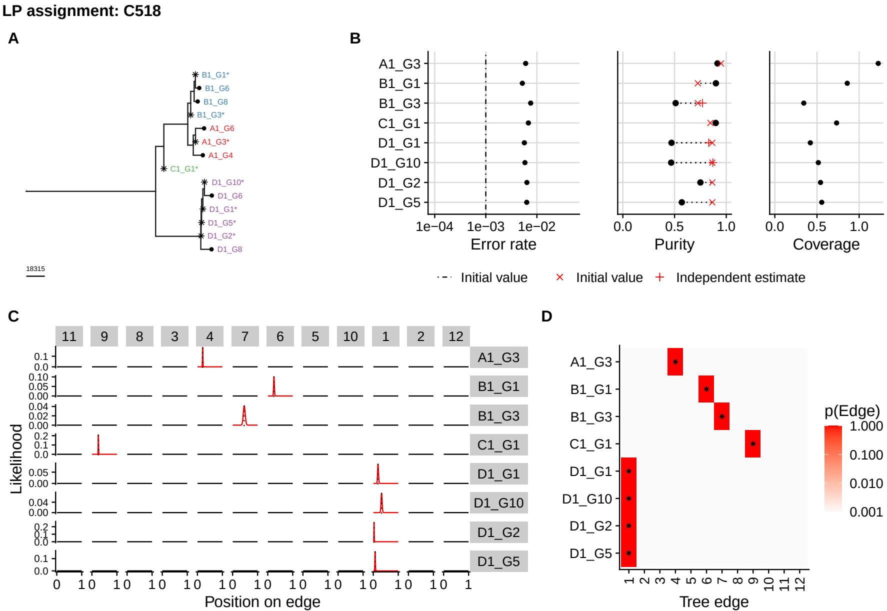
Supplementary Figure S.66: Correlation of CPM values of reads in regions of open chromatin in EPICC (normal & tumour), ENCODE (normal) and TCGA (tumour) datasets.

S.4 LP Assignment to Trees

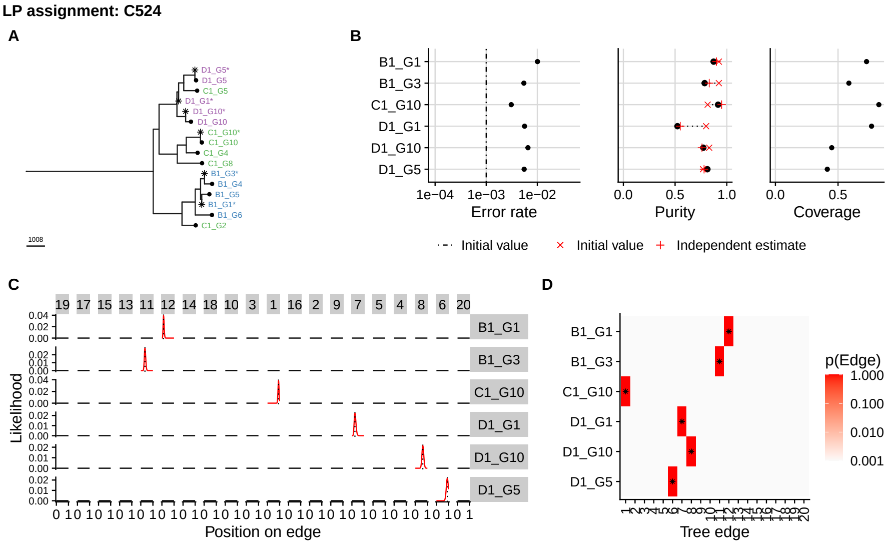
S.4.1 ML LP Trees of the EPICC Cohort



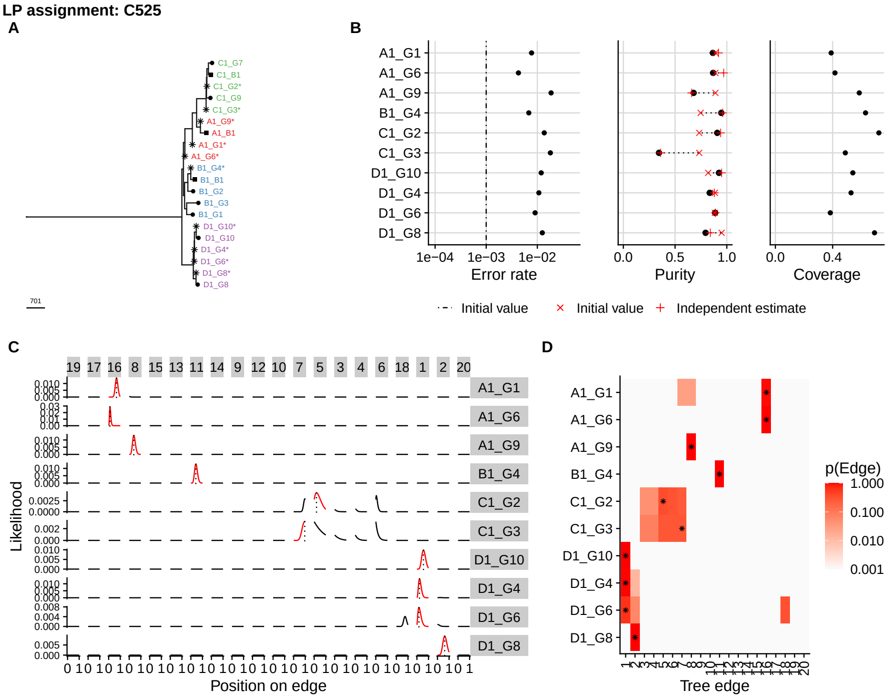
Supplementary Figure S.67: ML LP-WGS assignment results for case C516. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



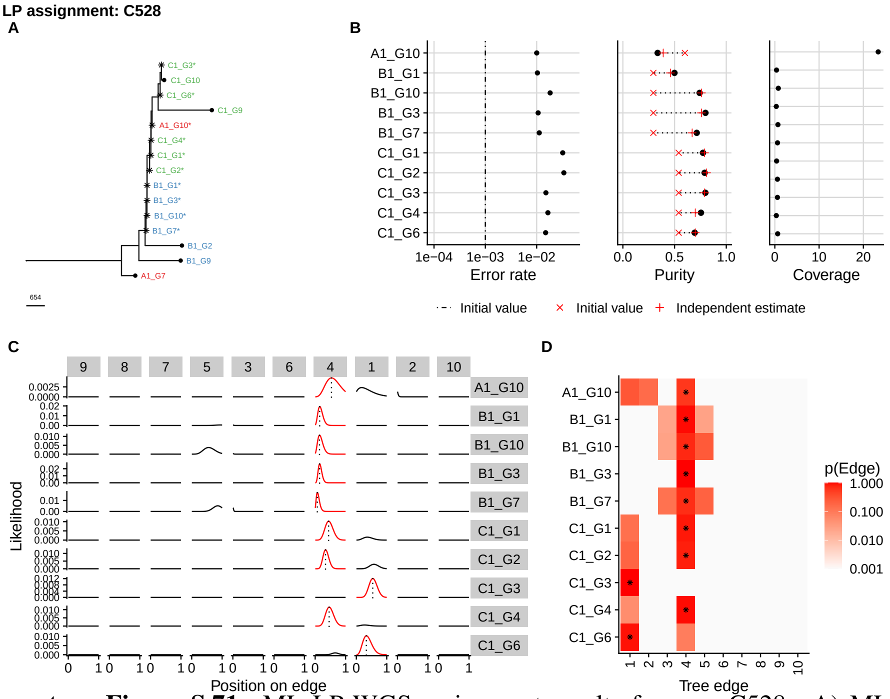
Supplementary Figure S.68: ML LP-WGS assignment results for case C518. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



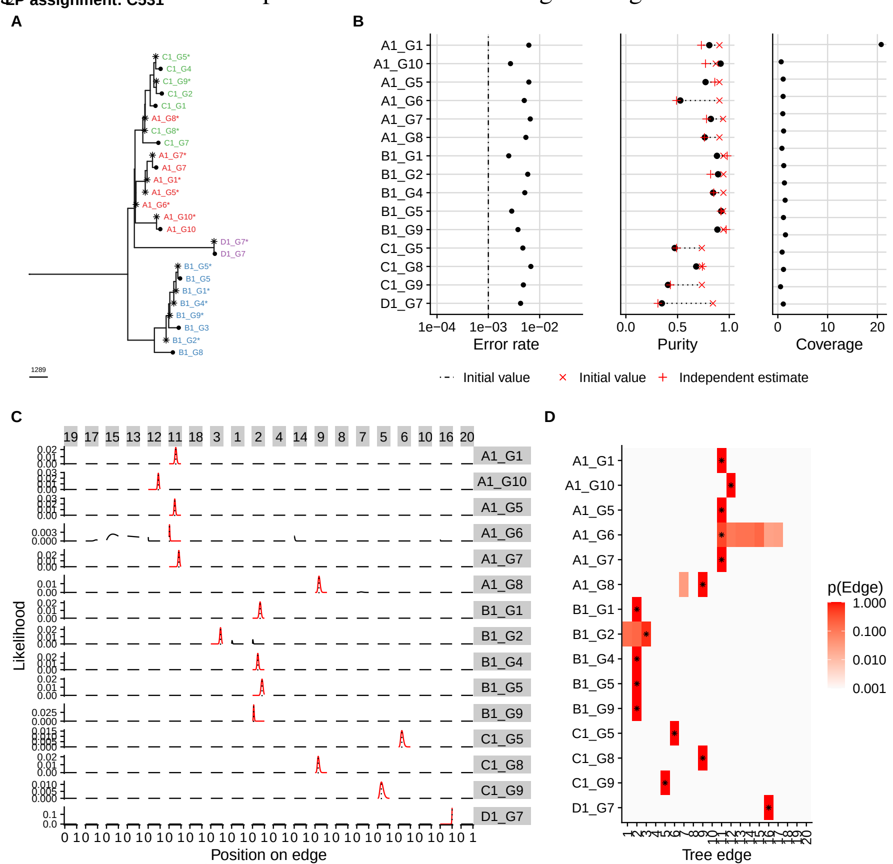
Supplementary Figure S.69: ML LP-WGS assignment results for case C524. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



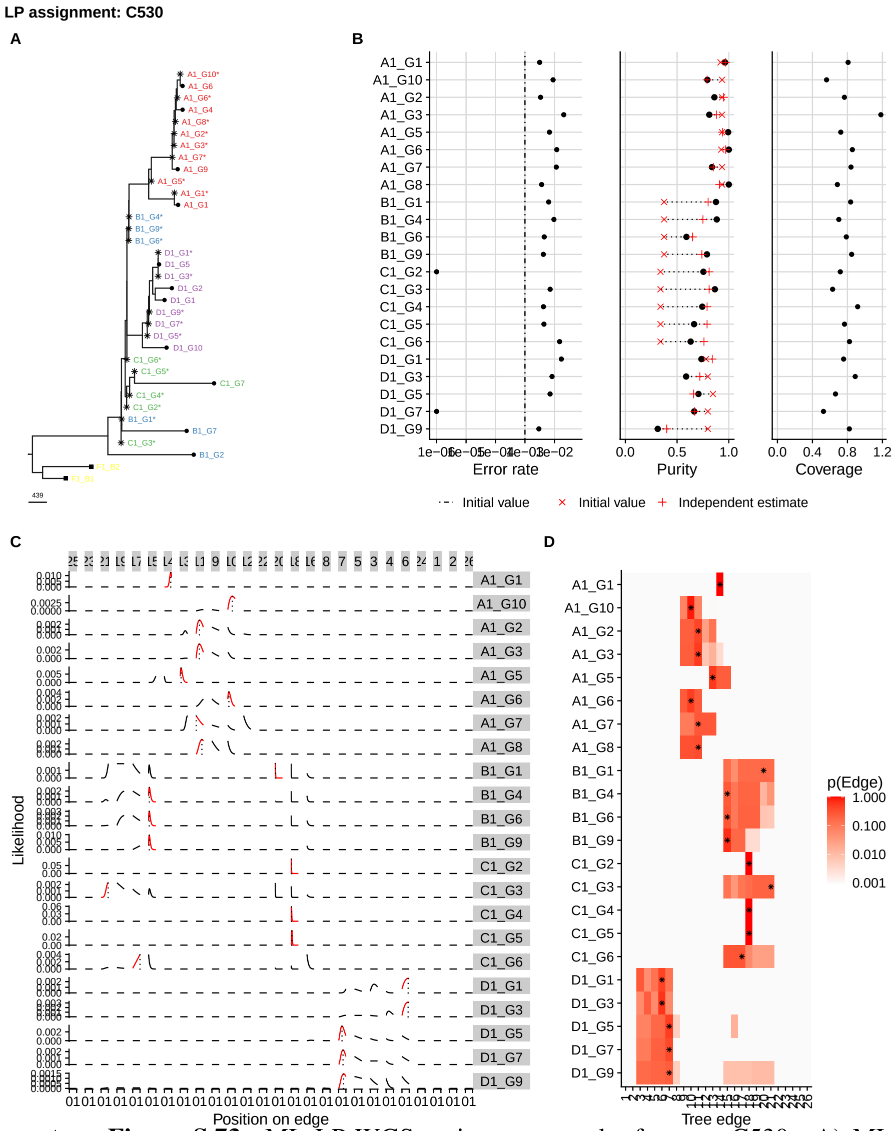
Supplementary Figure S.70: ML LP-WGS assignment results for case C525. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



Supplementary Figure S.71: ML LP-WGS assignment results for case C528. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.

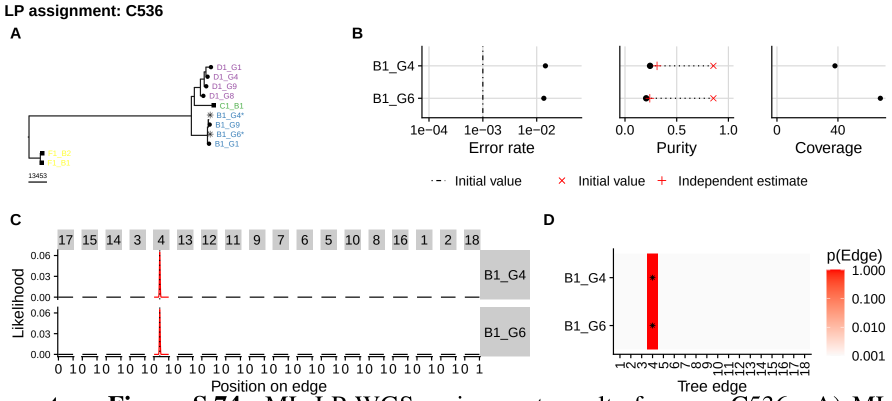


Supplementary Figure S.72: ML LP-WGS assignment results for case C531. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



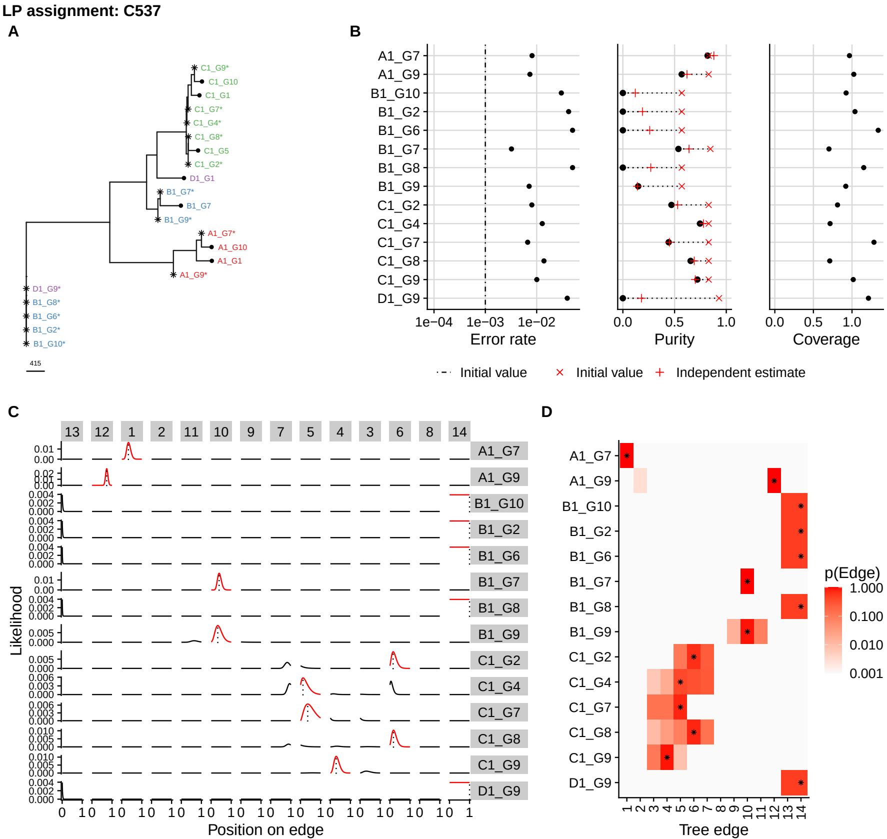
LP assignment: C530

Supplementary Figure S.73: ML LP-WGS assignment results for case C530. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.

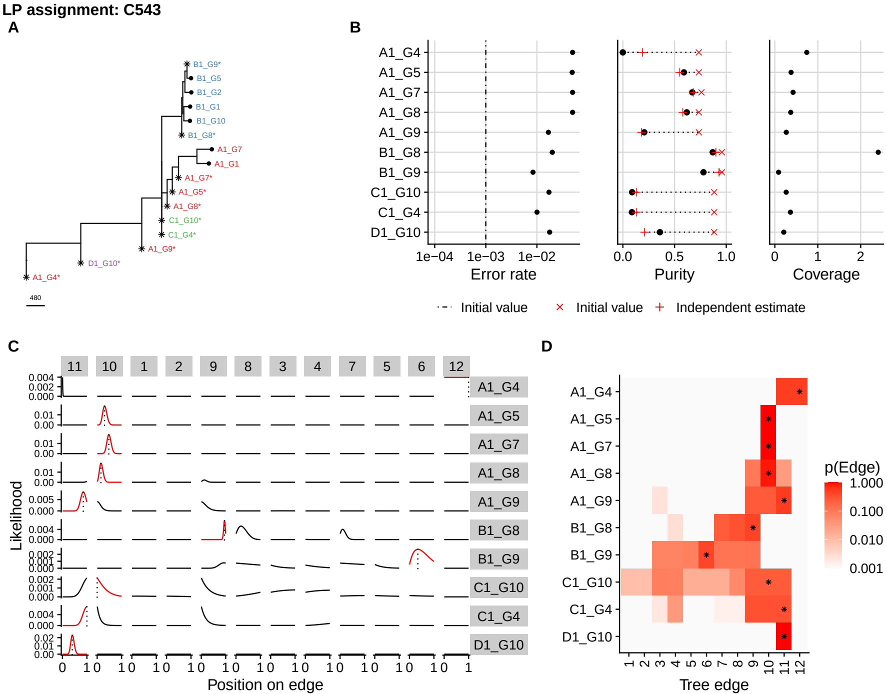


LP assignment: C536

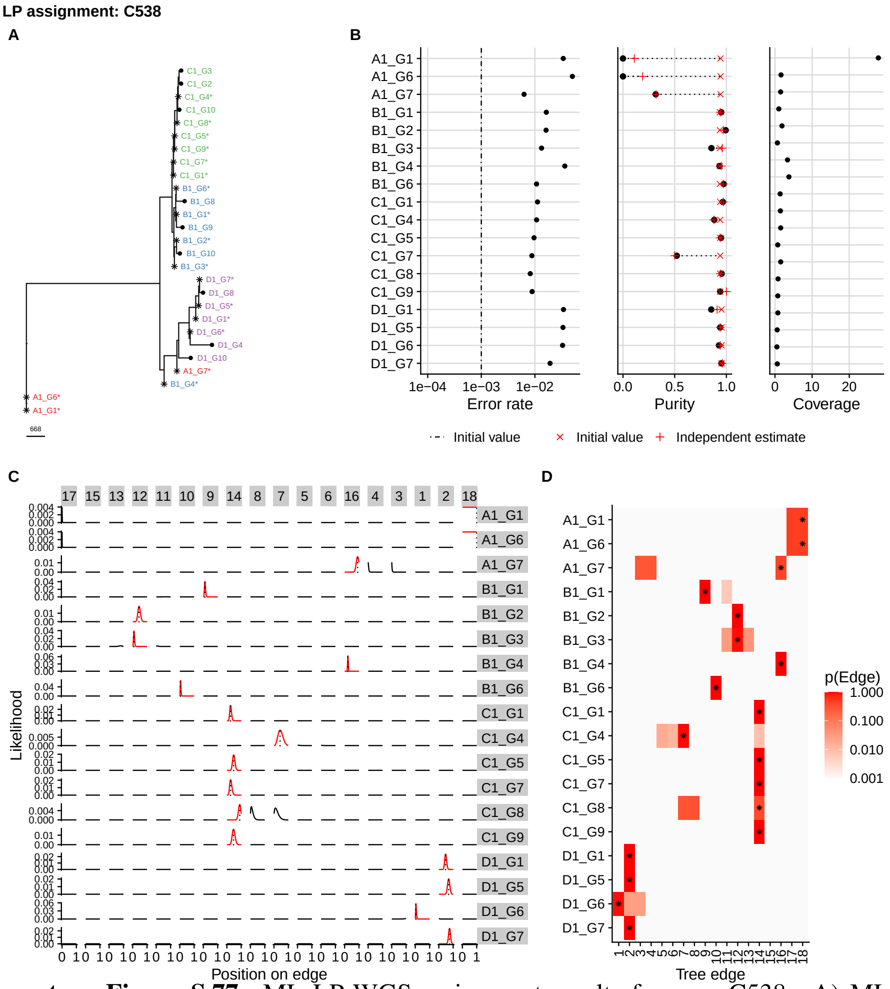
Supplementary Figure S.74: ML LP-WGS assignment results for case C536. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



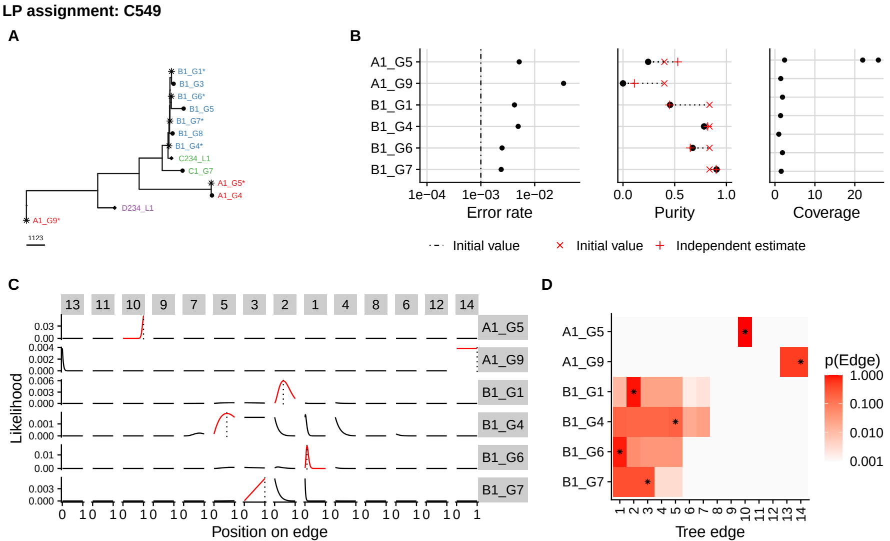
Supplementary Figure S.75: ML LP-WGS assignment results for case C537. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



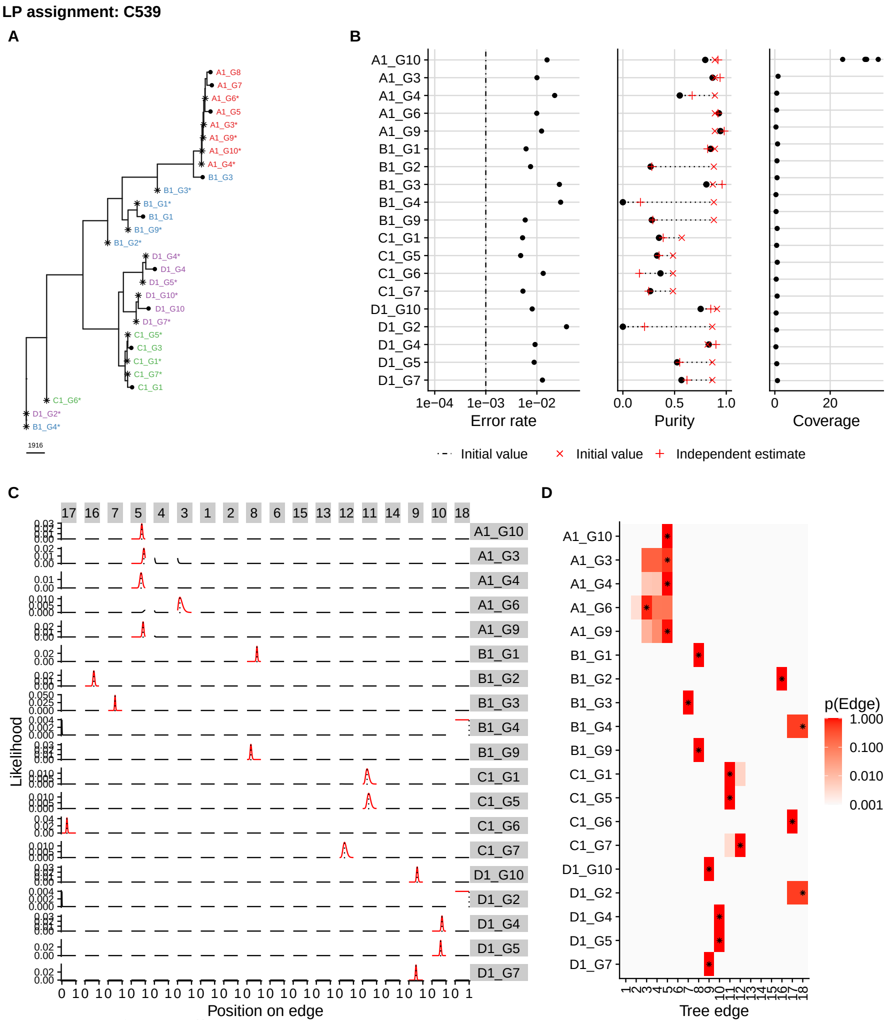
Supplementary Figure S.76: ML LP-WGS assignment results for case C543. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



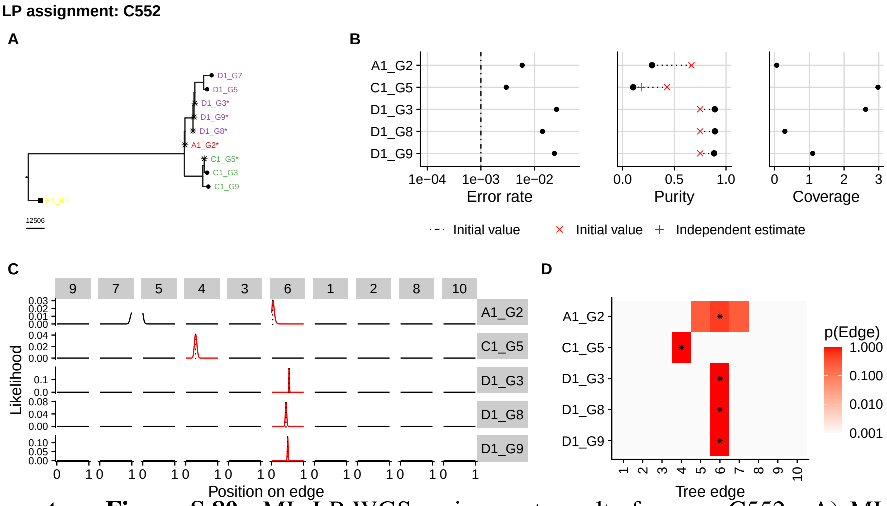
Supplementary Figure S.77: ML LP-WGS assignment results for case C538. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



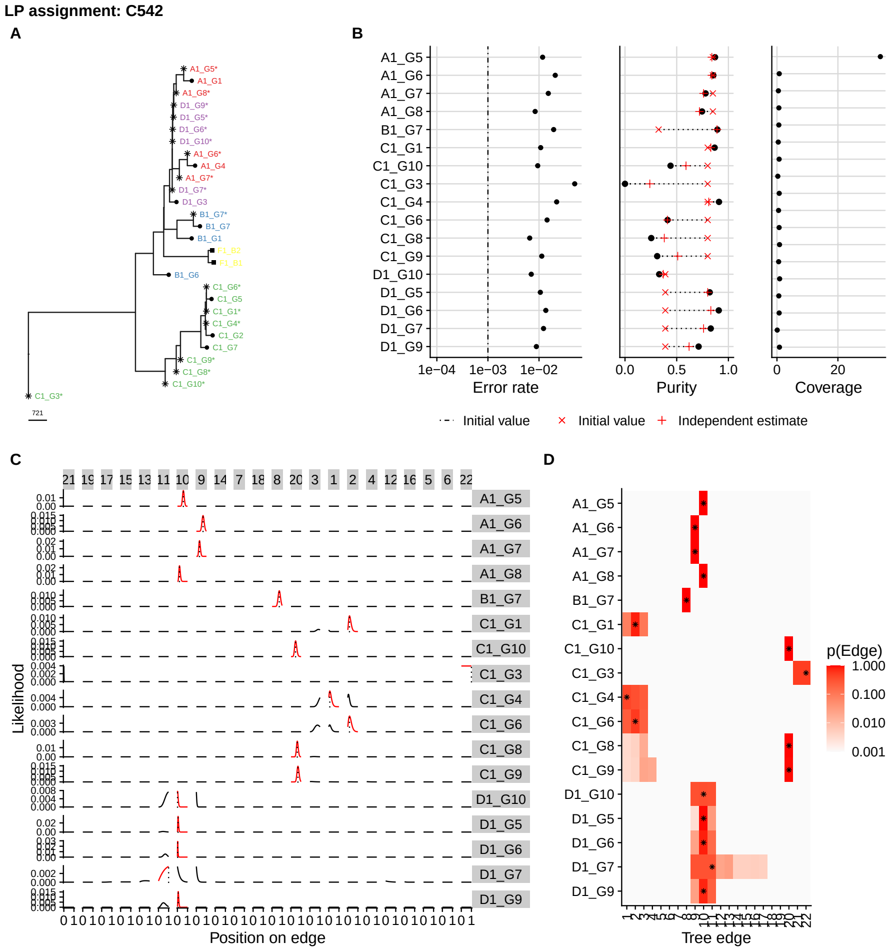
Supplementary Figure S.78: ML LP-WGS assignment results for case C549. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



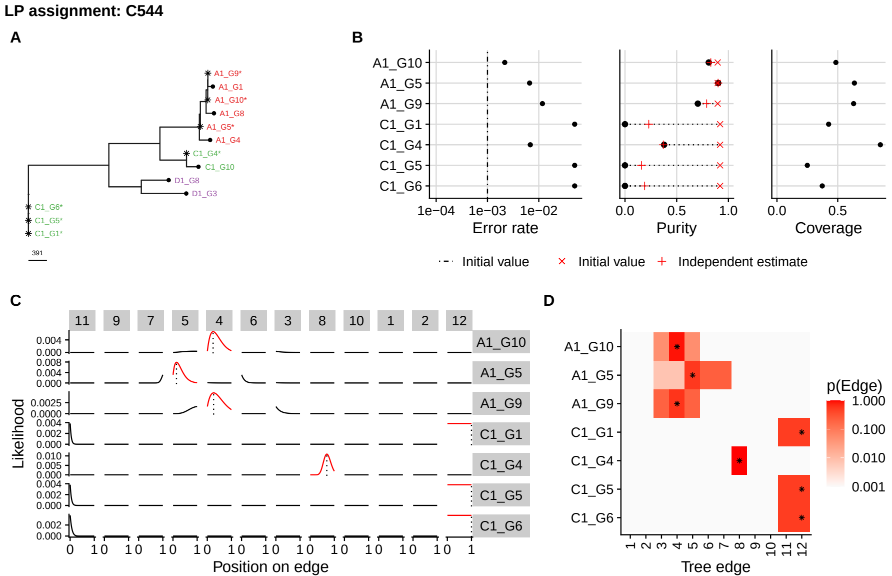
Supplementary Figure S.79: ML LP-WGS assignment results for case C539. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



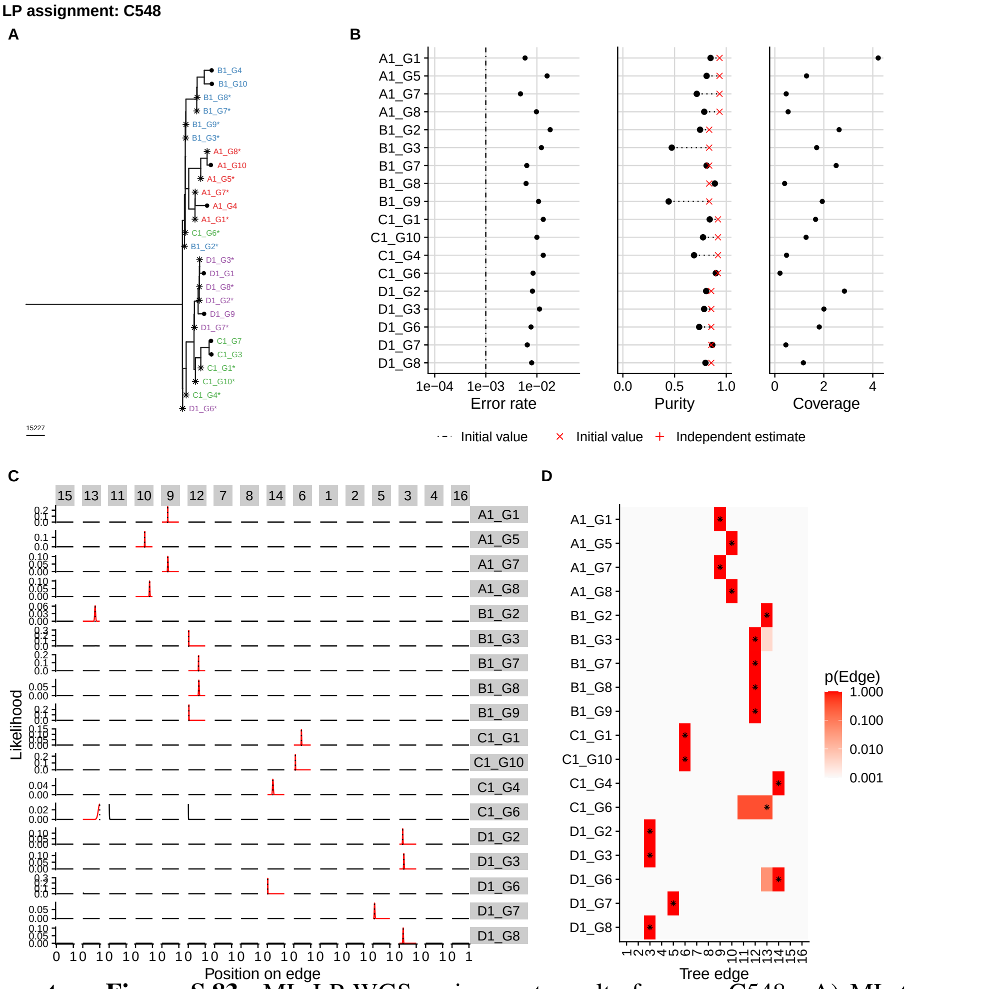
Supplementary Figure S.80: ML LP-WGS assignment results for case C552. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



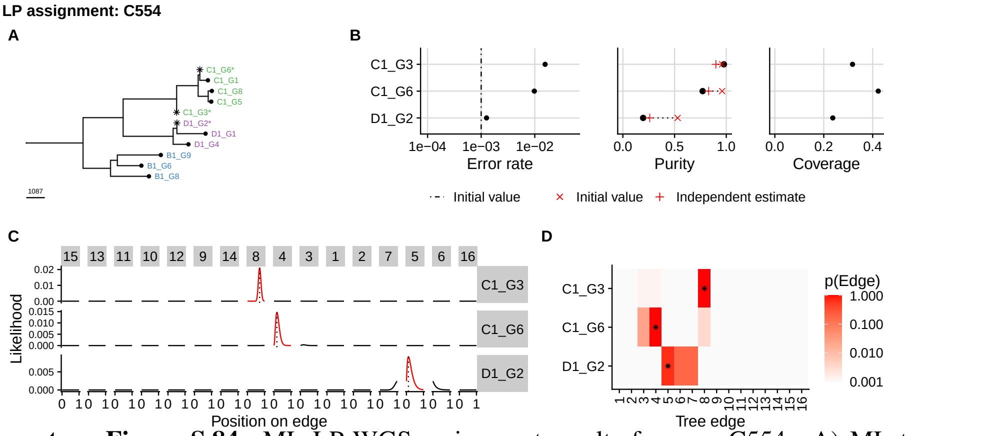
Supplementary Figure S.81: ML LP-WGS assignment results for case C542. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



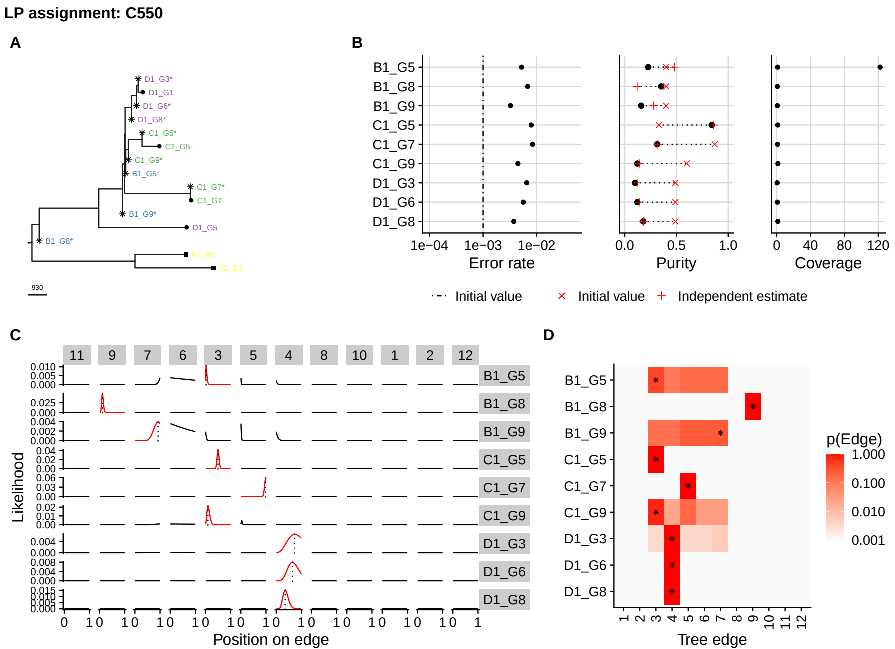
Supplementary Figure S.82: ML LP-WGS assignment results for case C544. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



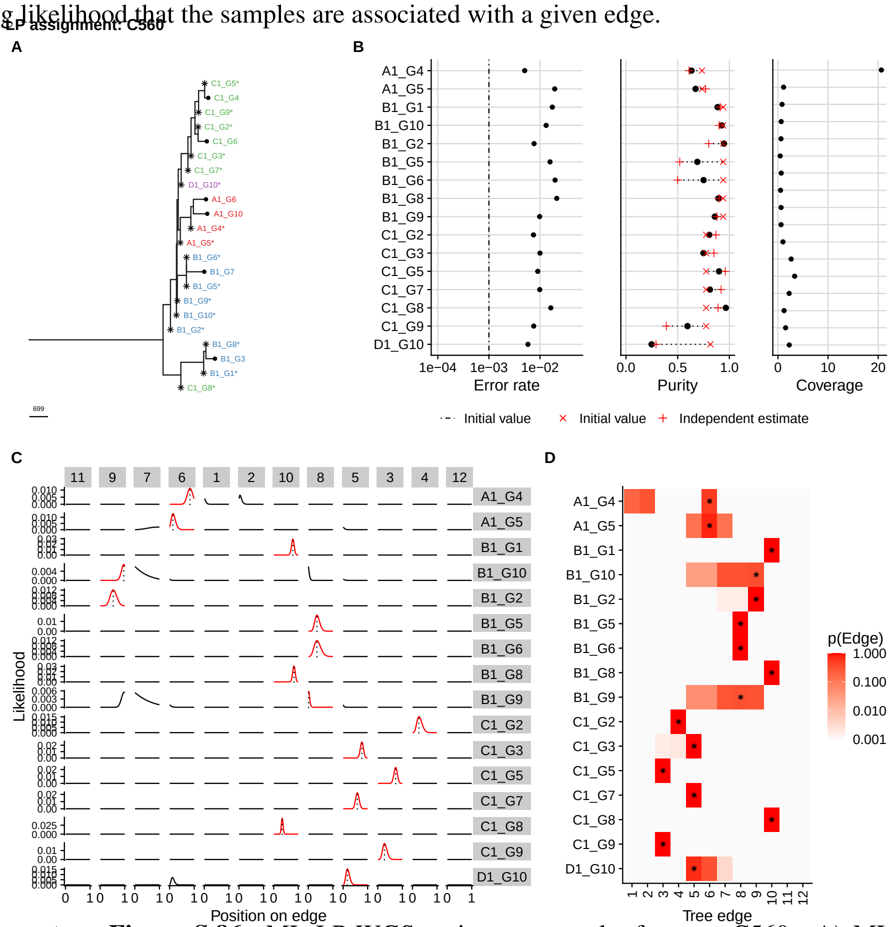
Supplementary Figure S.83: ML LP-WGS assignment results for case C548. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



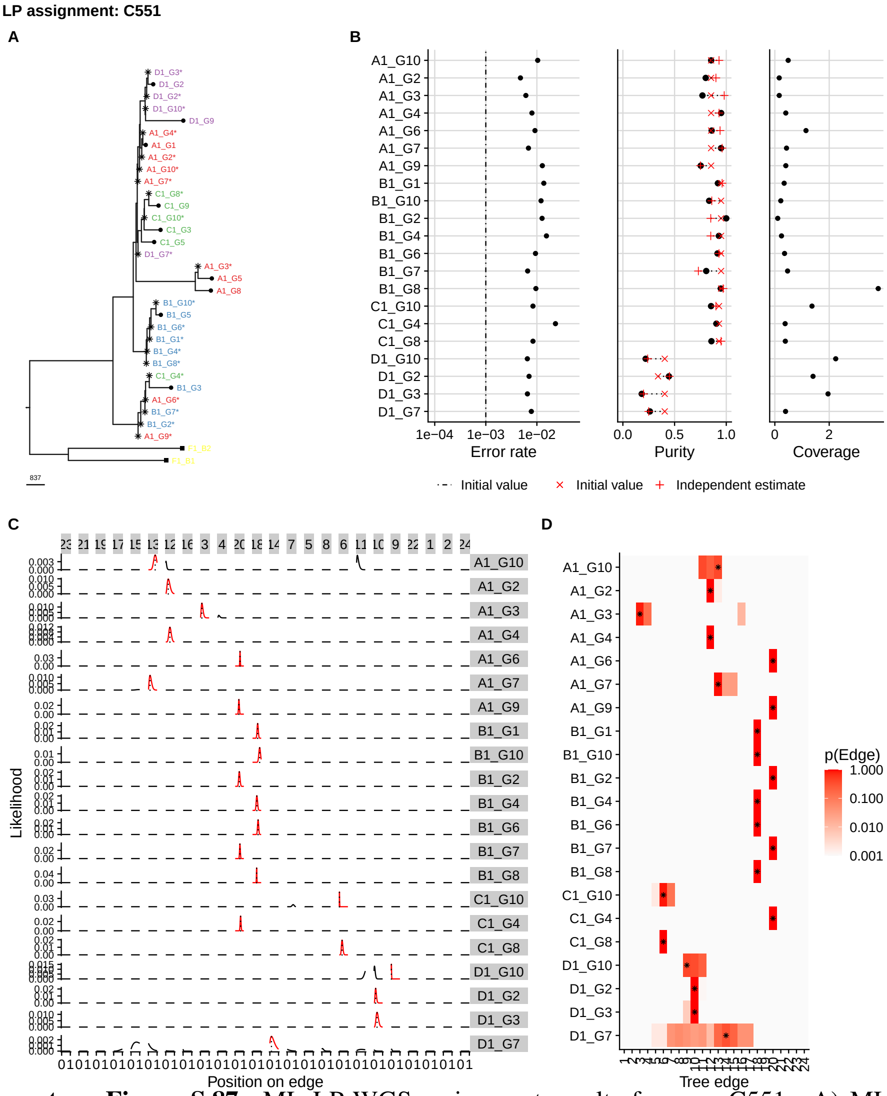
Supplementary Figure S.84: ML LP-WGS assignment results for case C554. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



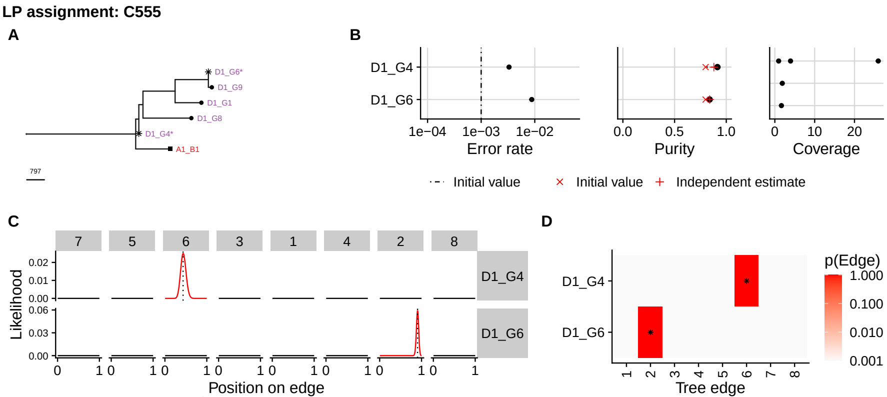
Supplementary Figure S.85: ML LP-WGS assignment results for case C550. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



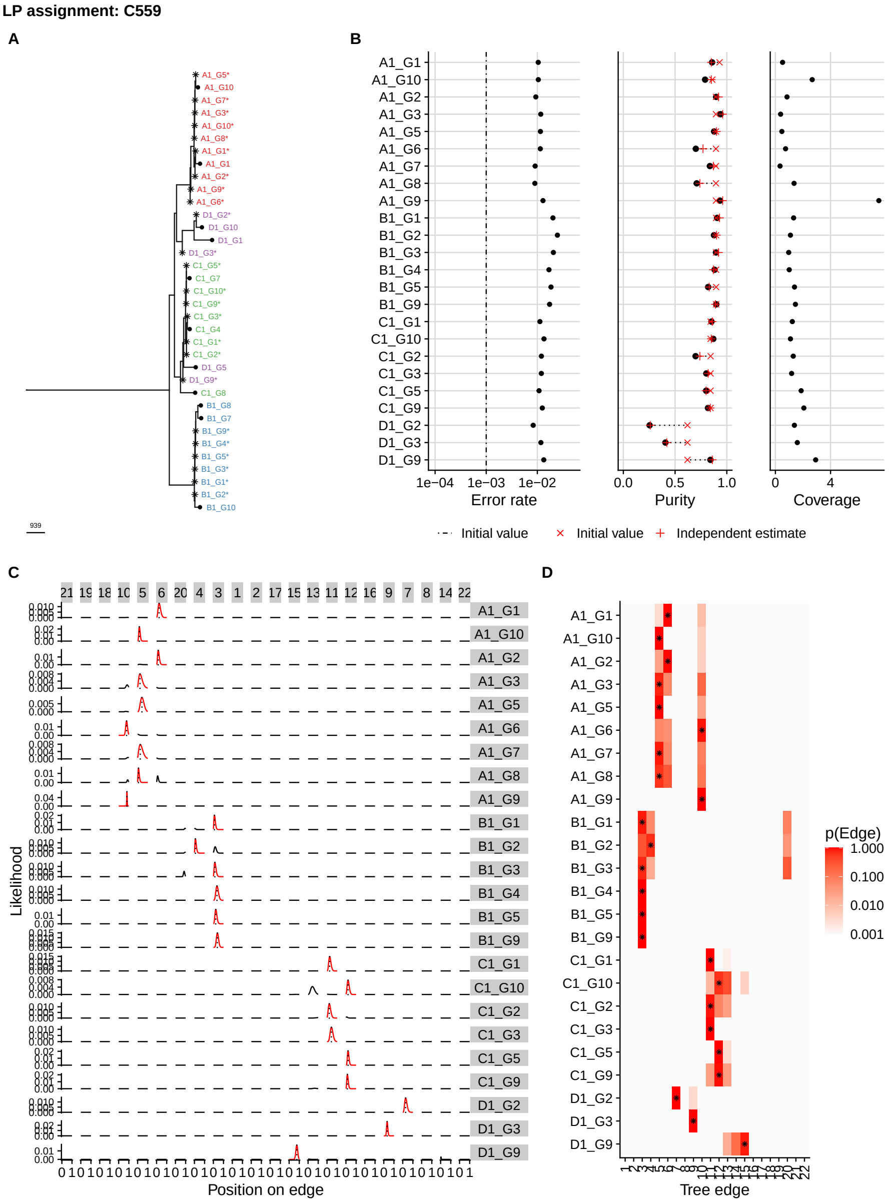
Supplementary Figure S.86: ML LP-WGS assignment results for case C560. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



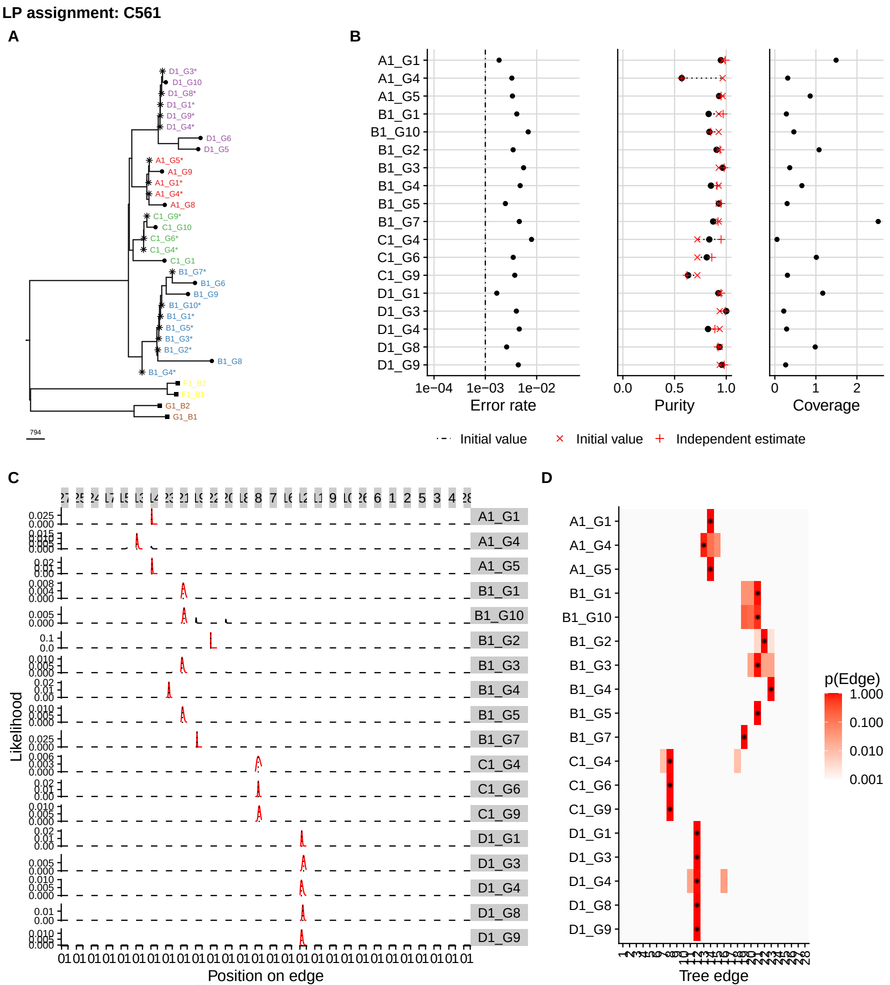
Supplementary Figure S.87: ML LP-WGS assignment results for case C551. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



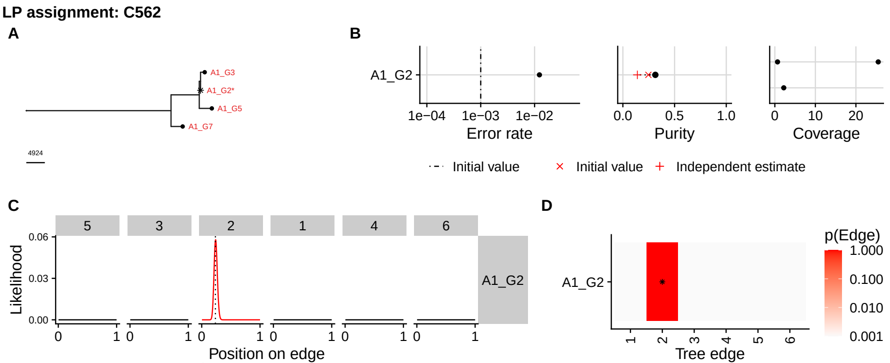
Supplementary Figure S.88: ML LP-WGS assignment results for case C555. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.



Supplementary Figure S.89: ML LP-WGS assignment results for case C559. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.

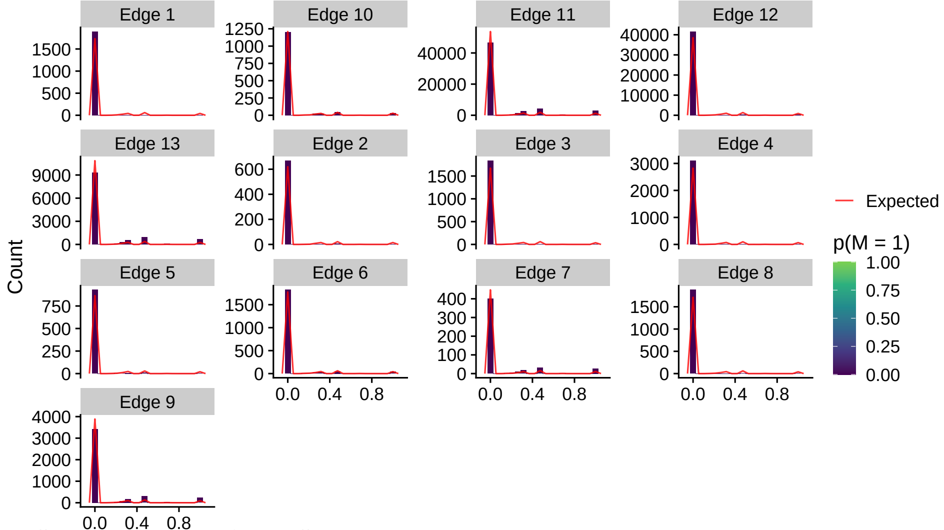


Supplementary Figure S.90: ML LP-WGS assignment results for case C561. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.

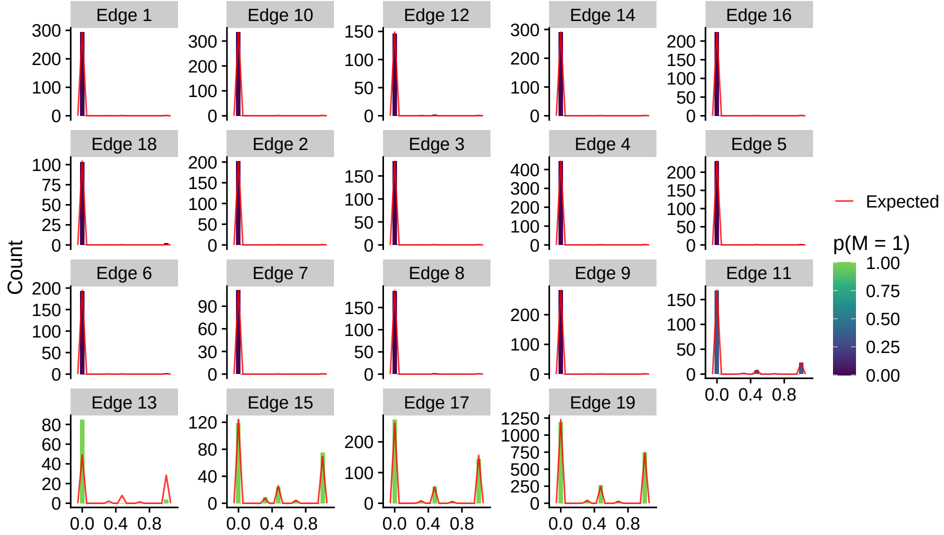


Supplementary Figure S.91: ML LP-WGS assignment results for case C562. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing likelihood that the samples are associated with a given edge.

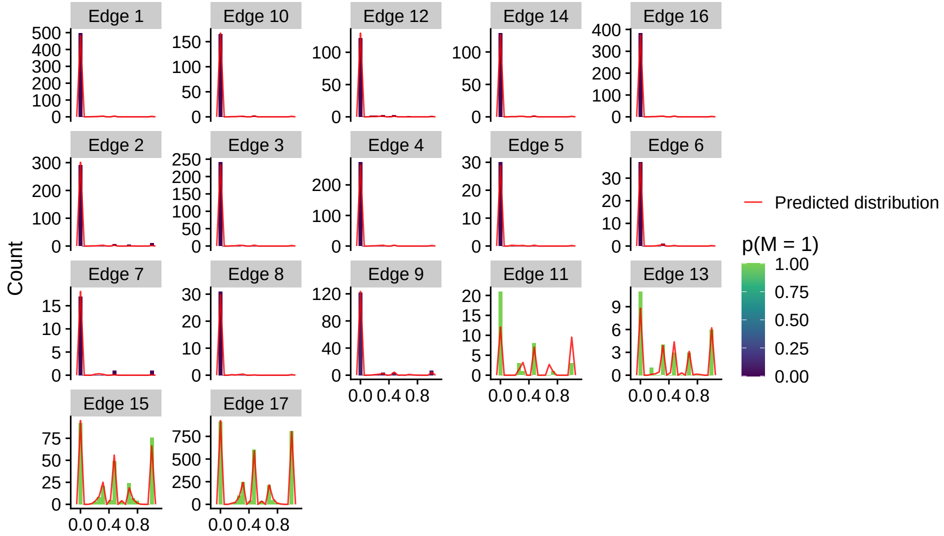
S.4.2 VAF Plots of LP Samples From the EPICC Cohort



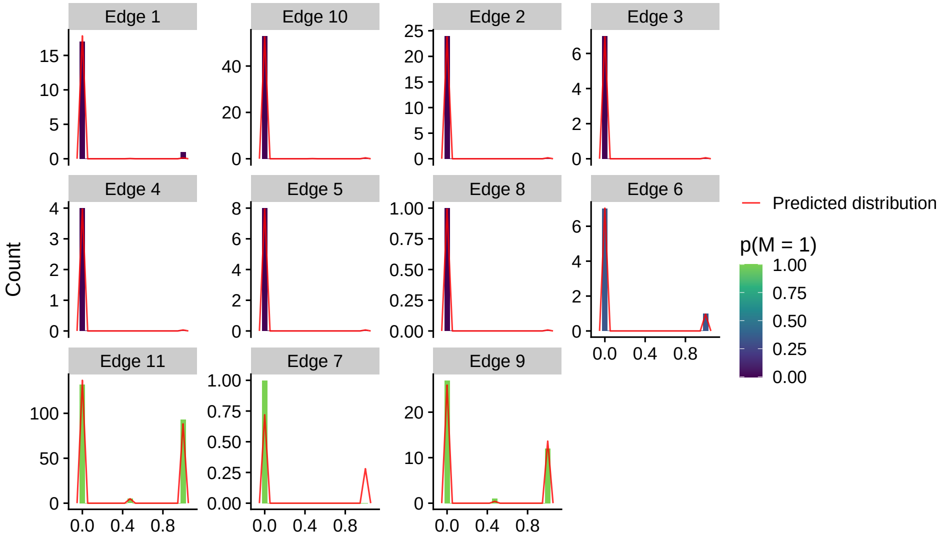
Supplementary Figure S.92: VAF distribution of the LP sample C516_A1_G3_L1.



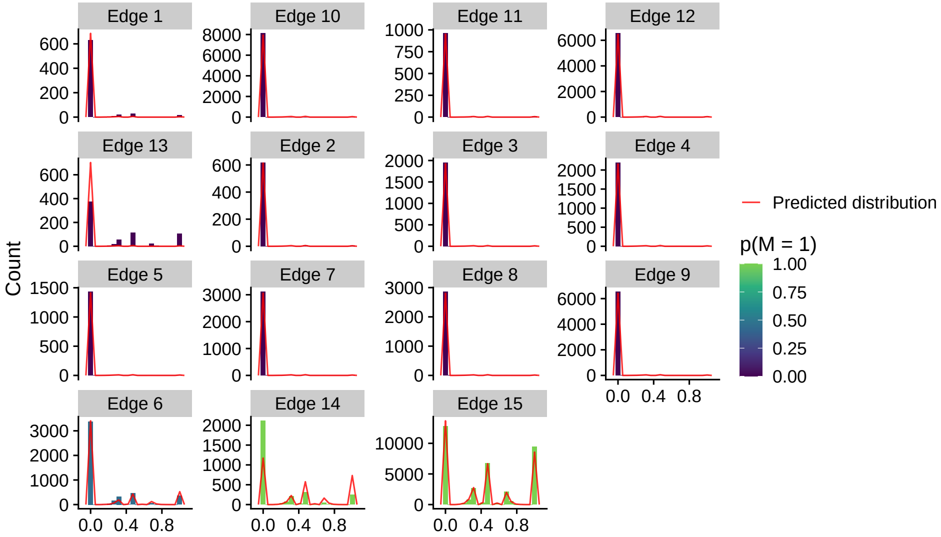
Supplementary Figure S.93: VAF distribution of the LP sample C524_B1_G3_L1.



Supplementary Figure S.94: VAF distribution of the LP sample C538_B1_G1_L1

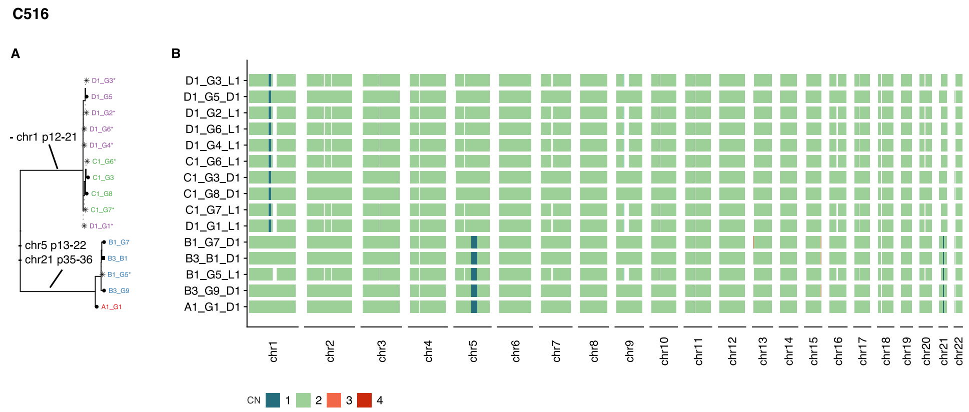


Supplementary Figure S.95: VAF distribution of the LP sample C543_B1_G9_L1.

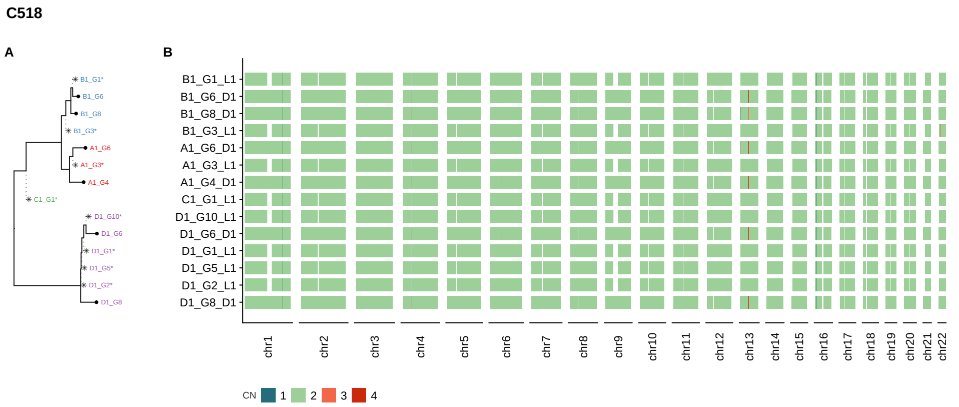


Supplementary Figure S.96: VAF distribution of the LP sample C548_C1_G1_L1

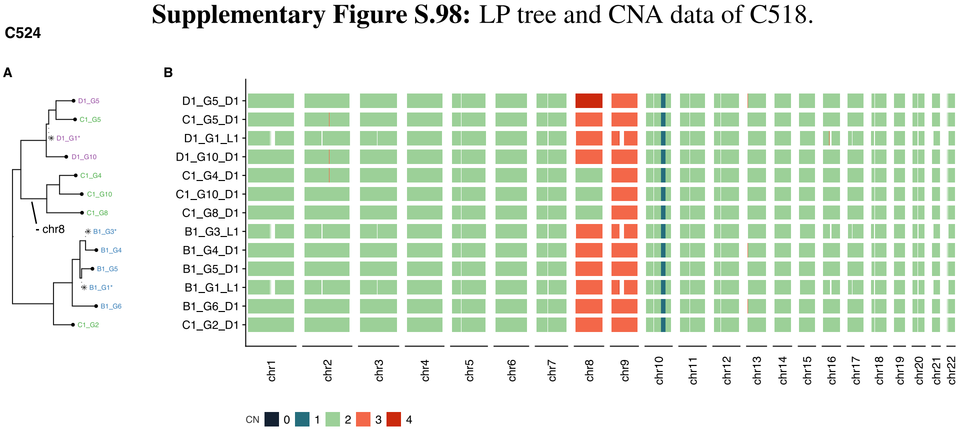
S.4.3 CN Plots of LP Samples From the EPICC Cohort



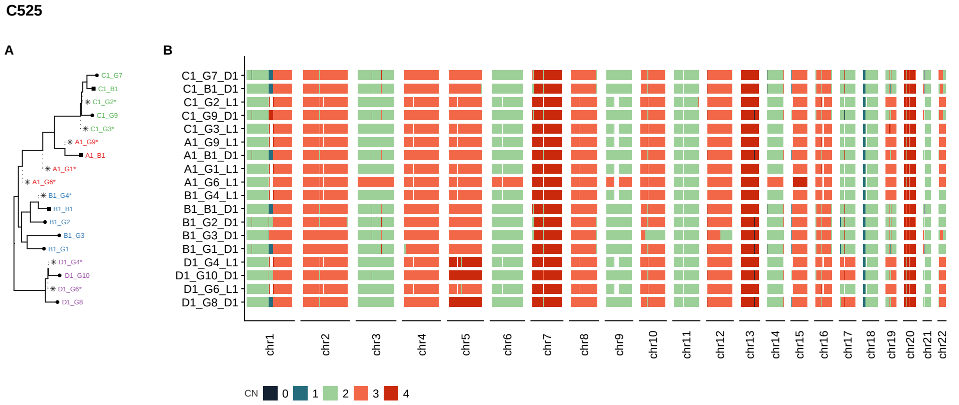
Supplementary Figure S.97: LP tree and CNA data of C516.



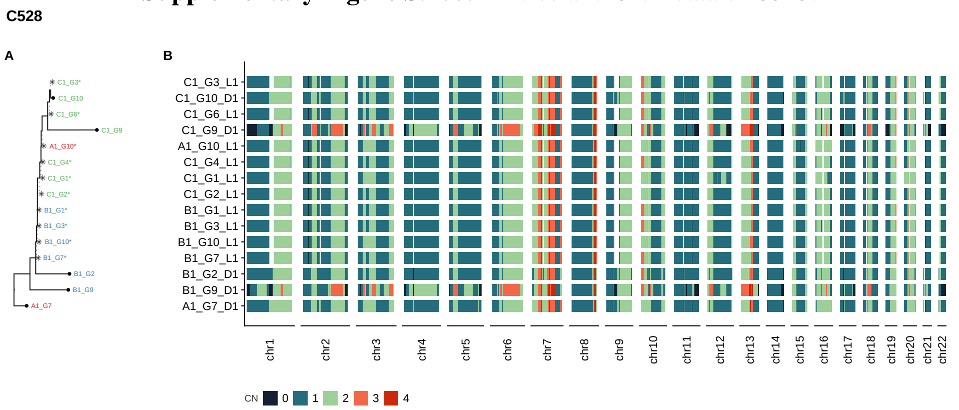
Supplementary Figure S.98: LP tree and CNA data of C518.



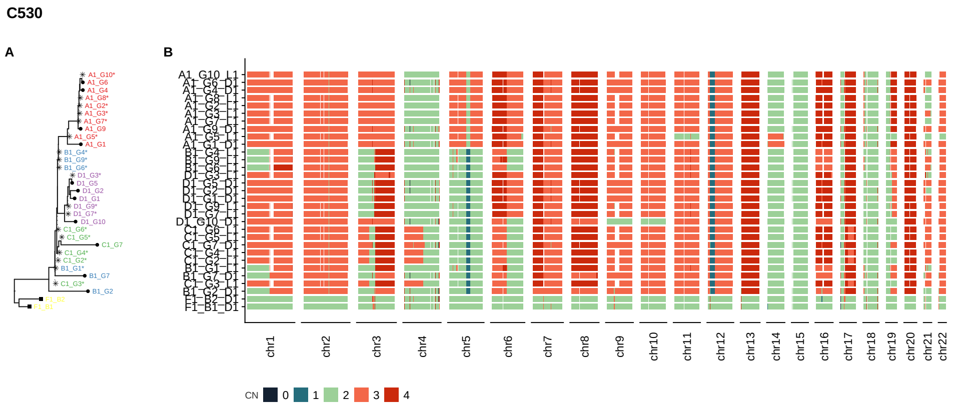
Supplementary Figure S.99: LP tree and CNA data of C524.



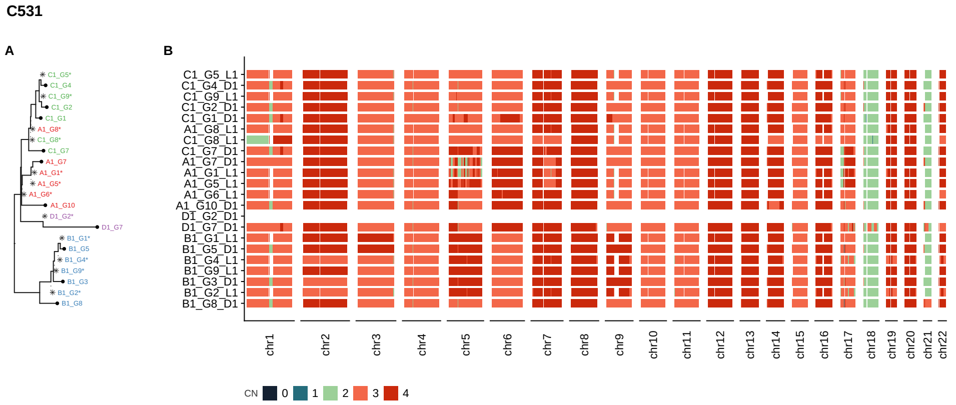
Supplementary Figure S.100: LP tree and CNA data of C525.



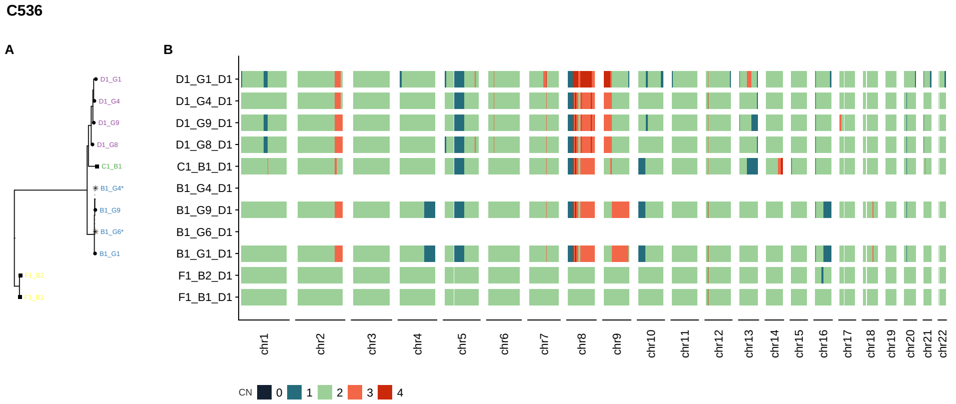
Supplementary Figure S.101: LP tree and CNA data of C528.



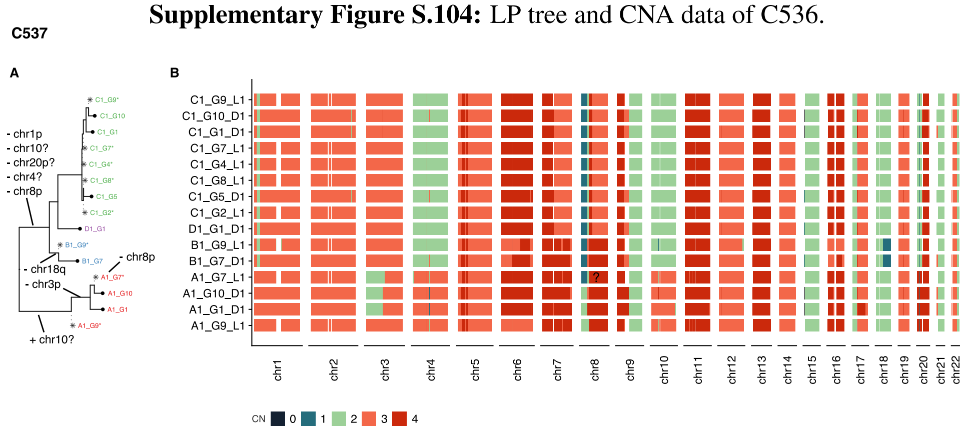
Supplementary Figure S.102: LP tree and CNA data of C530.



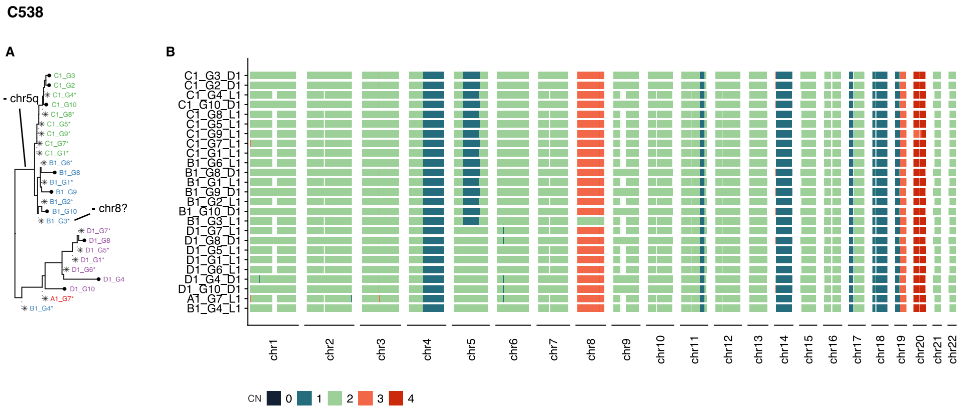
Supplementary Figure S.103: LP tree and CNA data of C531.



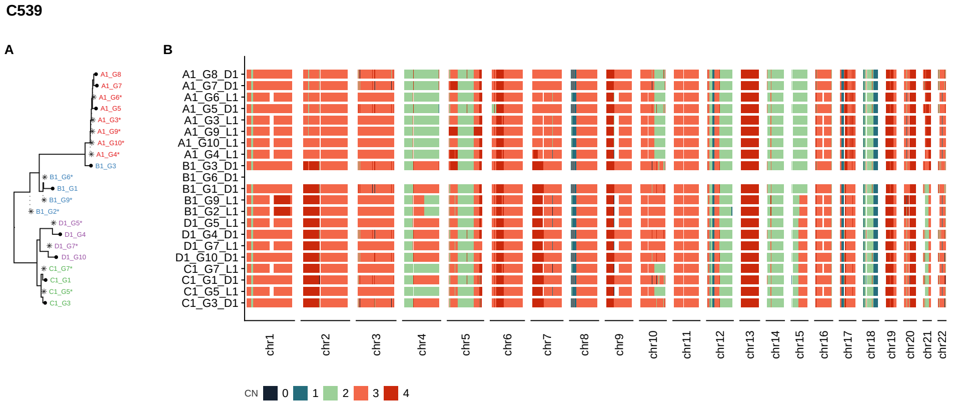
Supplementary Figure S.104: LP tree and CNA data of C536.



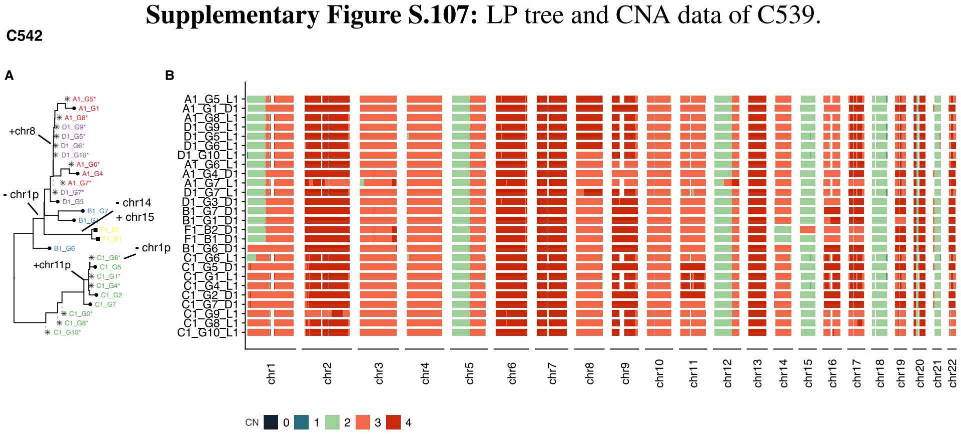
Supplementary Figure S.105: LP tree and CNA data of C537.



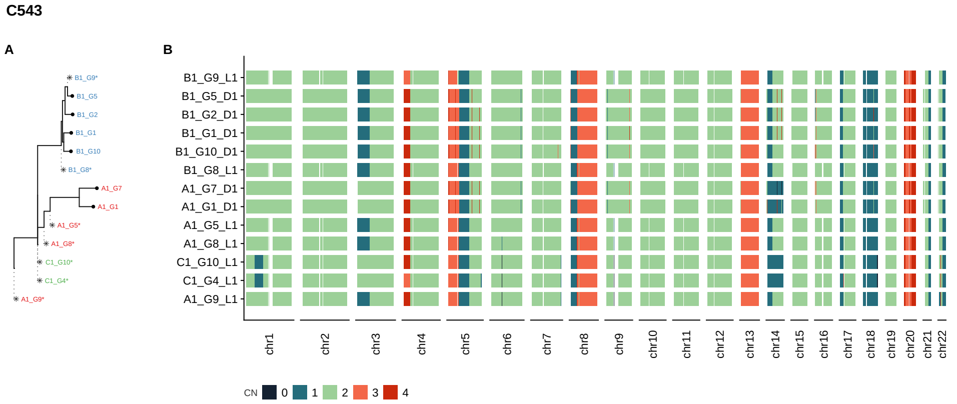
Supplementary Figure S.106: LP tree and CNA data of C538.



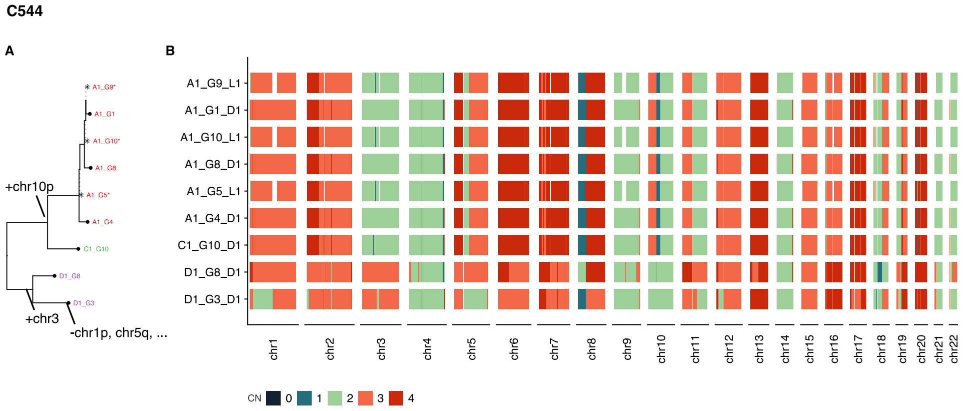
Supplementary Figure S.107: LP tree and CNA data of C539.



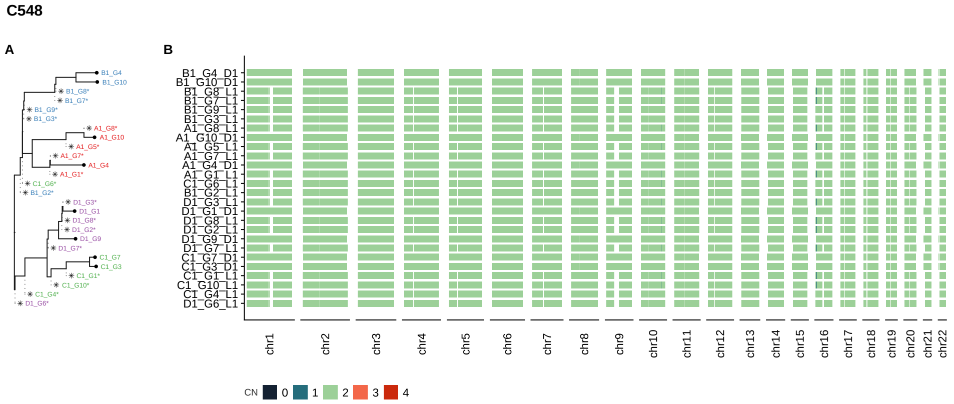
Supplementary Figure S.108: LP tree and CNA data of C542.



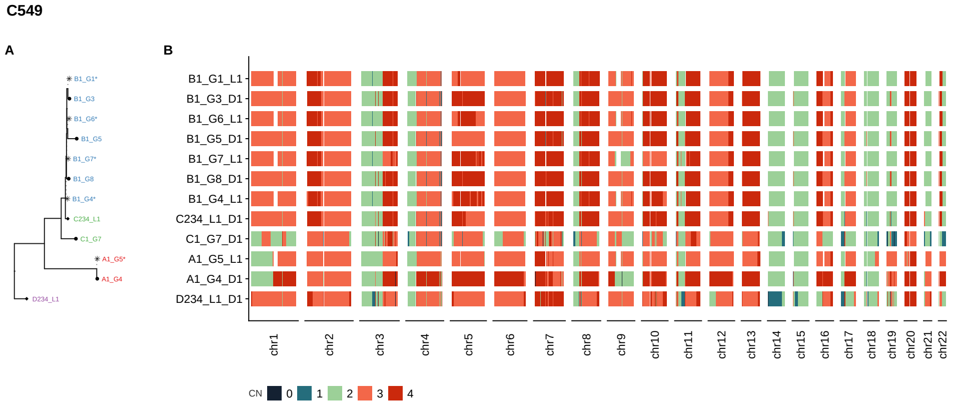
Supplementary Figure S.109: LP tree and CNA data of C543.



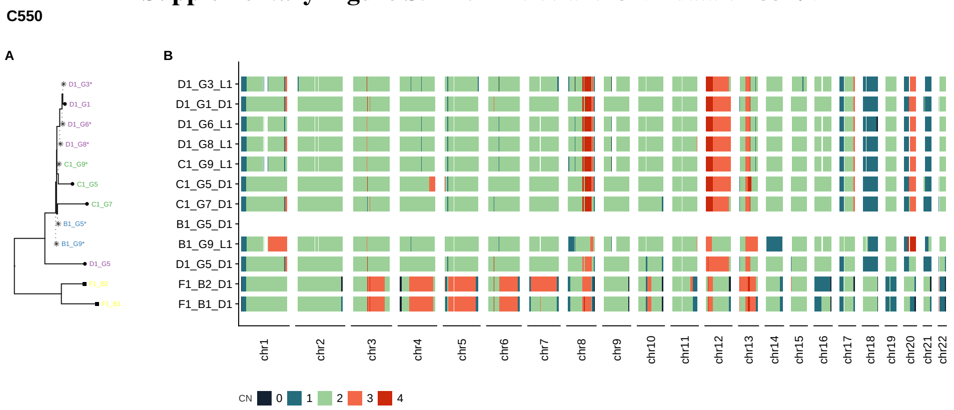
Supplementary Figure S.110: LP tree and CNA data of C544.



Supplementary Figure S.111: LP tree and CNA data of C548.

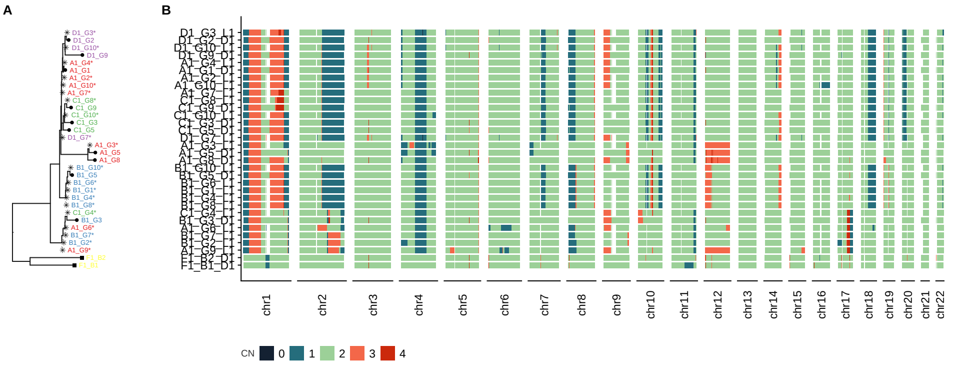


Supplementary Figure S.112: LP tree and CNA data of C549.



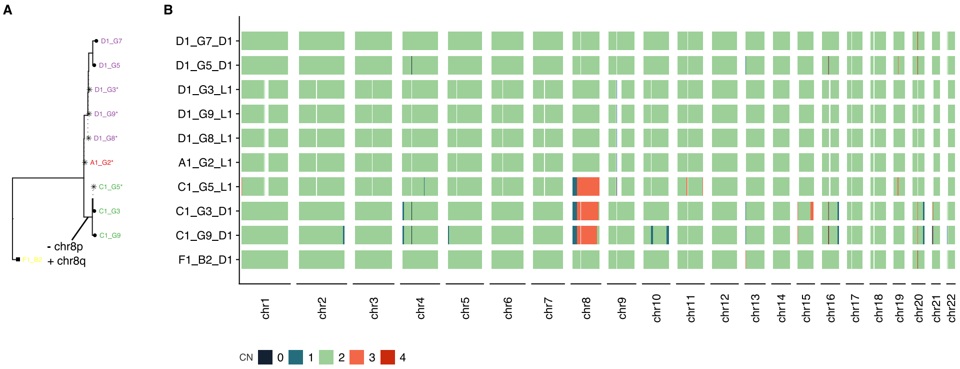
Supplementary Figure S.113: LP tree and CNA data of C550.

C551



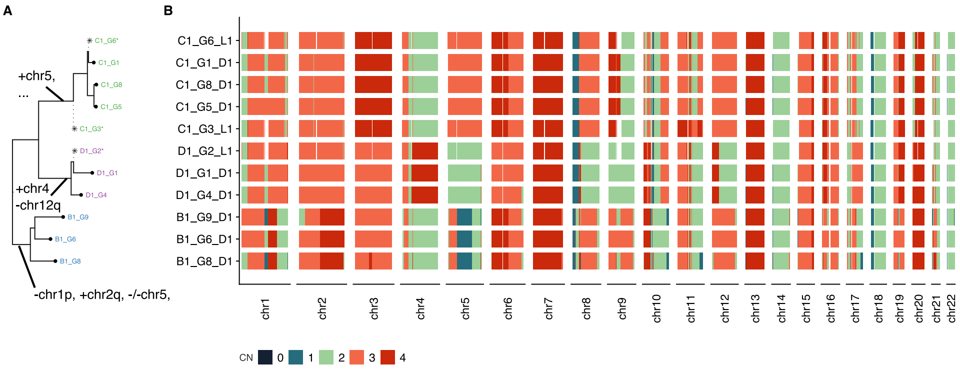
Supplementary Figure S.114: LP tree and CNA data of C551.

C552



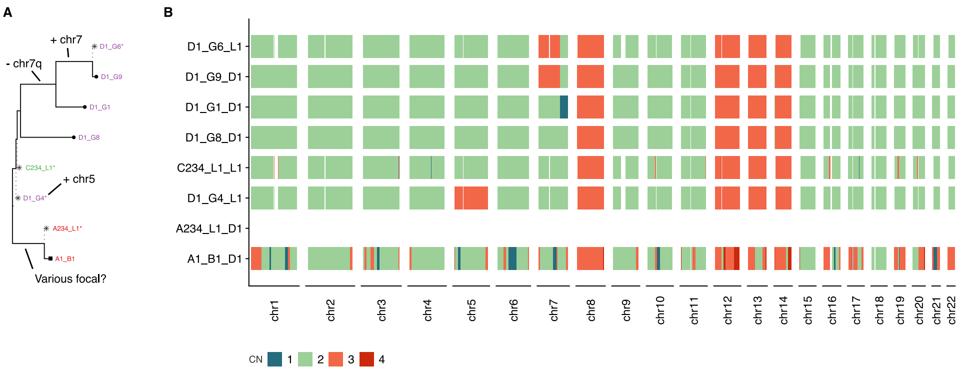
Supplementary Figure S.115: LP tree and CNA data of C552.

C554

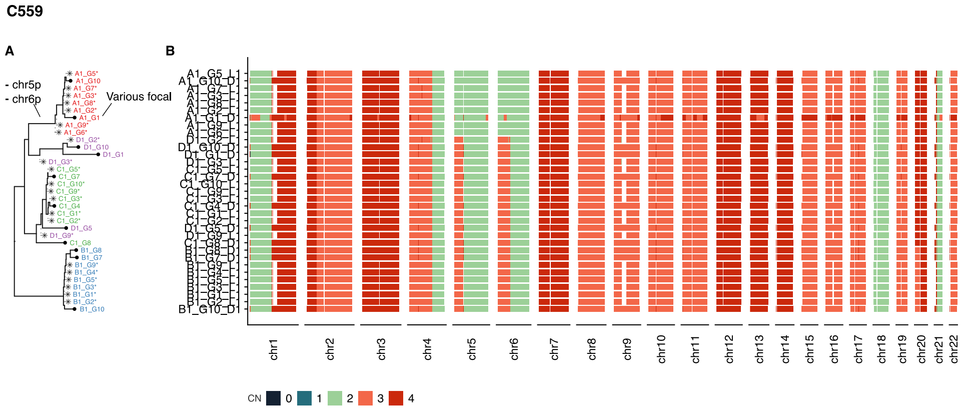


Supplementary Figure S.116: LP tree and CNA data of C554.

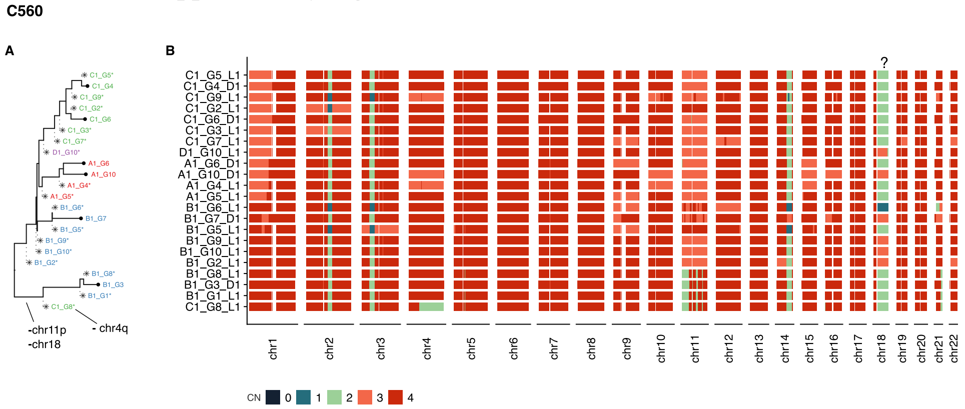
C555



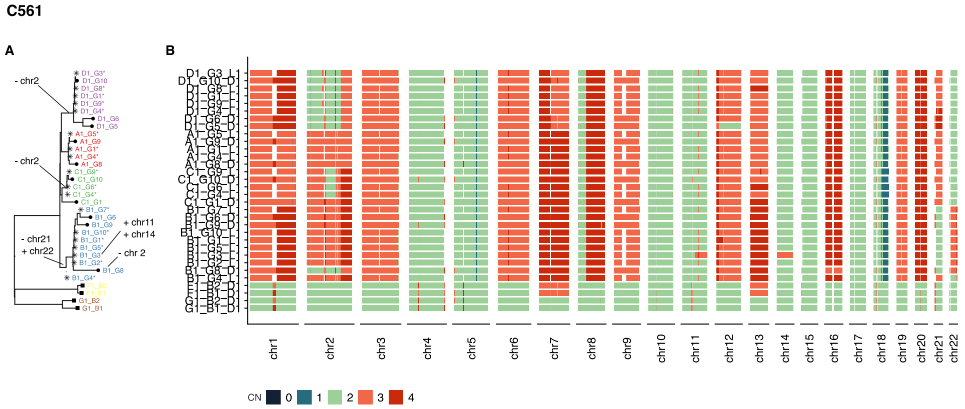
Supplementary Figure S.117: LP tree and CNA data of C555.



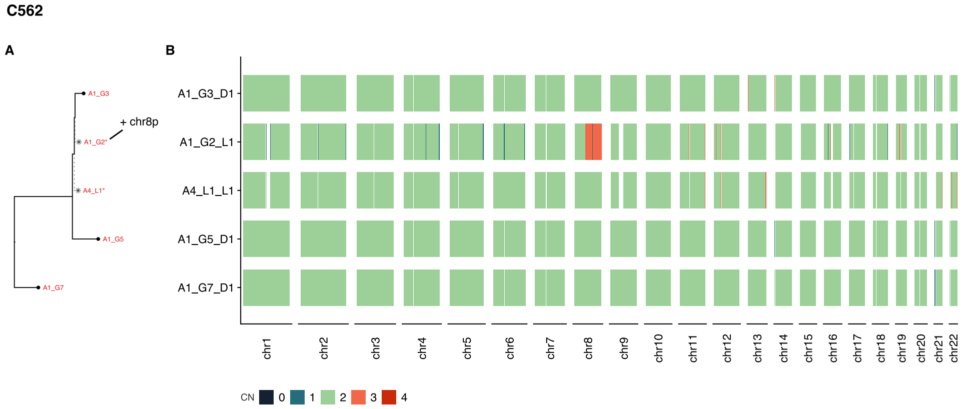
Supplementary Figure S.118: LP tree and CNA data of C559.



Supplementary Figure S.119: LP tree and CNA data of C560.

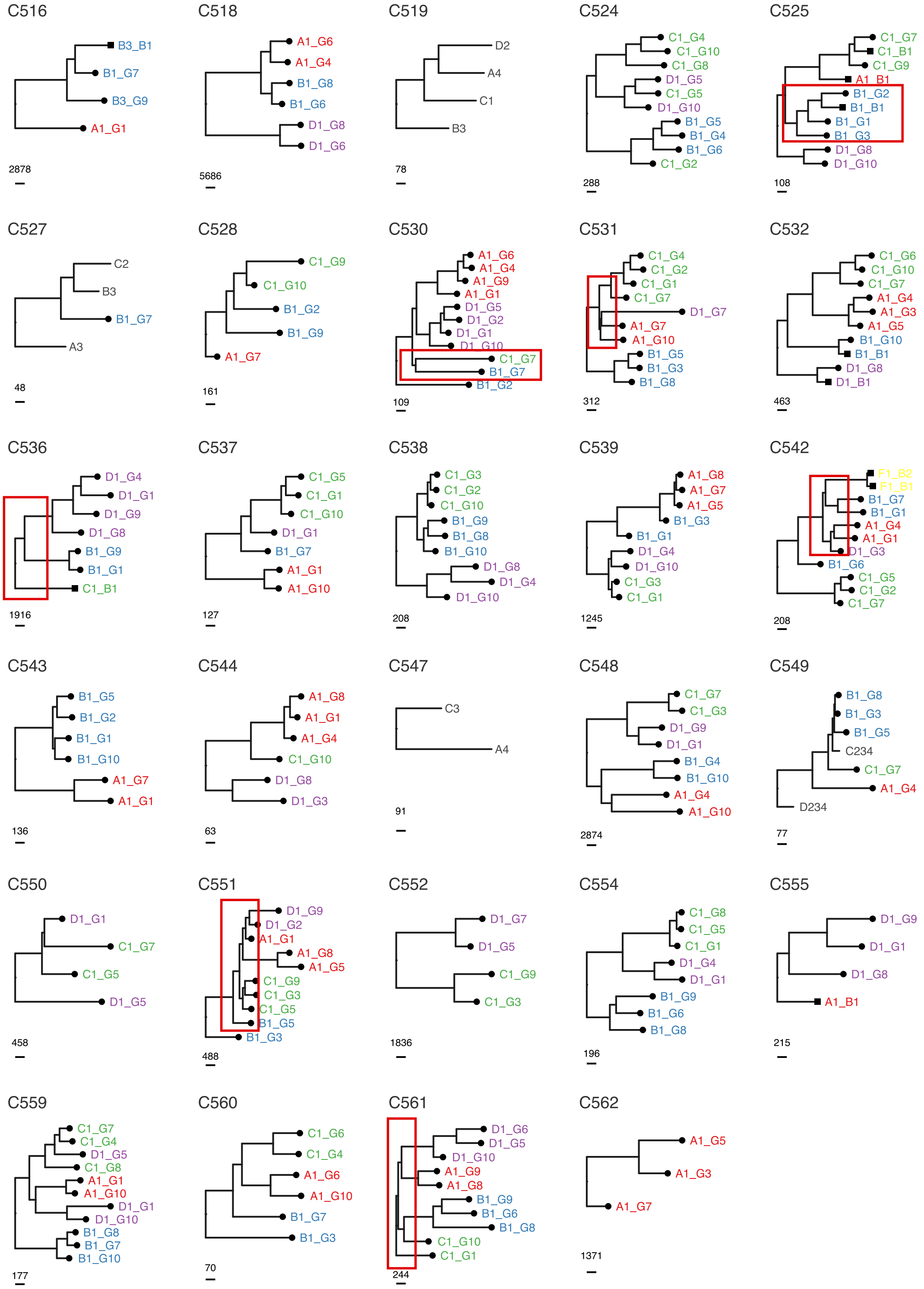


Supplementary Figure S.120: LP tree and CNA data of C561.

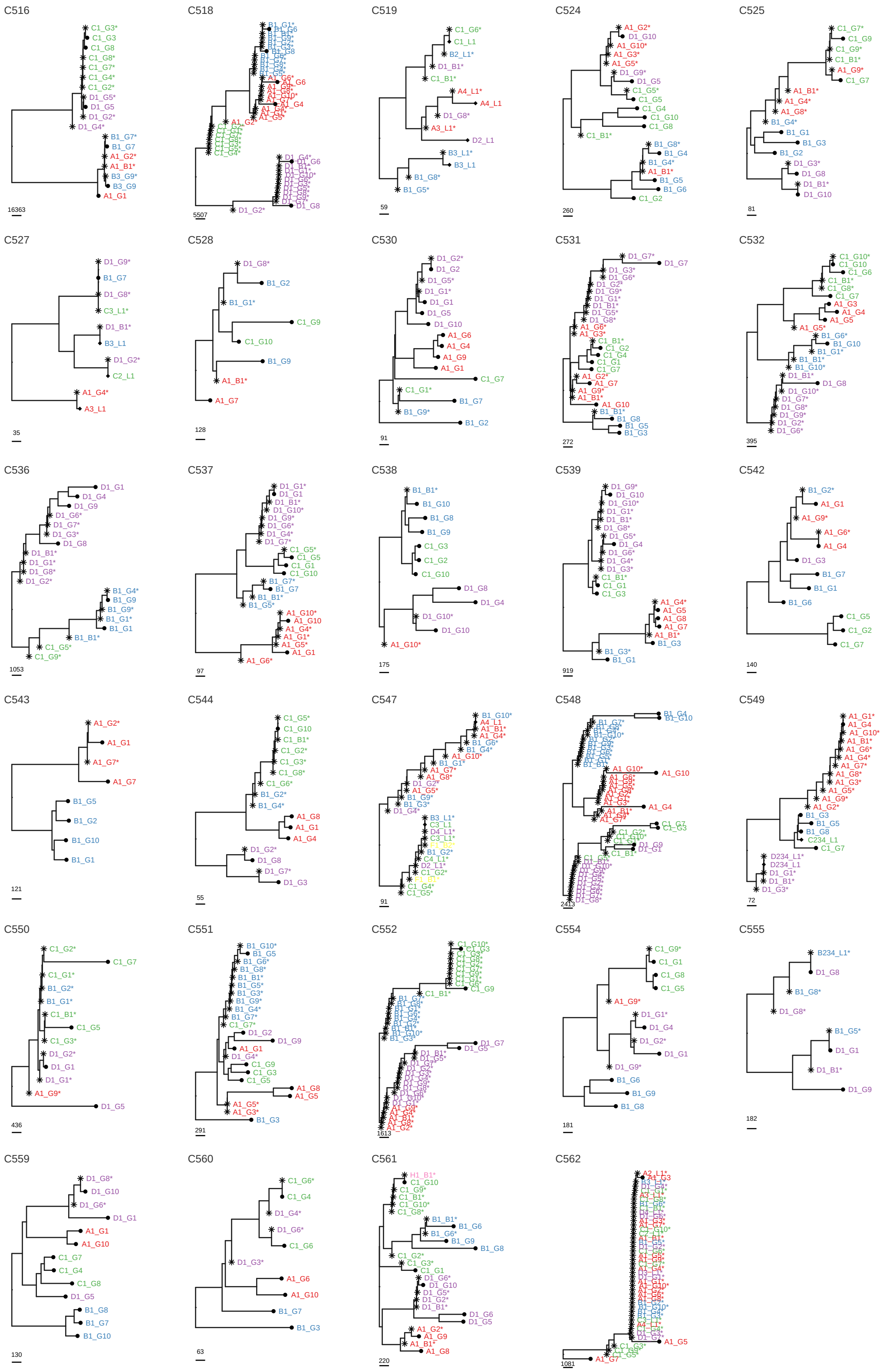


Supplementary Figure S.121: LP tree and CNA data of C562.

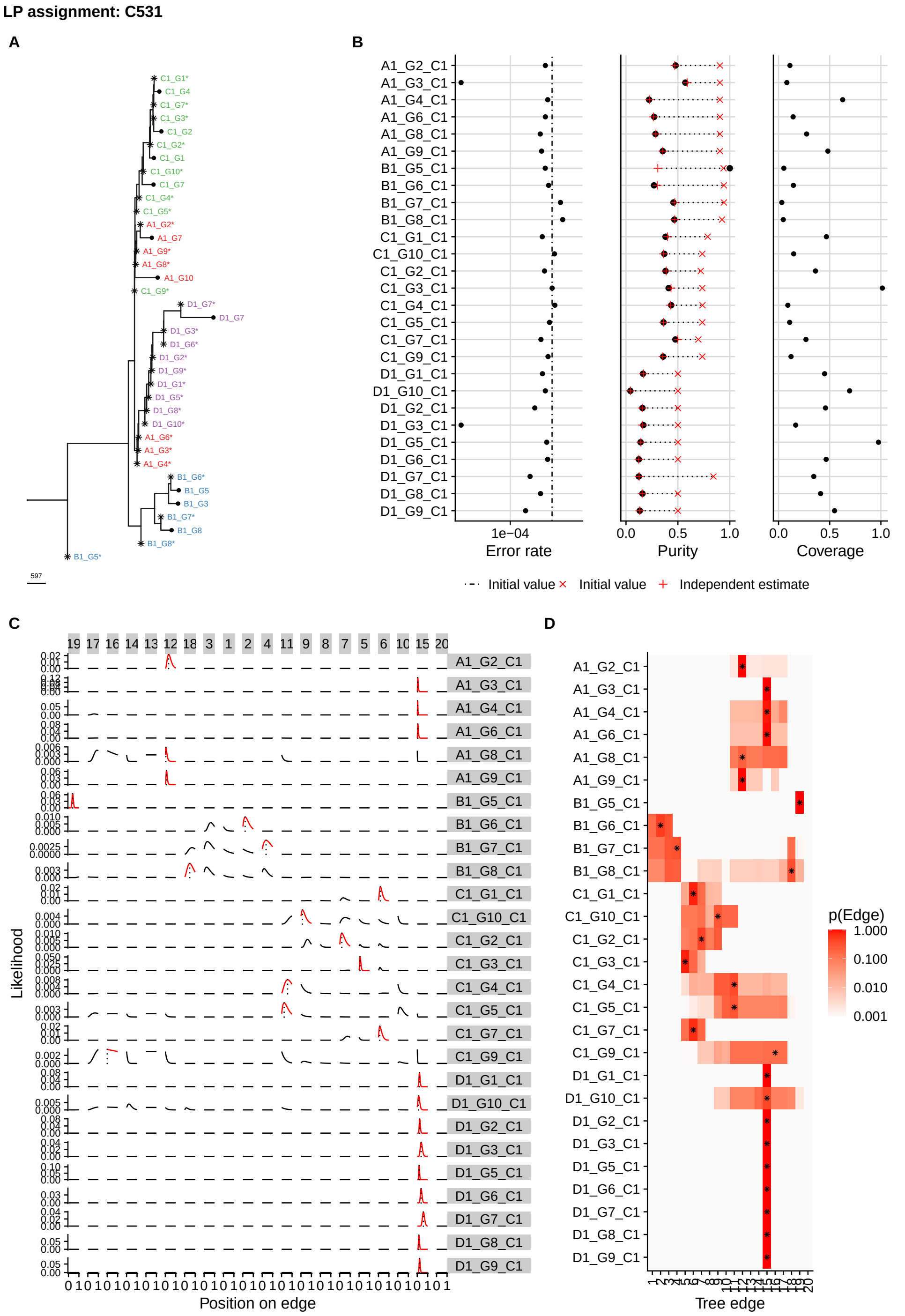
S.4.4 Various Other Figures



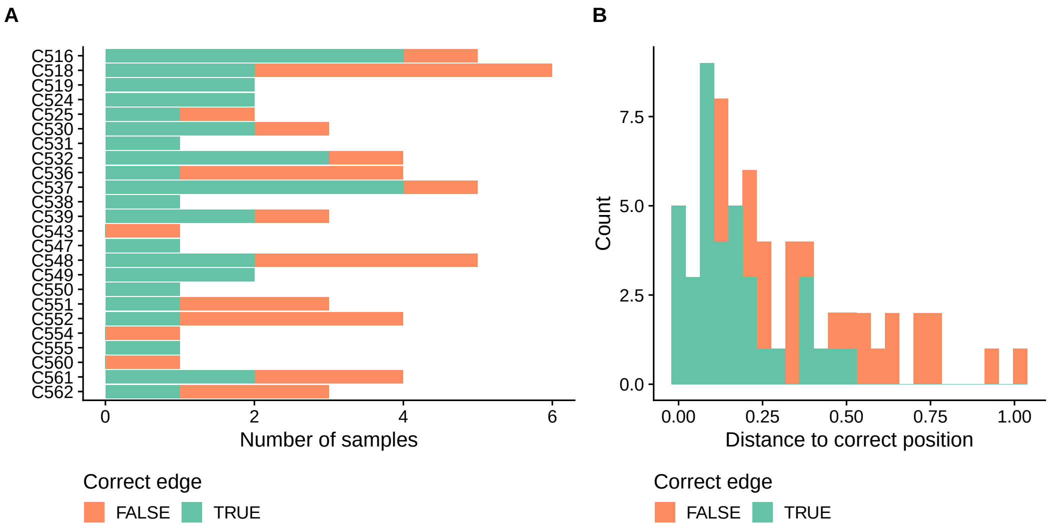
Supplementary Figure S.122: MP trees reconstructed for sites with equal CN (A&B allele state) in cancer samples. Differences in the tree topology compared to the equivalent tree reconstructed from all mutation (Figure S.51) are highlighted with red boxes.



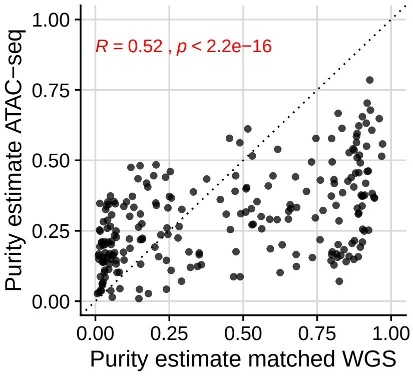
Supplementary Figure S.123: EPICC: ATAC-seq ML trees reconstructed from all mutations.



Supplementary Figure S.124: ML ATAC-seq assignment results for case C531. A) ML tree reconstructed from data. B) Maximum-likelihood estimate of per sample parameters. C) Plot of the likelihood along the different edges (panels on the x-axis) of the tree for each sample. D) Heatmap showing the distribution of likelihood that the samples are associated with a given edge.

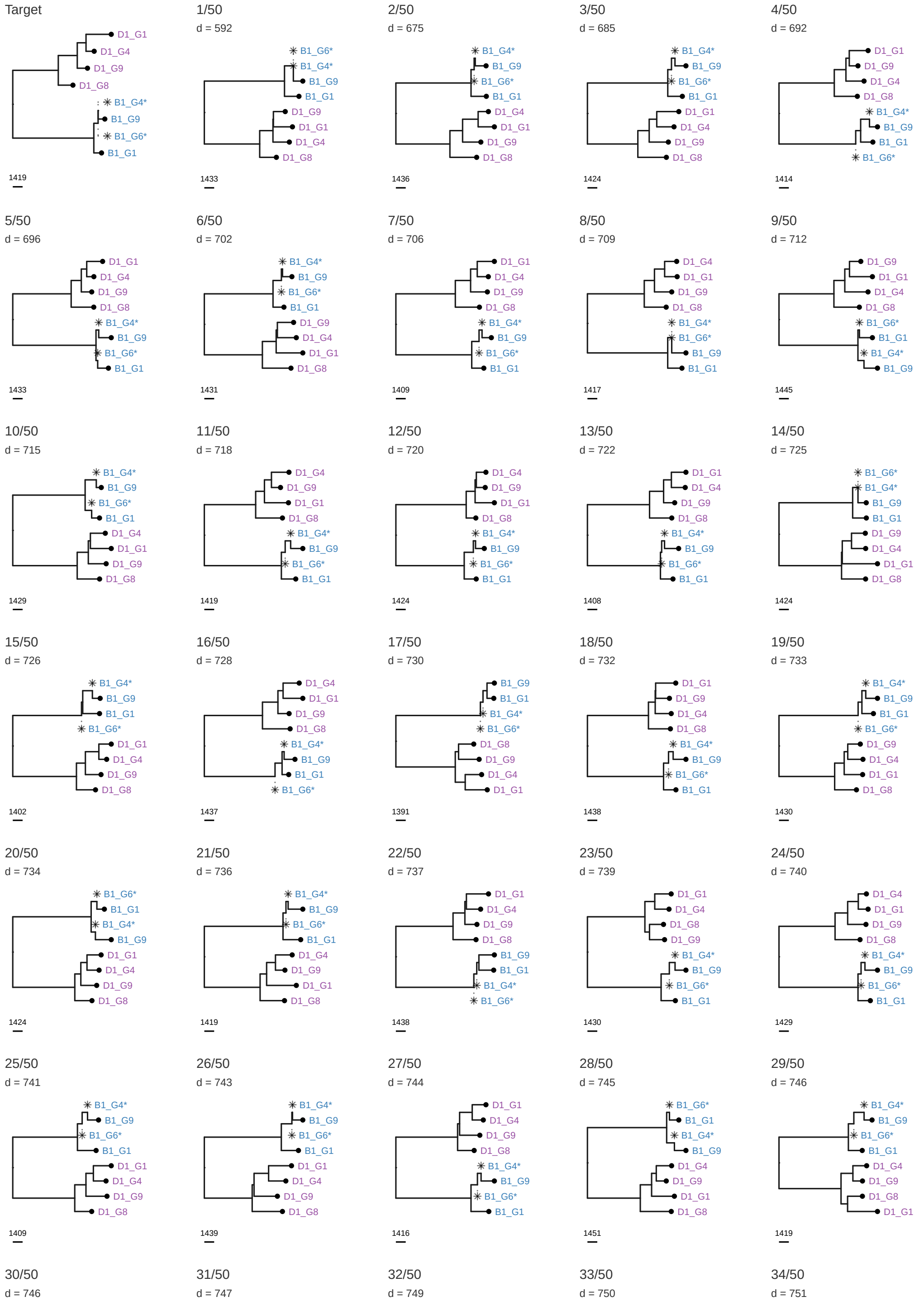


Supplementary Figure S.125: EPICC: ATAC-seq ML trees correctness of sample positions.



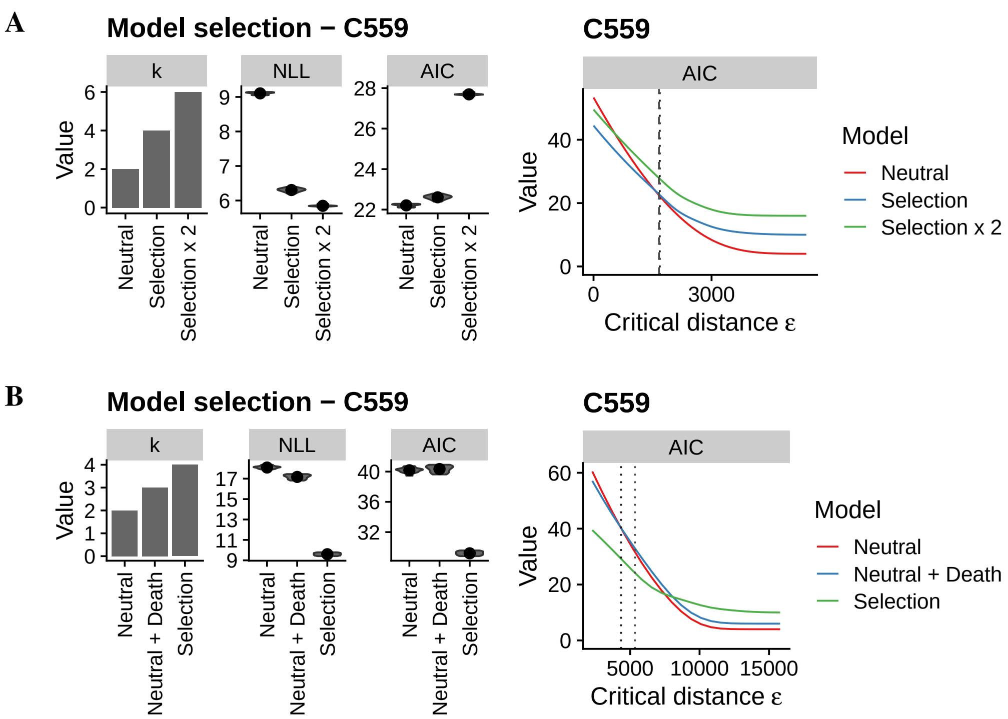
Supplementary Figure S.126: ML ATAC-seq trees: WGS vs. ATAC-seq purity.

S.5 SMC-ABC Inference

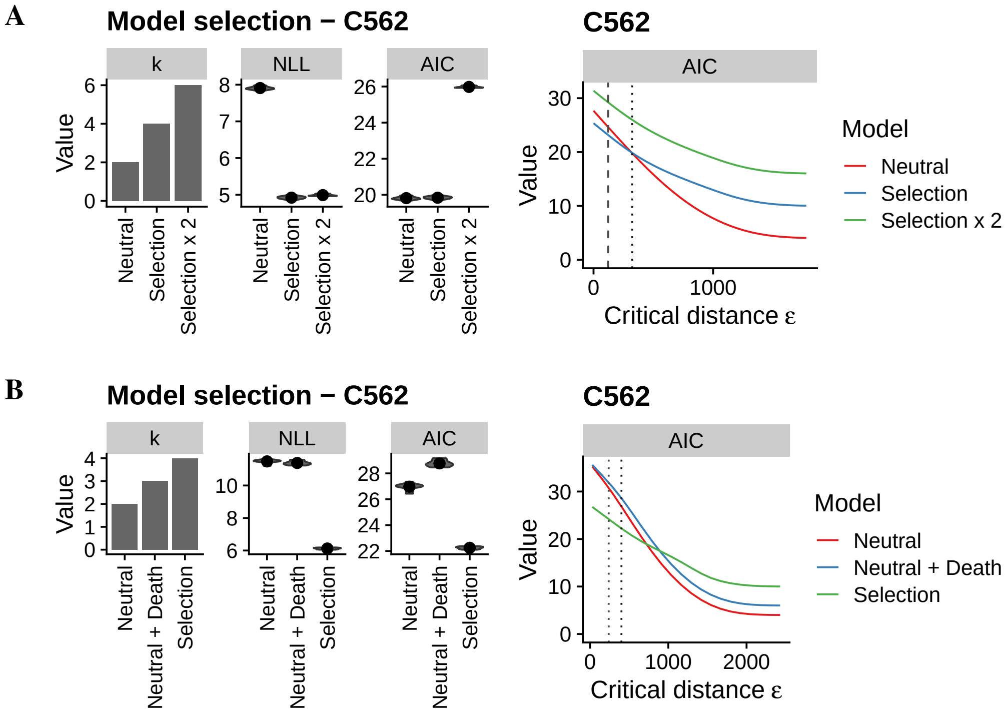


Supplementary Figure S.127: Range of accepted trees ABC-SMC inference on the LP-WGS tree of C536.

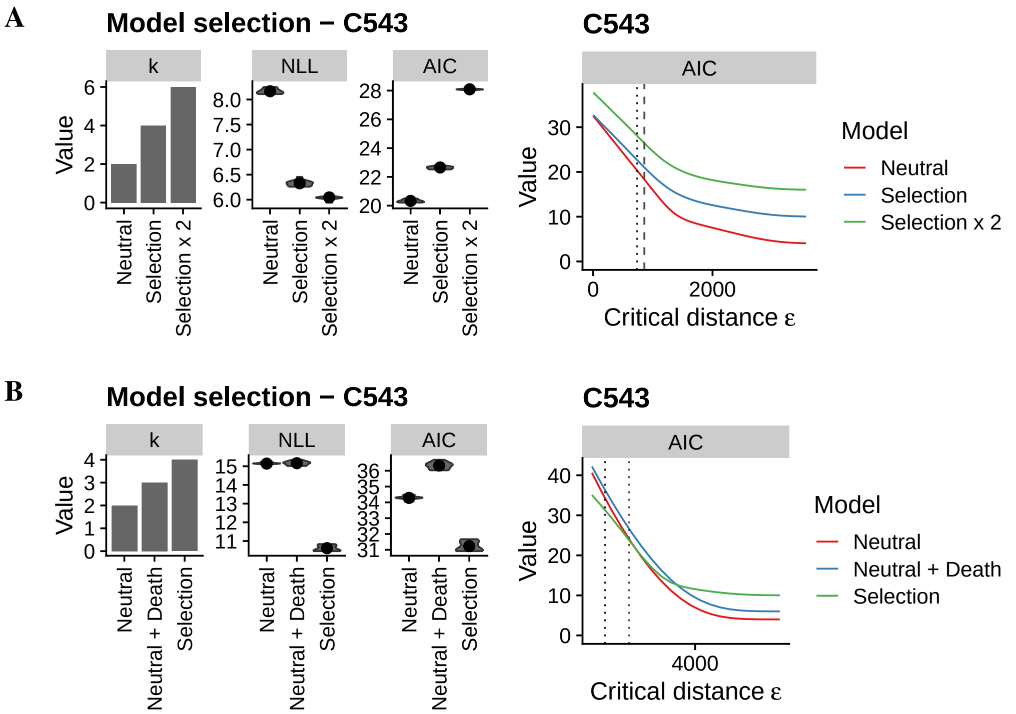
S.5.1 ABC-SMC Model Selection Results



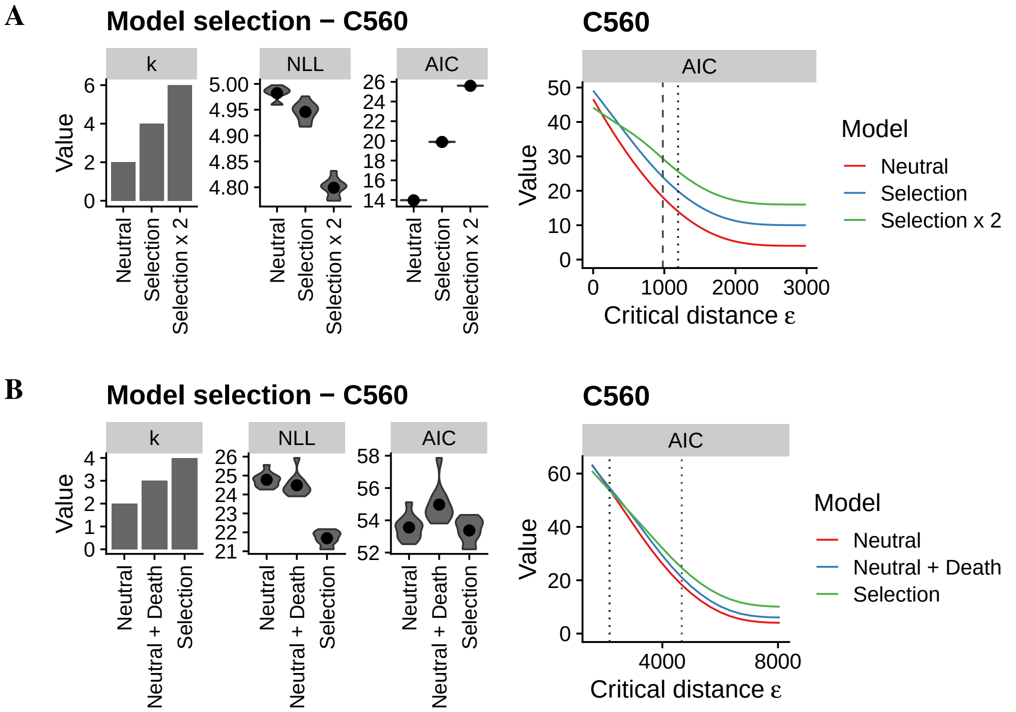
Supplementary Figure S.128: Model selection results for C559 using fixed sampling schema. A) WGS trees. B) ML LP-WGS trees.



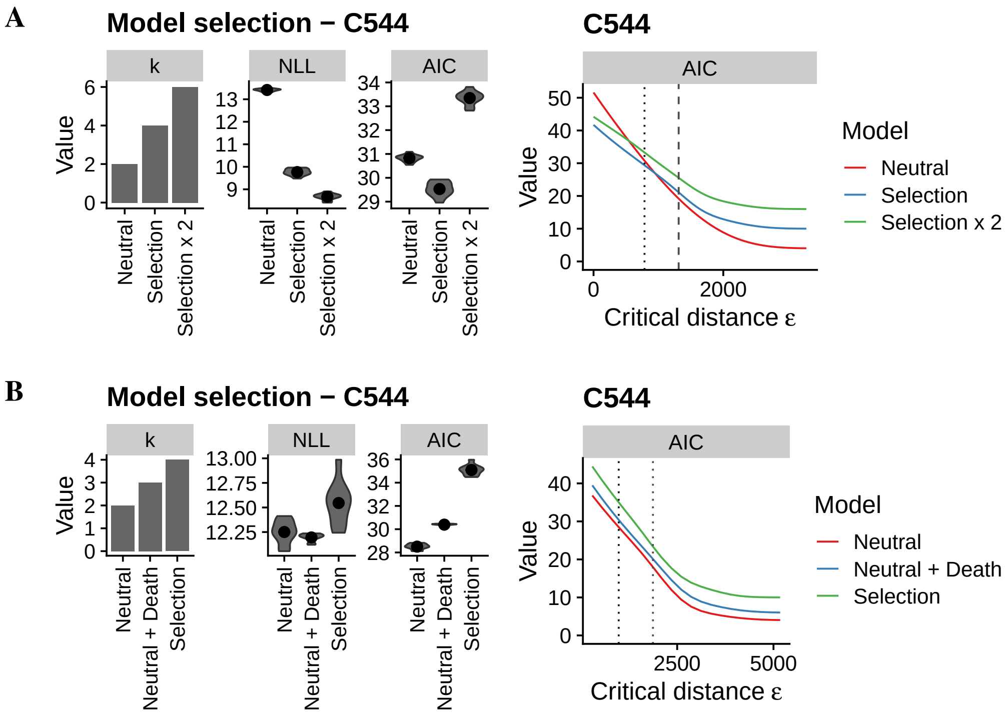
Supplementary Figure S.129: Model selection results for C562 using fixed sampling schema. A) WGS trees. B) ML LP-WGS trees.



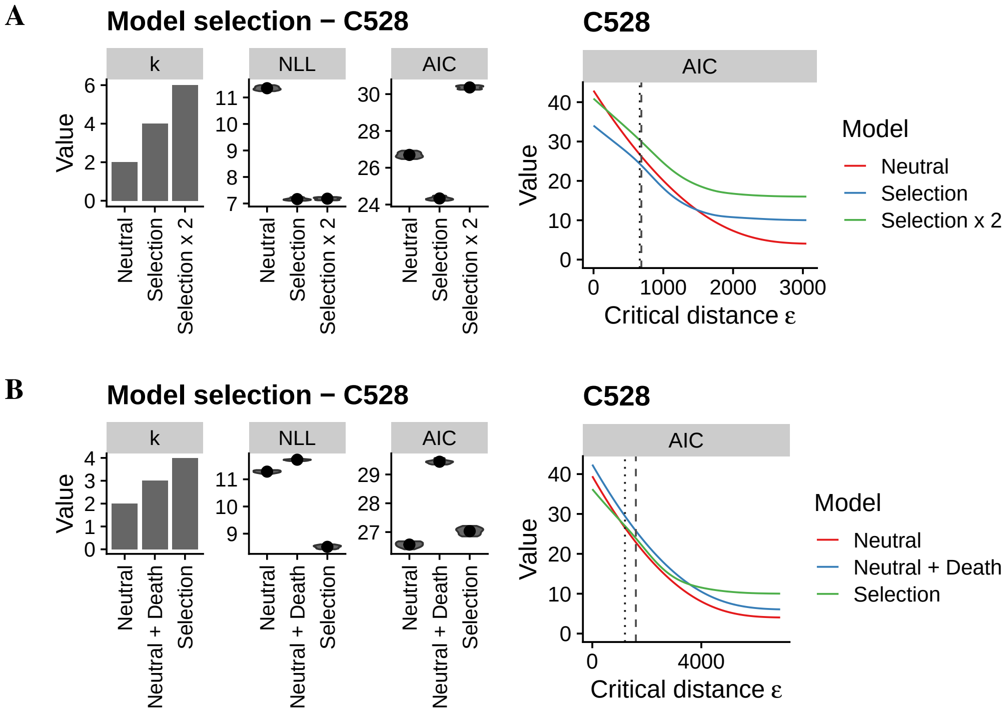
Supplementary Figure S.130: Model selection results for C543 using fixed sampling schema. A) WGS trees. B) ML LP-WGS trees.



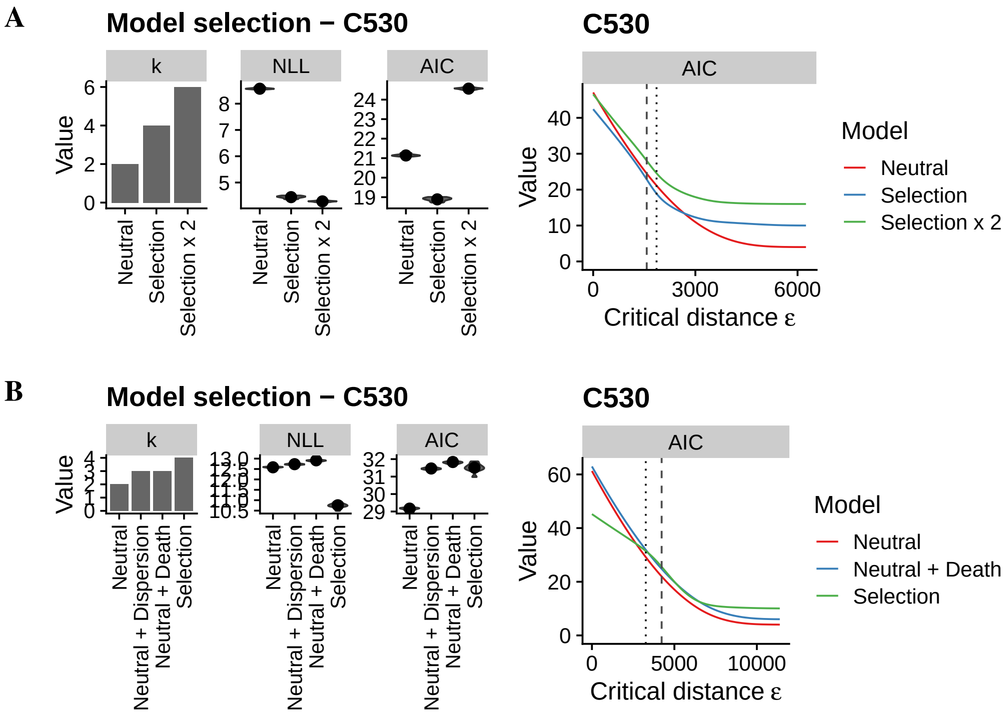
Supplementary Figure S.131: Model selection results for C560 using fixed sampling schema. A) WGS trees. B) ML LP-WGS trees.



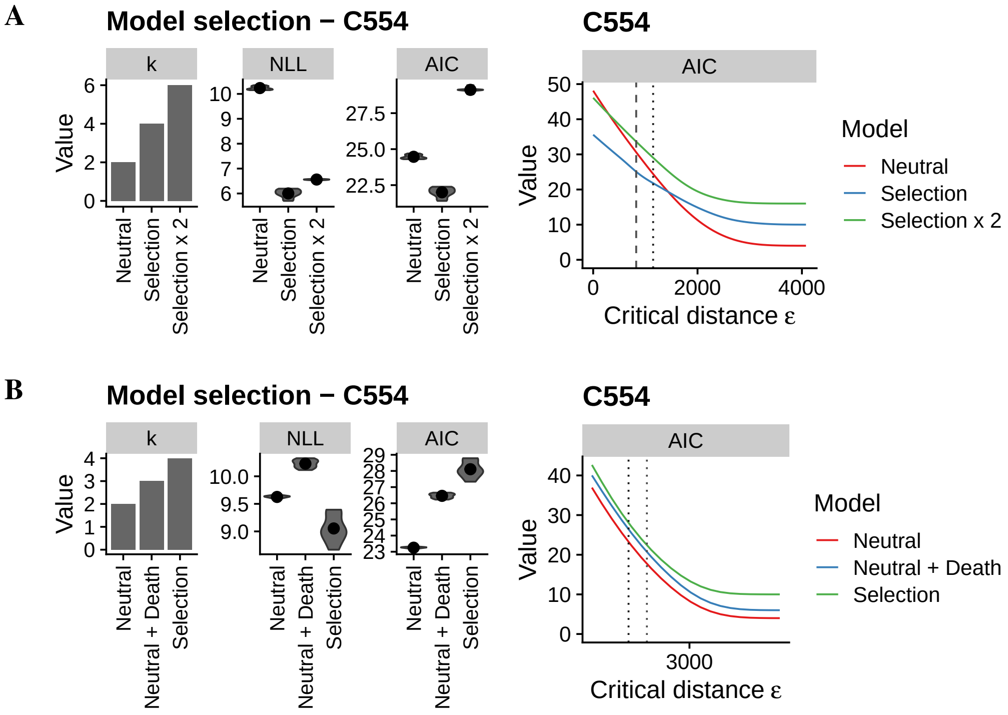
Supplementary Figure S.132: Model selection results for C544 using fixed sampling schema. A) WGS trees. B) ML LP-WGS trees.



Supplementary Figure S.133: Model selection results for C528 using fixed sampling schema. A) WGS trees. B) ML LP-WGS trees.

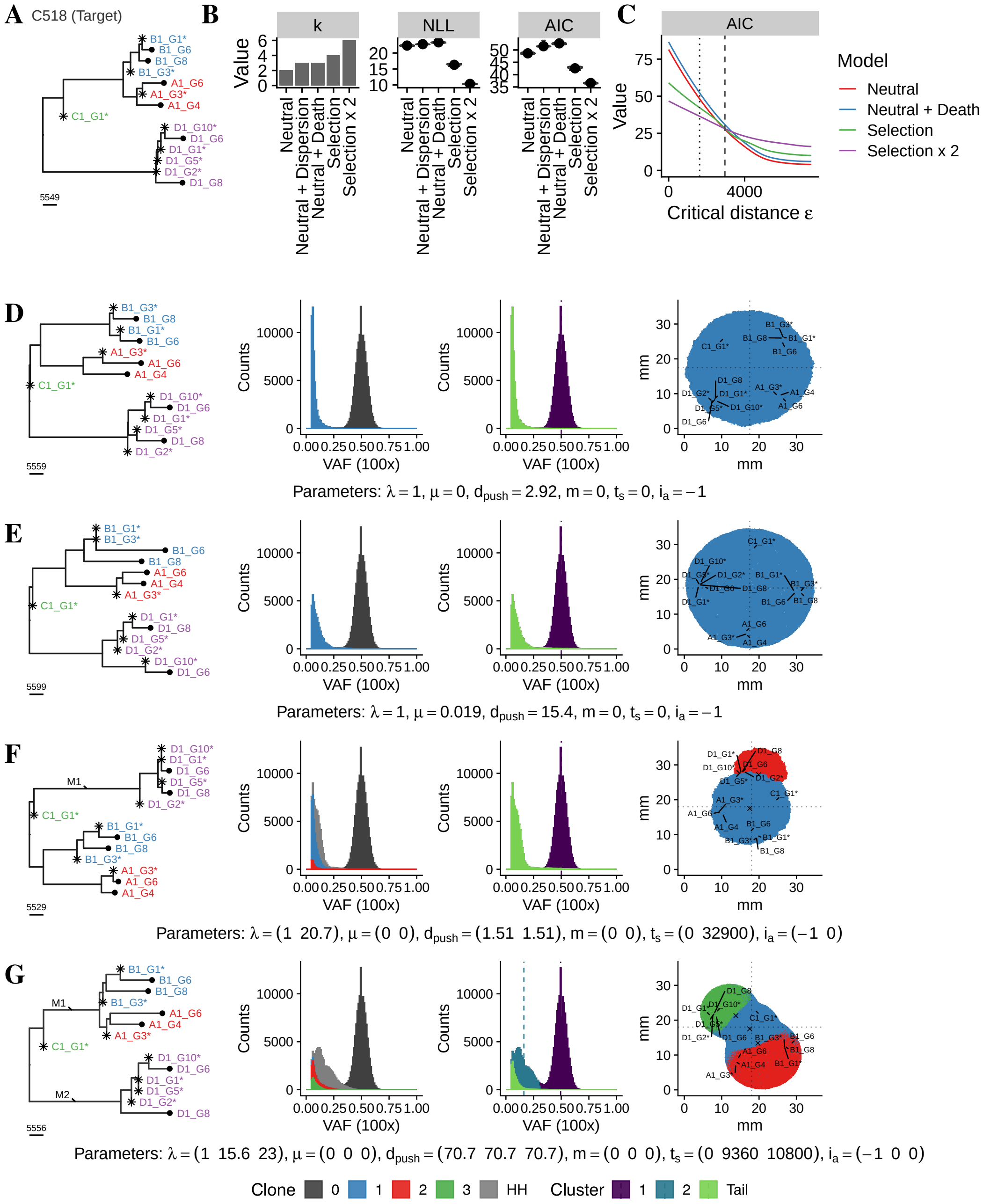


Supplementary Figure S.134: Model selection results for C530 using fixed sampling schema. A) WGS trees. B) ML LP-WGS trees.



Supplementary Figure S.135: Model selection results for C554 using fixed sampling schema. A) WGS trees. B) ML LP-WGS trees.

S.5.2 ABC-SMC Model Fits



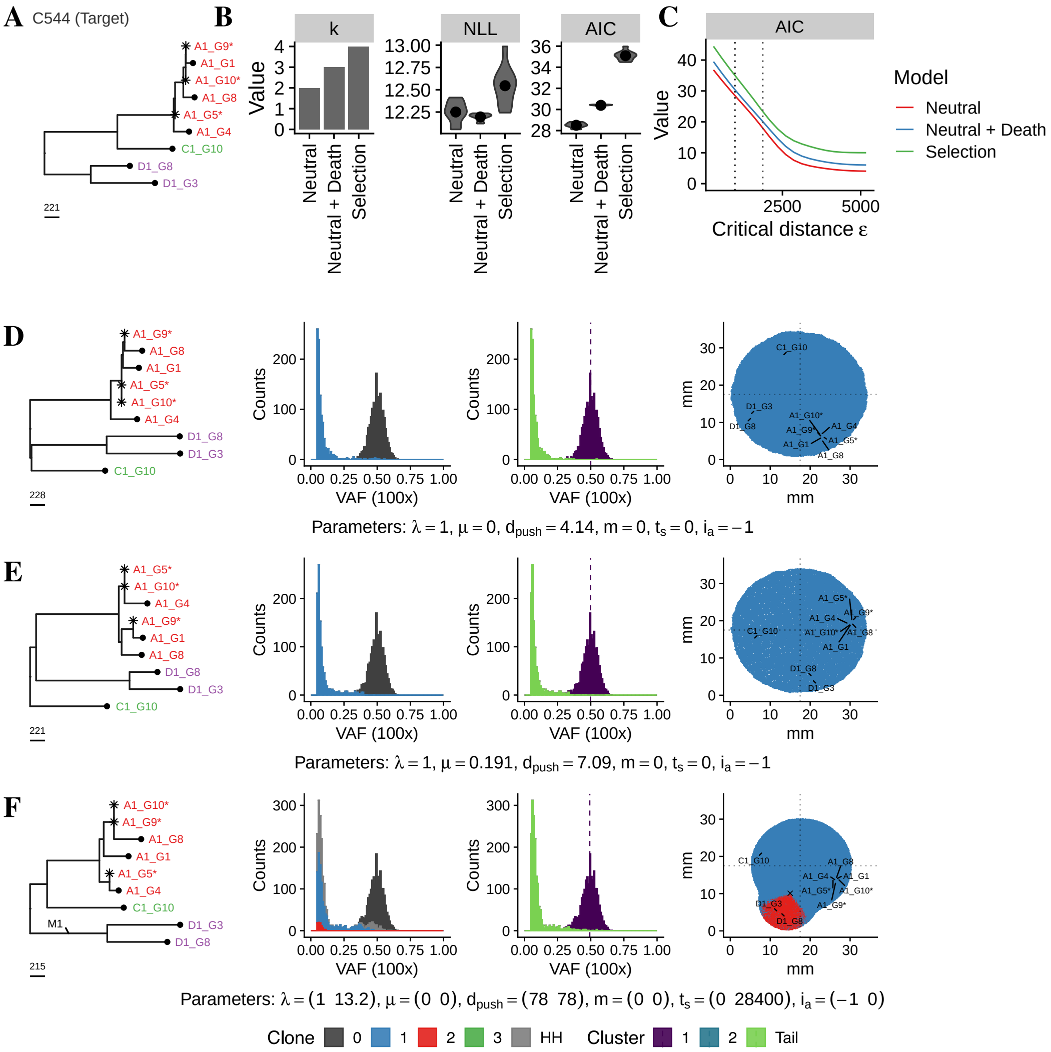
Parameters: $\lambda = 1$, $\mu = 0$, $d_{push} = 2.92$, $m = 0$, $t_s = 0$, $i_a = -1$

Parameters: $\lambda = 1$, $\mu = 0.019$, $d_{push} = 15.4$, $m = 0$, $t_s = 0$, $i_a = -1$

Parameters: $\lambda = (1\ \ 20.7)$, $\mu = (0\ \ 0)$, $d_{push} = (1.51\ \ 1.51)$, $m = (0\ \ 0)$, $t_s = (0\ \ 32900)$, $i_a = (-1\ \ 0)$

Parameters: $\lambda = (1\ \ 15.6\ \ 23)$, $\mu = (0\ \ 0\ \ 0)$, $d_{push} = (70.7\ \ 70.7\ \ 70.7)$, $m = (0\ \ 0\ \ 0)$, $t_s = (0\ \ 9360\ \ 10800)$, $i_a = (-1\ \ 0\ \ 0)$

Supplementary Figure S.136: SMC-ABC inference framework applied to ML-WGS tree of C518. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D-F) The best fitting simulated neutral tree for the 'Neutral', 'Neutral+Death' and 'Selection' model respectively.

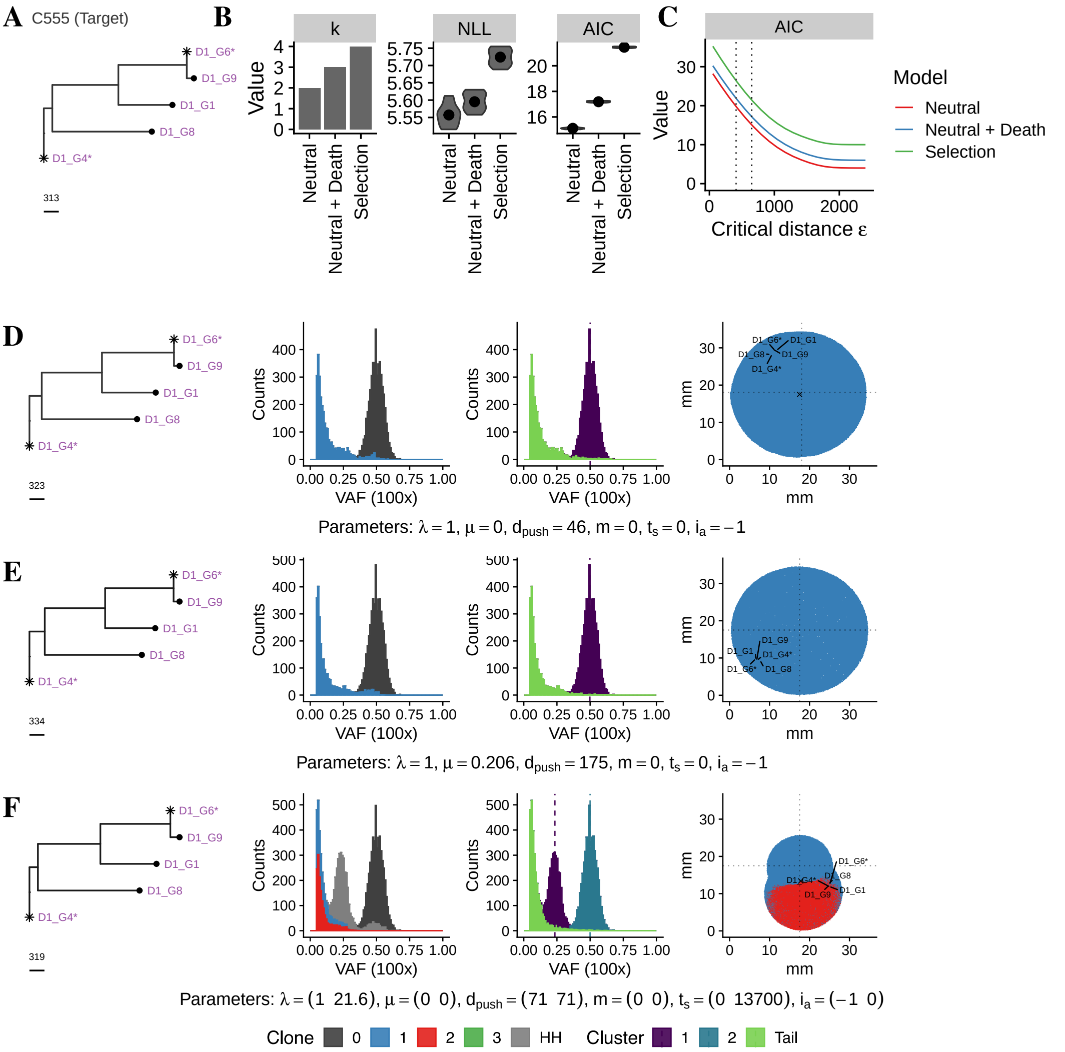


Parameters: $\lambda = 1$, $\mu = 0$, $d_{push} = 4.14$, $m = 0$, $t_s = 0$, $i_a = -1$

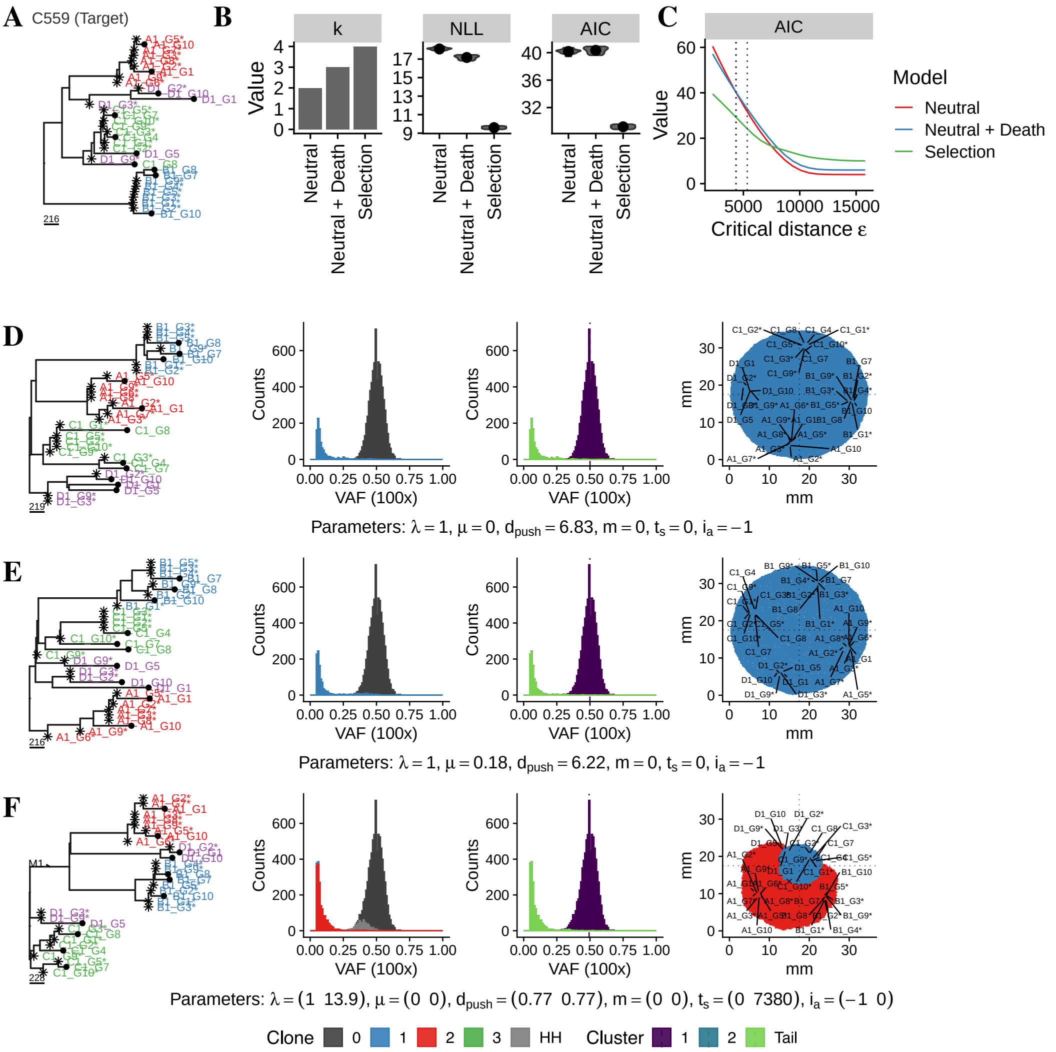
Parameters: $\lambda = 1$, $\mu = 0.191$, $d_{push} = 7.09$, $m = 0$, $t_s = 0$, $i_a = -1$

Parameters: $\lambda = (1 \;\; 13.2)$, $\mu = (0 \;\; 0)$, $d_{push} = (78 \;\; 78)$, $m = (0 \;\; 0)$, $t_s = (0 \;\; 28400)$, $i_a = (-1 \;\; 0)$

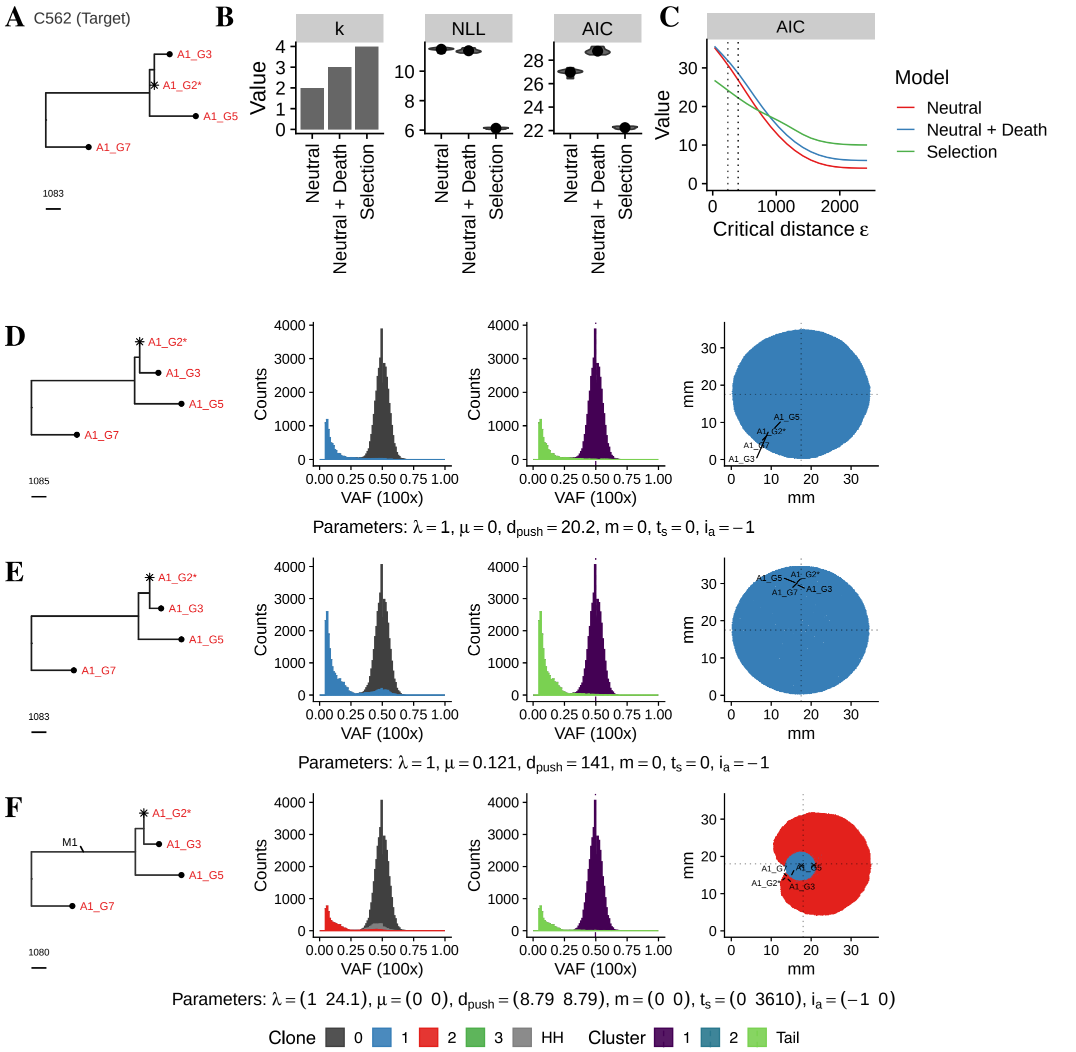
Supplementary Figure S.137: SMC-ABC inference framework applied to ML-WGS tree of C544. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D-G) The best fitting simulated neutral tree for the 'Neutral', 'Neutral+Death', 'Selection' and 'Selection x 2' model respectively.



Supplementary Figure S.138: SMC-ABC inference framework applied to ML-WGS tree of C555. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D-G) The best fitting simulated neutral tree for the ‘Neutral’, ‘Neutral+Death’, ‘Selection’ and ‘Selection x 2’ model respectively.

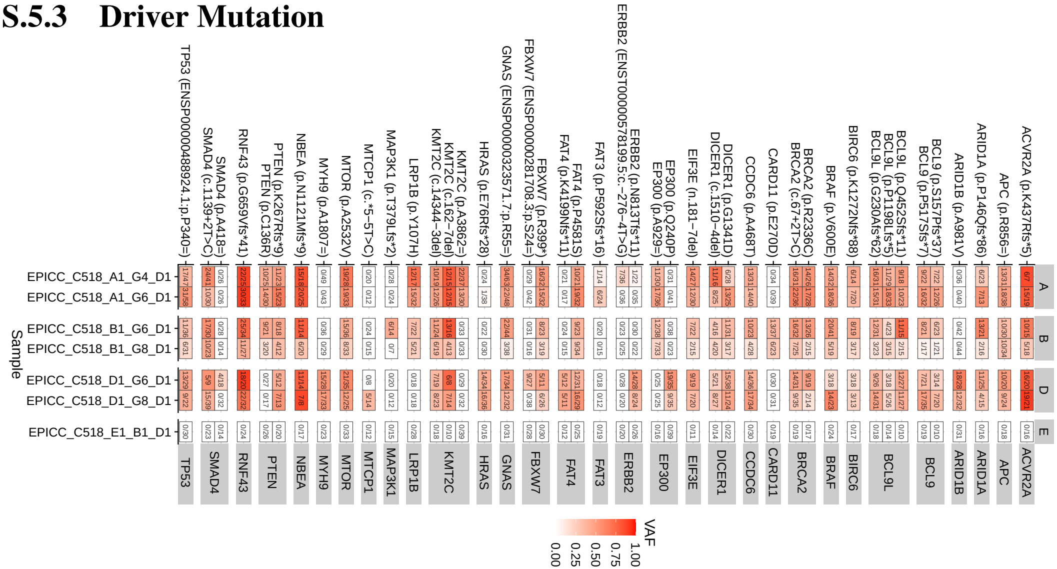


Supplementary Figure S.139: SMC-ABC inference framework applied to ML-WGS tree of C559. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D-G) The best fitting simulated neutral tree for the 'Neutral', 'Neutral+Death', 'Selection' and 'Selection x 2' model respectively.

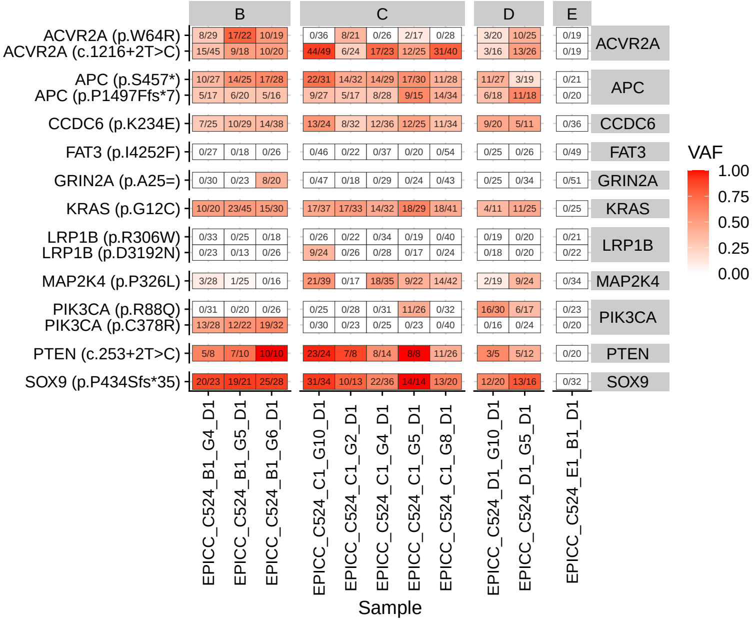


Supplementary Figure S.140: SMC-ABC inference framework applied to ML-WGS tree of C562. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D-G) The best fitting simulated neutral tree for the 'Neutral', 'Neutral+Death', 'Selection' and 'Selection x 2' model respectively.

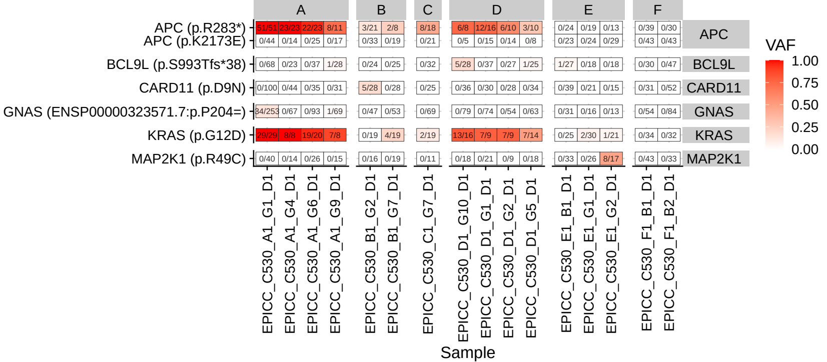
S.5.3 Driver Mutation



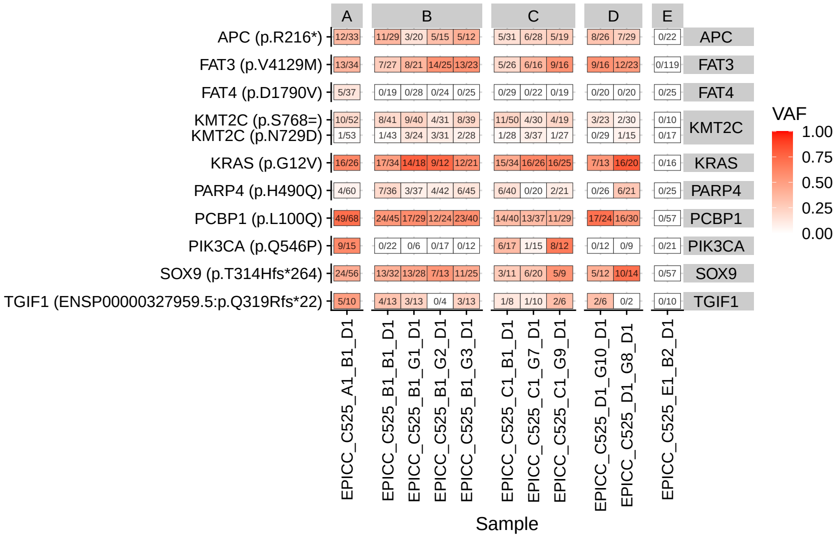
Supplementary Figure S.141: Driver mutations of C518.



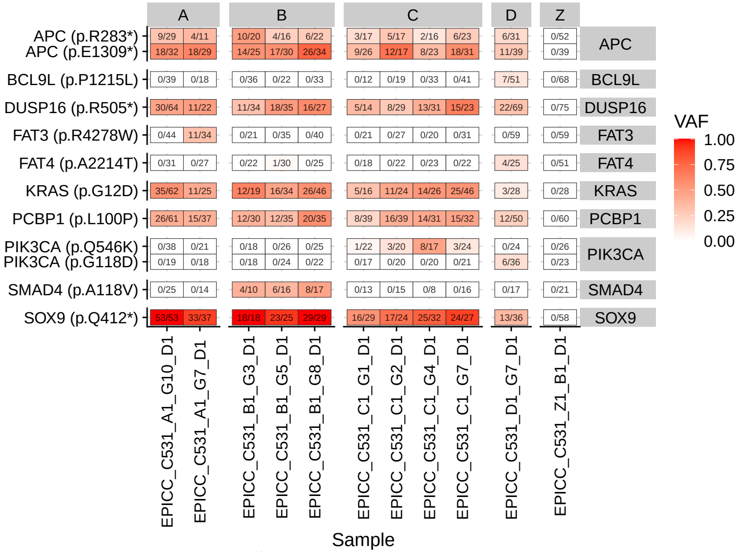
Supplementary Figure S.142: Driver mutations of C524.



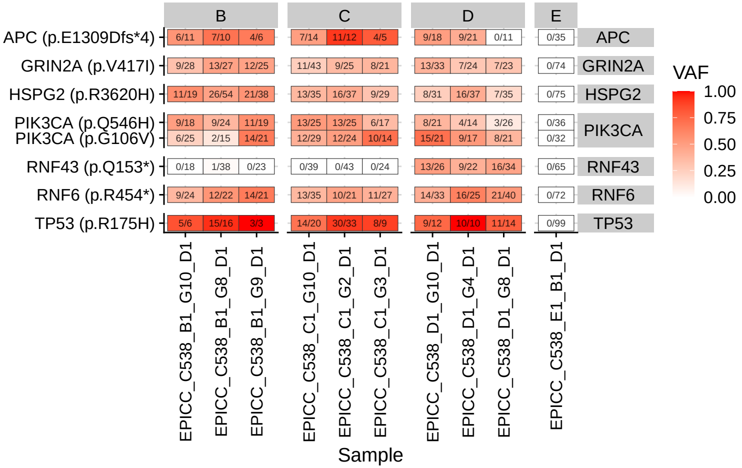
Supplementary Figure S.143: Driver mutations of C530.



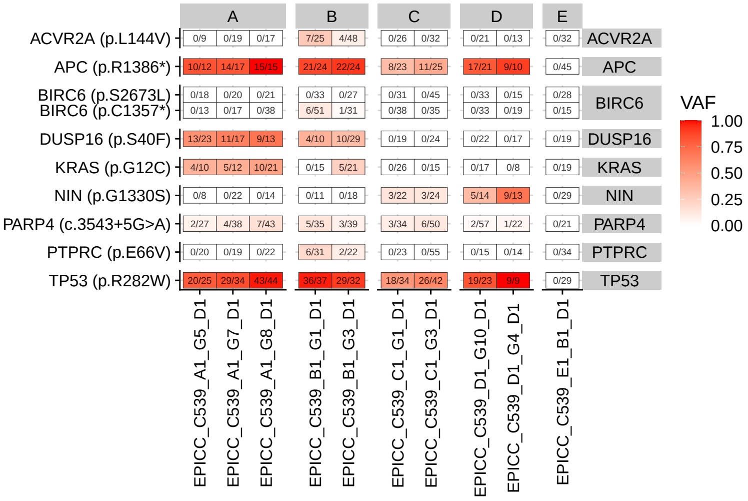
Supplementary Figure S.144: Driver mutations of C525.



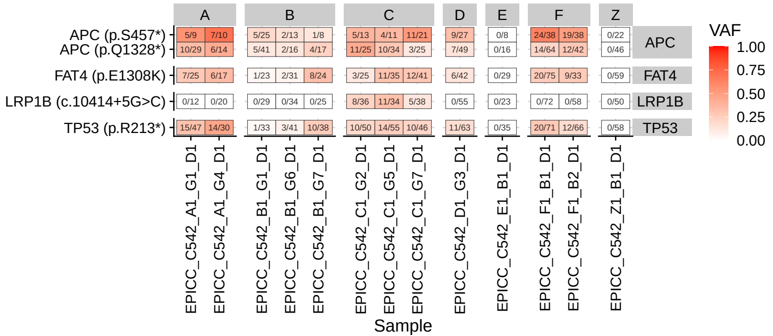
Supplementary Figure S.145: Driver mutations of C531.



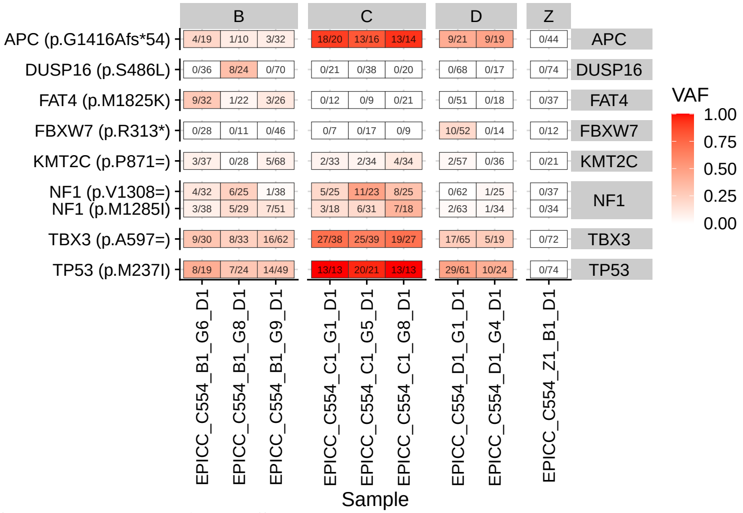
Supplementary Figure S.146: Driver mutations of C538.



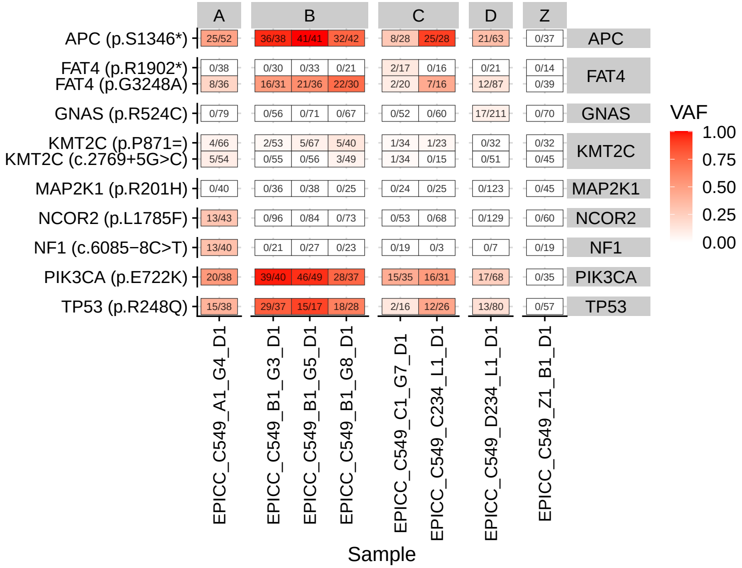
Supplementary Figure S.147: Driver mutations of C539.



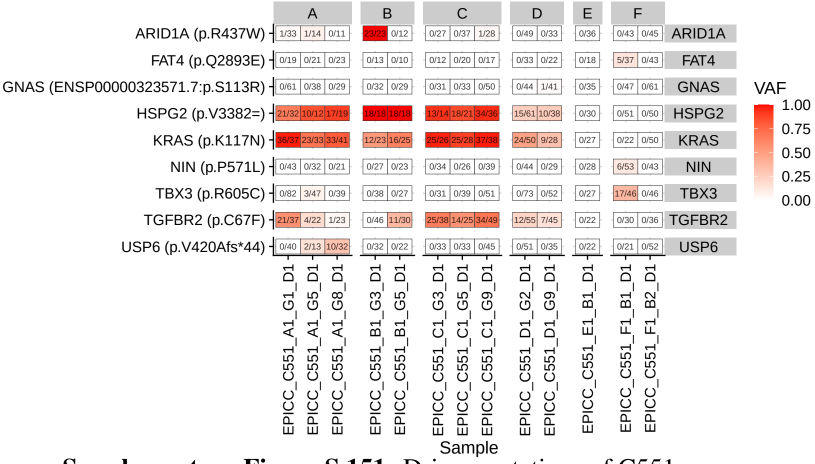
Supplementary Figure S.148: Driver mutations of C542.



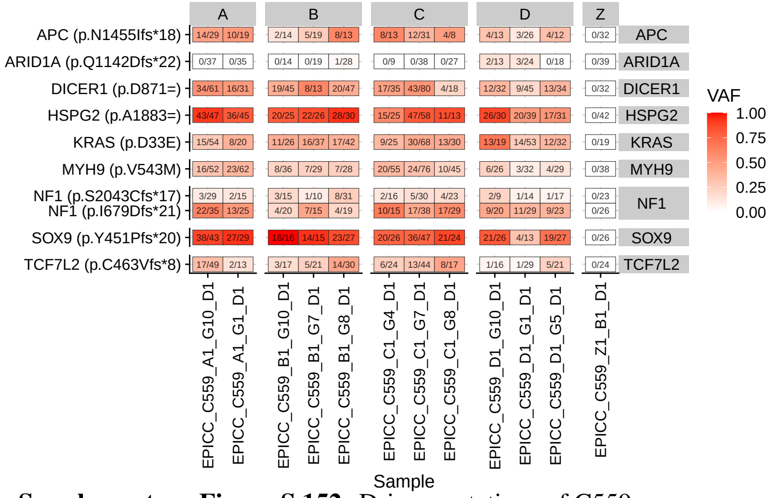
Supplementary Figure S.149: Driver mutations of C554.



Supplementary Figure S.150: Driver mutations of C549.

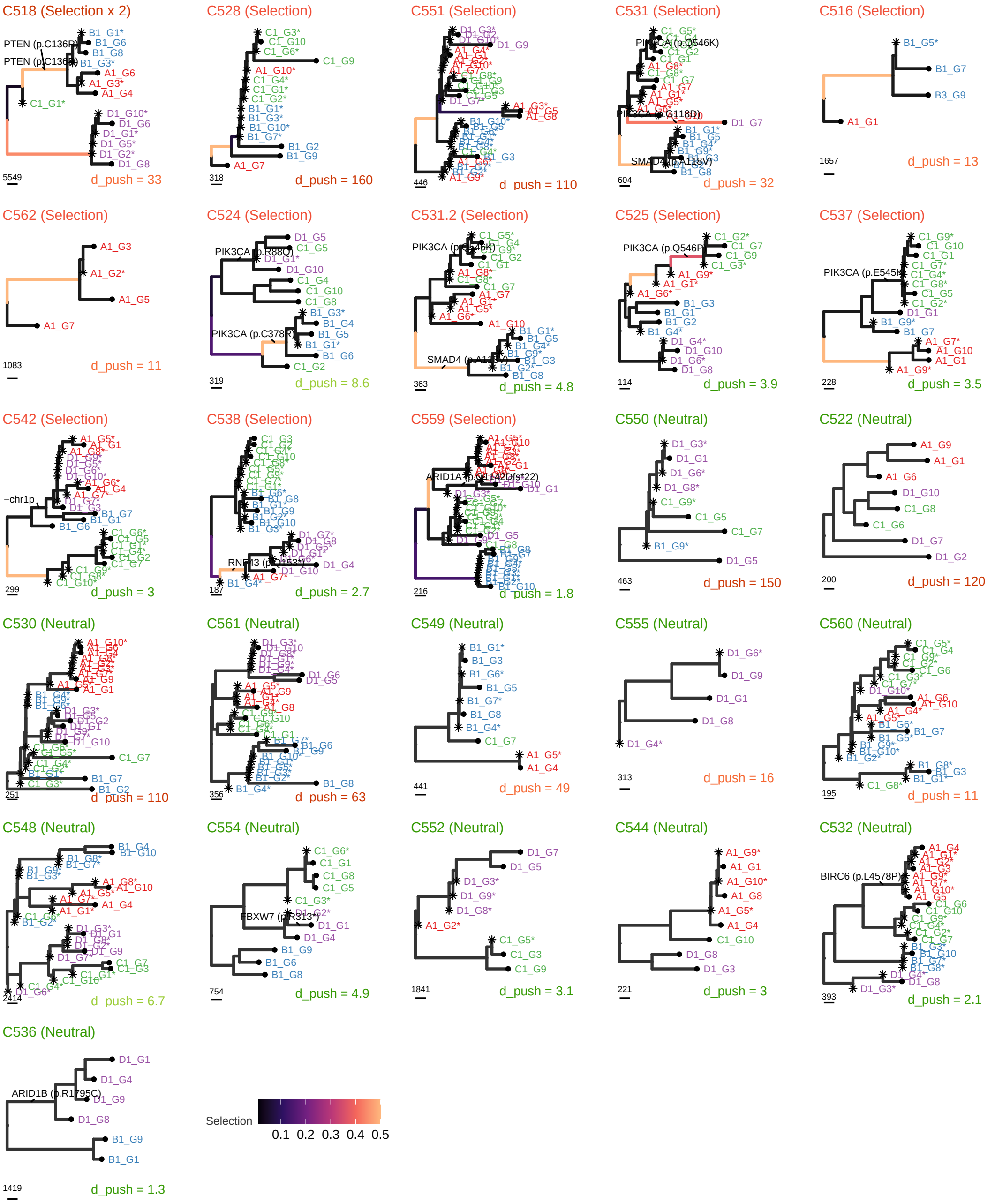


Supplementary Figure S.151: Driver mutations of C551.

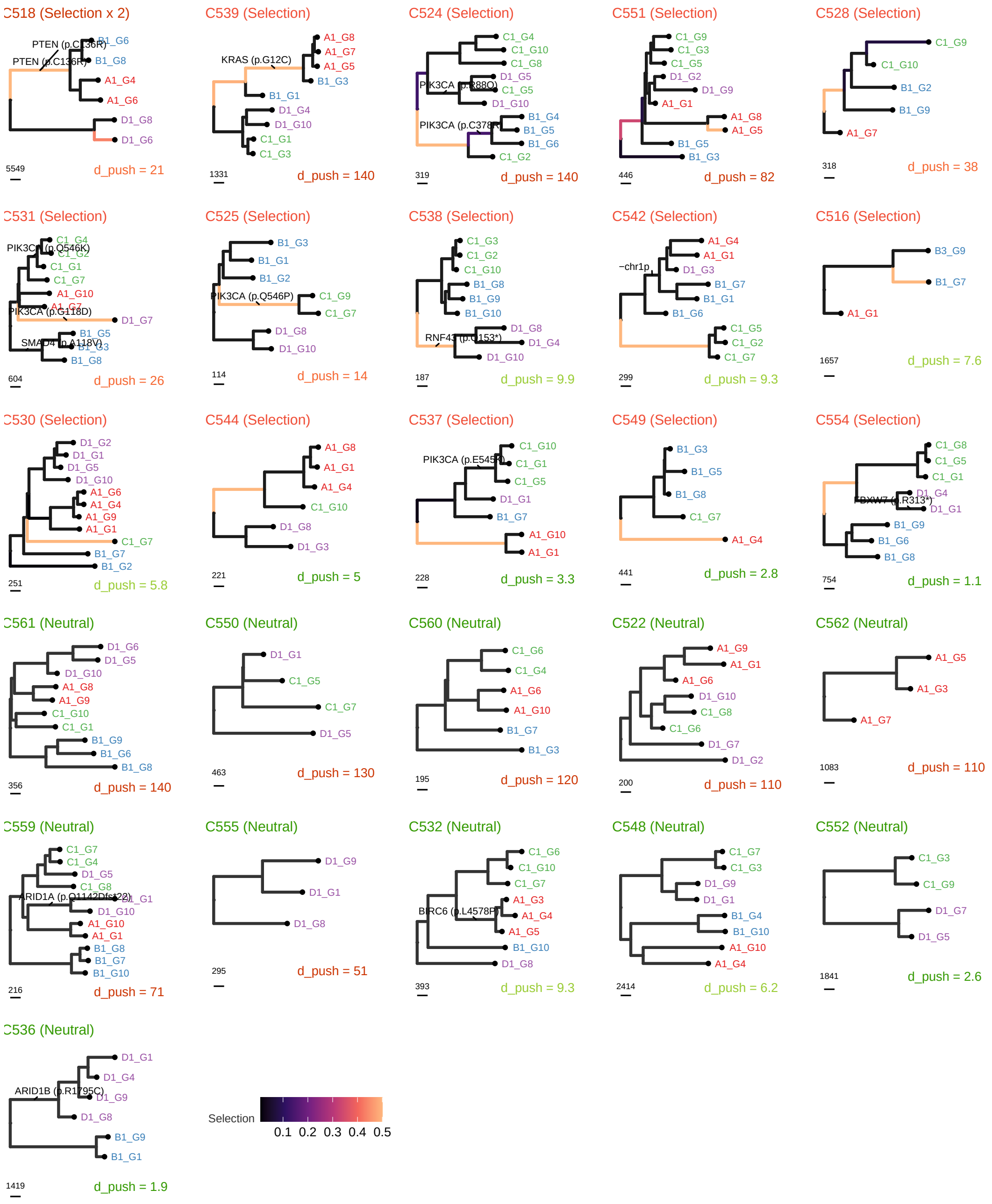


Supplementary Figure S.152: Driver mutations of C559.

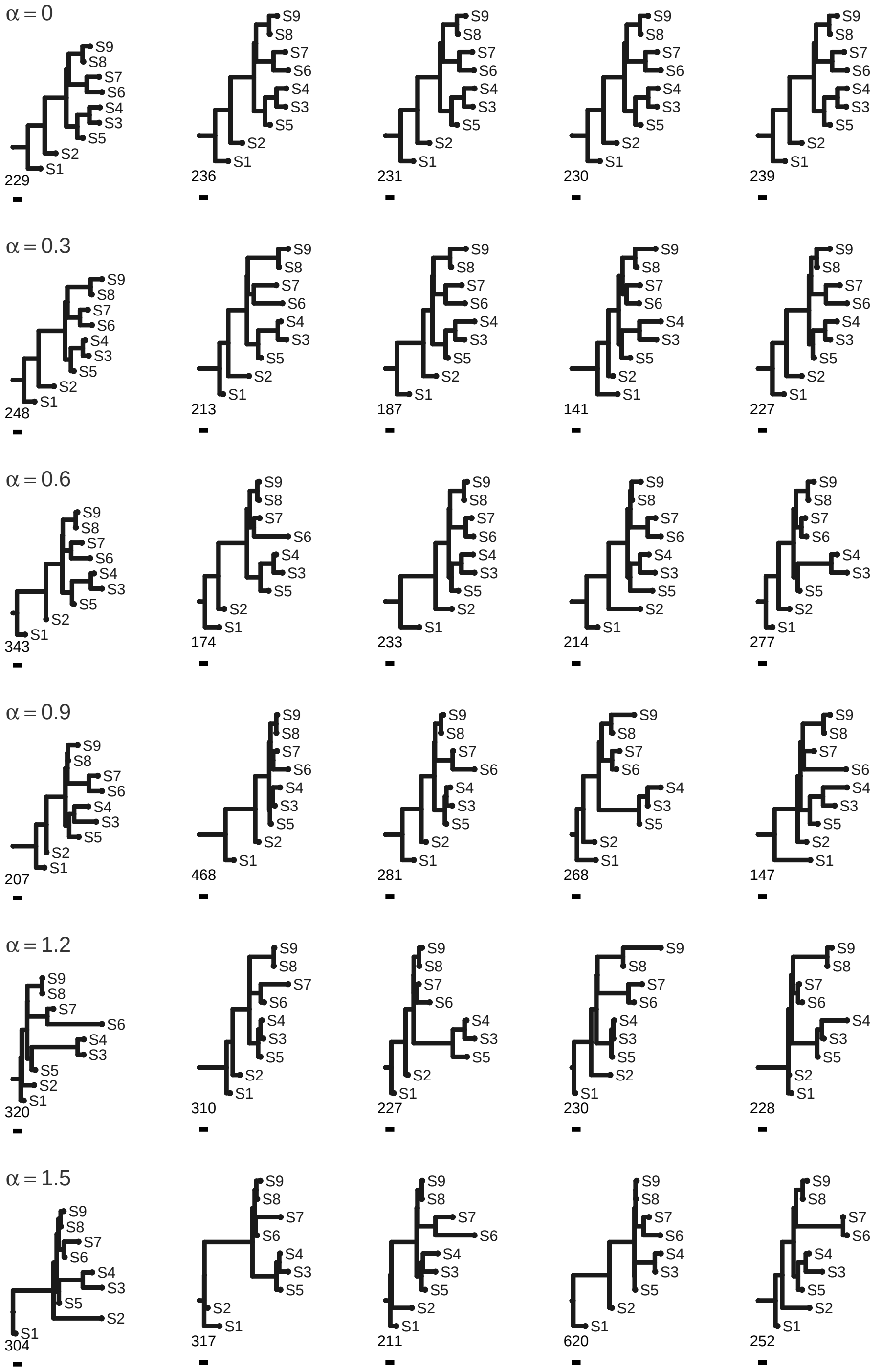
S.5.4 Other Figures



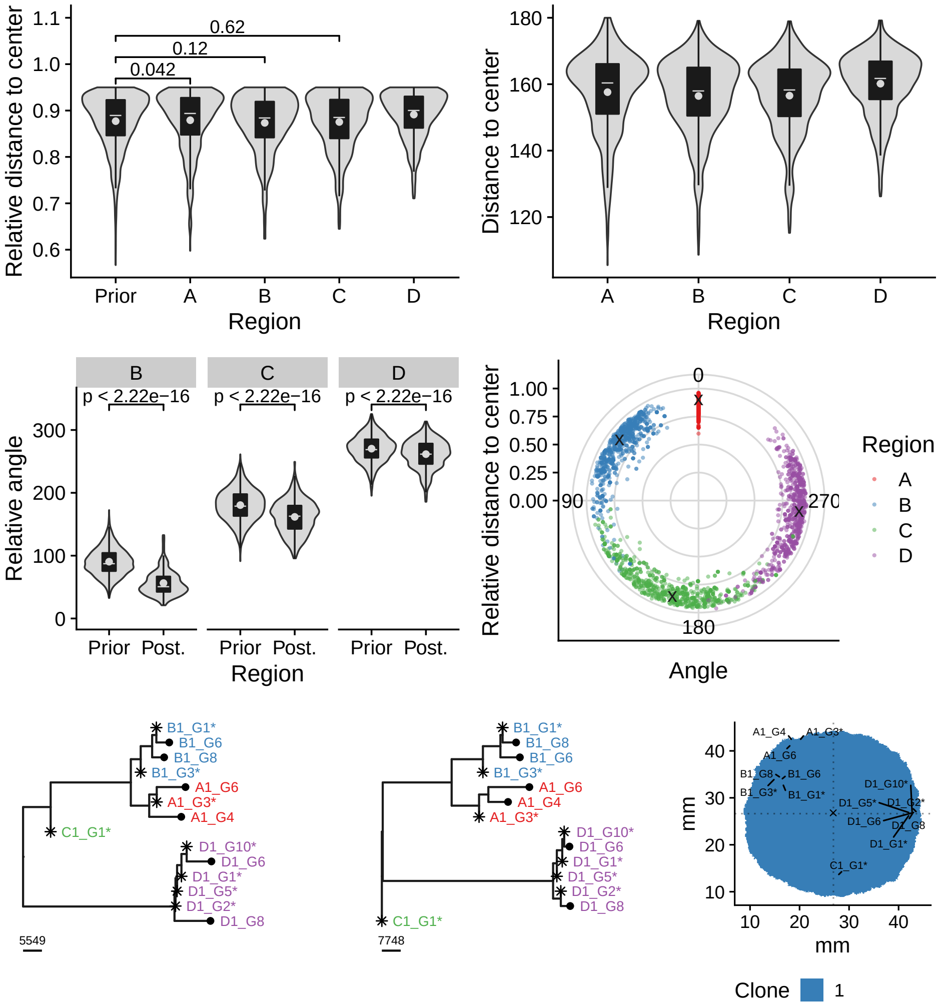
Supplementary Figure S.153: Inferred subclones mapped to WGS trees. Shown are the



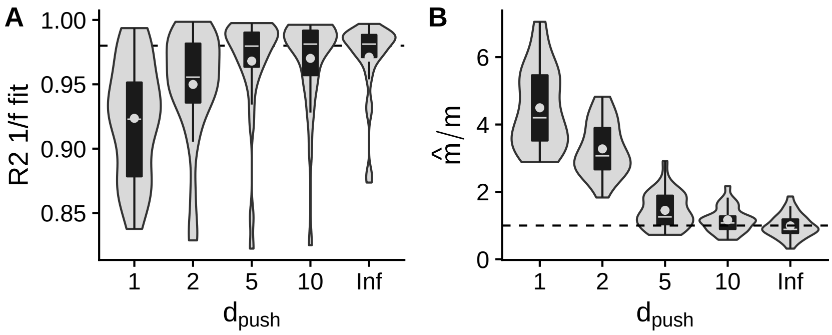
Supplementary Figure S.154: Inferred subclones mapped to ML LP-WGS trees.



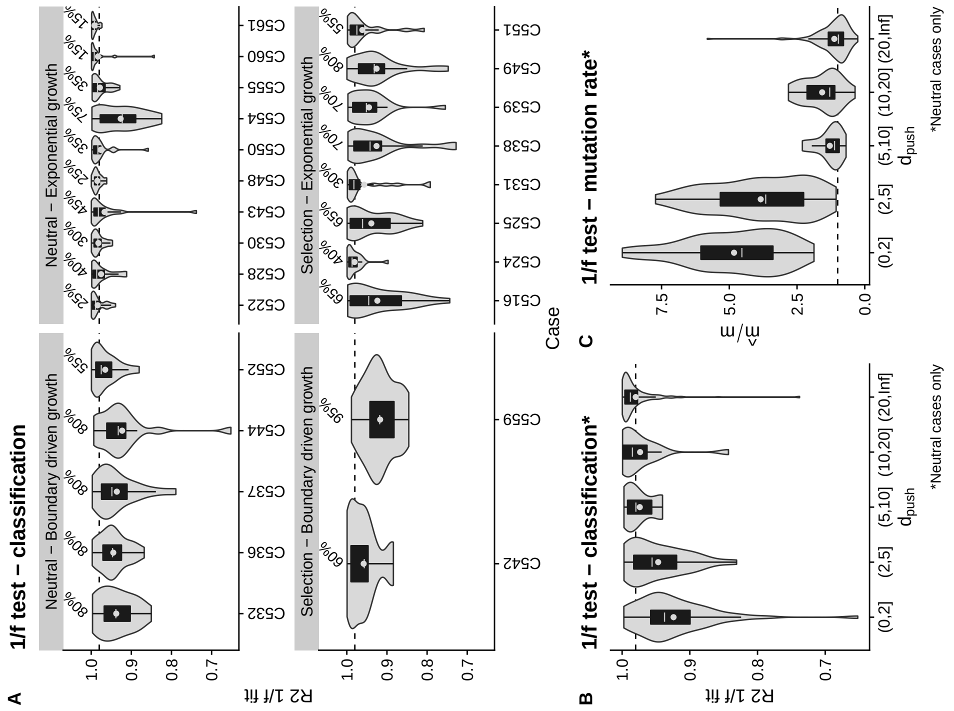
Supplementary Figure S.155: Effect of the overdispersion parameter on tree shapes. Each row shows four trees generated from the same target tree under different amounts of overdispersion α.



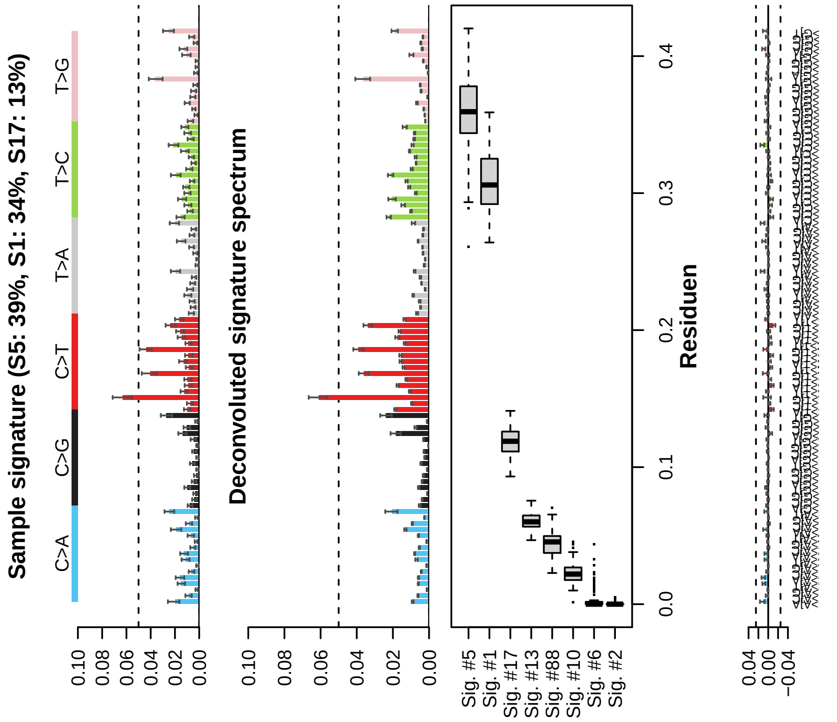
Supplementary Figure S.156: Regional sample positioning in case C518.



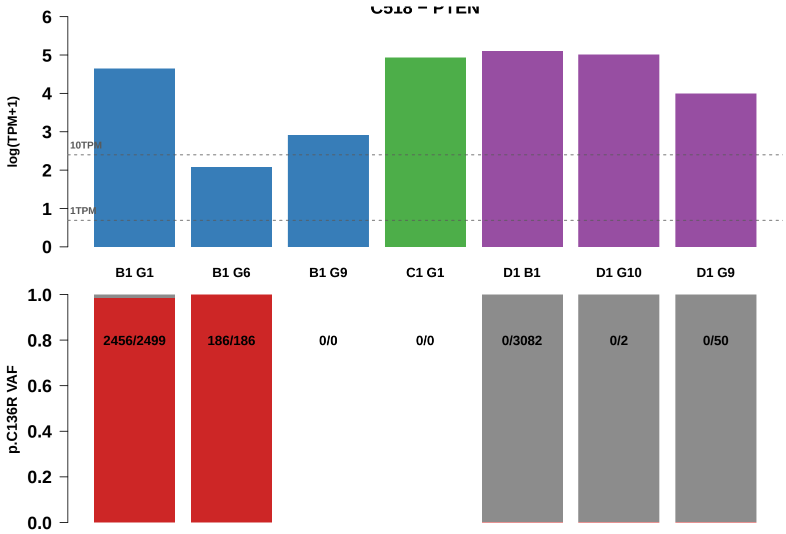
Supplementary Figure S.157: $1/f$ to neutral spatially simulations (2D) with variable degrees of boundary driven growth. A) Spatially limited growth leads to a deviation from the expected 1/f fit from Williams et al. (2016). B) Coefficient of the linear fit causes overestimation of the mutation rates.



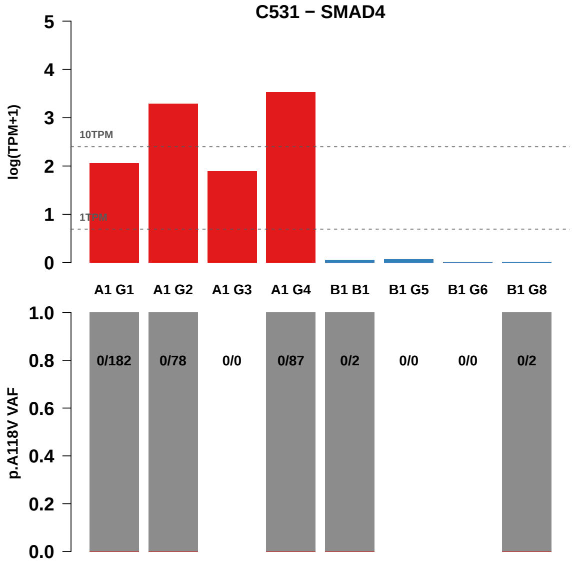
Supplementary Figure S.158: 1/f test applied to simulations from the posterior. For each the 20 best spatial simulations were generated and the 1/f test was applied to simulated global VAF data.



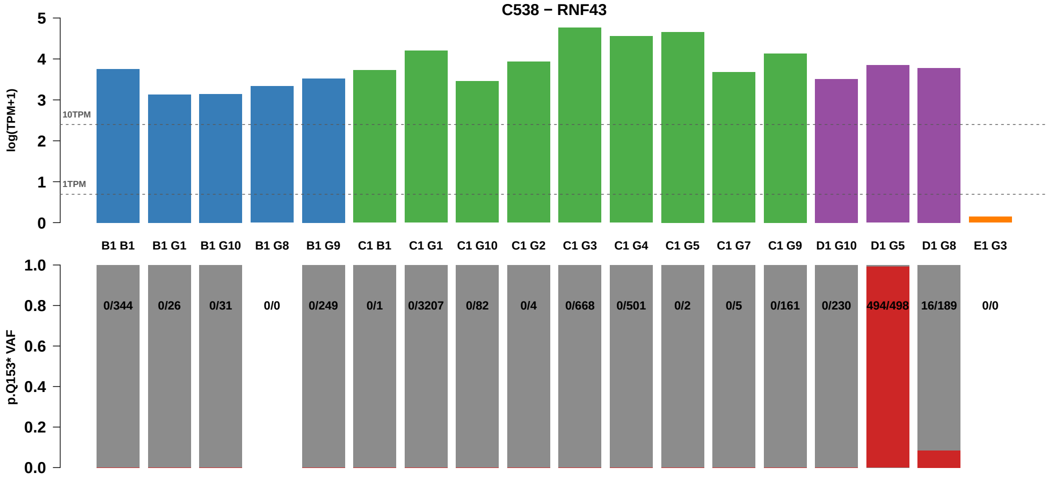
Supplementary Figure S.159: Mutation signature spectrum of clonal variants in C539.



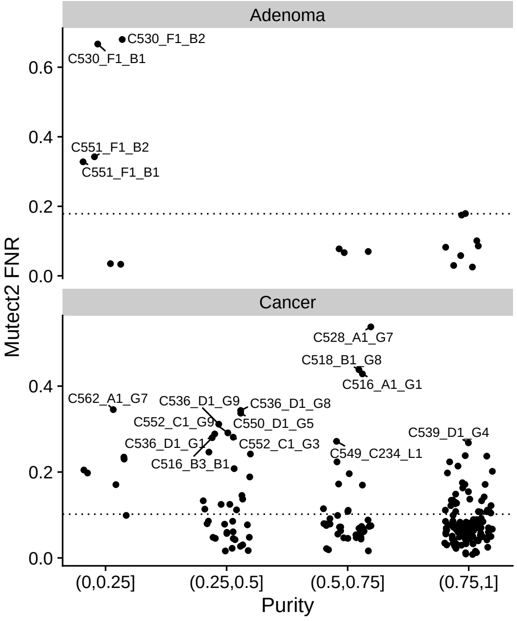
Supplementary Figure S.160: Expression analysis of PTEN p.C136R variant in C518 demonstrating the complete loss of the alternate PTEN allele in samples from region A&B.



Supplementary Figure S.161: Expression analysis of SMAD4 demonstrating the complete loss of SMAD4 expression in samples from region B, which harboured a subclonal p.A118V mutation.

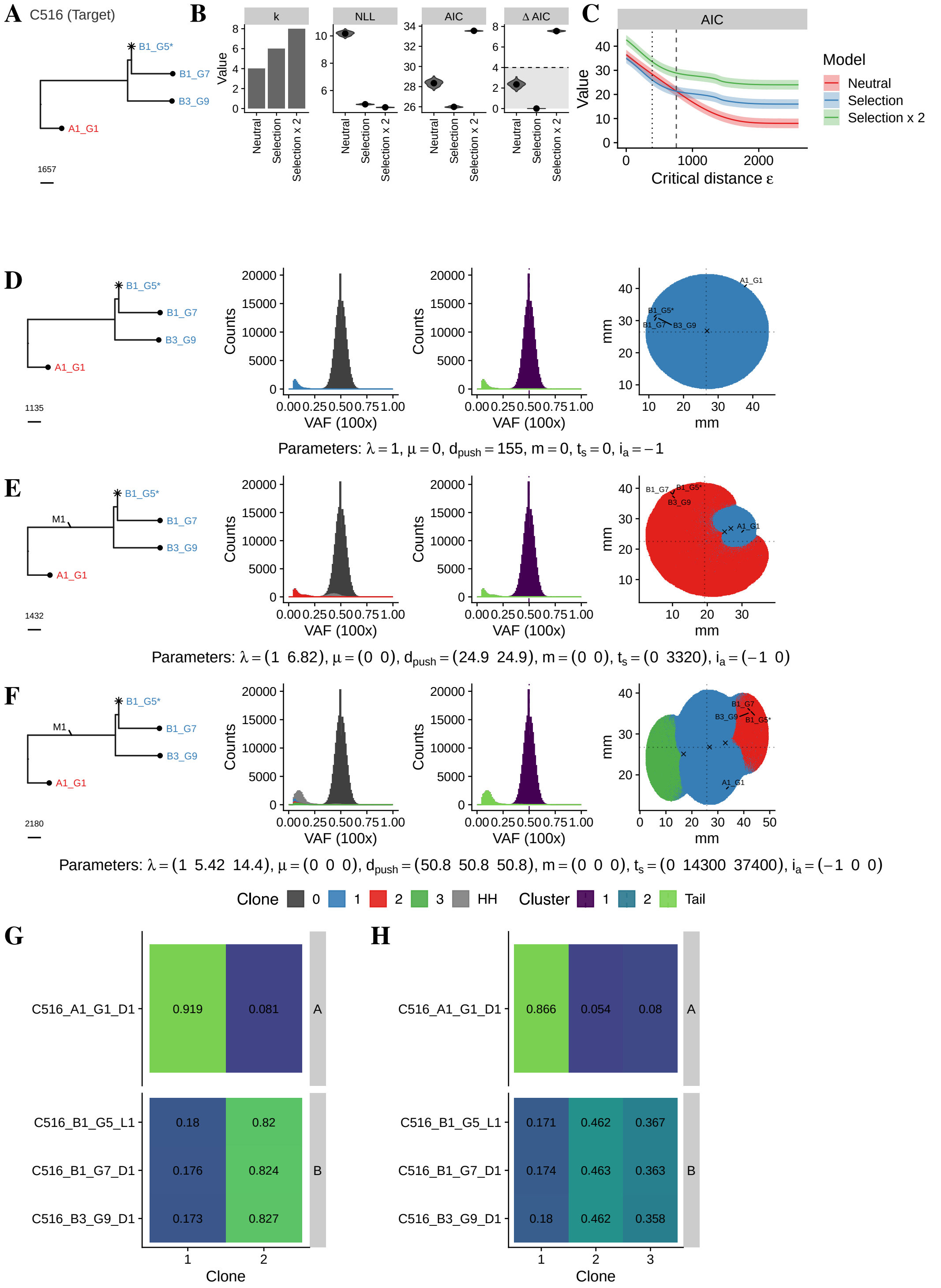


Supplementary Figure S.162: Expression of RNF43 in case C538.

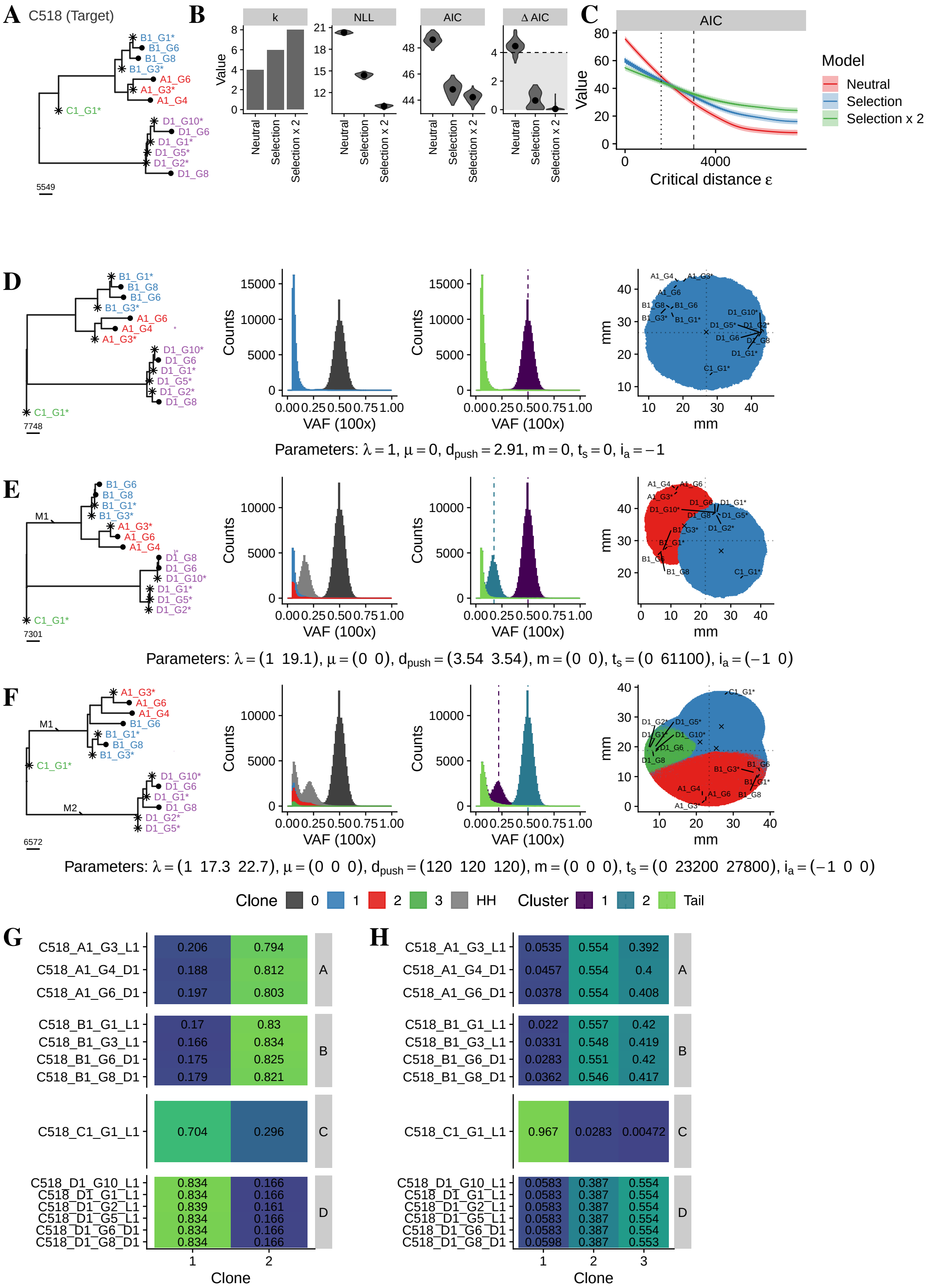


Supplementary Figure S.163: Estimated false negative rates (FNR) of Mutect2. For the estimation of the FNR of the Mutect2 variant calling in each sample (dots), variants present in all samples were identified and the fraction of these present in the filtered Mutect2 VCF files of each sample were determined.

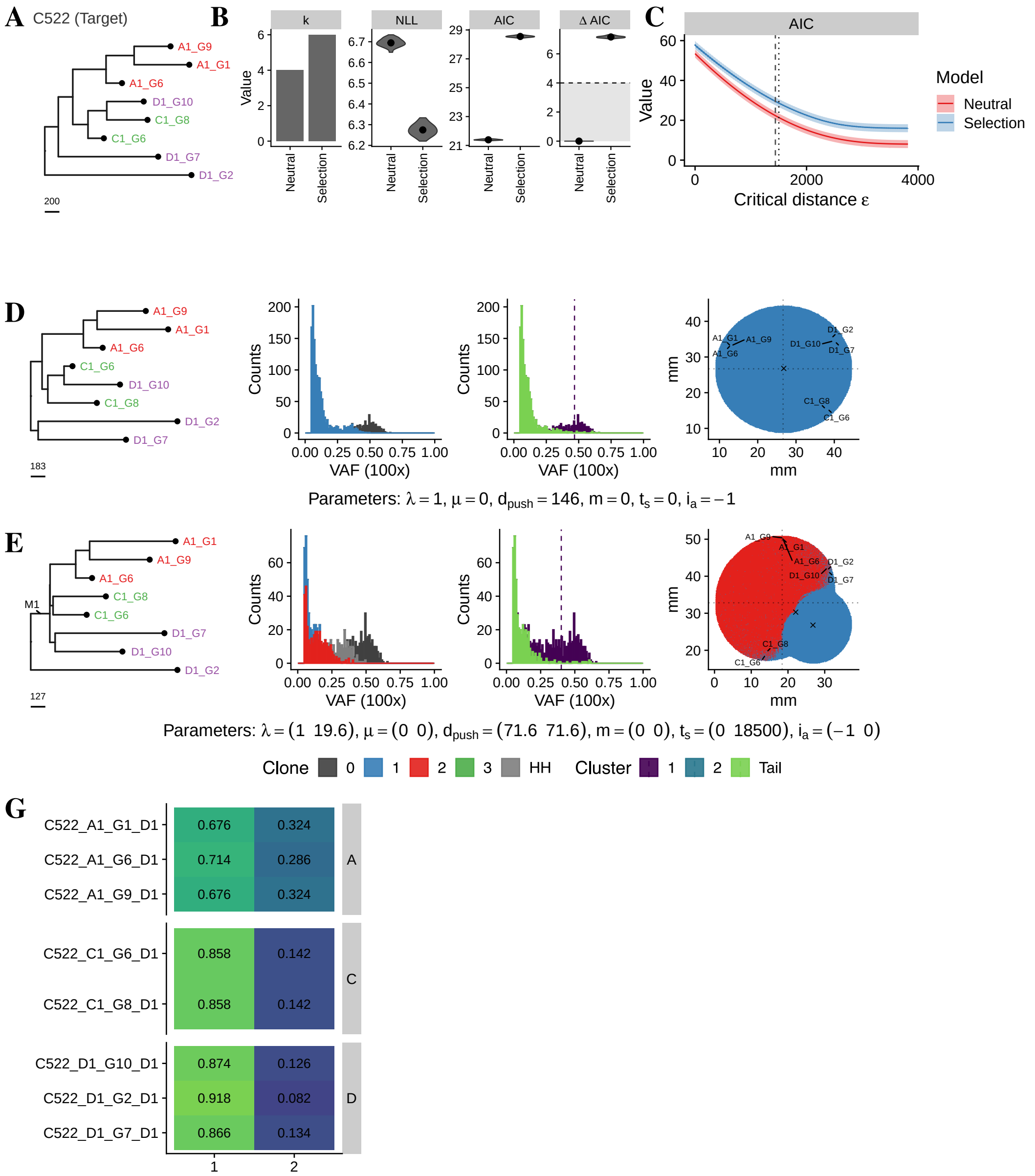
S.6 SMC-ABC Fits



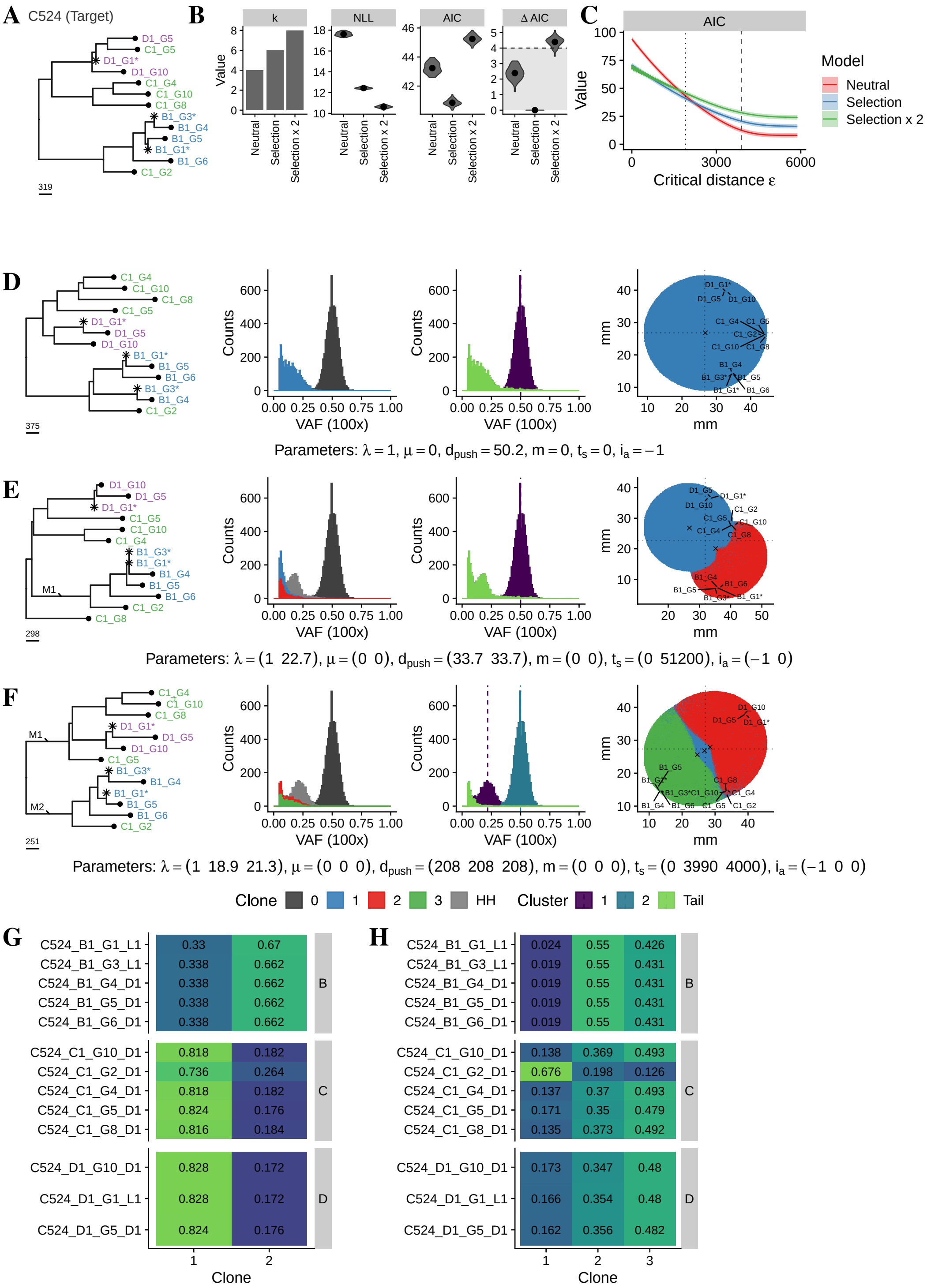
Supplementary Figure S.164: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C516. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



Supplementary Figure S.165: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C518. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



Supplementary Figure S.166: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C522. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.

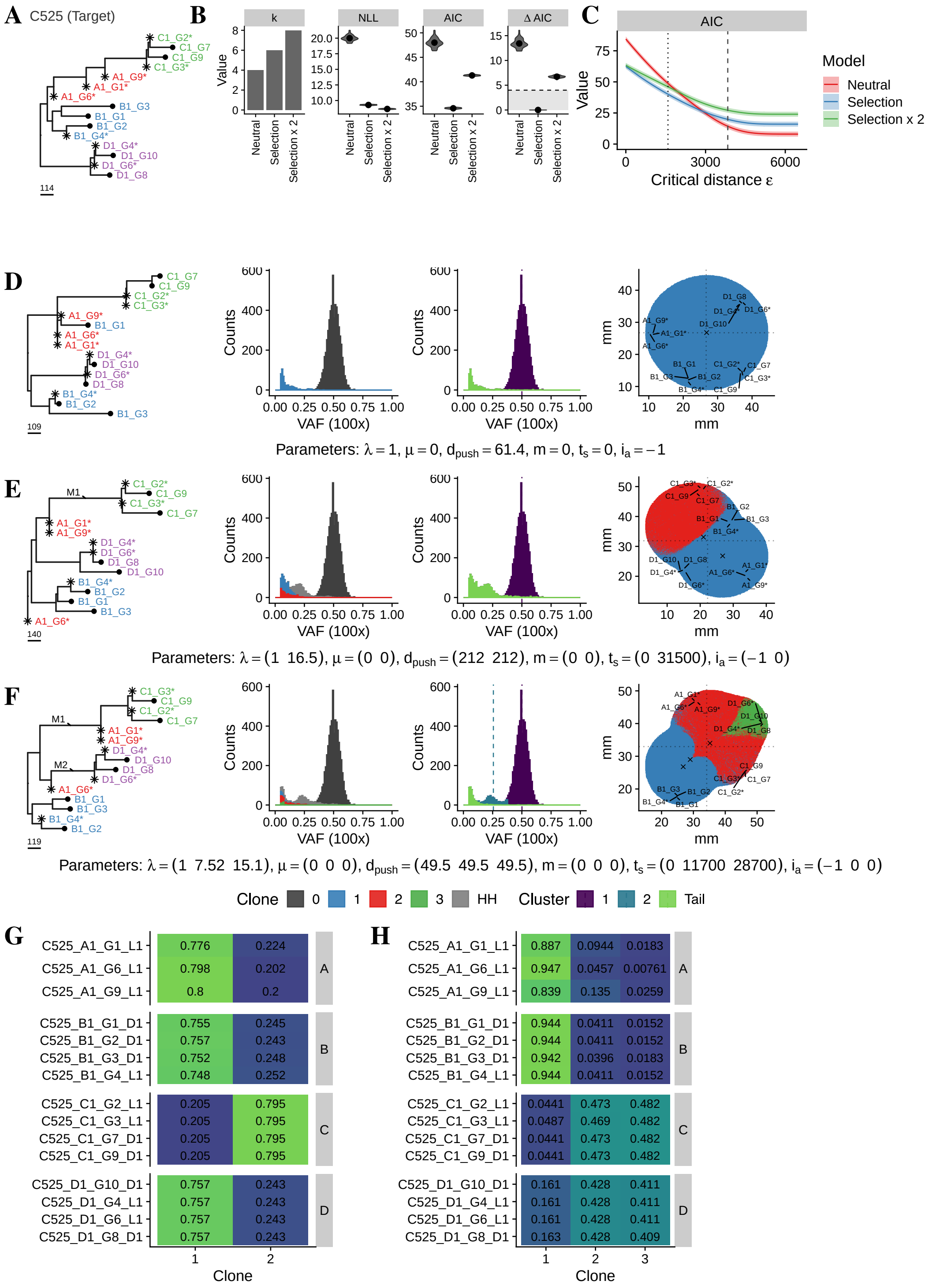


Parameters: $\lambda = 1$, $\mu = 0$, $d_{push} = 50.2$, $m = 0$, $t_s = 0$, $i_a = -1$

Parameters: $\lambda = (1\ \ 22.7)$, $\mu = (0\ \ 0)$, $d_{push} = (33.7\ \ 33.7)$, $m = (0\ \ 0)$, $t_s = (0\ \ 51200)$, $i_a = (-1\ \ 0)$

Parameters: $\lambda = (1\ \ 18.9\ \ 21.3)$, $\mu = (0\ \ 0\ \ 0)$, $d_{push} = (208\ \ 208\ \ 208)$, $m = (0\ \ 0\ \ 0)$, $t_s = (0\ \ 3990\ \ 4000)$, $i_a = (-1\ \ 0\ \ 0)$

Supplementary Figure S.167: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C524. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the ‘Neutral’, ‘Selection’ and ‘Selection x 2’ model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the ‘Selection’ and ‘Selection x 2’ model respectively. Figure parts F) and H) can be missing if the ‘Selection x 2’ model was not considered.

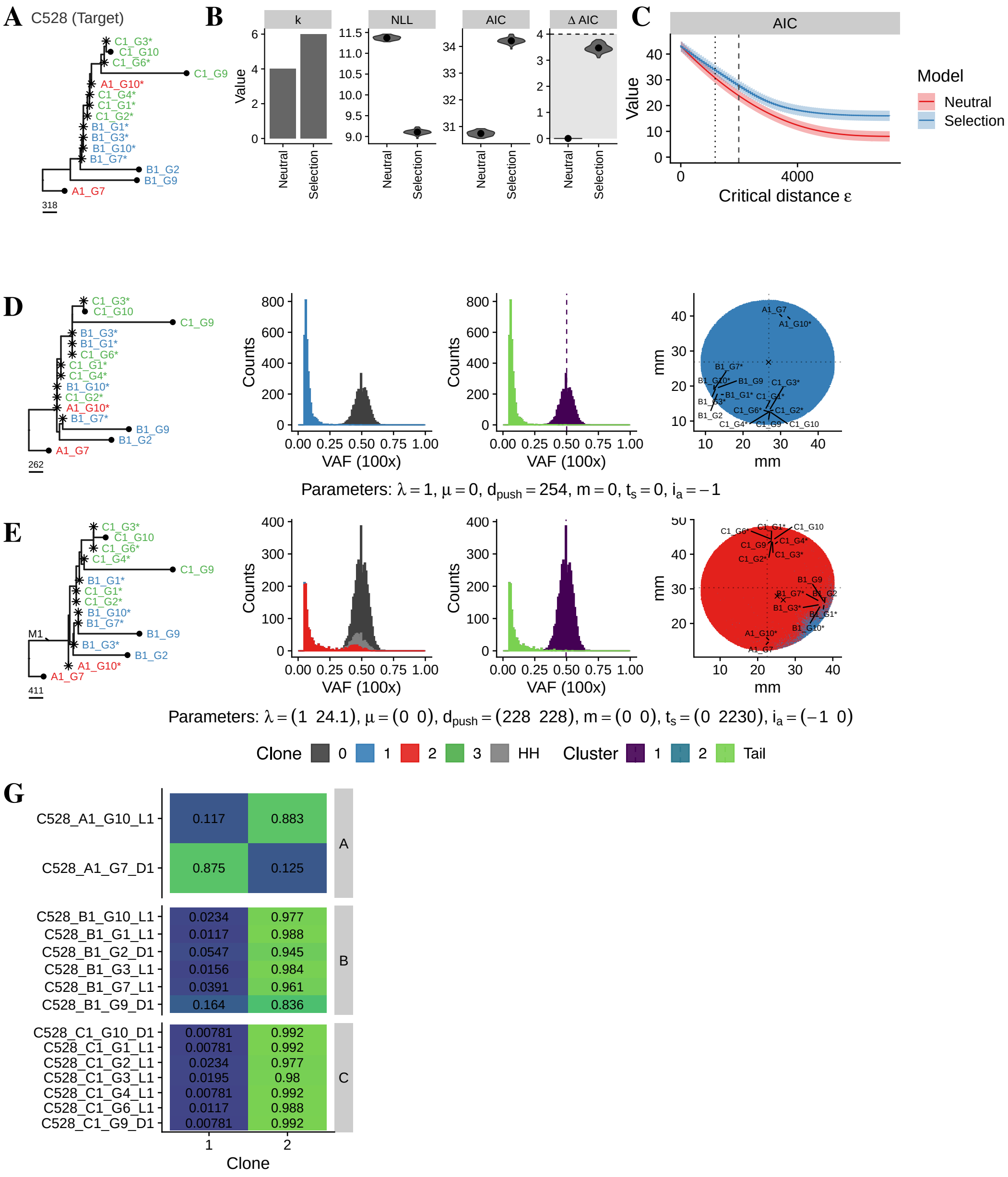


Parameters: $\lambda = 1$, $\mu = 0$, $d_{push} = 61.4$, $m = 0$, $t_s = 0$, $i_a = -1$

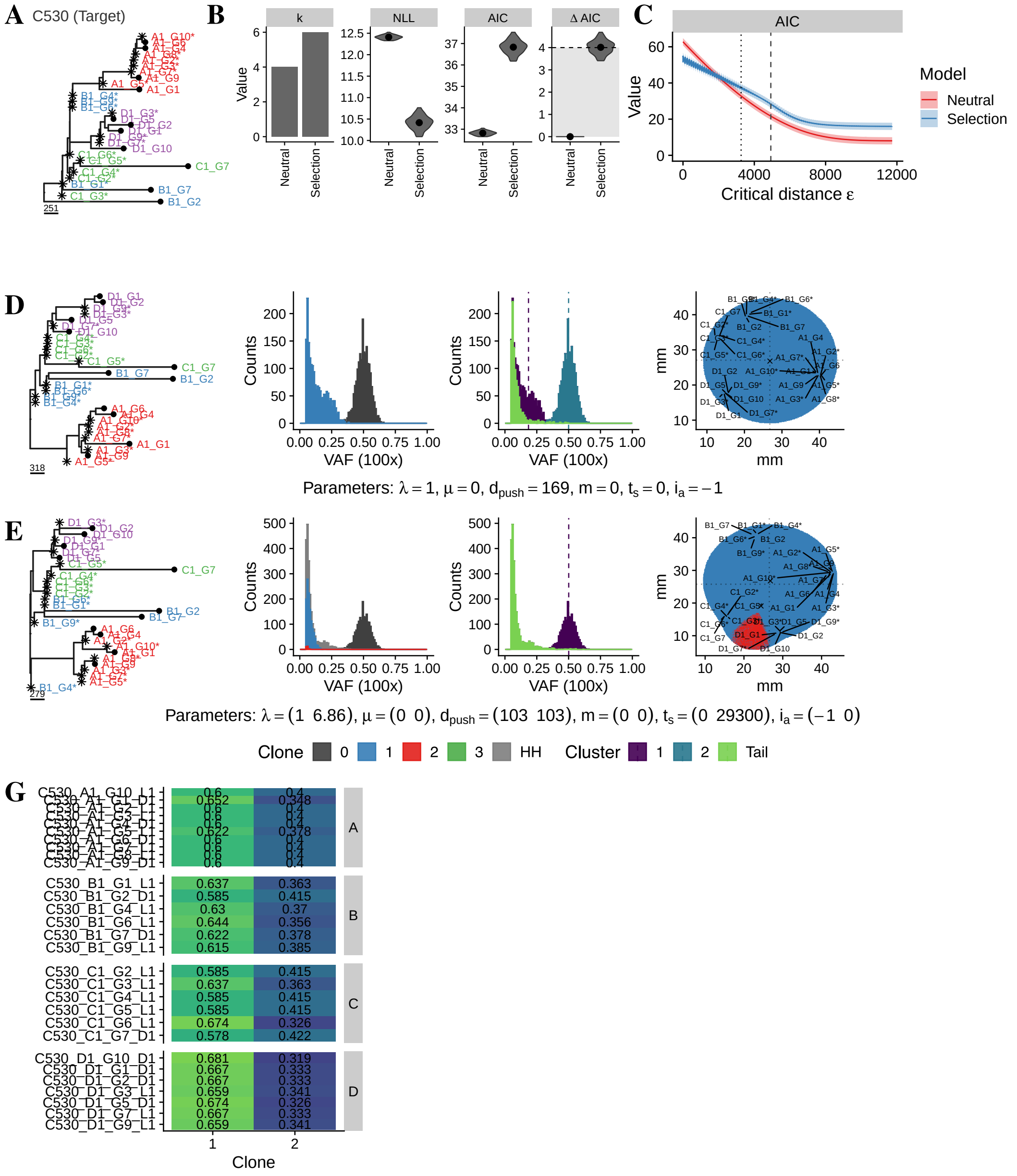
Parameters: $\lambda = (1\ \ 16.5)$, $\mu = (0\ \ 0)$, $d_{push} = (212\ \ 212)$, $m = (0\ \ 0)$, $t_s = (0\ \ 31500)$, $i_a = (-1\ \ 0)$

Parameters: $\lambda = (1\ \ 7.52\ \ 15.1)$, $\mu = (0\ \ 0\ \ 0)$, $d_{push} = (49.5\ \ 49.5\ \ 49.5)$, $m = (0\ \ 0\ \ 0)$, $t_s = (0\ \ 11700\ \ 28700)$, $i_a = (-1\ \ 0\ \ 0)$

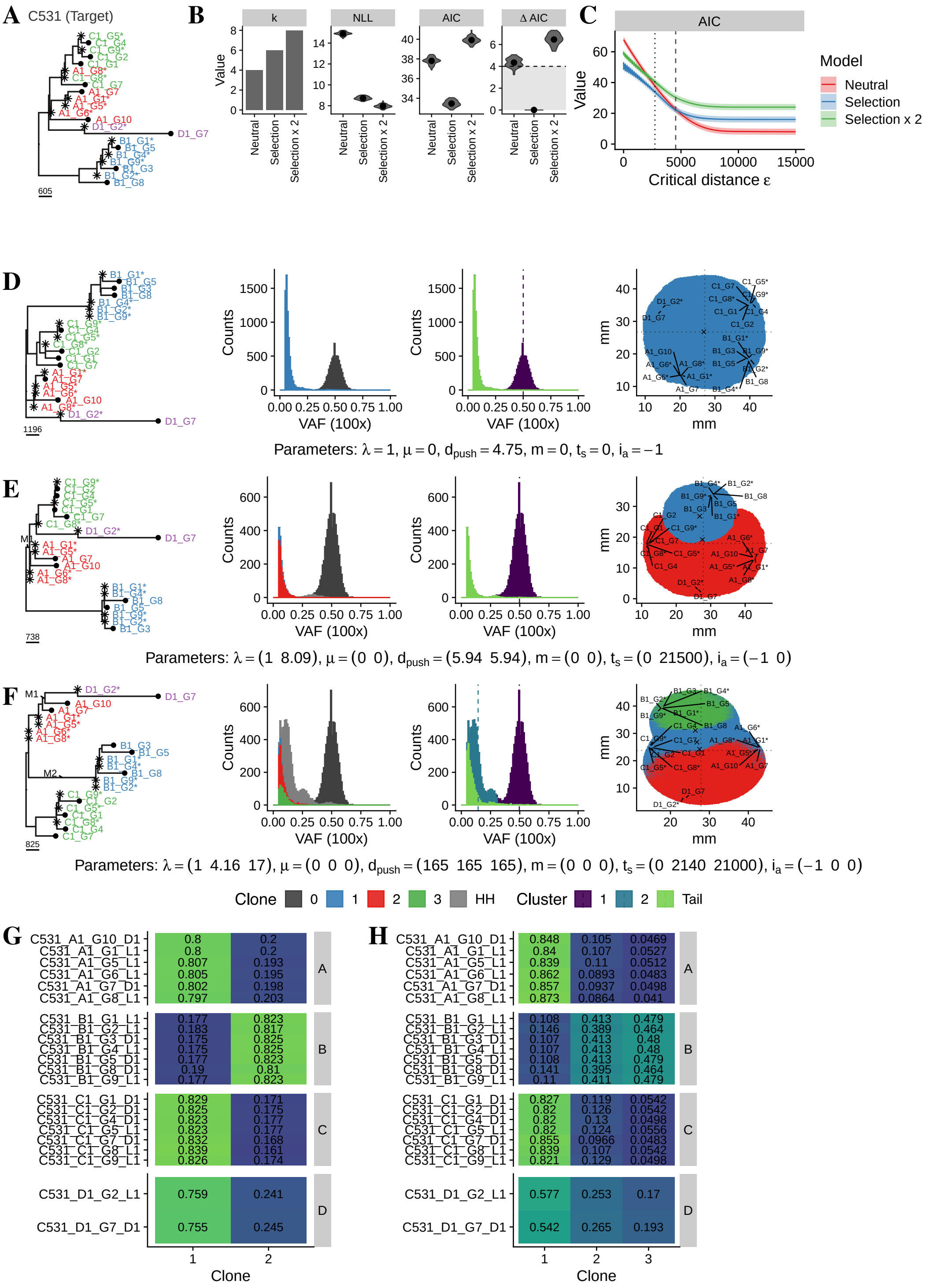
Supplementary Figure S.168: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C525. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the ‘Neutral’, ‘Selection’ and ‘Selection x 2’ model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the ‘Selection’ and ‘Selection x 2’ model respectively. Figure parts F) and H) can be missing if the ‘Selection x 2’ model was not considered.



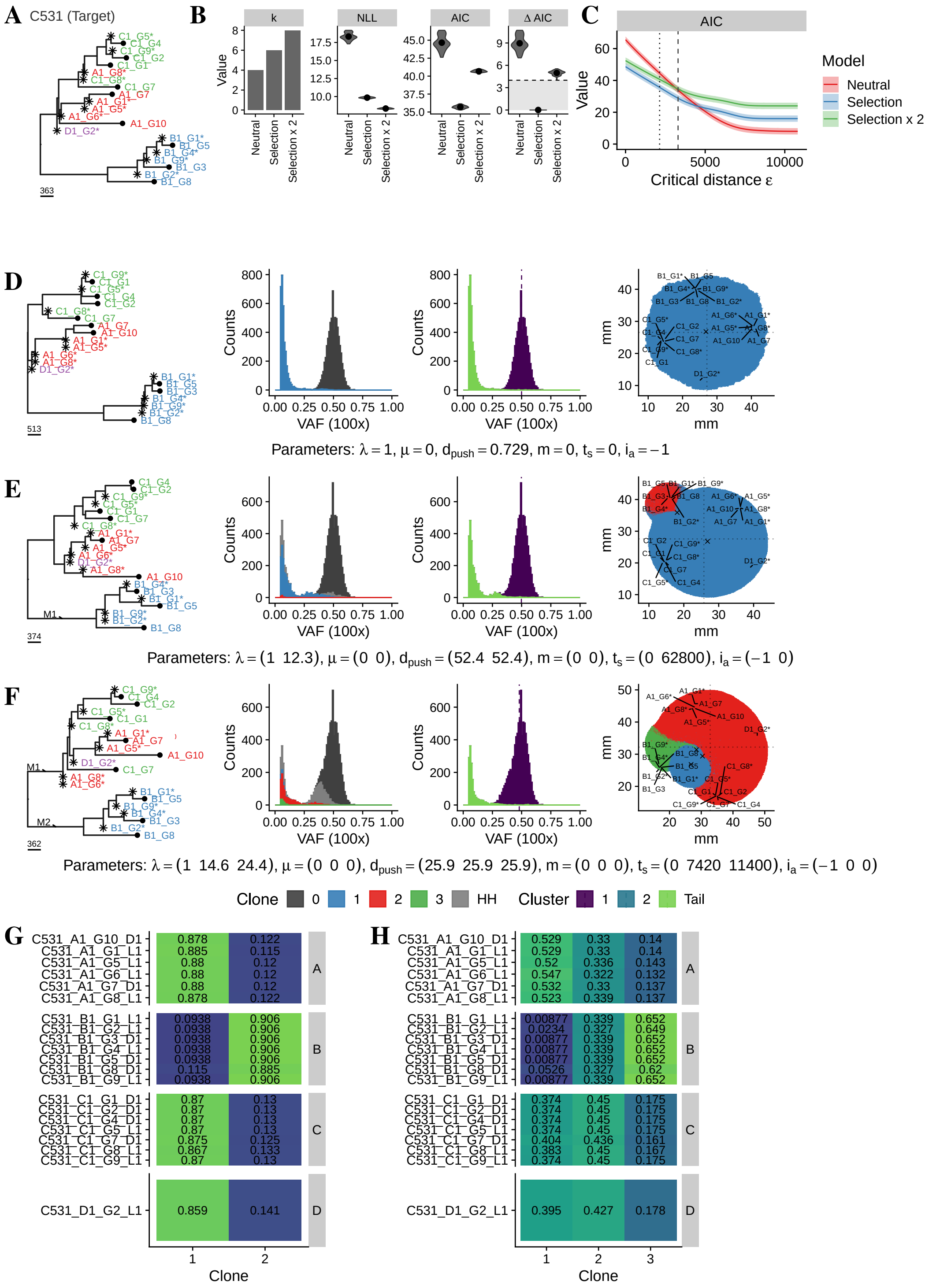
Supplementary Figure S.169: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C528. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the ‘Neutral’, ‘Selection’ and ‘Selection x 2’ model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the ‘Selection’ and ‘Selection x 2’ model respectively. Figure parts F) and H) can be missing if the ‘Selection x 2’ model was not considered.



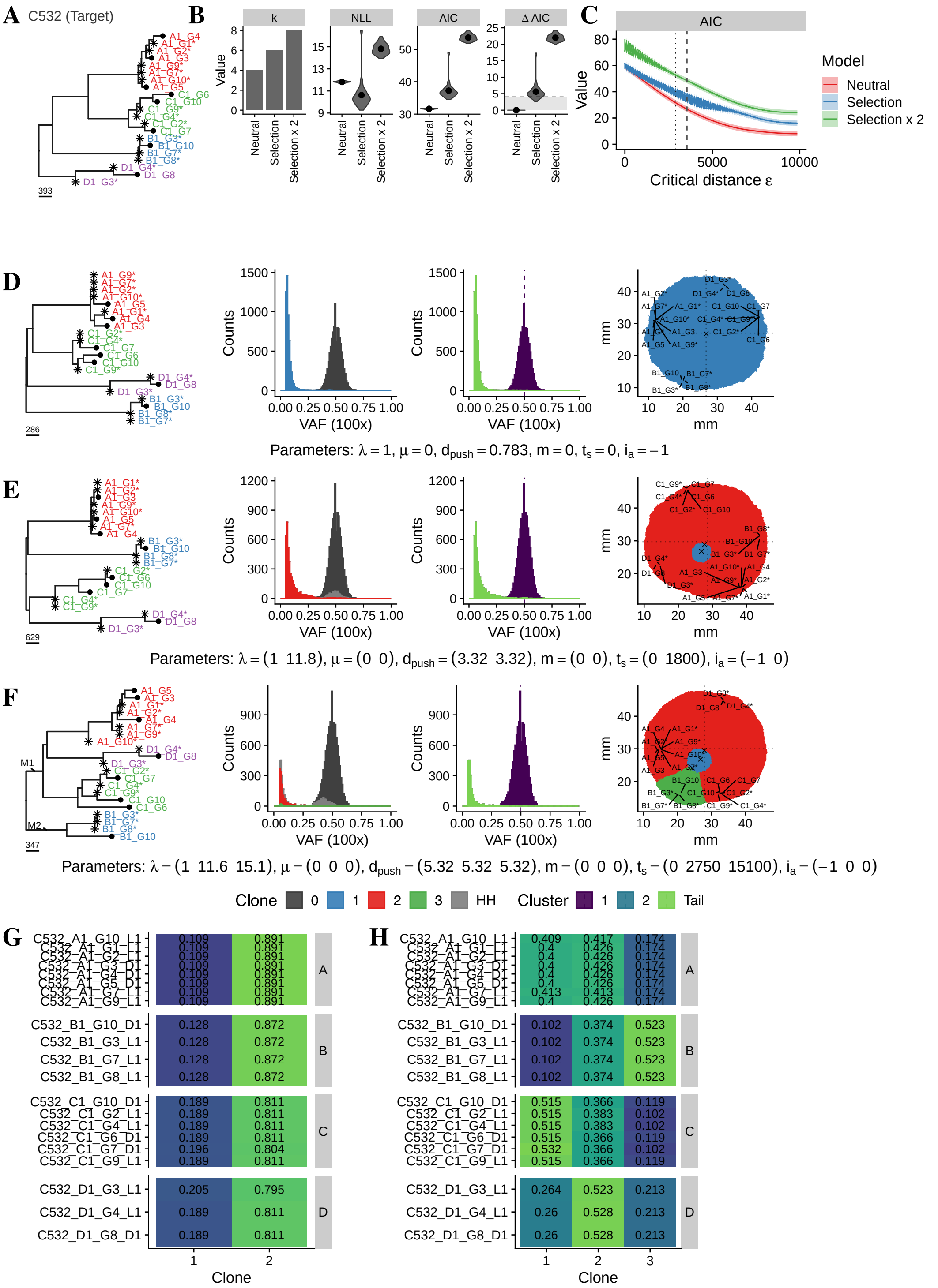
Supplementary Figure S.170: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C530. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



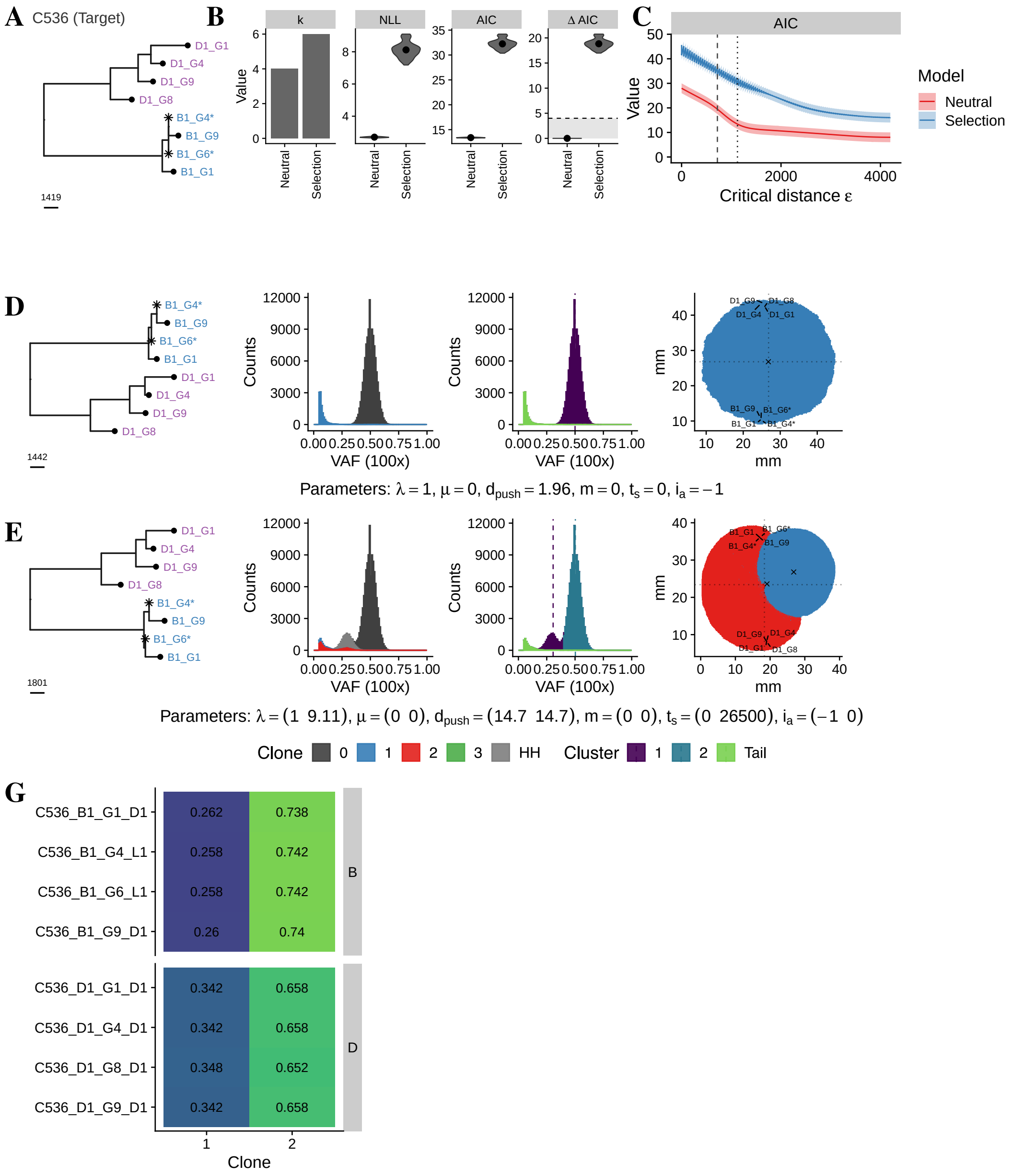
Supplementary Figure S.171: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C531. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



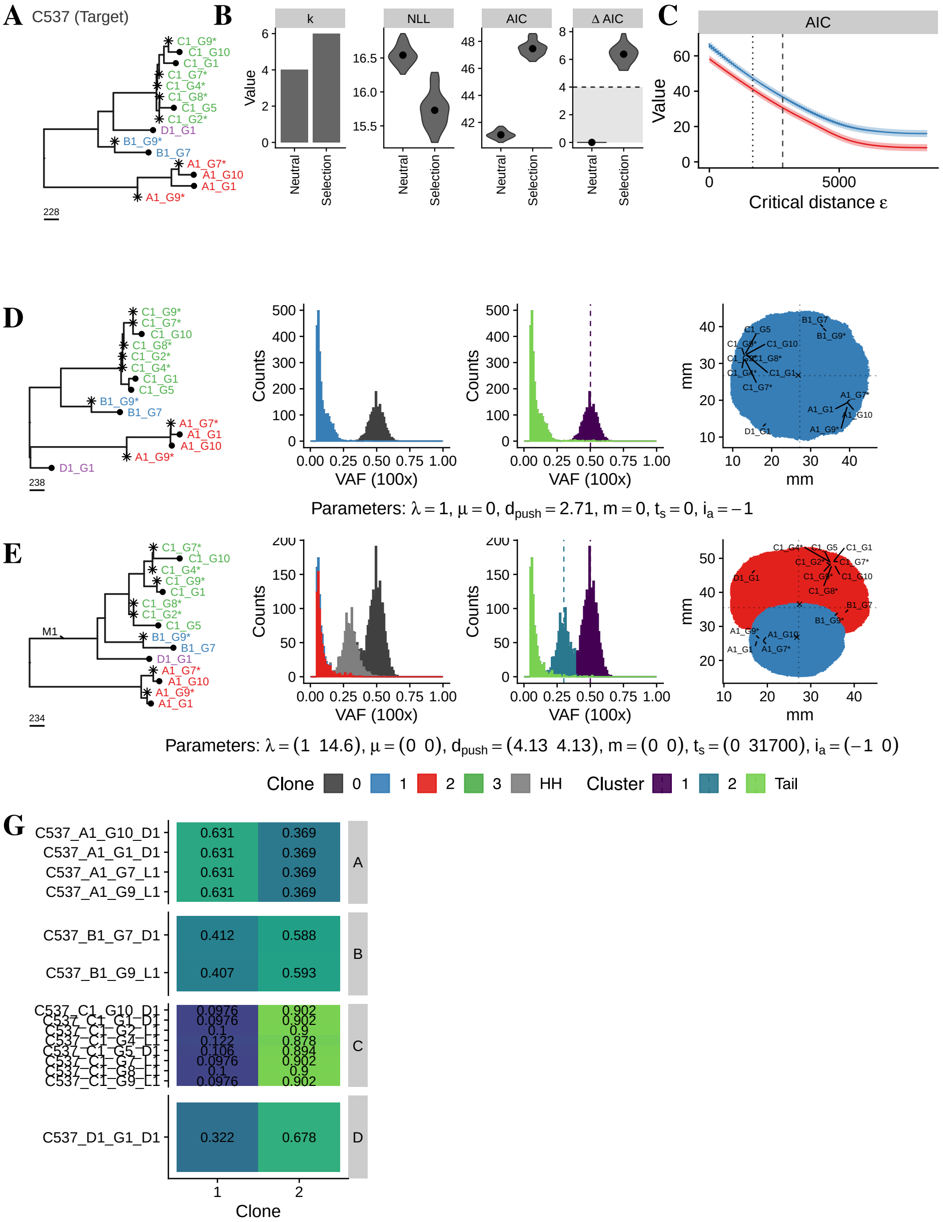
Supplementary Figure S.172: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C531 (excl. D1-G7). A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of *ε*. D), E) and F) The best fitting simulated neutral tree for the ‘Neutral’, ‘Selection’ and ‘Selection x 2’ model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the ‘Selection’ and ‘Selection x 2’ model respectively. Figure parts F) and H) can be missing if the ‘Selection x 2’ model was not considered.



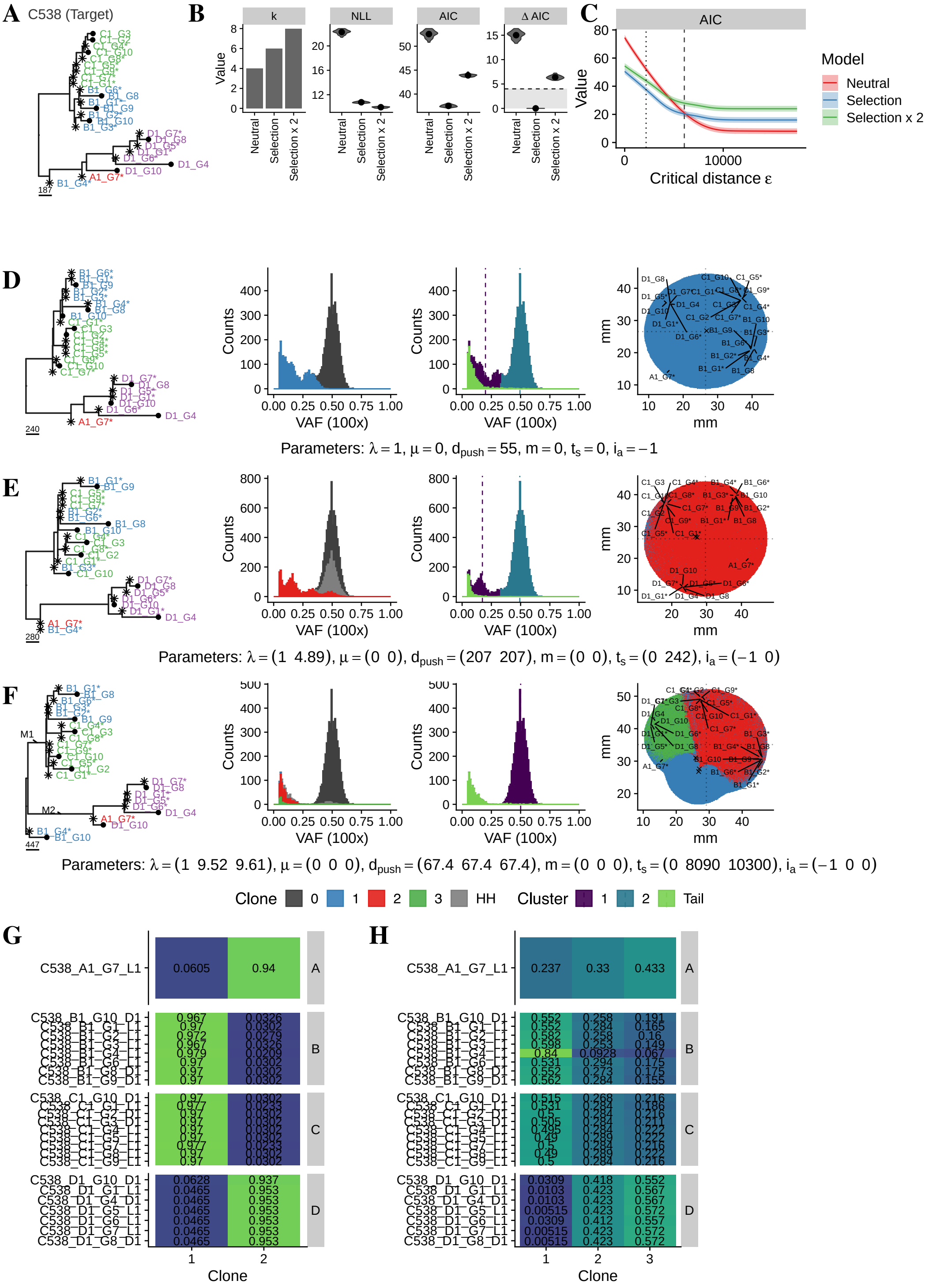
Supplementary Figure S.173: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C532. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



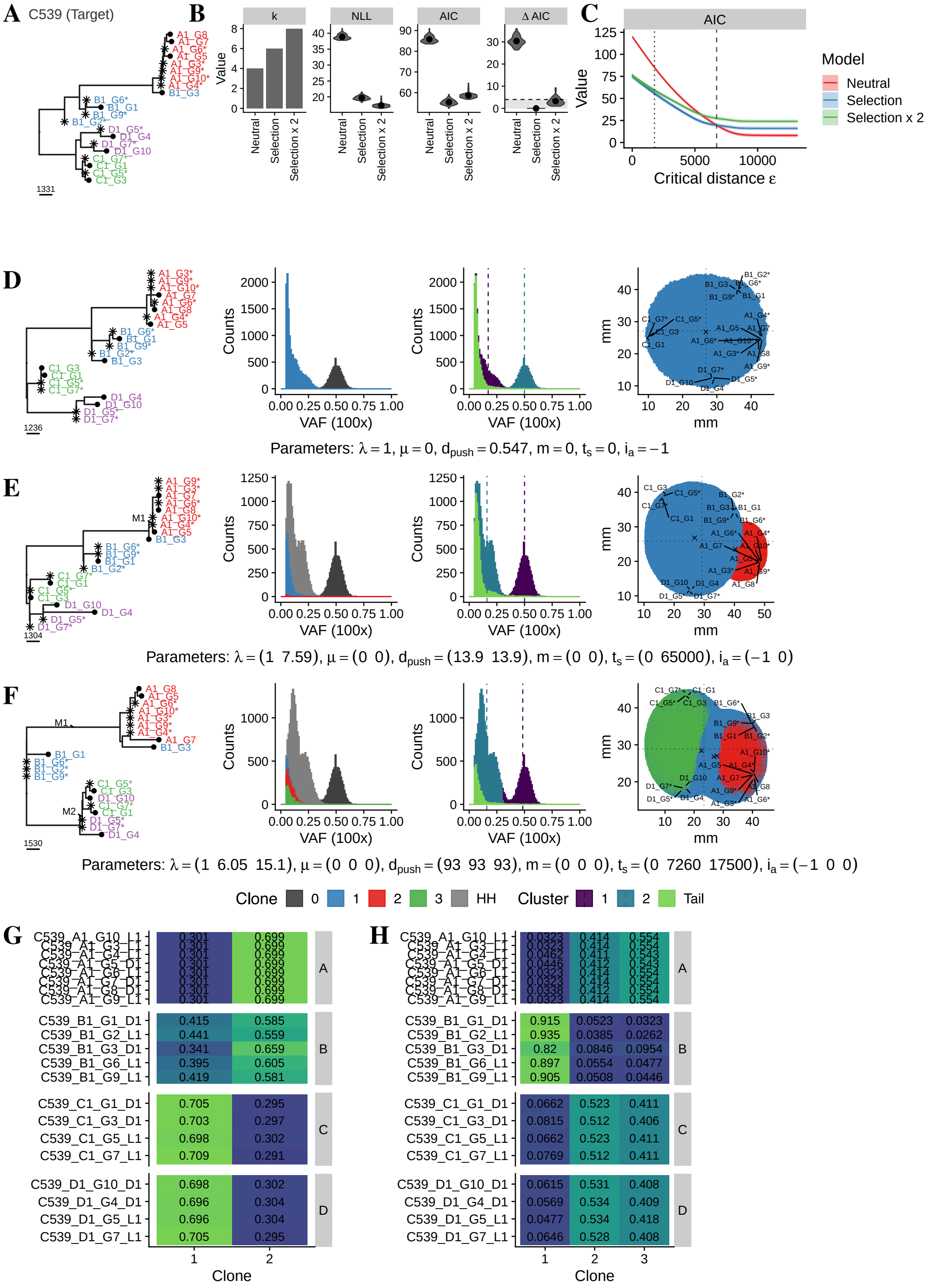
Supplementary Figure S.174: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C536. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



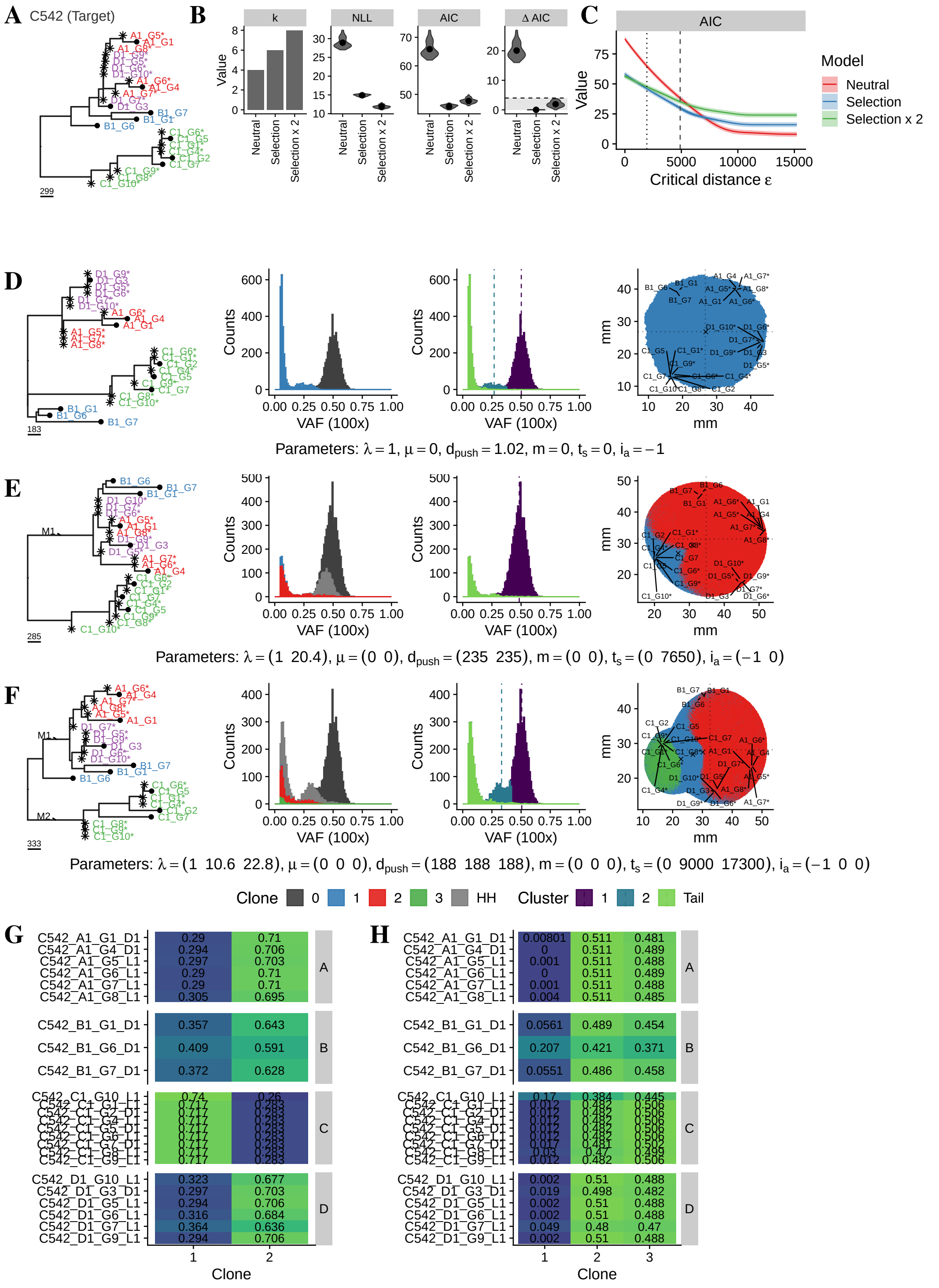
Supplementary Figure S.175: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C537. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



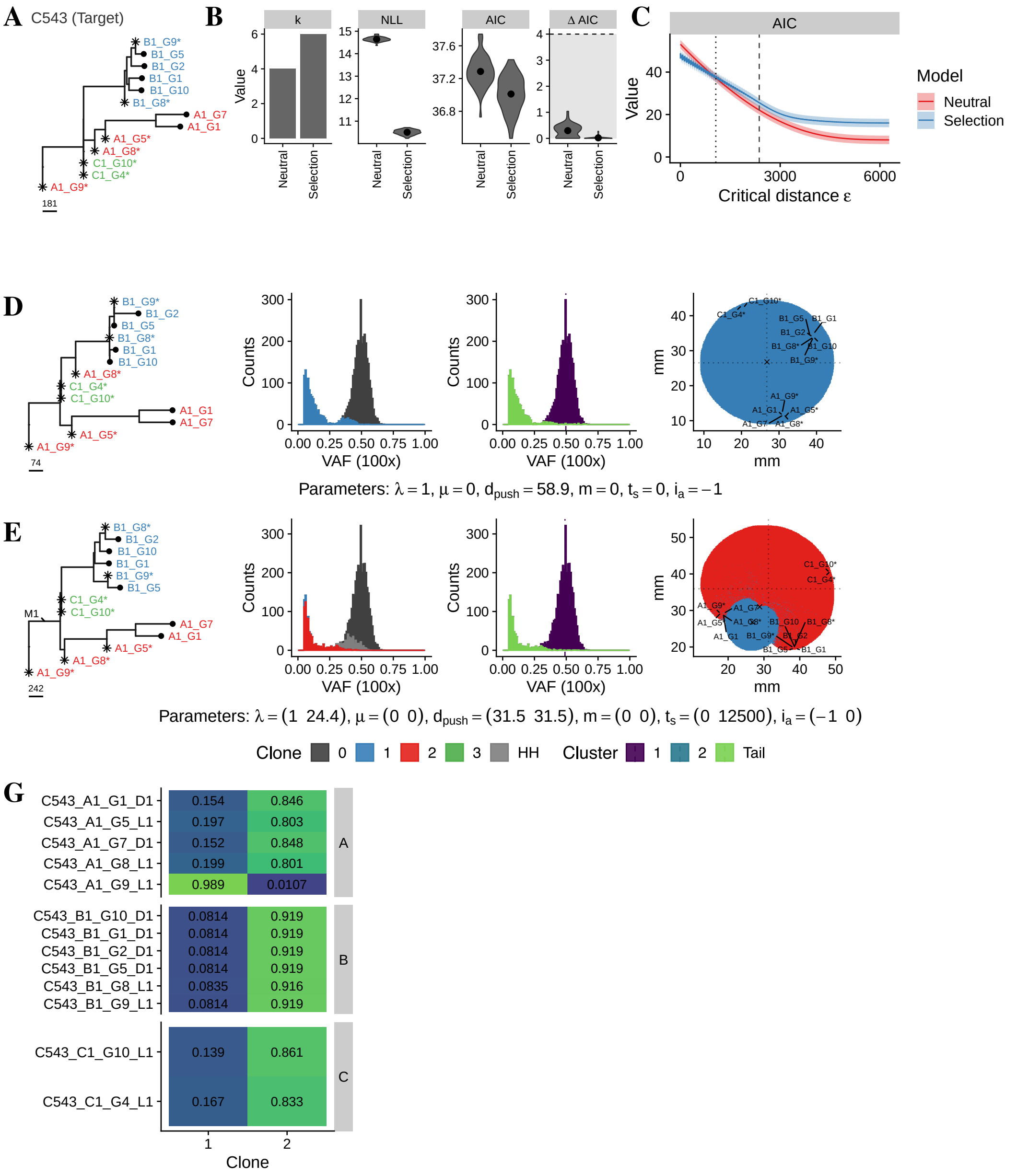
Supplementary Figure S.176: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C538. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



Supplementary Figure S.177: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C539. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



Supplementary Figure S.178: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C542. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the ‘Neutral’, ‘Selection’ and ‘Selection x 2’ model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the ‘Selection’ and ‘Selection x 2’ model respectively. Figure parts F) and H) can be missing if the ‘Selection x 2’ model was not considered.

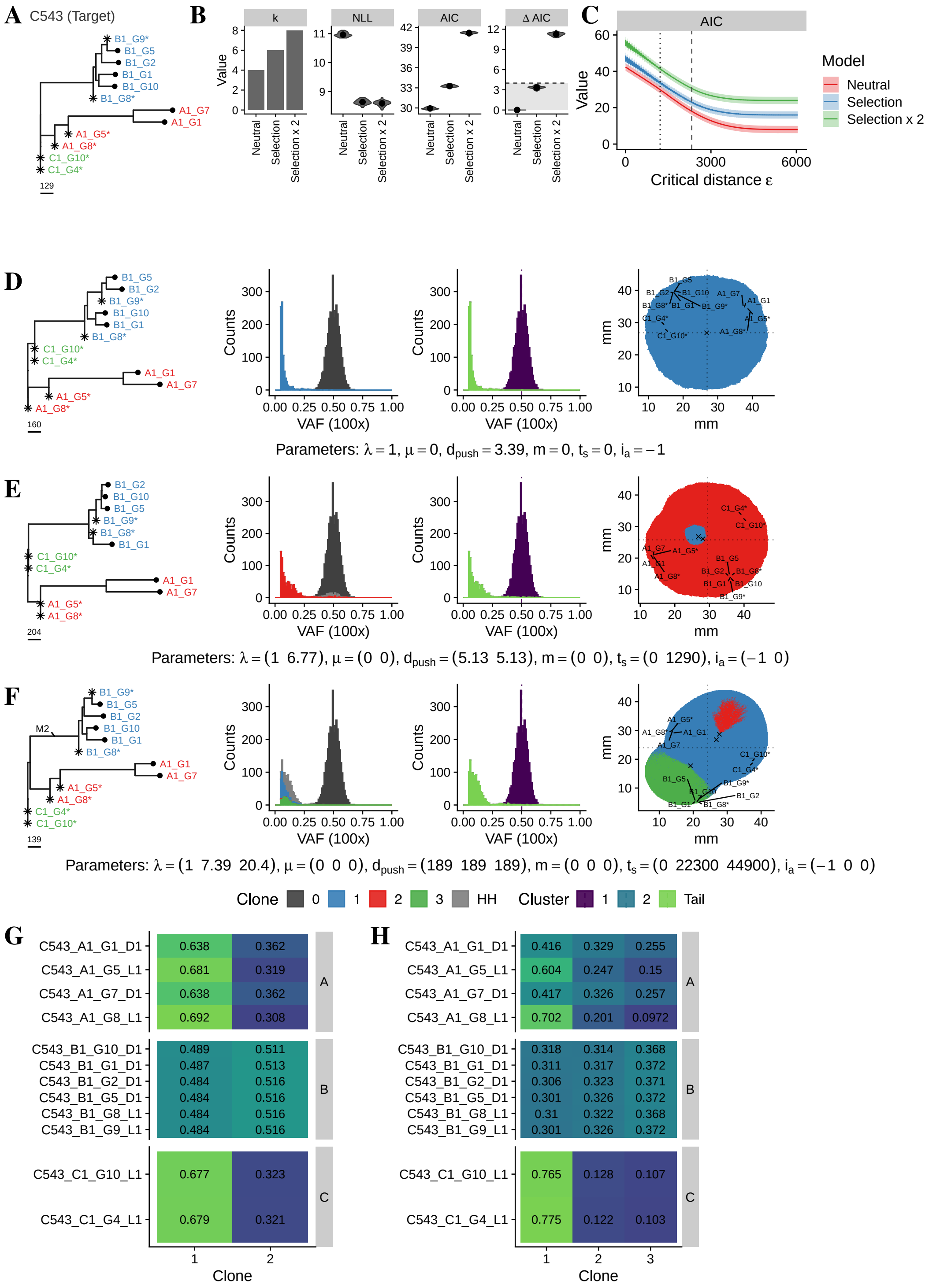


Parameters: $\lambda = 1$, $\mu = 0$, $d_{push} = 58.9$, $m = 0$, $t_s = 0$, $i_a = -1$

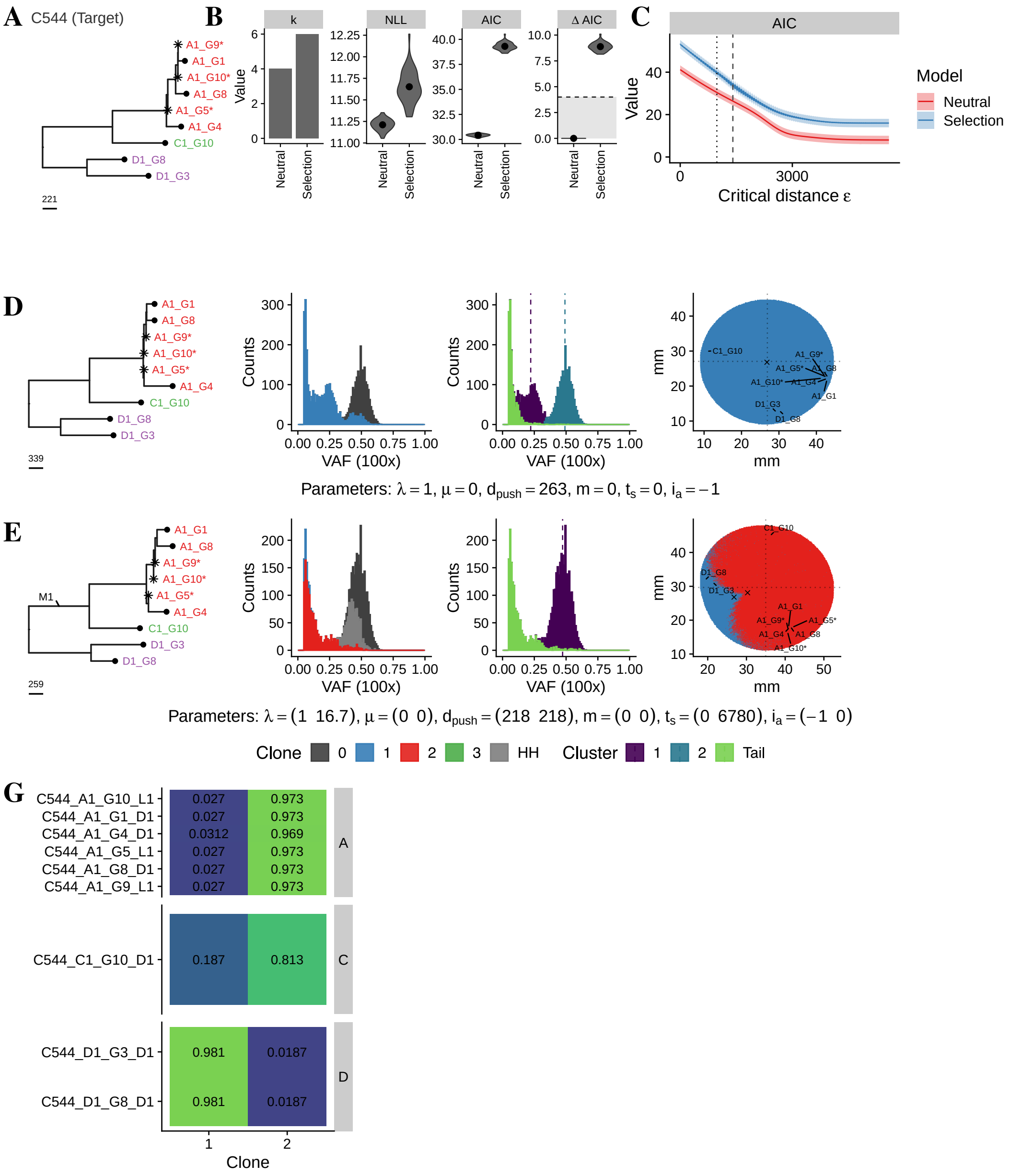
Parameters: $\lambda = (1\ \ 24.4)$, $\mu = (0\ \ 0)$, $d_{push} = (31.5\ \ 31.5)$, $m = (0\ \ 0)$, $t_s = (0\ \ 12500)$, $i_a = (-1\ \ 0)$

Clone ■ 0 ■ 1 ■ 2 ■ 3 ■ HH Cluster ■ 1 ■ 2 ■ Tail

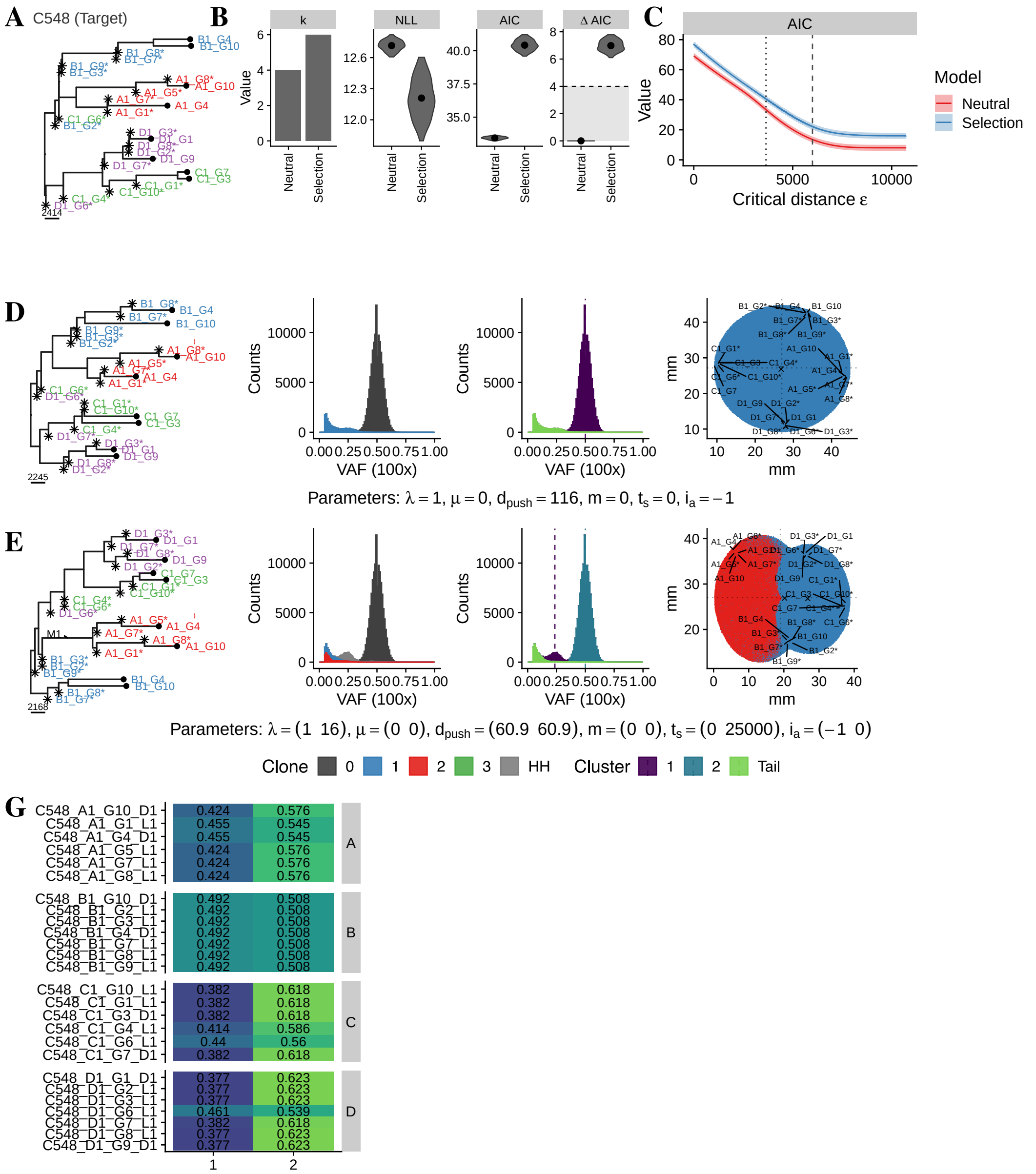
Supplementary Figure S.179: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C543. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the ‘Neutral’, ‘Selection’ and ‘Selection x 2’ model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the ‘Selection’ and ‘Selection x 2’ model respectively. Figure parts F) and H) can be missing if the ‘Selection x 2’ model was not considered.



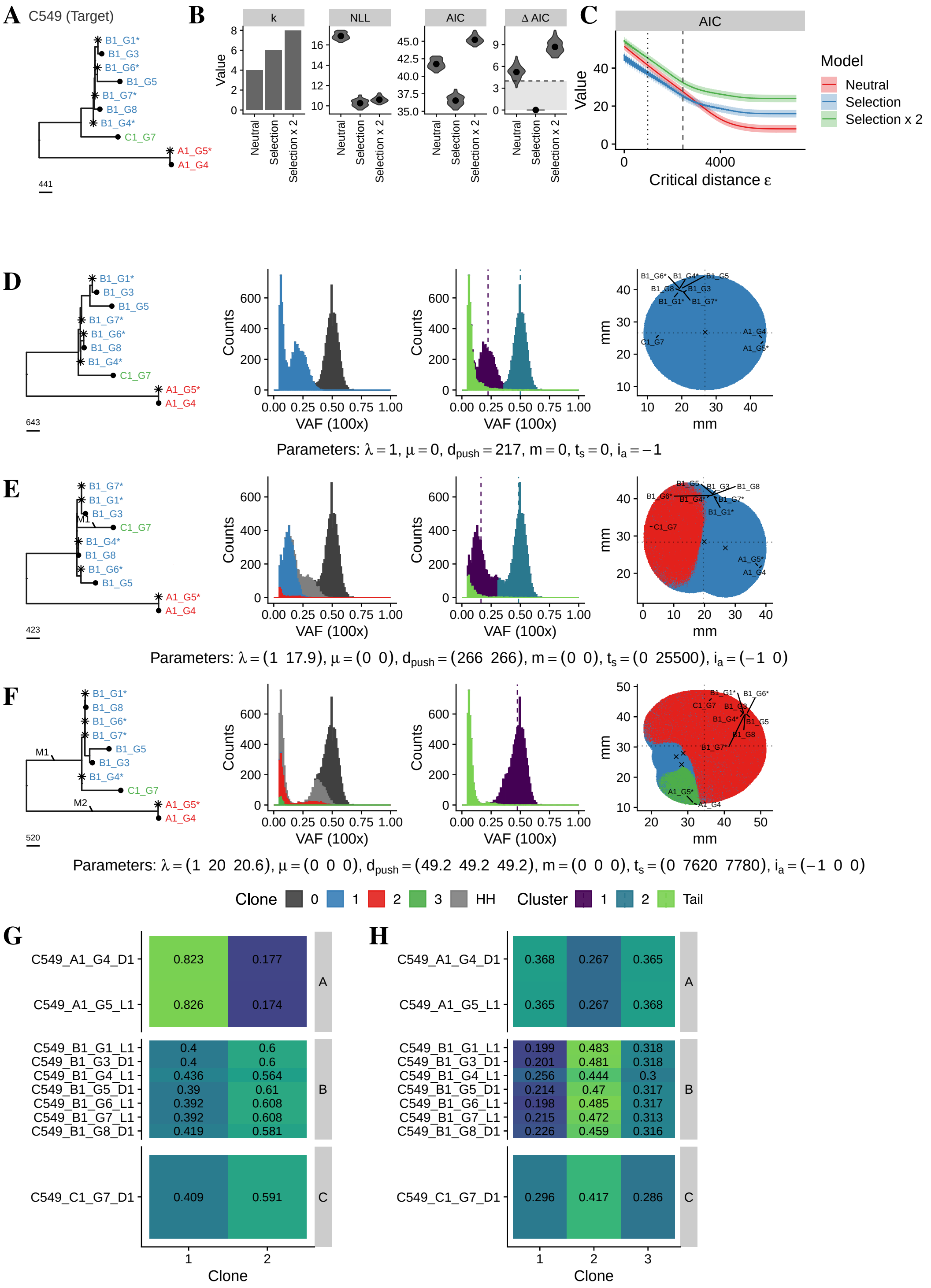
Supplementary Figure S.180: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C543 (excl. A1-G9). A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



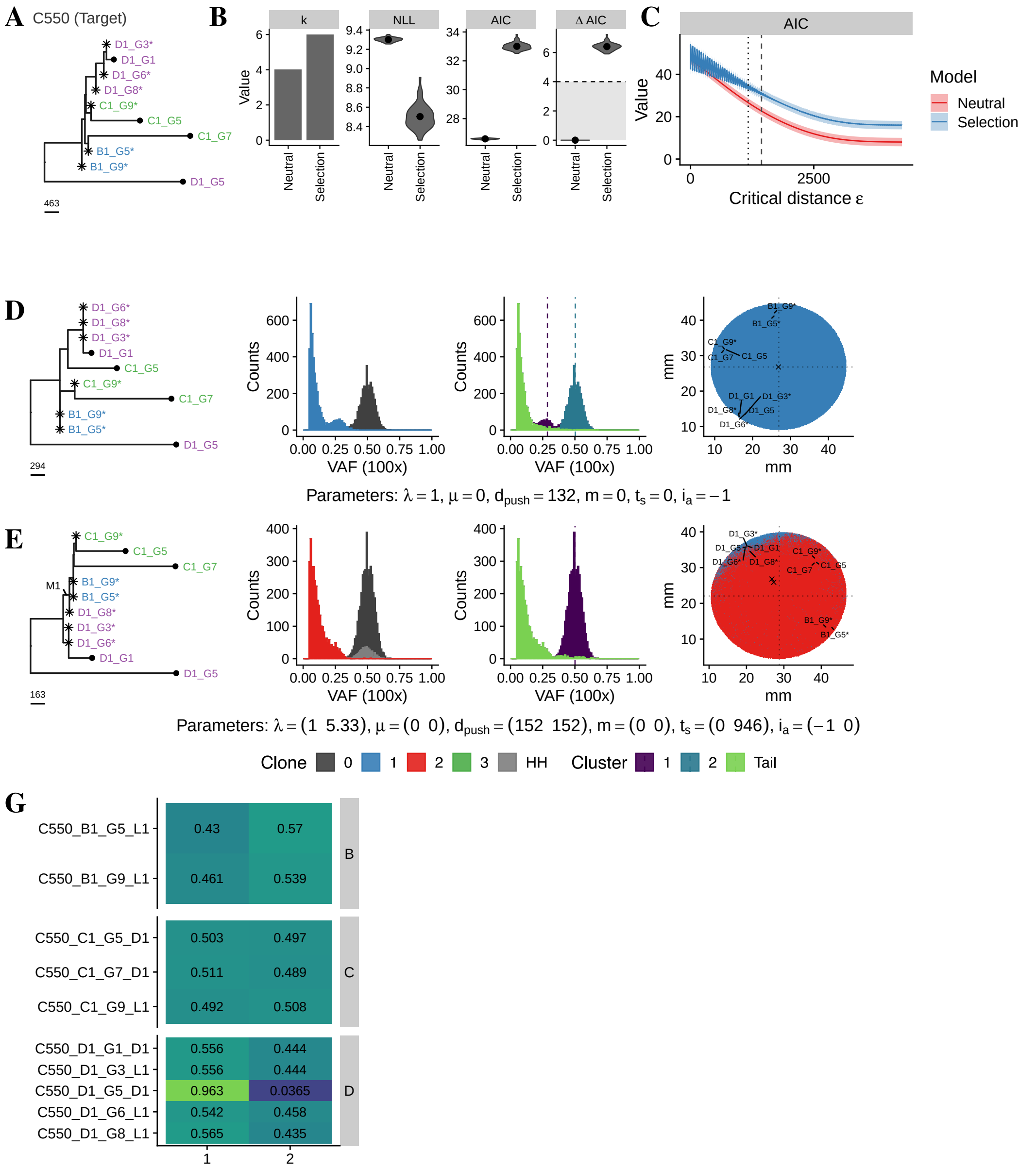
Supplementary Figure S.181: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C544. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the ‘Neutral’, ‘Selection’ and ‘Selection x 2’ model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the ‘Selection’ and ‘Selection x 2’ model respectively. Figure parts F) and H) can be missing if the ‘Selection x 2’ model was not considered.



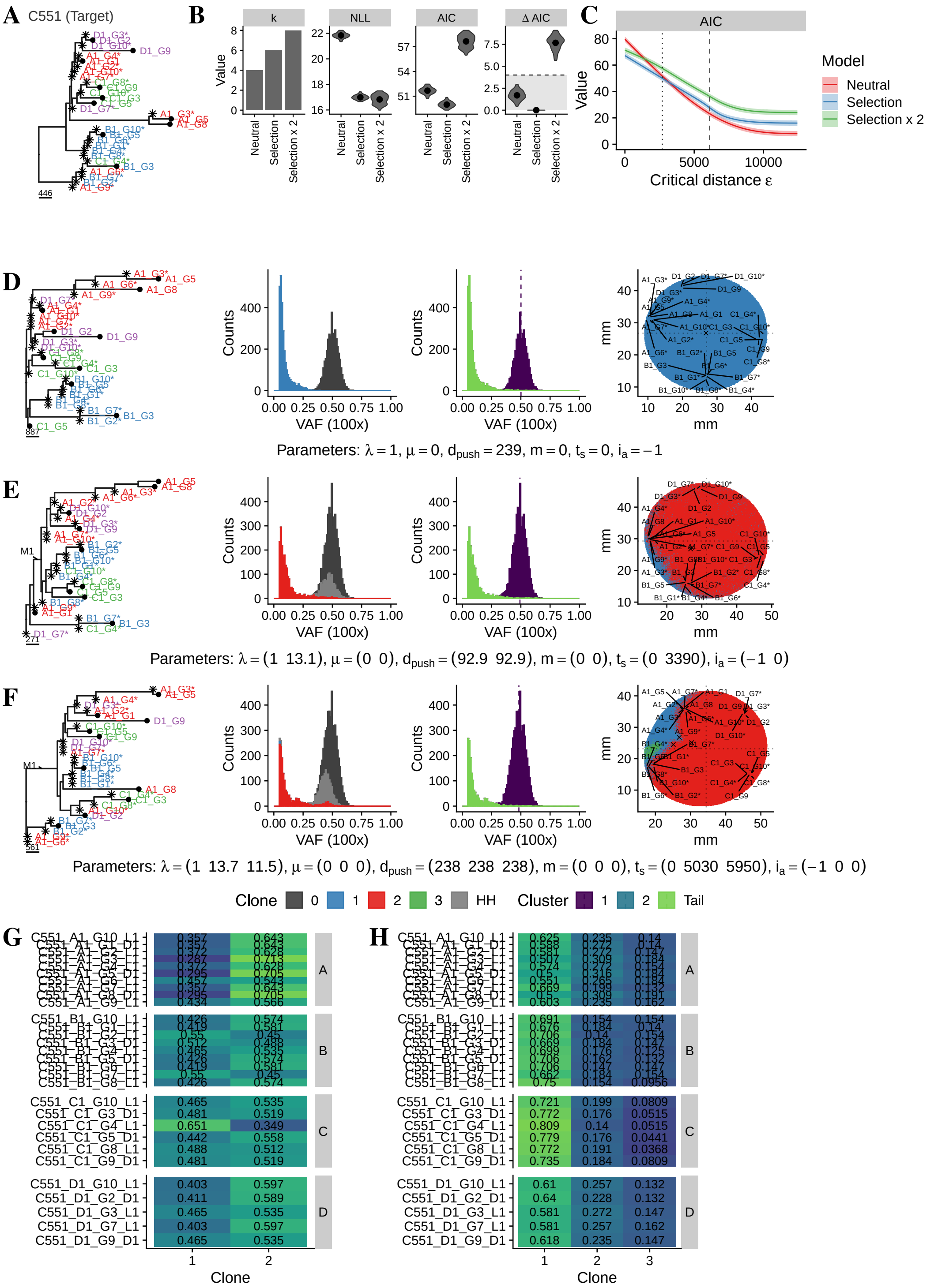
Supplementary Figure S.182: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C548. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the ‘Neutral’, ‘Selection’ and ‘Selection x 2’ model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the ‘Selection’ and ‘Selection x 2’ model respectively. Figure parts F) and H) can be missing if the ‘Selection x 2’ model was not considered.



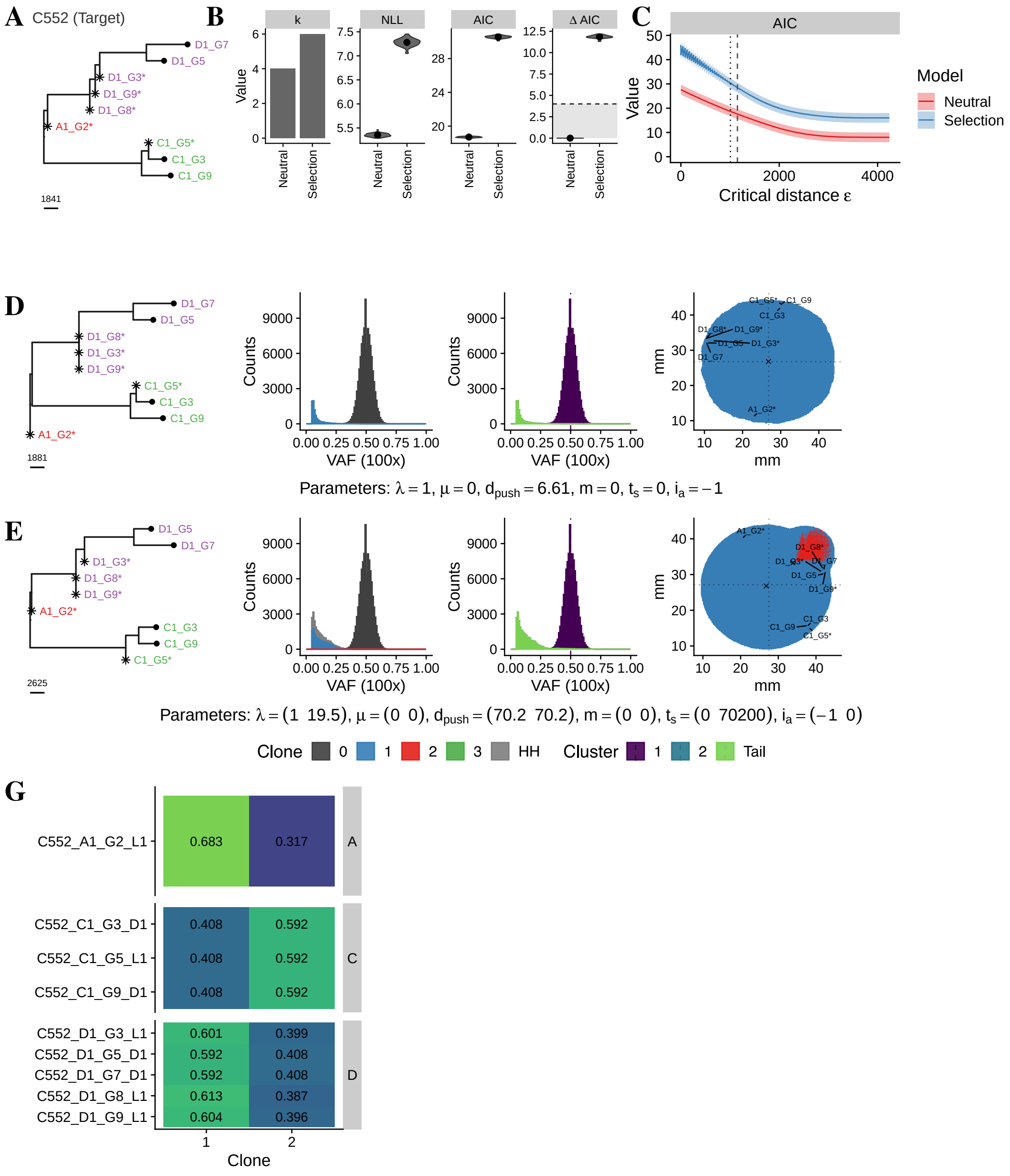
Supplementary Figure S.183: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C549. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ϵ. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



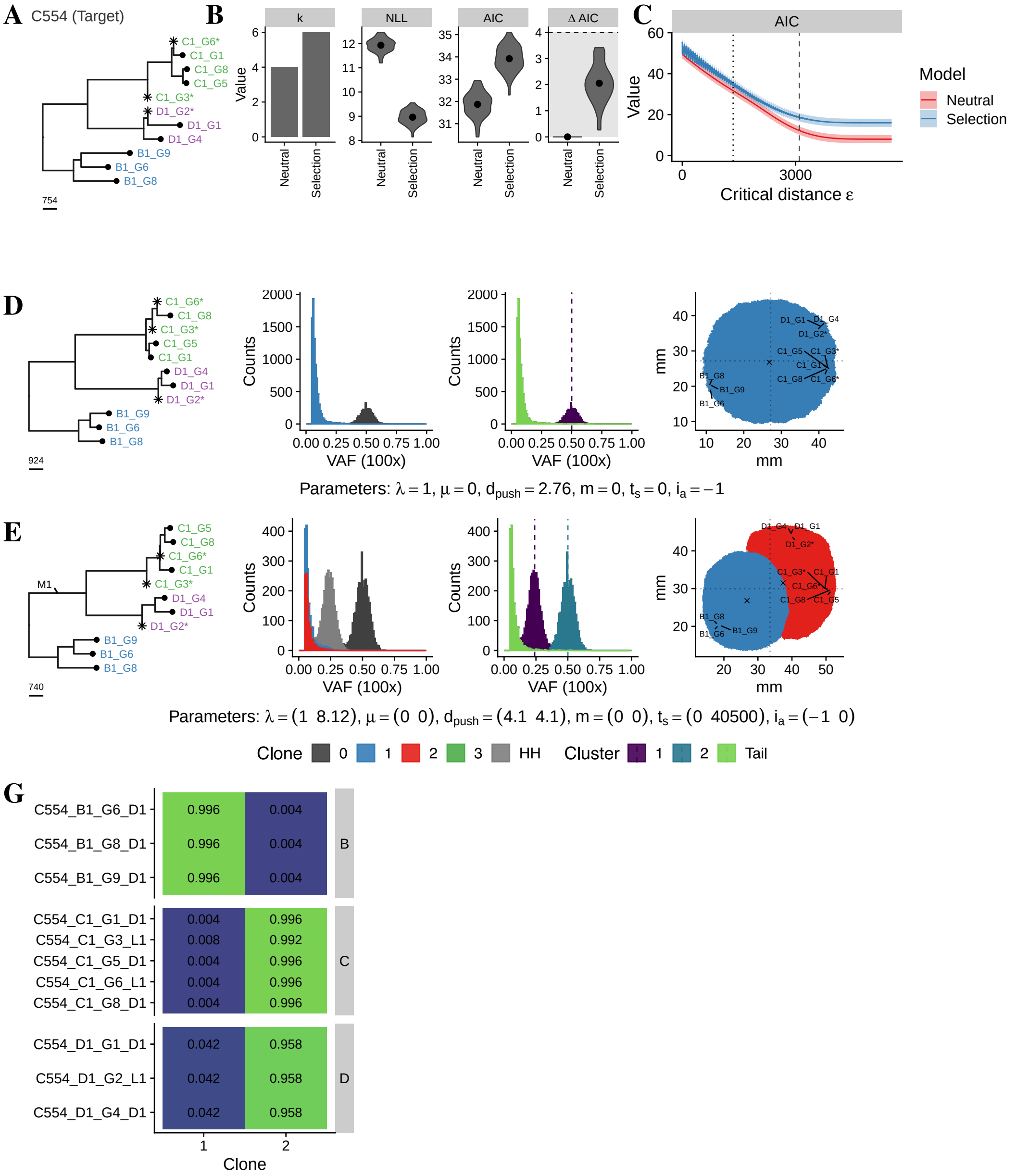
Supplementary Figure S.184: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C550. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



Supplementary Figure S.185: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C551. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



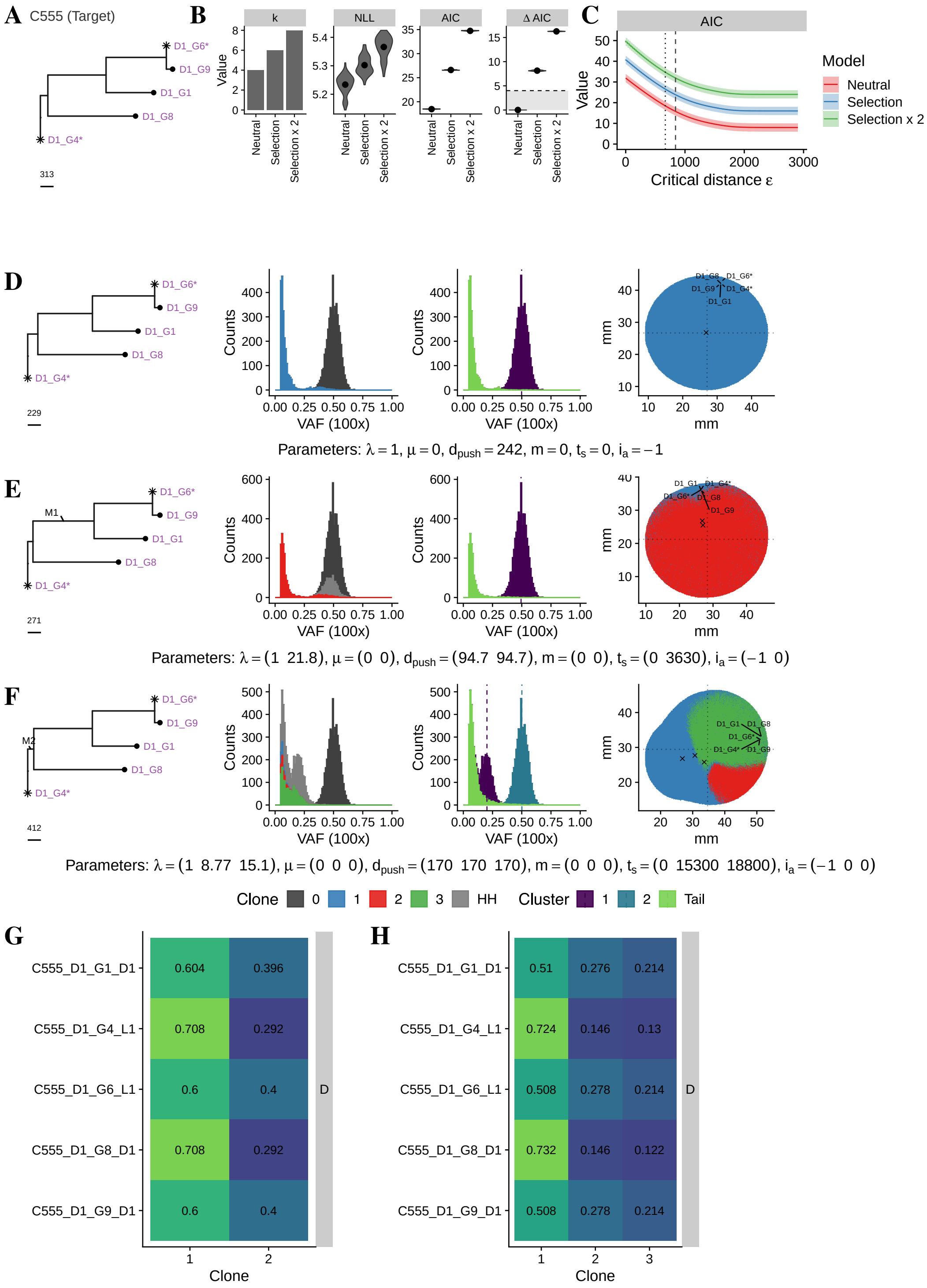
Supplementary Figure S.186: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C552. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



Parameters: $\lambda = 1$, $\mu = 0$, $d_{push} = 2.76$, $m = 0$, $t_s = 0$, $i_a = -1$

Parameters: $\lambda = (1\ \ 8.12)$, $\mu = (0\ \ 0)$, $d_{push} = (4.1\ \ 4.1)$, $m = (0\ \ 0)$, $t_s = (0\ \ 40500)$, $i_a = (-1\ \ 0)$

Supplementary Figure S.187: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C554. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.

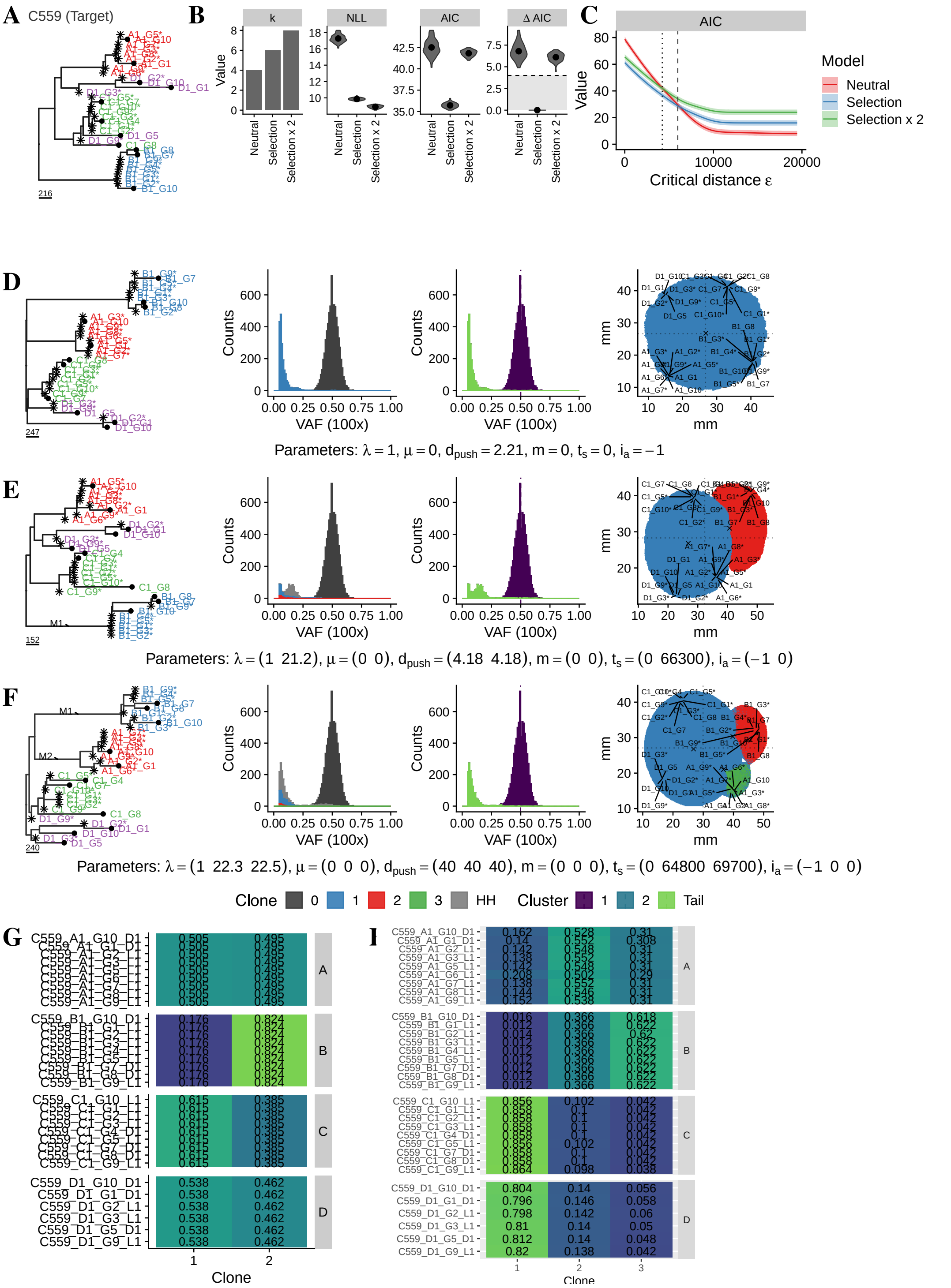


Parameters: $\lambda = 1$, $\mu = 0$, $d_{push} = 242$, $m = 0$, $t_s = 0$, $i_a = -1$

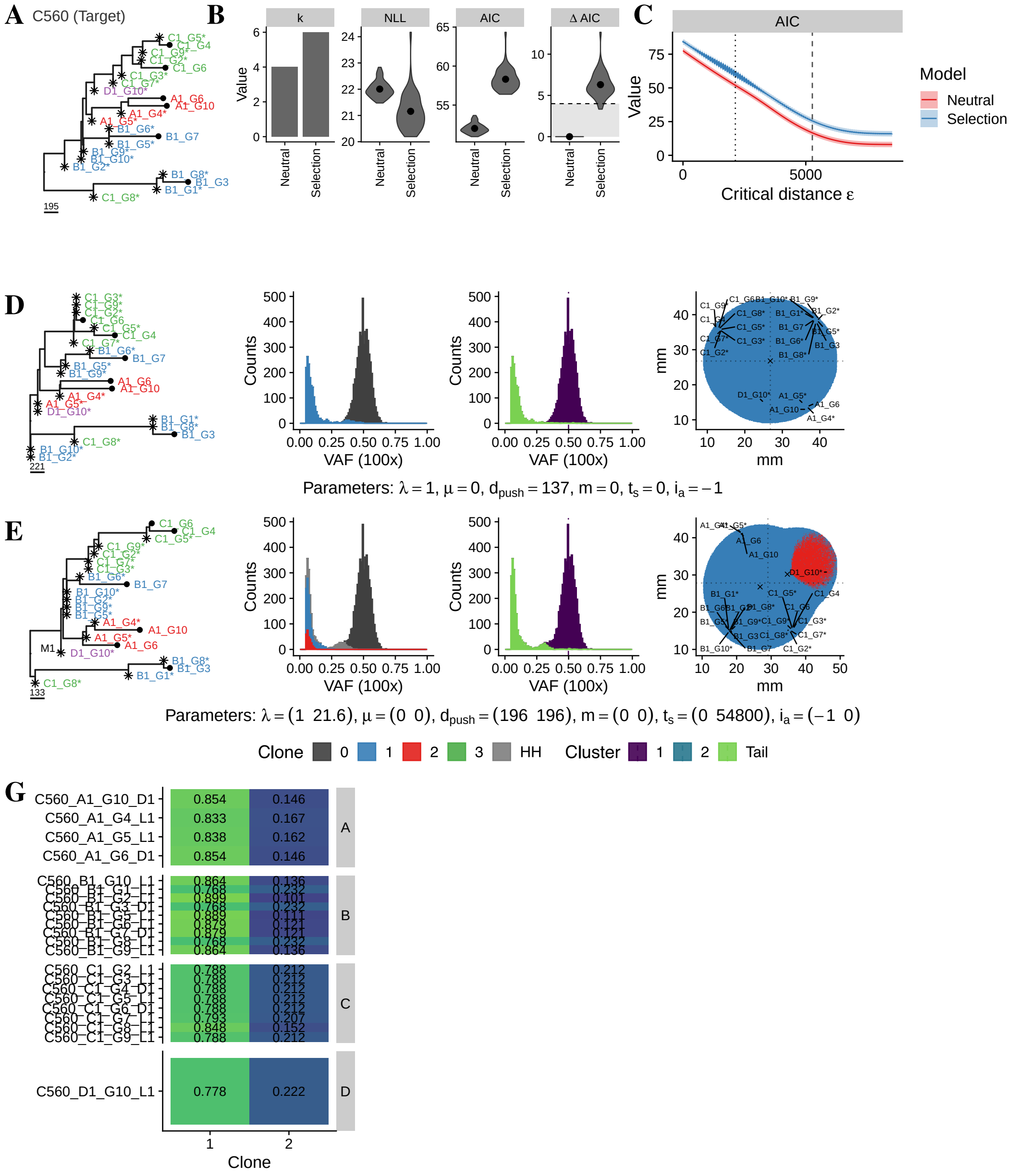
Parameters: $\lambda = (1\ \ 21.8)$, $\mu = (0\ \ 0)$, $d_{push} = (94.7\ \ 94.7)$, $m = (0\ \ 0)$, $t_s = (0\ \ 3630)$, $i_a = (-1\ \ 0)$

Parameters: $\lambda = (1\ \ 8.77\ \ 15.1)$, $\mu = (0\ \ 0\ \ 0)$, $d_{push} = (170\ \ 170\ \ 170)$, $m = (0\ \ 0\ \ 0)$, $t_s = (0\ \ 15300\ \ 18800)$, $i_a = (-1\ \ 0\ \ 0)$

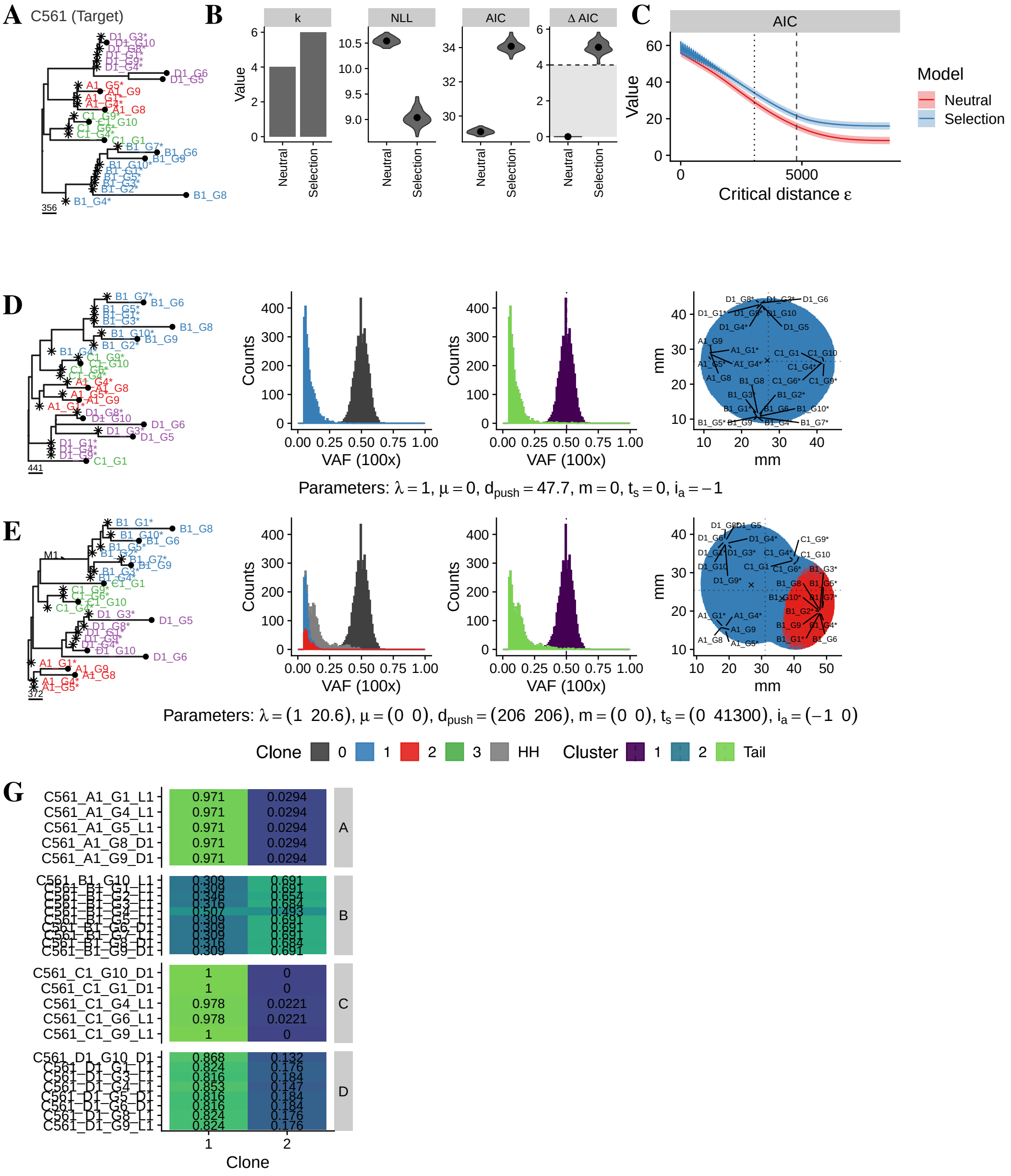
Supplementary Figure S.188: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C555. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



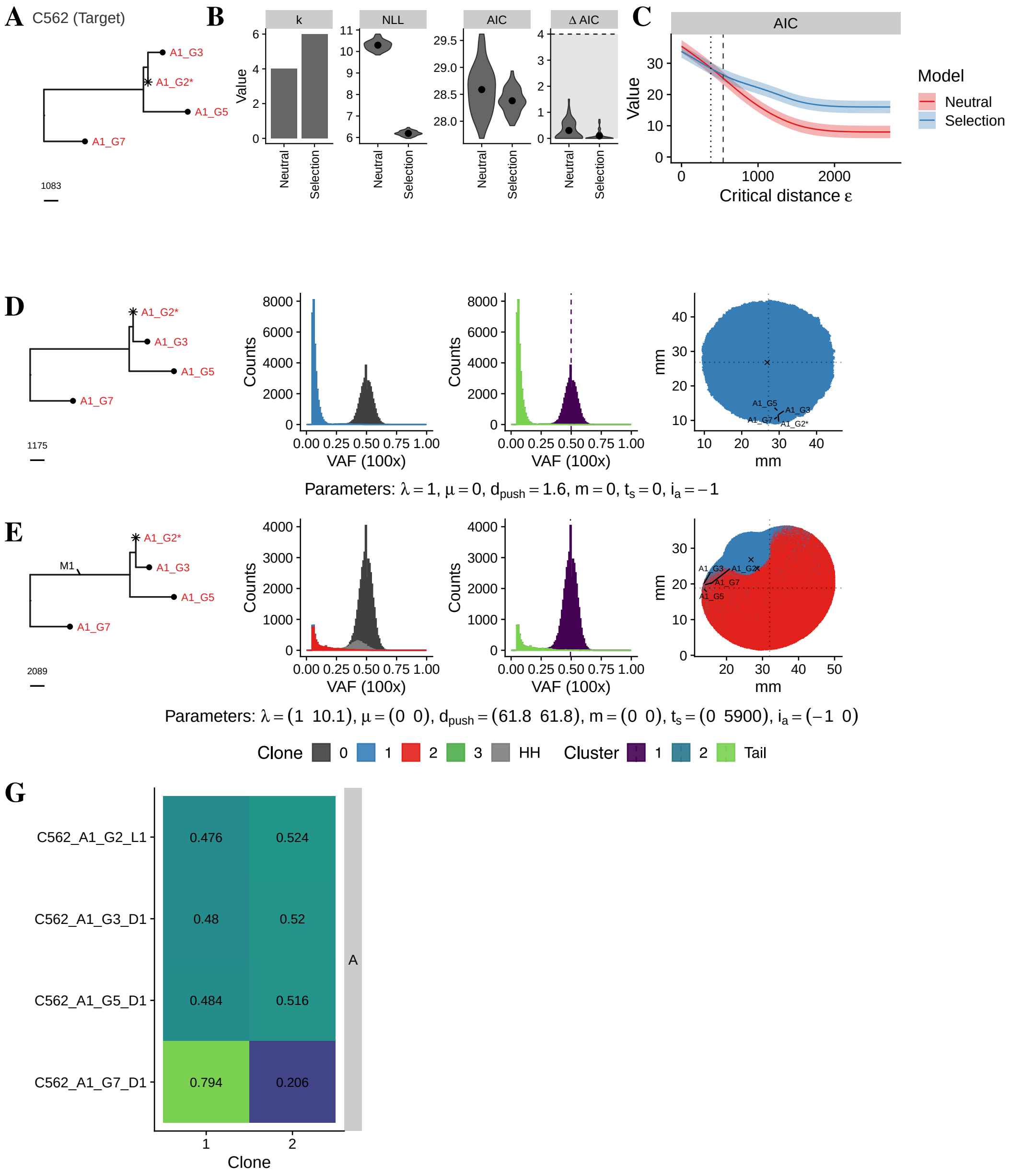
Supplementary Figure S.189: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C559. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



Supplementary Figure S.190: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C560. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



Supplementary Figure S.191: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C561. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the 'Neutral', 'Selection' and 'Selection x 2' model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the 'Selection' and 'Selection x 2' model respectively. Figure parts F) and H) can be missing if the 'Selection x 2' model was not considered.



Supplementary Figure S.192: SMC-ABC inference framework with variable sampling schema applied to ML-WGS tree of C562. A) The target tree. B) Model selection results for all tested models. C) Corresponding AIC values for a wide range of ε. D), E) and F) The best fitting simulated neutral tree for the ‘Neutral’, ‘Selection’ and ‘Selection x 2’ model respectively. G) and H) The fraction of times samples were located in the corresponding clones for the ‘Selection’ and ‘Selection x 2’ model respectively. Figure parts F) and H) can be missing if the ‘Selection x 2’ model was not considered.

Bibliography

Abascal, Federico, Luke M. R. Harvey, Emily Mitchell, Andrew R. J. Lawson, Stefanie V. Lensing, Peter Ellis, Andrew J. C. Russell, Raul E. Alcantara, Adrian Baez-Ortega, Yichen Wang, Eugene Jing Kwa, Henry Lee-Six, Alex Cagan, Tim H. H. Coorens, Michael Spencer Chapman, Sigurgeir Olafsson, Steven Leonard, David Jones, Heather E. Machado, Megan Davies, Nina F. Øbro, Krishnaa T. Mahubani, Kieren Allinson, Moritz Gerstung, Kourosh Saeb-Parsy, David G. Kent, Elisa Laurenti, Michael R. Stratton, Raheleh Rahbari, Peter J. Campbell, Robert J. Osborne, and Iñigo Martincorena (May 2021). "Somatic mutation landscapes at single-molecule resolution". In: *Nature* 593.7859, pp. 405–410.

Acar, Ahmet, Daniel Nichol, Javier Fernandez-Mateos, George D. Cresswell, Iros Barozzi, Sung Pil Hong, Nicholas Trahearn, Inmaculada Spiteri, Mark Stubbs, Rosemary Burke, Adam Stewart, Giulio Caravagna, Benjamin Werner, Georgios Vlachogiannis, Carlo C. Maley, Luca Magnani, Nicola Valeri, Udai Banerji, and Andrea Sottoriva (Apr. 21, 2020). "Exploiting evolutionary steering to induce collateral drug sensitivity in cancer". In: *Nature Communications* 11.1, p. 1923.

Adessi, Céline, Gilles Matton, Guidon Ayala, Gerardo Turcatti, Jean-Jacques Mermod, Pascal Mayer, and Eric Kawashima (Oct. 15, 2000). "Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms". In: *Nucleic Acids Research* 28.20, e87.

Akhtar-Zaidi, Batool, Richard Cowper-Sal·lari, Olivia Corradin, Alina Saiakhova, Cynthia F. Bartels, Dheepa Balasubramanian, Lois Myeroff, James Lutterbaugh, Awad Jarrar, Matthew F. Kalady, Joseph Willis, Jason H. Moore, Paul J. Tesar, Thomas Laframboise, Sanford Markowitz, Mathieu Lupien, and Peter C. Scacheri (May 11, 2012). "Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer". In: *Science* 336.6082, pp. 736–739.

Alazzouzi, Hafid, Pia Alhopuro, Reijo Salovaara, Heli Sammalkorpi, Heikki Järvinen, Jukka-Pekka Mecklin, Akeseli Hemminki, Simo Schwartz, Lauri A. Aaltonen, and Diego Arango (Apr. 1, 2005). "SMAD4 as a Prognostic Marker in Colorectal Cancer". In: *Clinical Cancer Research* 11.7, pp. 2606–2611.

Alberici, P., S. Jagmohan-Changur, E. De Pater, M. Van Der Valk, R. Smits, P. Hohenstein, and R. Fodde (Mar. 23, 2006). "Smad4 haploinsufficiency in mouse models for intestinal cancer". In: *Oncogene* 25.13, pp. 1841–1851.

Alexandrov, Ludmil B., Philip H. Jones, David C. Wedge, Julian E. Sale, Peter J. Campbell, Serena Nik-Zainal, and Michael R. Stratton (Dec. 2015). "Clock-like mutational processes in human somatic cells". In: *Nature genetics* 47.12, pp. 1402–1407.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R. Covington, Dmitry A. Gordenin, Erik N. Bergstrom, S. M. Ashiqul Islam, Nuria Lopez-Bigas, Leszek J. Klimczak, John R. McPherson, Sandro Morganella, Radhakrishnan Sabarinathan, David A. Wheeler, Ville Mustonen, Gad Getz, Steven G. Rozen, and Michael R. Stratton (Feb. 2020). “The repertoire of mutational signatures in human cancer”. In: *Nature* 578.7793, pp. 94–101.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Peter J. Campbell, and Michael R. Stratton (Jan. 31, 2013a). “Deciphering Signatures of Mutational Processes Operative in Human Cancer”. In: *Cell Reports* 3.1, pp. 246–259.

Alexandrov, Ludmil B. and Michael R. Stratton (Feb. 2014). “Mutational signatures: the patterns of somatic mutations hidden in cancer genomes”. In: *Current Opinion in Genetics & Development* 24, pp. 52–60.

Alexandrov, Ludmil B. et al. (Aug. 22, 2013b). “Signatures of mutational processes in human cancer”. In: *Nature* 500.7463, pp. 415–421.

Alexeyev, Mikhail, Inna Shokolenko, Glenn Wilson, and Susan LeDoux (May 2013). “The Maintenance of Mitochondrial DNA Integrity—Critical Analysis and Update”. In: *Cold Spring Harbor Perspectives in Biology* 5.5.

Allis, C. David and Thomas Jenuwein (Aug. 2016). “The molecular hallmarks of epigenetic control”. In: *Nature Reviews Genetics* 17.8, pp. 487–500.

Alsner, Jan, Vibeke Jensen, Marianne Kyndi, Birgitte Vrou Offersen, Phuong Vu, Anne-Lise Børresen-Dale, and Jens Overgaard (2008). “A comparison between p53 accumulation determined by immunohistochemistry and TP53 mutations as prognostic variables in tumours from breast cancer patients”. In: *Acta Oncologica (Stockholm, Sweden)* 47.4, pp. 600–607.

Alves, João M., Tamara Prieto, and David Posada (Aug. 2017). “Multiregional Tumor Trees Are Not Phylogenies”. In: *Trends in Cancer* 3.8, pp. 546–550.

Ames, Bruce N., William E. Durston, Edith Yamasaki, and Frank D. Lee (Aug. 1973). “Carcinogens are Mutagens: A Simple Test System Combining Liver Homogenates for Activation and Bacteria for Detection”. In: *Proceedings of the National Academy of Sciences of the United States of America* 70.8, pp. 2281–2285.

An, Ziwen, David J. Nott, and Christopher Drovandi (Oct. 3, 2019). “Robust Bayesian Synthetic Likelihood via a Semi-Parametric Approach”. In: *arXiv:1809.05800 [stat]*. arXiv: 1809.05800.

Anderson, Alexander R. A., Alissa M. Weaver, Peter T. Cummings, and Vito Quaranta (Dec. 1, 2006). “Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment”. In: *Cell* 127.5, pp. 905–915.

Anderson, Kristina, Christoph Lutz, Frederik W. van Delft, Caroline M. Bateman, Yanping Guo, Susan M. Colman, Helena Kempski, Anthony V. Moorman, Ian Titley, John Swansbury, Lyndal Kearney, Tariq Enver, and Mel Greaves (Jan. 20, 2011). “Genetic variegation of clonal architecture and propagating cells in leukaemia”. In: *Nature* 469.7330, pp. 356–361.

Anderson, Marti J. (2001). “A new method for non-parametric multivariate analysis of variance”. In: *Austral Ecology* 26.1, pp. 32–46.

Andreas Heger, Tildon Grant Belgard, Florian Finkernagel, Leo Goodstadt, Martin Goodson, Kevin B. Jacobs, Gerton Lunter, Marcel Martin, and Ben Schiller (Mar. 24, 2021). *pysam-developers/pysam*.

Antal, Tibor and P L Krapivsky (2011). “Exact solution of a two-type branching process: models of tumor progression”. In: p. 22.

Antelo, Marina, Francesc Balaguer, Jinru Shia, Yan Shen, Keun Hur, Leticia Moreira, Miriam Cuatrecasas, Luis Bujanda, Maria Dolores Giraldez, Masanobu Takahashi, Ana Cabanne, Mario Edmundo Barugel, Mildred Arnold, Enrique Luis Roca, Montserrat Andreu, Sergi Castellvi-Bel, Xavier Llor, Rodrigo Jover, Antoni Castells, C. Richard Boland, and Ajay Goel (Sept. 25, 2012). "A High Degree of LINE-1 Hypomethylation Is a Unique Feature of Early-Onset Colorectal Cancer". In: *PLOS ONE* 7.9, e45357.

Aran, Dvir, Marina Sirota, and Atul J. Butte (Dec. 4, 2015). "Systematic pan-cancer analysis of tumour purity". In: *Nature Communications* 6.1, p. 8971.

Ardaševa, Aleksandra, Alexander R. A. Anderson, Robert A. Gatenby, Helen M. Byrne, Philip K. Maini, and Tommaso Lorenzi (Oct. 2020). "Comparative study between discrete and continuum models for the evolution of competing phenotype-structured cell populations in dynamical environments". In: *Physical Review. E* 102.4, p. 042404.

Armitage, P. and R. Doll (Mar. 1954). "The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis". In: *British Journal of Cancer* 8.1, pp. 1–12.

— (June 1957). "A two-stage theory of carcinogenesis in relation to the age distribution of human cancer". In: *British Journal of Cancer* 11.2, pp. 161–169.

Armstrong, Bruce K and Anne Kricker (Oct. 1, 2001). "The epidemiology of UV induced skin cancer". In: *Journal of Photochemistry and Photobiology B: Biology*. Consequences of exposure to sunlight:elements to assess protection 63.1, pp. 8–18.

Arnedo-Pac, Claudia, Loris Mularoni, Ferran Muiños, Abel Gonzalez-Perez, and Nuria Lopez-Bigas (Nov. 1, 2019). "OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers". In: *Bioinformatics* 35.22, pp. 4788–4790.

Ashley, D. J. (June 1969). "The two "hit" and multiple "hit" theories of carcinogenesis." In: *British Journal of Cancer* 23.2, pp. 313–328.

Atlasi, Yaser and Hendrik G. Stunnenberg (Nov. 2017). "The interplay of epigenetic marks during stem cell differentiation and development". In: *Nature Reviews. Genetics* 18.11, pp. 643–658.

Auton, Adam, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, and Gonçalo R. Abecasis (Oct. 1, 2015). "A global reference for human genetic variation." In: *Nature* 526.7571, pp. 68–74.

Baba, Yoshifumi, Taisuke Yagi, Hiroshi Sawayama, Yukiharu Hiyoshi, Takatsugu Ishimoto, Masaaki Iwatsuki, Yuji Miyamoto, Naoya Yoshida, and Hideo Baba (2018). "Long Interspersed Element-1 Methylation Level as a Prognostic Biomarker in Gastrointestinal Cancers". In: *Digestion* 97.1, pp. 26–30.

Bache, Stefan Milton and Hadley Wickham (2014). *magrittr: A Forward-Pipe Operator for R*.

Bader, Andreas G., Sohye Kang, and Peter K. Vogt (Jan. 31, 2006). "Cancer-specific mutations in PIK3CA are oncogenic in vivo". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.5, pp. 1475–1479.

Bagasra, Omar (Nov. 2007). "Protocols for the in situ PCR-amplification and detection of mRNA and DNA sequences". In: *Nature Protocols* 2.11, pp. 2782–2795.

Bailey, Matthew H. et al. (Apr. 5, 2018). "Comprehensive Characterization of Cancer Driver Genes and Mutations". In: *Cell* 173.2, 371–385.e18.

Baker, Ann-Marie, Biancastella Cereser, Samuel Melton, Alexander G. Fletcher, Manuel Rodriguez-Justo, Paul J. Tadrous, Adam Humphries, George Elia, Stuart A. C. McDonald, Nicholas A. Wright, Benjamin D. Simons, Marnix Jansen, and Trevor A. Graham (Aug. 21, 2014). "Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon". In: *Cell Reports* 8.4, pp. 940–947.

Baker, Ann-Marie, William Cross, Kit Curtius, Ibrahim Al Bakir, Chang-Ho Ryan Choi, Hayley Louise Davis, Daniel Temko, Sujata Biswas, Pierre Martinez, Marc J. Williams, James O. Lindsay, Roger Feakins, Roser Vega, Stephen J. Hayes, Ian P. M. Tomlinson, Stuart A. C. McDonald, Morgan Moorghen, Andrew Silver, James E. East, Nicholas A. Wright, Lai Mun Wang, Manuel Rodriguez-Justo, Marnix Jansen, Ailsa L. Hart, Simon J. Leedham, and Trevor A. Graham (June 1, 2019). “Evolutionary history of human colitis-associated colorectal cancer”. In: *Gut* 68.6, pp. 985–995.

Balaparya, Abdul and Subhajyoti De (Dec. 2018). “Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data”. In: *Nature Genetics* 50.12, pp. 1626–1628.

Baltimore, D. (June 27, 1970). “RNA-dependent DNA polymerase in virions of RNA tumour viruses”. In: *Nature* 226.5252, pp. 1209–1211.

Baragatti, Meili, Agnès Grimaud, and Denys Pommeret (Apr. 30, 2012). “Likelihood-Free Parallel Tempering”. In: *arXiv:1108.3423 [stat]*. arXiv: 1108.3423.

Barber, Thomas D, Bert Vogelstein, Kenneth W Kinzler, and Victor E Velculescu (2004). “Somatic mutations of EGFR in colorectal cancers and glioblastomas”. In: *N Engl J Med* 351.27, p. 2883.

Bartek, J., R. Iggo, J. Gannon, and D. P. Lane (June 1990). “Genetic and immunochemical analysis of mutant p53 in human breast cancer cell lines”. In: *Oncogene* 5.6, pp. 893–899.

Bártek, J., J. Bártková, B. Vojtěsek, Z. Stasková, J. Lukás, A. Rejthar, J. Kovarík, C. A. Midgley, J. V. Gannon, and D. P. Lane (Sept. 1, 1991). “Aberrant expression of the p53 oncoprotein is a common feature of a wide spectrum of human malignancies”. In: *Oncogene* 6.9, pp. 1699–1703.

Bartolini, Alice, Sabrina Cardaci, Simona Lamba, Daniele Oddo, Caterina Marchiò, Paola Cassoni, Carla Azzurra Amoreo, Giorgio Corti, Alessandro Testori, Federico Bussolino, Renata Pasqualini, Wadih Arap, Davide Corà, Federica Di Nicolantonio, and Serena Marchiò (Oct. 1, 2016). “BCAM and LAMA5 Mediate the Recognition between Tumor Cells and the Endothelium in the Metastatic Spreading of KRAS-Mutant Colorectal Cancer”. In: *Clinical Cancer Research* 22.19, pp. 4923–4933.

Baylin, Stephen B. (Mar. 2011). “Resistance, epigenetics and the cancer ecosystem”. In: *Nature Medicine* 17.3, pp. 288–289.

Beaumont, Mark A., Jean-Marie Cornuet, Jean-Michel Marin, and Christian P. Robert (Dec. 1, 2009). “Adaptive approximate Bayesian computation”. In: *Biometrika* 96.4, pp. 983–990. arXiv: 0805.2256.

Beaumont, Mark A., Christian P. Robert, Université Paris Dauphine, Jean-michel Marin, and Jean-marie Cornuet (2008). “Adaptivity for ABC algorithms: the ABC-PMC scheme”. In: *In submission*.

Beerenwinkel, Niko, Roland F. Schwarz, Moritz Gerstung, and Florian Markowetz (Jan. 1, 2015). “Cancer Evolution: Mathematical Models and Computational Inference”. In: *Systematic Biology* 64.1, e1–e25.

Benjamini, Yoav and Yosef Hochberg (1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300.

Bennett, Dominic J., Mark D. Sutton, and Samuel T. Turvey (Jan. 7, 2017). “treeman: an R package for efficient and intuitive manipulation of phylogenetic trees”. In: *BMC Research Notes* 10.1, p. 30.

Berger, Alice H., Alfred G. Knudson, and Pier Paolo Pandolfi (Aug. 2011). “A continuum model for tumour suppression”. In: *Nature* 476.7359, pp. 163–169.

Bernstein, Carol, Harris Bernstein, Claire M. Payne, Katerina Dvorak, and Harinder Garewal (Feb. 18, 2008). “Field defects in progression to gastrointestinal tract cancers”. In: *Cancer letters* 260.1, pp. 1–10.

Bierie, Brian and Harold L. Moses (July 2006). “Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer”. In: *Nature Reviews. Cancer* 6.7, pp. 506–520.

Billingsley, Caroline C., David E. Cohn, David G. Mutch, Julie A. Stephens, Adrian A. Suarez, and Paul J. Goodfellow (2015). “Polymerase eta (POLE) mutations in endometrial cancer: Clinical outcomes and implications for Lynch syndrome testing”. In: *Cancer* 121.3, pp. 386–394.

Biswas, Subhankar and C. Mallikarjuna Rao (May 1, 2017). “Epigenetics in cancer: Fundamentals and Beyond”. In: *Pharmacology & Therapeutics* 173, pp. 118–134.

Black, James R. M. and Nicholas McGranahan (Mar. 16, 2021). “Genetic and non-genetic clonal diversity in cancer evolution”. In: *Nature Reviews Cancer*, pp. 1–14.

Blokzijl, Francis, Joep de Ligt, Myrthe Jager, Valentina Sasselli, Sophie Roerink, Nobuo Sasaki, Meritxell Huch, Sander Boymans, Ewart Kuijk, Pjotr Prins, Isaac J. Nijman, Inigo Martincorena, Michal Mokry, Caroline L. Wiegerinck, Sabine Middendorp, Toshiro Sato, Gerald Schwank, Edward E. S. Nieuwenhuis, Monique M. A. Verstegen, Luc J. W. van der Laan, Jeroen de Jonge, Jan N. M. IJzermans, Robert G. Vries, Marc van de Wetering, Michael R. Stratton, Hans Clevers, Edwin Cuppen, and Ruben van Boxtel (Oct. 2016). “Tissue-specific mutation accumulation in human adult stem cells during life”. In: *Nature* 538.7624, pp. 260–264.

Blomberg, Simon P., Theodore Garland, and Anthony R. Ives (Apr. 2003). “Testing for phylogenetic signal in comparative data: behavioral traits are more labile”. In: *Evolution; International Journal of Organic Evolution* 57.4, pp. 717–745.

Boettcher, Steffen, Peter G. Miller, Rohan Sharma, Marie McConkey, Matthew Leventhal, Andrei V. Krivtsov, Andrew O. Giacomelli, Waihay Wong, Jesi Kim, Sherry Chao, Kari J. Kurppa, Xiaoping Yang, Kirsten Milenkowic, Federica Piccioni, David E. Root, Frank G. Rücker, Yael Flamand, Donna Neuberg, R. Coleman Lindsley, Pasi A. Jänne, William C. Hahn, Tyler Jacks, Hartmut Döhner, Scott A. Armstrong, and Benjamin L. Ebert (Aug. 9, 2019). “A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies”. In: *Science* 365.6453, pp. 599–604.

Bonneau, D. and M. Longy (2000). “Mutations of the human PTEN gene”. In: *Human Mutation* 16.2, pp. 109–122.

Bos, J. L. (Sept. 1, 1989). “ras oncogenes in human cancer: a review”. In: *Cancer Research* 49.17, pp. 4682–4689.

Bos, Johannes L., Eric R. Fearon, Stanley R. Hamilton, Matty Verlaan-de Vries, Jacques H. van Boom, Alex J. van der Eb, and Bert Vogelstein (May 1987). “Prevalence of ras gene mutations in human colorectal cancers”. In: *Nature* 327.6120, pp. 293–297.

Boveri, Theodor (1914). *Zur frage der entstehung maligner tumoren.*

Box, George EP, RL Launer, and GN Wilkinson (1979). “Robustness in statistics”. In: *Robustness in the strategy of scientific model building*, pp. 201–236.

Boyko, Adam R., Scott H. Williamson, Amit R. Indap, Jeremiah D. Degenhardt, Ryan D. Hernandez, Kirk E. Lohmueller, Mark D. Adams, Steffen Schmidt, John J. Sninsky, Shamil R. Sunyaev, Thomas J. White, Rasmus Nielsen, Andrew G. Clark, and Carlos

D. Bustamante (May 30, 2008). “Assessing the evolutionary impact of amino acid mutations in the human genome”. In: *PLoS genetics* 4.5, e1000083.

Bozic, Ivana, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W. Kinzler, Bert Vogelstein, and Martin A. Nowak (Oct. 26, 2010). “Accumulation of driver and passenger mutations during tumor progression”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.43, pp. 18545–18550.

Bozic, Ivana, Jeffrey M. Gerold, and Martin A. Nowak (Feb. 1, 2016). “Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution”. In: *PLOS Computational Biology* 12.2, e1004731.

Bresenham, J. E. (1965). “Algorithm for computer control of a digital plotter”. In: *IBM Systems Journal* 4.1, pp. 25–30.

Bruens, Lotte, Saskia I. J. Ellenbroek, Jacco van Rheenen, and Hugo J. Snippert (Sept. 1, 2017). “In Vivo Imaging Reveals Existence of Crypt Fission and Fusion in Adult Mouse Intestine”. In: *Gastroenterology* 153.3, 674–677.e3.

Bruin, Elza C. de, Nicholas McGranahan, Richard Mitter, Max Salm, David C. Wedge, Lucy Yates, Mariam Jamal-Hanjani, Seema Shafi, Nirupa Murugaesu, Andrew J. Rowan, Eva Grönroos, Madiha A. Muhammad, Stuart Horswell, Marco Gerlinger, Ignacio Varela, David Jones, John Marshall, Thierry Voet, Peter Van Loo, Doris M. Rassl, Robert C. Rintoul, Sam M. Janes, Siow-Ming Lee, Martin Forster, Tanya Ahmad, David Lawrence, Mary Falzon, Arrigo Capitanio, Timothy T. Harkins, Clarence C. Lee, Warren Tom, Enock Teefe, Shann-Ching Chen, Sharmin Begum, Adam Rabinowitz, Benjamin Phillimore, Bradley Spencer-Dene, Gordon Stamp, Zoltan Szallasi, Nik Matthews, Aengus Stewart, Peter Campbell, and Charles Swanton (Oct. 10, 2014). “Spatial and temporal diversity in genomic instability processes defines lung cancer evolution”. In: *Science (New York, N.Y.)* 346.6206, pp. 251–256.

Buenrostro, Jason, Beijing Wu, Howard Chang, and William Greenleaf (Jan. 5, 2015). “ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide”. In: *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* 109, pp. 21.29.1–21.29.9.

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf (Dec. 1, 2013). “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position”. In: *Nature Methods* 10.12, pp. 1213–1218.

Buettner-Janusch, John, Vina Buettner-Janusch, and George A. Mason (Aug. 1, 1969). “Amino acid compositions and amino-terminal end groups of α and β chains from polymorphic hemoglobins of Pongo pygmaeus”. In: *Archives of Biochemistry and Biophysics* 133.1, pp. 164–170.

Burke, J. R., P. Brown, A. Quyn, H. Lambie, D. Tolan, and P. Sagar (2020). “Tumour growth rate of carcinoma of the colon and rectum: retrospective cohort study”. In: *BJS Open* 4.6, pp. 1200–1207.

Cairns, John (May 1975). “Mutation selection and the natural history of cancer”. In: *Nature* 255.5505, pp. 197–200.

Campbell, Malcolm G. and Lisa M. Giocomo (Dec. 2019). “How a fly’s neural compass adapts to an ever-changing world”. In: *Nature* 576.7785, pp. 42–43.

Campbell, Peter J. et al. (Feb. 2020). “Pan-cancer analysis of whole genomes”. In: *Nature* 578.7793, pp. 82–93.

Canli, Özge, Adele M. Nicolas, Jalaj Gupta, Fabian Finkelmeier, Olga Goncharova, Marina Pesic, Tobias Neumann, David Horst, Martin Löwer, Ugur Sahin, and Florian R. Greten (Dec. 11, 2017). "Myeloid Cell-Derived Reactive Oxygen Species Induce Epithelial Mutagenesis". In: *Cancer Cell* 32.6, 869–883.e5.

Caravagna, Giulio, Timon Heide, Marc J. Williams, Luis Zapata, Daniel Nichol, Ketevan Chkhaidze, William Cross, George D. Cresswell, Benjamin Werner, Ahmet Acar, Louis Chesler, Chris P. Barnes, Guido Sanguinetti, Trevor A. Graham, and Andrea Sottoriva (Sept. 2020). "Subclonal reconstruction of tumors by using machine learning and population genetics". In: *Nature Genetics* 52.9, pp. 898–907.

Carneiro, Miguel, Frank W. Albert, José Melo-Ferreira, Nicolas Galtier, Philippe Gayral, Jose A. Blanco-Aguiar, Rafael Villafuerte, Michael W. Nachman, and Nuno Ferrand (July 2012). "Evidence for widespread positive and purifying selection across the European rabbit (Oryctolagus cuniculus) genome". In: *Molecular Biology and Evolution* 29.7, pp. 1837–1849.

Carter, Scott L., Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W. Laird, Robert C. Onofrio, Wendy Winckler, Barbara A. Weir, Rameen Beroukhim, David Pellman, Douglas A. Levine, Eric S. Lander, Matthew Meyerson, and Gad Getz (May 2012). "Absolute quantification of somatic DNA alterations in human cancer". In: *Nature Biotechnology* 30.5, pp. 413–421.

Cavalli, Giacomo and Edith Heard (July 2019). "Advances in epigenetics link genetics to the environment and disease". In: *Nature* 571.7766, pp. 489–499.

Chang, C., D. T. Simmons, M. A. Martin, and P. T. Mora (Aug. 1979). "Identification and partial characterization of new antigens from simian virus 40-transformed mouse cells". In: *Journal of Virology* 31.2, pp. 463–471.

Chen, Ru, Peter S. Rabinovitch, David A. Crispin, Mary J. Emond, Mary P. Bronner, and Teresa A. Brentnall (Sept. 2005). "The initiation of colon cancer in a chronic inflammatory setting". In: *Carcinogenesis* 26.9, pp. 1513–1519.

Chen, Yu, Jinsong Wang, Donghua Wang, Ting Kang, Jinghu Du, Zeqiang Yan, and Manyu Chen (2020). "TNNT1, negatively regulated by miR-873, promotes the progression of colorectal cancer". In: *The Journal of Gene Medicine* 22.2, e3152.

Chi, Ping, C. David Allis, and Gang Greg Wang (July 2010). "Covalent histone modifications — miswritten, misinterpreted and mis-erased in human cancers". In: *Nature Reviews Cancer* 10.7, pp. 457–469.

Chignola, R. and R.I. Foroni (May 2005). "Estimating the growth kinetics of experimental tumors from as few as two determinations of tumor size: implications for clinical oncology". In: *IEEE Transactions on Biomedical Engineering* 52.5, pp. 808–815.

Chkhaidze, Ketevan, Timon Heide, Benjamin Werner, Marc J. Williams, Weini Huang, Giulio Caravagna, Trevor A. Graham, and Andrea Sottoriva (July 29, 2019). "Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data". In: *PLOS Computational Biology* 15.7, e1007243.

Chowdhury, Salim Akhter, Stanley E. Shackney, Kerstin Heselmeyer-Haddad, Thomas Ried, Alejandro A. Schäffer, and Russell Schwartz (July 31, 2014). "Algorithms to Model Single Gene, Single Chromosome, and Whole Genome Copy Number Changes Jointly in Tumor Phylogenetics". In: *PLOS Computational Biology* 10.7, e1003740.

Christensen, Sharon, Bastiaan Van der Roest, Nicolle Besselink, Roel Janssen, Sander Boymans, John W. M. Martens, Marie-Laure Yaspo, Peter Priestley, Ewart Kuijk, Edwin Cuppen, and Arne Van Hoeck (Oct. 8, 2019). "5-Fluorouracil treatment induces characteristic T¿G mutations in human cancer". In: *Nature Communications* 10.1, p. 4571.

Chumakov, P. M., V. S. Iotsova, and G. P. Georgiev (1982). "[Isolation of a plasmid clone containing the mRNA sequence for mouse nonviral T-antigen]". In: *Doklady Akademii nauk SSSR* 267.5, pp. 1272–1275.

Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz (Mar. 2013). "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples". In: *Nature Biotechnology* 31.3, pp. 213–219.

Claude, Albert, Keith R. Porter, and Edward G. Pickels (July 1, 1947). "Electron Microscope Study of Chicken Tumor Cells". In: *Cancer Research* 7.7, pp. 421–430.

Coldman, A. J. and J. H. Goldie (Jan. 1, 1986). "A stochastic model for the origin and treatment of tumors containing drug-resistant cells". In: *Bulletin of Mathematical Biology*. Simulation in Cancer Research 48.3, pp. 279–292.

Colless, Donald H (1982). *Phylogenetics: The Theory and Practice of Phylogenetic Systematics.*

Comen, Elizabeth, Patrick G. Morris, and Larry Norton (Dec. 2012). "Translating mathematical modeling of tumor growth patterns into novel therapeutic approaches for breast cancer". In: *Journal of Mammary Gland Biology and Neoplasia* 17.3, pp. 241–249.

Comeron, Josep M. (Dec. 1, 1995). "A method for estimating the numbers of synonymous and nonsynonymous substitutions per site". In: *Journal of Molecular Evolution* 41.6, pp. 1152–1159.

Conger, Alan D. and Marvin C. Ziskin (Feb. 1, 1983). "Growth of Mammalian Multicellular Tumor Spheroids". In: *Cancer Research* 43.2, pp. 556–560.

Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, Clarice Groeneveld, Christopher K. Wong, Seung Woo Cho, Ansuman T. Satpathy, Maxwell R. Mumbach, Katherine A. Hoadley, A. Gordon Robertson, Nathan C. Sheffield, Ina Felau, Mauro A. A. Castro, Benjamin P. Berman, Louis M. Staudt, Jean C. Zenklusen, Peter W. Laird, Christina Curtis, The Cancer Genome Atlas Analysis Network†, William J. Greenleaf, and Howard Y. Chang (Oct. 26, 2018). "The chromatin accessibility landscape of primary human cancers". In: *Science* 362.6413.

Coronado, Tomás M., Mareike Fischer, Lina Herbst, Francesc Rosselló, and Kristina Wicke (Feb. 17, 2020). "On the minimum value of the Colless index and the bifurcating trees that achieve it". In: *arXiv:1907.05064 [cs, math, q-bio]*. arXiv: `1907.05064`.

Costa, Ana, Alix Scholer-Dahirel, and Fatima Mechta-Grigoriou (Apr. 2014). "The role of reactive oxygen species and metabolism on cancer cells and their microenvironment". In: *Seminars in Cancer Biology* 25, pp. 23–32.

Crick, F. H. C., Leslie Barnett, S. Brenner, and R. J. Watts-Tobin (Dec. 1961). "General Nature of the Genetic Code for Proteins". In: *Nature* 192.4809, pp. 1227–1232.

Cross, William, Michal Kovac, Ville Mustonen, Daniel Temko, Hayley Davis, Ann-Marie Baker, Sujata Biswas, Roland Arnold, Laura Chegwidden, Chandler Gatenbee, Alexander R. Anderson, Viktor H. Koelzer, Pierre Martinez, Xiaowei Jiang, Enric Domingo, Dan J. Woodcock, Yun Feng, Monika Kovacova, Tim Maughan, S:CORT Consortium, Marnix Jansen, Manuel Rodriguez-Justo, Shazad Ashraf, Richard Guy, Christopher Cunningham, James E. East, David C. Wedge, Lai Mun Wang, Claire Palles, Karl Heinimann, Andrea Sottoriva, Simon J. Leedham, Trevor A. Graham, and Ian P. M. Tomlinson (Oct. 2018). "The evolutionary landscape of colorectal tumorigenesis". In: *Nature Ecology & Evolution* 2.10, pp. 1661–1672.

Cross, William, Maximilian Mossner, Salpie Nowinski, George Cresswell, Abhirup Banerjee, Marc Williams, Laura Gay, Ann-Marie Baker, Christopher Kimberley, Hayley Davis, Pierre Martinez, Maria Traki, Viola Walther, Kane Smith, Giulio Caravagna, Sasikumar Amarasingam, George Elia, Alison Berner, Ryan Changho Choi, Pradeep Ramagiri, Ritika Chauhan, Nik Matthews, Jamie Murphy, Anthony Antoniou, Susan Clark, Jo-Anne Chin Aleong, Enric Domingo, Inmaculada Spiteri, Stuart AC McDonald, Darryl Shibata, Miangela M. Lacle, Lai Mun Wang, Morgan Moorghen, Ian PM Tomlinson, Marco Novelli, Marnix Jansen, Alan Watson, Nicholas A. Wright, John Bridgewater, Manuel Rodriguez-Justo, Hemant Kocher, Simon J. Leedham, Andrea Sottoriva, and Trevor A. Graham (Mar. 29, 2020). “Stabilising selection causes grossly altered but stable karyotypes in metastatic colorectal cancer”. In: *bioRxiv*, p. 2020.03.26.007138.

Darwin 1809-1882, Charles (1859). *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life.*

Darwin, Charles and Alfred Wallace (Aug. 1, 1858). “On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection”. In: *Zoological Journal of the Linnean Society* 3.9, pp. 45–62.

Davis, Alexander, Ruli Gao, and Nicholas Navin (Apr. 2017). “Tumor evolution: Linear, branching, neutral or punctuated?” In: *Biochimica Et Biophysica Acta. Reviews on Cancer* 1867.2, pp. 151–161.

Del Monte, Ugo (Feb. 2009). “Does the cell number 10 ^{9} still really fit one gram of tumor tissue?” In: *Cell Cycle* 8.3, pp. 505–506.

Del Moral, Pierre, Arnaud Doucet, and Ajay Jasra (Sept. 1, 2012). “An adaptive sequential Monte Carlo method for approximate Bayesian computation”. In: *Statistics and Computing* 22.5, pp. 1009–1020.

DeLeo, A. B., G. Jay, E. Appella, G. C. Dubois, L. W. Law, and L. J. Old (May 1, 1979). “Detection of a transformation-related antigen in chemically induced sarcomas and other transformed cells of the mouse”. In: *Proceedings of the National Academy of Sciences* 76.5, pp. 2420–2424.

Denissenko, M. F., A. Pao, M. Tang, and G. P. Pfeifer (Oct. 18, 1996). “Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53”. In: *Science (New York, N.Y.)* 274.5286, pp. 430–432.

Dentro, Stefan C., David C. Wedge, and Peter Van Loo (Aug. 1, 2017). “Principles of Reconstructing the Subclonal Architecture of Cancers”. In: *Cold Spring Harbor Perspectives in Medicine* 7.8.

Dentro, Stefan C. et al. (Apr. 7, 2021). “Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes”. In: *Cell.*

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, Guillermo del Angel, Manuel A. Rivas, Matt Hanna, Aaron McKenna, Tim J. Fennell, Andrew M. Kernytsky, Andrey Y. Sivachenko, Kristian Cibulskis, Stacey B. Gabriel, David Altshuler, and Mark J. Daly (May 2011). “A framework for variation discovery and genotyping using next-generation DNA sequencing data”. In: *Nature Genetics* 43.5, pp. 491–498.

Desper, R., F. Jiang, O. P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer (1999). “Inferring tree models for oncogenesis from comparative genome hybridization data”. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 6.1, pp. 37–51.

Deutsch-Wenzel, Reintraud P., Horst Brune, Gernot Grimmer, Gerhard Dettbarn, and Jürgen Misfeld (Sept. 1, 1983). "Experimental Studies in Rat Lungs on the Carcinogenicity and Dose-Response Relationships of Eight Frequently Occurring Environmental Polycyclic Aromatic Hydrocarbons23". In: *JNCI: Journal of the National Cancer Institute* 71.3, pp. 539–544.

DeVita, V. T., R. C. Young, and G. P. Canellos (Jan. 1975). "Combination versus single agent chemotherapy: a review of the basis for selection of drug treatment of cancer". In: *Cancer* 35.1, pp. 98–110.

Diamond, Benjamin, Venkata Yellapantula, Even H. Rustad, Kylee H. Maclachlan, Marius Mayerhoefer, Martin Kaiser, Gareth Morgan, Ola Landgren, and Francesco Maura (Jan. 22, 2021). "Positive selection as the unifying force for clonal evolution in multiple myeloma". In: *Leukemia*, pp. 1–5.

Diaz Jr, Luis A., Richard T. Williams, Jian Wu, Isaac Kinde, J. Randolph Hecht, Jordan Berlin, Benjamin Allen, Ivana Bozic, Johannes G. Reiter, Martin A. Nowak, Kenneth W. Kinzler, Kelly S. Oliner, and Bert Vogelstein (June 2012). "The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers". In: *Nature* 486.7404, pp. 537–540.

Dietlein, Felix, Donate Weghorn, Amaro Taylor-Weiner, André Richters, Brendan Reardon, David Liu, Eric S. Lander, Eliezer M. Van Allen, and Shamil R. Sunyaev (Feb. 2020). "Identification of cancer driver genes based on nucleotide context". In: *Nature Genetics* 52.2, pp. 208–218.

Dietmaier, W., S. Wallinger, T. Bocker, F. Kullmann, R. Fishel, and J. Rüschoff (Nov. 1, 1997). "Diagnostic microsatellite instability: definition and correlation with mismatch repair protein expression". In: *Cancer Research* 57.21, pp. 4749–4756.

Dillon, Lloye M. and Todd W. Miller (Jan. 2014). "Therapeutic targeting of cancers with loss of PTEN function". In: *Current Drug Targets* 15.1, pp. 65–79.

Dingli, David, Franziska Michor, Tibor Antal, and Jorge M. Pacheco (Mar. 2007). "The emergence of tumor metastases". In: *Cancer Biology & Therapy* 6.3, pp. 383–390.

Dogruluk, Turgut, Yiu Huen Tsang, Maribel Espitia, Fengju Chen, Tenghui Chen, Zechen Chong, Vivek Appadurai, Armel Dogruluk, Agna Karina Eterovic, Penelope E. Bonnen, Chad J. Creighton, Ken Chen, Gordon B. Mills, and Kenneth L. Scott (Dec. 15, 2015). "Identification of Variant-Specific Functions of *PIK3CA* by Rapid Phenotyping of Rare Mutations". In: *Cancer Research* 75.24, pp. 5341–5354.

Domazet-Loso, Tomislav and Diethard Tautz (May 21, 2010). "Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa". In: *BMC biology* 8, p. 66.

Dunham, Ian et al. (Sept. 2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74.

Dunson, David B (2010). "Nonparametric Bayes applications to biostatistics". In: *Bayesian nonparametrics* 28, pp. 223–273.

Durrett, Rick (Feb. 2013). "Population genetics of neutral mutations in exponentially growing cancer cell populations". In: *Annals of Applied Probability* 23.1, pp. 230–250.

Durrett, Rick, Jasmine Foo, Kevin Leder, John Mayberry, and Franziska Michor (Aug. 2010). "Evolutionary dynamics of tumor progression with random fitness values". In: *Theoretical Population Biology* 78.1, pp. 54–66.

Eddelbuettel, Dirk (2013). *Seamless R and C++ Integration with Rcpp*. New York.

Eddelbuettel, Dirk and Romain Francois (Apr. 13, 2011). “Rcpp: Seamless R and C++ Integration”. In: *Journal of Statistical Software* 40.1, pp. 1–18.

Edwards, Jack, Andriy Marusyk, and David Basanta (Oct. 13, 2020). “Selection-driven tumor evolution involving non-cell growth promotion leads to patterns of clonal expansion consistent with neutrality interpretation”. In: *bioRxiv*, p. 2020.02.11.944843.

Efron, Bradley (1992). “Bootstrap methods: another look at the jackknife”. In: *Breakthroughs in statistics*, pp. 569–593.

Enard, David, Le Cai, Carina Gwennap, and Dmitri A. Petrov (May 17, 2016). “Viruses are a dominant driver of protein adaptation in mammals”. In: *eLife* 5.

ENCODE Project Consortium (Sept. 6, 2012). “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414, pp. 57–74.

Everitt, Richard G. (Jan. 17, 2018). “Bootstrapped synthetic likelihood”. In: *arXiv:1711.05825 [physics, stat]*. arXiv: `1711.05825`.

Evrard, Camille, Gaëlle Tachon, Violaine Randrian, Lucie Karayan-Tapon, and David Tougeron (Oct. 15, 2019). “Microsatellite Instability: Diagnosis, Heterogeneity, Discordance, and Clinical Impact in Colorectal Cancer”. In: *Cancers* 11.10.

Fang, Celestia, Zhenjia Wang, Cuijuan Han, Stephanie L. Safgren, Kathryn A. Helmin, Emmalee R. Adelman, Valentina Serafin, Giuseppe Basso, Kyle P. Eagen, Alexandre Gaspar-Maia, Maria E. Figueroa, Benjamin D. Singer, Aakrosh Ratan, Panagiotis Ntziachristos, and Chongzhi Zang (Sept. 15, 2020). “Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation”. In: *Genome Biology* 21.1, p. 247.

Farris, James S. (1970). “Methods for Computing Wagner Trees”. In: *Systematic Zoology* 19.1, pp. 83–92.

Favero, F., T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund (Jan. 2015). “Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data”. In: *Annals of Oncology: Official Journal of the European Society for Medical Oncology* 26.1, pp. 64–70.

Fay, J. C. and C. I. Wu (July 2000). “Hitchhiking under positive Darwinian selection”. In: *Genetics* 155.3, pp. 1405–1413.

Fearon, Eric R. and Bert Vogelstein (June 1, 1990). “A genetic model for colorectal tumorigenesis”. In: *Cell* 61.5, pp. 759–767.

Fedurco, Milan, Anthony Romieu, Scott Williams, Isabelle Lawrence, and Gerardo Turcatti (Feb. 9, 2006). “BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies”. In: *Nucleic Acids Research* 34.3, e22.

Felsenstein, Joseph and Joseph Felenstein (2004). *Inferring phylogenies*. Vol. 2.

Fialkow, Philip J. (Oct. 12, 1976). “Clonal origin of human tumors”. In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 458.3, pp. 283–321.

Field, Adam E., Neil A. Robertson, Tina Wang, Aaron Havas, Trey Ideker, and Peter D. Adams (Sept. 20, 2018). “DNA Methylation Clocks in Aging: Categories, Causes, and Consequences”. In: *Molecular cell* 71.6, pp. 882–895.

Filippi, Sarah, Chris P. Barnes, Julien Cornebise, and Michael P. H. Stumpf (Mar. 26, 2013). “On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo”. In: *Statistical Applications in Genetics and Molecular Biology* 12.1, pp. 87–107.

Filippova, Galina N., Annika Lindblom, Linda J. Meincke, Elena M. Klenova, Paul E. Neiman, Steve J. Collins, Norman A. Doggett, and Victor V. Lobanenkov (1998). "A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers". In: *Genes, Chromosomes and Cancer* 22.1, pp. 26–36.

Fischer, Mareike (Dec. 16, 2020). "Extremal values of the Sackin tree balance index". In: *arXiv:1801.10418 [math, q-bio]*. arXiv: `1801.10418`.

Fisher, R. A. (1923). "XXI.—On the Dominance Ratio". In: *Proceedings of the Royal Society of Edinburgh* 42, pp. 321–341.

Fisher, Ronald Aylmer (1958). *The genetical theory of natural selection*.

Fishilevich, Simon, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, Doron Lancet, and Dana Cohen (2017). "GeneHancer: genome-wide integration of enhancers and target genes in GeneCards". In: *Database: The Journal of Biological Databases and Curation* 2017.

Fitch, Walter M. (1971). "Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology". In: *Systematic Zoology* 20.4, pp. 406–416.

Flavahan, William A., Elizabeth Gaskell, and Bradley E. Bernstein (July 21, 2017). "Epigenetic plasticity and the hallmarks of cancer". In: *Science (New York, N.Y.)* 357.6348.

Fleming, Matthew, Sreelakshmi Ravula, Sergei F. Tatishchev, and Hanlin L. Wang (Sept. 2012). "Colorectal carcinoma: Pathologic aspects". In: *Journal of Gastrointestinal Oncology* 3.3, pp. 153–173.

Fleming, Nicholas I., Robert N. Jorissen, Dmitri Mouradov, Michael Christie, Anuratha Sakthianandeswaren, Michelle Palmieri, Fiona Day, Shan Li, Cary Tsui, Lara Lipton, Jayesh Desai, Ian T. Jones, Stephen McLaughlin, Robyn L. Ward, Nicholas J. Hawkins, Andrew R. Ruszkiewicz, James Moore, Hong-Jian Zhu, John M. Mariadason, Antony W. Burgess, Dana Busam, Qi Zhao, Robert L. Strausberg, Peter Gibbs, and Oliver M. Sieber (Jan. 15, 2013). "SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer". In: *Cancer Research* 73.2, pp. 725–735.

Fontana, Gabriele A and Hailey L Gahlon (Nov. 18, 2020). "Mechanisms of replication and repair in mitochondrial DNA deletion formation". In: *Nucleic Acids Research* 48.20, pp. 11244–11258.

Fousteri, Maria and Leon HF Mullenders (Jan. 2008). "Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects". In: *Cell Research* 18.1, pp. 73–84.

Fraley, Chris and Adrian E. Raftery (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation". In: *Journal of the American Statistical Association* 97.458, pp. 611–631.

Fraley, Chris, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca (2012). *mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation*. Technical report.

Frazer, Kelly A. et al. (Oct. 2007). "A second generation human haplotype map of over 3.1 million SNPs". In: *Nature* 449.7164, pp. 851–861.

Friberg, S. and S. Mattson (Aug. 1997). "On the growth rates of human malignant tumors: implications for medical decision making". In: *Journal of Surgical Oncology* 65.4, pp. 284–297.

Friedewald, William F. and Peyton Rous (Aug. 1, 1944). "THE INITIATING AND PROMOTING ELEMENTS IN TUMOR PRODUCTION : AN ANALYSIS OF THE EFFECTS OF TAR, BENZPYRENE, AND METHYLCHOLANTHRENE ON RABBIT SKIN". In: *Journal of Experimental Medicine* 80.2, pp. 101–126.

Friend, S. H., R. Bernards, S. Rogelj, R. A. Weinberg, J. M. Rapaport, D. M. Albert, and T. P. Dryja (Oct. 16, 1986). "A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma". In: *Nature* 323.6089, pp. 643–646.

Fu, Y. X. and W. H. Li (Mar. 1993). "Statistical Tests of Neutrality of Mutations". In: *Genetics* 133.3, pp. 693–709.

Fujimoto, Akihiro, Masashi Fujita, Takanori Hasegawa, Jing Hao Wong, Kazuhiro Maejima, Aya Oku-Sasaki, Kaoru Nakano, Yuichi Shiraishi, Satoru Miyano, Go Yamamoto, Kiwamu Akagi, Seiya Imoto, and Hidewaki Nakagawa (Mar. 2020). "Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types". In: *Genome Research* 30.3, pp. 334–346.

Fusco, Diana, Matti Gralka, Jona Kayser, Alex Anderson, and Oskar Hallatschek (Oct. 3, 2016). "Excess of mutational jackpot events in expanding populations revealed by spatial Luria–Delbrück experiments". In: *Nature Communications* 7.

Galatenko, Vladimir V., Diana V. Maltseva, Alexey V. Galatenko, Sergey Rodin, and Alexander G. Tonevitsky (Feb. 13, 2018). "Cumulative prognostic power of laminin genes in colorectal cancer". In: *BMC Medical Genomics* 11.1, p. 9.

Garcia, S. B., H. S. Park, M. Novelli, and N. A. Wright (Jan. 1999). "Field cancerization, clonality, and epithelial stem cells: the spread of mutated clones in epithelial sheets". In: *The Journal of Pathology* 187.1, pp. 61–81.

Gawad, Charles, Winston Koh, and Stephen R. Quake (Mar. 2016). "Single-cell genome sequencing: current state of the science". In: *Nature Reviews. Genetics* 17.3, pp. 175–188.

Gerlee, Philip (Apr. 15, 2013). "The model muddle: in search of tumor growth laws". In: *Cancer Research* 73.8, pp. 2407–2411.

Gerlinger, Marco, Stuart Horswell, James Larkin, Andrew J. Rowan, Max P. Salm, Ignacio Varela, Rosalie Fisher, Nicholas McGranahan, Nicholas Matthews, Claudio R. Santos, Pierre Martinez, Benjamin Phillimore, Sharmin Begum, Adam Rabinowitz, Bradley Spencer-Dene, Sakshi Gulati, Paul A. Bates, Gordon Stamp, Lisa Pickering, Martin Gore, David L. Nicol, Steven Hazell, P. Andrew Futreal, Aengus Stewart, and Charles Swanton (Mar. 2014). "Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing". In: *Nature Genetics* 46.3, pp. 225–233.

Gerlinger, Marco, Andrew J. Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q. McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R. Santos, Mahrokh Nohadani, Aron C. Eklund, Bradley Spencer-Dene, Graham Clark, Lisa Pickering, Gordon Stamp, Martin Gore, Zoltan Szallasi, Julian Downward, P. Andrew Futreal, and Charles Swanton (Mar. 8, 2012). "Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing". In: *New England Journal of Medicine* 366.10, pp. 883–892.

Gerstung, Moritz, Christian Beisel, Markus Rechsteiner, Peter Wild, Peter Schraml, Holger Moch, and Niko Beerenwinkel (May 1, 2012). "Reliable detection of subclonal single-nucleotide variants in tumour cell populations". In: *Nature Communications* 3.1, p. 811.

Gerstung, Moritz, Elli Papaemmanuil, and Peter J. Campbell (May 1, 2014). “Subclonal variant calling with multiple samples and prior knowledge”. In: *Bioinformatics* 30.9, pp. 1198–1204.

Ghadimi, B. Michael, Marian Grade, Carsten Mönkemeyer, Bettina Kulle, Jochen Gaedcke, Bastian Gunawan, Claus Langer, Torsten Liersch, and Heinz Becker (2006). “Distinct chromosomal profiles in metastasizing and non-metastasizing colorectal carcinomas”. In: *Cellular Oncology: The Official Journal of the International Society for Cellular Oncology* 28.5, pp. 273–281.

Ghirlando, Rodolfo and Gary Felsenfeld (Apr. 15, 2016). “CTCF: making the right connections”. In: *Genes & Development* 30.8, pp. 881–891.

Giannakis, Marios, Eran Hodis, Xinmeng Jasmine Mu, Mai Yamauchi, Joseph Rosenbluh, Kristian Cibulskis, Gordon Saksena, Michael S. Lawrence, Zhi Rong Qian, Reiko Nishihara, Eliezer M. Van Allen, William C. Hahn, Stacey B. Gabriel, Eric S. Lander, Gad Getz, Shuji Ogino, Charles S. Fuchs, and Levi A. Garraway (Dec. 2014). “RNF43 is frequently mutated in colorectal and endometrial cancers”. In: *Nature Genetics* 46.12, pp. 1264–1266.

Gillespie, Daniel T (Dec. 1, 1976). “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”. In: *Journal of Computational Physics* 22.4, pp. 403–434.

— (Dec. 1, 1977). “Exact stochastic simulation of coupled chemical reactions”. In: *The Journal of Physical Chemistry* 81.25, pp. 2340–2361.

— (2007). “Stochastic Simulation of Chemical Kinetics”. In: *Annual Review of Physical Chemistry* 58.1, pp. 35–55.

Gillespie, John H (1984). “The status of the neutral theory: the neutral theory of molecular evolution.” In: *Science* 224.4650, pp. 732–733.

Glover, Kevin Alan, Cino Pertoldi, Francois Besnier, Vidar Wennevik, Matthew Kent, and Øystein Skaala (Aug. 23, 2013). “Atlantic salmon populations invaded by farmed escapees: quantifying genetic introgression with a Bayesian approach and SNPs”. In: *BMC Genetics* 14.1, p. 74.

Goh, Amanda M, Cynthia R Coffill, and David P Lane (2011). “The role of mutant p53 in human cancer”. In: *The Journal of Pathology* 223.2, pp. 116–126.

Goldman, N and Z Yang (Sept. 1, 1994). “A codon-based model of nucleotide substitution for protein-coding DNA sequences.” In: *Molecular Biology and Evolution* 11.5, pp. 725–736.

Gong, Yixiao, Charalampos Lazaris, Theodore Sakellaropoulos, Aurelie Lozano, Prabhanjan Kambadur, Panagiotis Ntziachristos, Iannis Aifantis, and Aristotelis Tsirigos (Feb. 7, 2018). “Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries”. In: *Nature Communications* 9.1, p. 542.

Gonzalez-Perez, Abel, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P. Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas (Nov. 2013). “IntOGen-mutations identifies cancer drivers across tumor types”. In: *Nature Methods* 10.11, pp. 1081–1082.

González-González, María, Celia Fontanillo, María M. Abad, María L. Gutiérrez, Ines Mota, Oscar Bengoechea, Ángel Santos-Briz, Oscar Blanco, Emilio Fonseca, Juana Ciudad, Manuel Fuentes, Javier De Las Rivas, José A. Alcazar, Jacinto García, Luís Muñoz-Bellvis, Alberto Orfao, and José M. Sayagués (2014). “Identification of a characteristic copy number alteration profile by high-resolution single nucleotide polymorphism arrays associated with metastatic sporadic colorectal cancer". In: *Cancer* 120.13, pp. 1948–1959.

Gordon-Weeks, Alex, Su Yin Lim, Arseniy Yuzhalin, Serena Lucotti, Jenny Adriana Francisca Vermeer, Keaton Jones, Jianzhou Chen, and Ruth J. Muschel (May 6, 2019). "Tumour-Derived Laminin α5 (LAMA5) Promotes Colorectal Liver Metastasis Growth, Branching Angiogenesis and Notch Pathway Inhibition". In: *Cancers* 11.5.

Graham, Trevor A., Adam Humphries, Theodore Sanders, Manuel Rodriguez–Justo, Paul J. Tadrous, Sean L. Preston, Marco R. Novelli, Simon J. Leedham, Stuart A. C. McDonald, and Nicholas A. Wright (Apr. 1, 2011). "Use of Methylation Patterns to Determine Expansion of Stem Cell Clones in Human Colon Tissue". In: *Gastroenterology* 140.4, 1241–1250.e9.

Grazian, Clara and Yanan Fan (Sept. 6, 2019). "A review of Approximate Bayesian Computation methods via density estimation: inference for simulator-models". In: *arXiv:1909.02736 [stat]*. arXiv: 1909.02736.

Greaves, Laura C., Sean L. Preston, Paul J. Tadrous, Robert W. Taylor, Martin J. Barron, Dahmane Oukrif, Simon J. Leedham, Maesha Deheragoda, Peter Sasieni, Marco R. Novelli, Janusz A. Z. Jankowski, Douglass M. Turnbull, Nicholas A. Wright, and Stuart A. C. McDonald (Jan. 17, 2006). "Mitochondrial DNA mutations are established in human colonic stem cells, and mutated clones expand by crypt fission". In: *Proceedings of the National Academy of Sciences* 103.3, pp. 714–719.

Greaves, Mel (Aug. 2015). "Evolutionary Determinants of Cancer". In: *Cancer discovery* 5.8, pp. 806–820.

Greaves, Mel and Carlo C. Maley (Jan. 2012). "Clonal evolution in cancer". In: *Nature* 481.7381, pp. 306–313.

Green, Shon, Christy L. Trejo, and Martin McMahon (Dec. 15, 2015). "PIK3CA(H1047R) Accelerates and Enhances KRAS(G12D)-Driven Lung Tumorigenesis". In: *Cancer Research* 75.24, pp. 5378–5391.

Griffiths, R. C. and Simon Tavaré (Jan. 1, 1998). "The age of a mutation in a general coalescent tree". In: *Communications in Statistics. Stochastic Models* 14.1, pp. 273–295.

Grivennikov, Sergei I., Florian R. Greten, and Michael Karin (Mar. 19, 2010). "Immunity, Inflammation, and Cancer". In: *Cell* 140.6, pp. 883–899.

Grossman, Robert L., Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt (Sept. 22, 2016). "Toward a Shared Vision for Cancer Genomic Data". In: *New England Journal of Medicine* 375.12, pp. 1109–1112.

Gundem, Gunes, Peter Van Loo, Barbara Kremeyer, Ludmil B. Alexandrov, Jose M. C. Tubio, Elli Papaemmanuil, Daniel S. Brewer, Heini M. L. Kallio, Gunilla Högnäs, Matti Annala, Kati Kivinummi, Victoria Goody, Calli Latimer, Sarah O'Meara, Kevin J. Dawson, William Isaacs, Michael R. Emmert-Buck, Matti Nykter, Christopher Foster, Zsofia Kote-Jarai, Douglas Easton, Hayley C. Whitaker, David E. Neal, Colin S. Cooper, Rosalind A. Eeles, Tapio Visakorpi, Peter J. Campbell, Ultan McDermott, David C. Wedge, and G. Steven Bova (Apr. 2015). "The evolutionary history of lethal metastatic prostate cancer". In: *Nature* 520.7547, pp. 353–357.

Habano, Wataru, Shin-ichi Nakamura, and Tamotsu Sugai (Oct. 1998). "Microsatellite instability in the mitochondrial DNA of colorectal carcinomas: Evidence for mismatch repair systems in mitochondrial genome". In: *Oncogene* 17.15, pp. 1931–1937.

Haddrill, Penelope R., Kevin R. Thornton, Brian Charlesworth, and Peter Andolfatto (June 2005). "Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations". In: *Genome Research* 15.6, pp. 790–799.

Haeckel, Ernst (1866). *Generelle Morphologie der Organismen. Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Descendenz-Theorie, etc*. Vol. 1.

Haeno, Hiroshi and Franziska Michor (Mar. 7, 2010). "The evolution of tumor metastases during clonal expansion". In: *Journal of theoretical biology* 263.1, pp. 30–44.

Haeussler, Maximilian, Ann S. Zweig, Cath Tyner, Matthew L. Speir, Kate R. Rosenbloom, Brian J. Raney, Christopher M. Lee, Brian T. Lee, Angie S. Hinrichs, Jairo Navarro Gonzalez, David Gibson, Mark Diekhans, Hiram Clawson, Jonathan Casper, Galt P. Barber, David Haussler, Robert M. Kuhn, and W. James Kent (Jan. 8, 2019). "The UCSC Genome Browser database: 2019 update". In: *Nucleic Acids Research* 47 (D1), pp. D853–D858.

Hahne, Florian and Robert Ivanek (2016). "Visualizing Genomic Data Using Gviz and Bioconductor". In: *Statistical Genomics: Methods and Protocols*. Ed. by Ewy Mathé and Sean Davis. Methods in Molecular Biology. New York, NY, pp. 335–351.

Hajdu, Steven I. (2011). "A note from history: Landmarks in history of cancer, part 1". In: *Cancer* 117.5, pp. 1097–1102.

Haldane, J. B. S. (Dec. 1, 1957). "The cost of natural selection". In: *Journal of Genetics* 55.3, p. 511.

Halligan, Daniel L., Fiona Oliver, Adam Eyre-Walker, Bettina Harr, and Peter D. Keightley (Jan. 22, 2010). "Evidence for pervasive adaptive protein evolution in wild mice". In: *PLoS genetics* 6.1, e1000825.

Hamilton, Nicholas E. and Michael Ferry (2018). "ggtern: Ternary Diagrams Using ggplot2". In: *Journal of Statistical Software, Code Snippets* 87.3, pp. 1–17.

Hamilton, Stanley R, Lauri A Aaltonen, et al. (2000). *Pathology and genetics of tumours of the digestive system*. Vol. 2.

Han, Shuang-Yin, Hideaki Kato, Shunsuke Kato, Takao Suzuki, Hiroyuki Shibata, Seiichi Ishii, Ken-ichi Shiiba, Seiki Matsuno, Ryunosuke Kanamaru, and Chikashi Ishioka (June 15, 2000). "Functional Evaluation of PTEN Missense Mutations Using in Vitro Phosphoinositide Phosphatase Assay". In: *Cancer Research* 60.12, pp. 3147–3151.

Hanahan, Douglas and Robert A. Weinberg (Jan. 7, 2000). "The Hallmarks of Cancer". In: *Cell* 100.1, pp. 57–70.

Hansemann, David (1890). "Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung". In: *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin* 119.2, pp. 299–326.

Hao, Huai-Xiang, Yang Xie, Yue Zhang, Olga Charlat, Emma Oster, Monika Avello, Hong Lei, Craig Mickanin, Dong Liu, Heinz Ruffner, Xiaohong Mao, Qicheng Ma, Raffaella Zamponi, Tewis Bouwmeester, Peter M. Finan, Marc W. Kirschner, Jeffery A. Porter, Fabrizio C. Serluca, and Feng Cong (May 2012). "ZNRF3 promotes Wnt receptor turnover in an R-spondin-sensitive manner". In: *Nature* 485.7397, pp. 195–200.

Hao, Yun-He, Shu-Yong Yu, Rui-Sha Tu, and Yao-Qing Cai (Jan. 2, 2020). "TNNT1, a prognostic indicator in colon adenocarcinoma, regulates cell behaviors and mediates EMT process". In: *Bioscience, Biotechnology, and Biochemistry* 84.1, pp. 111–117.

He, Xin, Nicholas Arrotta, Deepa Radhakrishnan, Yu Wang, Todd Romigh, and Charis Eng (May 15, 2013). "Cowden Syndrome-Related Mutations in *PTEN* Associate with Enhanced Proteasome Activity". In: *Cancer Research* 73.10, pp. 3029–3040.

Heckbert, Paul S. (1990). “DIGITAL LINE DRAWING”. In: *Graphics Gems*. Ed. by ANDREW S. GLASSNER. San Diego, pp. 99–100.

Heide, Timon, Luis Zapata, Marc J. Williams, Benjamin Werner, Giulio Caravagna, Chris P. Barnes, Trevor A. Graham, and Andrea Sottoriva (Dec. 2018). “Reply to ‘Neutral tumor evolution?’” In: *Nature Genetics* 50.12, pp. 1633–1637.

Heijden, Maartje van der, Daniël M. Miedema, Bartlomiej Waclaw, Veronique L. Veenstra, Maria C. Lecca, Lisanne E. Nijman, Erik van Dijk, Sanne M. van Neerven, Sophie C. Lodestijn, Kristiaan J. Lenos, Nina E. de Groot, Pramudita R. Prasetyanti, Andrea Arricibita Varea, Douglas J. Winton, Jan Paul Medema, Edward Morrissey, Bauke Ylstra, Martin A. Nowak, Maarten F. Bijlsma, and Louis Vermeulen (Mar. 26, 2019). “Spatiotemporal regulation of clonogenicity in colorectal cancer xenografts”. In: *Proceedings of the National Academy of Sciences* 116.13, pp. 6140–6145.

Heitzer, Ellen and Ian Tomlinson (Feb. 2014). “Replicative DNA polymerase mutations in cancer”. In: *Current Opinion in Genetics & Development* 24, pp. 107–113.

Herman, J. G., A. Umar, K. Polyak, J. R. Graff, N. Ahuja, J. P. Issa, S. Markowitz, J. K. Willson, S. R. Hamilton, K. W. Kinzler, M. F. Kane, R. D. Kolodner, B. Vogelstein, T. A. Kunkel, and S. B. Baylin (June 9, 1998). “Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma”. In: *Proceedings of the National Academy of Sciences of the United States of America* 95.12, pp. 6870–6875.

Hinegardner, R. T. and J. Engelberg (Nov. 22, 1963). “RATIONALE FOR A UNIVERSAL GENETIC CODE”. In: *Science (New York, N.Y.)* 142.3595, pp. 1083–1085.

Hlubek, F., A. Jung, N. Kotzor, T. Kirchner, and T. Brabletz (Nov. 15, 2001). “Expression of the invasion factor laminin gamma2 in colorectal carcinomas is regulated by beta-catenin”. In: *Cancer Research* 61.22, pp. 8089–8093.

Hnisz, Denes, Abraham S. Weintraub, Daniel S. Day, Anne-Laure Valton, Rasmus O. Bak, Charles H. Li, Johanna Goldmann, Bryan R. Lajoie, Zi Peng Fan, Alla A. Sigova, Jessica Reddy, Diego Borges-Rivera, Tong Ihn Lee, Rudolf Jaenisch, Matthew H. Porteus, Job Dekker, and Richard A. Young (Mar. 25, 2016). “Activation of proto-oncogenes by disruption of chromosome neighborhoods”. In: *Science (New York, N.Y.)* 351.6280, pp. 1454–1458.

Hoodless, P. A., T. Tsukazaki, S. Nishimatsu, L. Attisano, J. L. Wrana, and G. H. Thomsen (Mar. 15, 1999). “Dominant-negative Smad2 mutants inhibit activin/Vg1 signaling and disrupt axis formation in Xenopus”. In: *Developmental Biology* 207.2, pp. 364–379.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp (Jan. 1, 2002). “The Ensembl genome database project”. In: *Nucleic Acids Research* 30.1, pp. 38–41.

Huebner, R. J. and G. J. Todaro (Nov. 1969). “Oncogenes of RNA tumor viruses as determinants of cancer”. In: *Proceedings of the National Academy of Sciences of the United States of America* 64.3, pp. 1087–1094.

Humphries, Adam, Biancastella Cereser, Laura J. Gay, Daniel S. J. Miller, Bibek Das, Alice Gutteridge, George Elia, Emma Nye, Rosemary Jeffery, Richard Poulsom, Marco R. Novelli, Manuel Rodriguez-Justo, Stuart A. C. McDonald, Nicholas A. Wright, and Trevor A. Graham (July 2, 2013). “Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution”. In: *Proceedings of the National Academy of Sciences* 110.27, E2490–E2499.

Humphries, Adam and Nicholas A. Wright (June 2008). “Colonic crypt organization and tumorigenesis”. In: *Nature Reviews Cancer* 8.6, pp. 415–424.

Iacobuzio-Donahue, Christine A., Jason Song, Giovanni Parmiagiani, Charles J. Yeo, Ralph H. Hruban, and Scott E. Kern (Mar. 1, 2004). “Missense Mutations of MADH4: Characterization of the Mutational Hot Spot and Functional Consequences in Human Tumors”. In: *Clinical Cancer Research* 10.5, pp. 1597–1604.

Isaksson-Mettävainio, Martin, Richard Palmqvist, Johan Forssell, Roger Stenling, and Åke Öberg (Jan. 1, 2006). “SMAD4/DPC4 Expression and Prognosis in Human Colorectal Cancer”. In: *Anticancer Research* 26.1, pp. 507–510.

Ives, Anthony R., Peter E. Midford, and Theodore Garland Jr. (Apr. 1, 2007). “Within-Species Variation and Measurement Error in Phylogenetic Comparative Methods”. In: *Systematic Biology* 56.2, pp. 252–270.

Iwasa, Yoh, Martin A. Nowak, and Franziska Michor (Apr. 2006). “Evolution of resistance during clonal expansion”. In: *Genetics* 172.4, pp. 2557–2566.

Iwatsuki, Masaaki, Koshi Mimori, Hideshi Ishii, Takehiko Yokobori, Yasushi Takatsuno, Tetsuya Sato, Hiroyuki Toh, Ichiro Onoyama, Keiichi I. Nakayama, Hideo Baba, and Masaki Mori (2010). “Loss of FBXW7, a cell cycle regulating gene, in colorectal cancer: Clinical significance”. In: *International Journal of Cancer* 126.8, pp. 1828–1837.

Jahn, Katharina, Jack Kuipers, and Niko Beerenwinkel (May 5, 2016). “Tree inference for single-cell data”. In: *Genome Biology* 17.1, p. 86.

Jass, J. R., J. Young, and B. A. Leggett (Feb. 2001). “Biological Significance of Microsatellite Instability-Low (MSI-L) Status in Colorectal Tumors”. In: *The American Journal of Pathology* 158.2, pp. 779–781.

Javed, Nauman, Yossi Farjoun, Tim Fennell, Charles Epstein, Bradley E. Bernstein, and Noam Shoresh (Jan. 1, 2020). “Detecting sample swaps in diverse NGS data types using linkage disequilibrium”. In: *bioRxiv*, p. 2020.03.15.992750.

Javier, Breanna M., Rona Yaeger, Lu Wang, Francisco Sanchez-Vega, Ahmet Zehir, Sumit Middha, Justyna Sadowska, Efsevia Vakiani, Jinru Shia, David Klimstra, Marc Ladanyi, Christine A. Iacobuzio-Donahue, and Jaclyn F. Hechtman (May 29, 2016). “Recurrent, truncating SOX9 mutations are associated with SOX9 overexpression, KRAS mutation, and TP53 wild type status in colorectal carcinoma”. In: *Oncotarget* 7.32, pp. 50875–50882.

Jensen, Jeffrey D., Yuseob Kim, Vanessa Bauer DuMont, Charles F. Aquadro, and Carlos D. Bustamante (July 2005). “Distinguishing between selective sweeps and demography using DNA polymorphism data”. In: *Genetics* 170.3, pp. 1401–1410.

Jensen, Jeffrey D., Bret A. Payseur, Wolfgang Stephan, Charles F. Aquadro, Michael Lynch, Deborah Charlesworth, and Brian Charlesworth (Jan. 2019). “The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018”. In: *Evolution; international journal of organic evolution* 73.1, pp. 111–114.

Jiang, Hongshan, Rong Lei, Shou-Wei Ding, and Shuifang Zhu (June 12, 2014). “Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads”. In: *BMC Bioinformatics* 15.1, p. 182.

Jiang, Xiaomo, Olga Charlat, Raffaella Zamponi, Yi Yang, and Feng Cong (May 7, 2015). “Dishevelled Promotes Wnt Receptor Degradation through Recruitment of ZNRF3/RNF43 E3 Ubiquitin Ligases”. In: *Molecular Cell* 58.3, pp. 522–533.

Jiang, Xiaomo, Huai-Xiang Hao, Joseph D. Growney, Steve Woolfenden, Cindy Bottiglio, Nicholas Ng, Bo Lu, Mindy H. Hsieh, Linda Bagdasarian, Ronald Meyer, Timothy R.

Smith, Monika Avello, Olga Charlat, Yang Xie, Jeffery A. Porter, Shifeng Pan, Jun Liu, Margaret E. McLaughlin, and Feng Cong (July 30, 2013). "Inactivating mutations of RNF43 confer Wnt dependency in pancreatic ductal adenocarcinoma". In: *Proceedings of the National Academy of Sciences of the United States of America* 110.31, pp. 12649–12654.

Jiao, Wei, Shankar Vembu, Amit G. Deshwar, Lincoln Stein, and Quaid Morris (Feb. 1, 2014). "Inferring clonal evolution of tumors from single nucleotide somatic mutations". In: *BMC bioinformatics* 15, p. 35.

Johnstone, Sarah E., Alejandro Reyes, Yifeng Qi, Carmen Adriaens, Esmat Hegazi, Karin Pelka, Jonathan H. Chen, Luli S. Zou, Yotam Drier, Vivian Hecht, Noam Shoresh, Martin K. Selig, Caleb A. Lareau, Sowmya Iyer, Son C. Nguyen, Eric F. Joyce, Nir Hacohen, Rafael A. Irizarry, Bin Zhang, Martin J. Aryee, and Bradley E. Bernstein (Sept. 17, 2020). "Large-Scale Topological Changes Restrain Malignant Progression in Colorectal Cancer". In: *Cell* 182.6, 1474–1489.e23.

Jolly, Mohit Kumar, Prakash Kulkarni, Keith Weninger, John Orban, and Herbert Levine (2018). "Phenotypic Plasticity, Bet-Hedging, and Androgen Independence in Prostate Cancer: Role of Non-Genetic Heterogeneity". In: *Frontiers in Oncology* 8.

Jones, David, Keiran M. Raine, Helen Davies, Patrick S. Tarpey, Adam P. Butler, Jon W. Teague, Serena Nik-Zainal, and Peter J. Campbell (Dec. 8, 2016). "cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data". In: *Current protocols in bioinformatics* 56, pp. 15.10.1–15.10.18.

Jones, Peter A. and Stephen B. Baylin (Feb. 23, 2007). "The Epigenomics of Cancer". In: *Cell* 128.4, pp. 683–692.

Jones, Siân, Xiaosong Zhang, D. Williams Parsons, Jimmy Cheng-Ho Lin, Rebecca J. Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, Hirohiko Kamiyama, Antonio Jimeno, Seung-Mo Hong, Baojin Fu, Ming-Tseh Lin, Eric S. Calhoun, Mihoko Kamiyama, Kimberly Walter, Tatiana Nikolskaya, Yuri Nikolsky, James Hartigan, Douglas R. Smith, Manuel Hidalgo, Steven D. Leach, Alison P. Klein, Elizabeth M. Jaffee, Michael Goggins, Anirban Maitra, Christine Iacobuzio-Donahue, James R. Eshleman, Scott E. Kern, Ralph H. Hruban, Rachel Karchin, Nickolas Papadopoulos, Giovanni Parmigiani, Bert Vogelstein, Victor E. Velculescu, and Kenneth W. Kinzler (Sept. 26, 2008). "Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses". In: *Science (New York, N.Y.)* 321.5897, pp. 1801–1806.

Joost, Patrick, Nynke Veurink, Susanne Holck, Louise Klarskov, Anders Bojesen, Maria Harbo, Bo Baldetorp, Eva Rambech, and Mef Nilbert (June 26, 2014). "Heterogenous mismatch-repair status in colorectal cancer". In: *Diagnostic Pathology* 9.1, p. 126.

Jung, Gerhard, Eva Hernández-Illán, Leticia Moreira, Francesc Balaguer, and Ajay Goel (Feb. 2020). "Epigenetics of colorectal cancer: biomarker and therapeutic potential". In: *Nature Reviews Gastroenterology & Hepatology* 17.2, pp. 111–130.

Kaessmann, Henrik, Victor Wiebe, Gunter Weiss, and Svante Pääbo (Feb. 2001). "Great ape DNA sequences reveal a reduced diversity and an expansion in humans". In: *Nature Genetics* 27.2, pp. 155–156.

Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F. McMichael, Matthew A. Wyczalkowski, Mark D. M. Leiserson, Christopher A. Miller, John S. Welch, Matthew J. Walter, Michael C. Wendl, Timothy J. Ley, Richard K. Wilson, Benjamin J. Raphael, and Li Ding (Oct. 2013). "Mutational landscape and significance across 12 major cancer types". In: *Nature* 502.7471, pp. 333–339.

Kaneda, Hiroyasu, Tokuzo Arao, Kaoru Tanaka, Daisuke Tamura, Keiichi Aomatsu, Kanae Kudo, Kazuko Sakai, Marco A. De Velasco, Kazuko Matsumoto, Yoshihiko Fujita, Yasuhide Yamada, Junji Tsurutani, Isamu Okamoto, Kazuhiko Nakagawa, and Kazuto Nishio (Mar. 1, 2010). "FOXQ1 Is Overexpressed in Colorectal Cancer and Enhances Tumorigenicity and Tumor Growth". In: *Cancer Research* 70.5, pp. 2053–2063.

Karabatsos, George and Fabrizio Leisen (May 8, 2018). "An Approximate Likelihood Perspective on ABC Methods". In: *arXiv:1708.05341 [stat]*. arXiv: `1708.05341`.

Katainen, Riku, Kashyap Dave, Esa Pitkänen, Kimmo Palin, Teemu Kivioja, Niko Välimäki, Alexandra E. Gylfe, Heikki Ristolainen, Ulrika A. Hänninen, Tatiana Cajuso, Johanna Kondelin, Tomas Tanskanen, Jukka-Pekka Mecklin, Heikki Järvinen, Laura Renkonen-Sinisalo, Anna Lepistö, Eevi Kaasinen, Outi Kilpivaara, Sari Tuupanen, Martin Enge, Jussi Taipale, and Lauri A. Aaltonen (July 2015). "CTCF/cohesin-binding sites are frequently mutated in cancer". In: *Nature Genetics* 47.7, pp. 818–821.

Kawaguchi, Yoshikuni, Scott Kopetz, Timothy E. Newhook, Mario De Bellis, Yun Shin Chun, Ching-Wei D. Tzeng, Thomas A. Aloia, and Jean-Nicolas Vauthey (Oct. 1, 2019). "Mutation Status of RAS, TP53, and SMAD4 is Superior to Mutation Status of RAS Alone for Predicting Prognosis after Resection of Colorectal Liver Metastases". In: *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 25.19, pp. 5843–5851.

Kemp, Christopher J., James M. Moore, Russell Moser, Brady Bernard, Matt Teater, Leslie E. Smith, Natalia A. Rabaia, Kay E. Gurley, Justin Guinney, Stephanie E. Busch, Rita Shaknovich, Victor V. Lobanenkov, Denny Liggitt, Ilya Shmulevich, Ari Melnick, and Galina N. Filippova (May 22, 2014). "CTCF Haploinsufficiency Destabilizes DNA Methylation and Predisposes to Cancer". In: *Cell Reports* 7.4, pp. 1020–1029.

El-Kenawi, Asmaa and Brian Ruffell (Dec. 2017). "Inflammation, ROS, and Mutagenesis". In: *Cancer Cell* 32.6, pp. 727–729.

Kendall, David G. (1960). "Birth-and-Death Processes, and the Theory of Carcinogenesis". In: *Biometrika* 47.1, pp. 13–21.

Kern, Andrew D and Matthew W Hahn (June 1, 2018). "The Neutral Theory in Light of Natural Selection". In: *Molecular Biology and Evolution* 35.6. Ed. by Sudhir Kumar, pp. 1366–1371.

Kessler, David A., Robert H. Austin, and Herbert Levine (Sept. 1, 2014). "Resistance to chemotherapy: patient variability and cellular heterogeneity". In: *Cancer Research* 74.17, pp. 4663–4670.

Kessler, David A. and Herbert Levine (July 16, 2013). "Large population solution of the stochastic Luria–Delbrück evolution model". In: *Proceedings of the National Academy of Sciences* 110.29, pp. 11682–11687.

— (Feb. 1, 2015). "Scaling Solution in the Large Population Limit of the General Asymmetric Stochastic Luria–Delbrück Evolution Process". In: *Journal of Statistical Physics* 158.4, pp. 783–805.

Khan, Khurum H., David Cunningham, Benjamin Werner, Georgios Vlachogiannis, Inmaculada Spiteri, Timon Heide, Javier Fernandez Mateos, Alexandra Vatsiou, Andrea Lampis, Mahnaz Darvish Damavandi, Hazel Lote, Ian Said Huntingford, Somaieh Hedayat, Ian Chau, Nina Tunariu, Giulia Mentrasti, Francesco Trevisani, Sheela Rao, Gayathri Anandappa, David Watkins, Naureen Starling, Janet Thomas, Clare Peckitt, Nasir Khan, Massimo Rugge, Ruwaida Begum, Blanka Hezelova, Annette Bryant, Thomas Jones, Paula Proszek, Matteo Fassan, Jens C. Hahne, Michael Hubank, Chiara Braconi, Andrea Sottoriva, and Nicola Valeri (Oct. 1, 2018). "Longitudinal Liquid

Biopsy and Mathematical Modeling of Clonal Evolution Forecast Time to Treatment Failure in the PROSPECT-C Phase II Colorectal Cancer Clinical Trial". In: *Cancer Discovery* 8.10, pp. 1270–1285.

Kim, Min Sun and Young Jin Park (Aug. 14, 2007). "Detection and treatment of synchronous lesions in colorectal cancer: The clinical implication of perioperative colonoscopy". In: *World Journal of Gastroenterology : WJG* 13.30, pp. 4108–4111.

Kim, Tae Hoon, Ziedulla K. Abdullaev, Andrew D. Smith, Keith A. Ching, Dmitri I. Loukinov, Roland D. Green, Michael Q. Zhang, Victor V. Lobanenkov, and Bing Ren (Mar. 23, 2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome". In: *Cell* 128.6, pp. 1231–1245.

Kim, Tae-Min, Seung-Hyun Jung, Chang Hyeok An, Sung Hak Lee, In-Pyo Baek, Min Sung Kim, Sung-Won Park, Je-Keun Rhee, Sug-Hyung Lee, and Yeun-Jun Chung (Oct. 1, 2015). "Subclonal Genomic Architectures of Primary and Metastatic Colorectal Cancer Based on Intratumoral Genetic Heterogeneity". In: *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 21.19, pp. 4461–4472.

Kimura, M. (Mar. 15, 1955). "Solution of a process of random genetic drift with a continuous model". In: *Proceedings of the National Academy of Sciences of the United States of America* 41.3, pp. 144–150.

— (June 1968a). "Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles". In: *Genetical Research* 11.3, pp. 247–269.

— (Feb. 17, 1968b). "Evolutionary rate at the molecular level". In: *Nature* 217.5129, pp. 624–626.

— (1989). "The neutral theory of molecular evolution and the world view of the neutralists". In: *Genome* 31.1, pp. 24–31.

— (Aug. 1991). "The neutral theory of molecular evolution: a review of recent evidence". In: *Idengaku Zasshi* 66.4, pp. 367–386.

Kimura, M. and J. F. Crow (Apr. 1964). "The number of alleles that can be maintained in a finite population". In: *Genetics* 49, pp. 725–738.

Kimura, M. and T. Ohta (Mar. 1969). "The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population". In: *Genetics* 61.3, pp. 763–771.

Kimura, Motoo (Apr. 1969). "The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations". In: *Genetics* 61.4, pp. 893–903.

— (1983). *The Neutral Theory of Molecular Evolution*. Cambridge.

King, C. R., M. H. Kraus, and S. A. Aaronson (Sept. 6, 1985). "Amplification of a novel v-erbB-related gene in a human mammary carcinoma". In: *Science (New York, N.Y.)* 229.4717, pp. 974–976.

King, Jack Lester and Thomas H Jukes (1969). "Non-Darwinian evolution." In: *University of California, Berkeley* 94720, p. 788.

Kinzler, Kenneth W and Bert Vogelstein (1996). "Lessons from hereditary colorectal cancer". In: *Cell* 87.2, pp. 159–170.

Knudson, Alfred G. (Apr. 1971). "Mutation and Cancer: Statistical Study of Retinoblastoma". In: *Proceedings of the National Academy of Sciences of the United States of America* 68.4, pp. 820–823.

Komarova, Natalia (Apr. 7, 2006). “Stochastic modeling of drug resistance in cancer”. In: *Journal of Theoretical Biology* 239.3, pp. 351–366.

Konishi, Hiroyuki, Bedri Karakas, Abde M. Abukhdeir, Josh Lauring, John P. Gustin, Joseph P. Garay, Yuko Konishi, Eike Gallmeier, Kurtis E. Bachman, and Ben Ho Park (Sept. 15, 2007). “Knock-in of mutant K-ras in nontumorigenic human epithelial cells as a new model for studying K-ras mediated transformation”. In: *Cancer Research* 67.18, pp. 8460–8467.

Koo, Bon-Kyoung, Maureen Spit, Ingrid Jordens, Teck Y. Low, Daniel E. Stange, Marc van de Wetering, Johan H. van Es, Shabaz Mohammed, Albert J. R. Heck, Madelon M. Maurice, and Hans Clevers (Aug. 2012). “Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors”. In: *Nature* 488.7413, pp. 665–669.

Korhonen, Janne, Petri Martinmäki, Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen (Dec. 1, 2009). “MOODS: fast search for position weight matrix matches in DNA sequences”. In: *Bioinformatics* 25.23, pp. 3181–3182.

Kosakovsky Pond, Sergei L. and Simon D. W. Frost (May 1, 2005). “Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection”. In: *Molecular Biology and Evolution* 22.5, pp. 1208–1222.

Köster, Johannes and Sven Rahmann (Oct. 1, 2012). “Snakemake—a scalable bioinformatics workflow engine”. In: *Bioinformatics* 28.19, pp. 2520–2522.

Kouzarides, Tony (Feb. 23, 2007). “Chromatin modifications and their function”. In: *Cell* 128.4, pp. 693–705.

Kreitman, Martin (1996). “The neutral theory is dead. Long live the neutral theory”. In: *BioEssays* 18.8, pp. 678–683.

Kress, M., E. May, R. Cassingena, and P. May (Aug. 1, 1979). “Simian virus 40-transformed cells express new species of proteins precipitable by anti-simian virus 40 tumor serum.” In: *Journal of Virology* 31.2, pp. 472–483.

Lakatos, Eszter, Marc J. Williams, Ryan O. Schenck, William C. H. Cross, Jacob Househam, Luis Zapata, Benjamin Werner, Chandler Gatenbee, Mark Robertson-Tessi, Chris P. Barnes, Alexander R. A. Anderson, Andrea Sottoriva, and Trevor A. Graham (Oct. 2020). “Evolutionary dynamics of neoantigens in growing tumors”. In: *Nature Genetics* 52.10, pp. 1057–1066.

Lal, Avantika, Keli Liu, Robert Tibshirani, Arend Sidow, and Daniele Ramazzotti (June 28, 2021). “De novo mutational signature discovery in tumor genomes using SparseSignatures”. In: *PLOS Computational Biology* 17.6, e1009119.

Lamprecht, Sebastian, Eva Marina Schmidt, Cristina Blaj, Heiko Hermeking, Andreas Jung, Thomas Kirchner, and David Horst (Nov. 10, 2017). “Multicolor lineage tracing reveals clonal architecture and dynamics in colon cancer”. In: *Nature Communications* 8.1, p. 1406.

Lander, Eric S. et al. (Feb. 1, 2001). “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822, pp. 860–921.

Lane, D. P. and L. V. Crawford (Mar. 1979). “T antigen is bound to a host protein in SY40-transformed cells”. In: *Nature* 278.5701, pp. 261–263.

Lang, Fengchao, Xin Li, Wenhai Zheng, Zhuoran Li, Danfeng Lu, Guijun Chen, Daohua Gong, Liping Yang, Jinlin Fu, Peng Shi, and Jumin Zhou (Oct. 10, 2017). “CTCF prevents genomic instability by promoting homologous recombination-directed DNA double-strand break repair”. In: *Proceedings of the National Academy of Sciences of the United States of America* 114.41, pp. 10912–10917.

Lang, Gene A., Tomoo Iwakuma, Young-Ah Suh, Geng Liu, V. Ashutosh Rao, John M. Parant, Yasmine A. Valentin-Vega, Tamara Terzian, Lisa C. Caldwell, Louise C. Strong, Adel K. El-Naggar, and Guillermina Lozano (Dec. 17, 2004). "Gain of function of a p53 hot spot mutation in a mouse model of Li-Fraumeni syndrome". In: *Cell* 119.6, pp. 861–872.

Langley, C. H. and W. M. Fitch (1974). "An examination of the constancy of the rate of molecular evolution". In: *Journal of Molecular Evolution* 3.3, pp. 161–177.

Langmead, Ben and Steven L. Salzberg (Apr. 2012). "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4, pp. 357–359.

Langmead, Ben, Christopher Wilks, Valentin Antonescu, and Rone Charles (Feb. 1, 2019). "Scaling read aligners to hundreds of threads on general-purpose processors". In: *Bioinformatics* 35.3, pp. 421–432.

Lannagan, Tamsin R. M., Young K. Lee, Tongtong Wang, Jatin Roper, Mark L. Bettington, Lochlan Fennell, Laura Vrbanac, Lisa Jonavicius, Roshini Somashekar, Krystyna Gieniec, Miao Yang, Jia Q. Ng, Nobumi Suzuki, Mari Ichinose, Josephine A. Wright, Hiroki Kobayashi, Tracey L. Putoczki, Yoku Hayakawa, Simon J. Leedham, Helen E. Abud, Ömer H. Yilmaz, Julie Marker, Sonja Klebe, Pratyaksha Wirapati, Siddhartha Mukherjee, Sabine Tejpar, Barbara A. Leggett, Vicki L. J. Whitehall, Daniel L. Worthley, and Susan L. Woods (Apr. 1, 2019). "Genetic editing of colonic organoids provides a molecularly distinct and orthotopic preclinical model of serrated carcinogenesis". In: *Gut* 68.4, pp. 684–692.

Lao, Victoria Valinluck and William M. Grady (Oct. 18, 2011). "Epigenetics and colorectal cancer". In: *Nature Reviews. Gastroenterology & Hepatology* 8.12, pp. 686–700.

Larsson, Chatarina, Jørn Koch, Anders Nygren, George Janssen, Anton K. Raap, Ulf Landegren, and Mats Nilsson (Dec. 2004). "In situ genotyping individual DNA molecules by target-primed rolling-circle amplification of padlock probes". In: *Nature Methods* 1.3, pp. 227–232.

Lawrence, Michael, Robert Gentleman, and Vincent Carey (2009). "rtracklayer: an R package for interfacing with genome browsers". In: *Bioinformatics* 25, pp. 1841–1842.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey (2013a). "Software for Computing and Annotating Genomic Ranges". In: *PLoS Computational Biology* 9.8.

Lawrence, Michael S., Petar Stojanov, Craig H. Mermel, Levi A. Garraway, Todd R. Golub, Matthew Meyerson, Stacey B. Gabriel, Eric S. Lander, and Gad Getz (Jan. 23, 2014). "Discovery and saturation analysis of cancer genes across 21 tumor types". In: *Nature* 505.7484, pp. 495–501.

Lawrence, Michael S. et al. (July 2013b). "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499.7457, pp. 214–218.

Lawson, Andrew R. J., Federico Abascal, Tim H. H. Coorens, Yvette Hooks, Laura O'Neill, Calli Latimer, Keiran Raine, Mathijs A. Sanders, Anne Y. Warren, Krishnaa T. A. Mahbubani, Bethany Bareham, Timothy M. Butler, Luke M. R. Harvey, Alex Cagan, Andrew Menzies, Luiza Moore, Alexandra J. Colquhoun, William Turner, Benjamin Thomas, Vincent Gnanapragasam, Nicholas Williams, Doris M. Rassl, Harald Vöhringer, Sonia Zumalave, Jyoti Nangalia, José M. C. Tubío, Moritz Gerstung, Kourosh Saeb-Parsy, Michael R. Stratton, Peter J. Campbell, Thomas J. Mitchell, and Iñigo Martincorena (Oct. 2, 2020). "Extensive heterogeneity in somatic mutation and selection in the human bladder". In: *Science* 370.6512, pp. 75–82.

Lee, Daniel D. and H. Seung (1999). “Learning the parts of objects by non-negative matrix factorization”. In: *Nature*.

Lee, Ming-Hsiang, Benjamin Siddoway, Gwendolyn E. Kaeser, Igor Segota, Richard Rivera, William Romanow, Grace Kennedy, Tao Long, and Jerold Chun (Nov. 2018). “Somatic APP gene recombination and mutations occur mosaically in normal and Alzheimer’s disease neurons”. In: *Nature* 563.7733, pp. 639–645.

Lee-Six, Henry, Sigurgeir Olafsson, Peter Ellis, Robert J. Osborne, Mathijs A. Sanders, Luiza Moore, Nikitas Georgakopoulos, Franco Torrente, Ayesha Noorani, Martin Goddard, Philip Robinson, Tim H. H. Coorens, Laura O’Neill, Christopher Alder, Jingwei Wang, Rebecca C. Fitzgerald, Matthias Zilbauer, Nicholas Coleman, Kourosh Saeb-Parsy, Inigo Martincorena, Peter J. Campbell, and Michael R. Stratton (Oct. 2019). “The landscape of somatic mutation in normal colorectal epithelial cells”. In: *Nature* 574.7779, pp. 532–537.

Legendre, Pierre, François-Joseph Lapointe, and Philippe Casgrain (1994). “Modeling Brain Evolution from Behavior: A Permutational Regression Approach”. In: *Evolution* 48.5, pp. 1487–1499.

Leigh, E. G. (2007). “Neutral theory: a historical perspective”. In: *Journal of Evolutionary Biology* 20.6, pp. 2075–2091.

Leroi, Armand M., Ben Lambert, James Rosindell, Xiangyu Zhang, and Giorgos D. Kokkoris (Aug. 2020). “Neutral syndrome”. In: *Nature Human Behaviour* 4.8, pp. 780–790.

Letouzé, Eric, Yves Allory, Marc A. Bollet, François Radvanyi, and Frédéric Guyon (2010). “Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis”. In: *Genome Biology* 11.7, R76.

Leung, Marco L., Alexander Davis, Ruli Gao, Anna Casasent, Yong Wang, Emi Sei, Eduardo Vilar, Dipen Maru, Scott Kopetz, and Nicholas E. Navin (Aug. 2017). “Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer”. In: *Genome Research* 27.8, pp. 1287–1299.

Levine, Arnold J. and Moshe Oren (Oct. 2009). “The first 30 years of p53: growing ever more complex”. In: *Nature Reviews. Cancer* 9.10, pp. 749–758.

Li, Heng (May 26, 2013). “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”. In: *arXiv:1303.3997 [q-bio]*. arXiv: `1303.3997`.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup (Aug. 15, 2009). “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25.16, pp. 2078–2079.

Li, Rui, Jingsi Dong, Hongwan Zhang, Qi Zhao, Xingyang Li, Xuefei Liu, Ying Ye, Shuang Deng, Dongxin Lin, Jian Zheng, and Zhixiang Zuo (Sept. 1, 2020). “Clinical and genomic characterization of neutral tumor evolution in Head and Neck Squamous Cell Carcinoma”. In: *Genomics* 112.5, pp. 3448–3454.

Li, W H, C I Wu, and C C Luo (Mar. 1, 1985). “A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes.” In: *Molecular Biology and Evolution* 2.2, pp. 150–174.

Liao, Xiaoyun, Paul Lochhead, Reiko Nishihara, Teppei Morikawa, Aya Kuchiba, Mai Yamauchi, Yu Imamura, Zhi Rong Qian, Yoshifumi Baba, Kaori Shima, Ruifang Sun, Katsuhiko Nosho, Jeffrey A. Meyerhardt, Edward Giovannucci, Charles S. Fuchs, Andrew T. Chan, and Shuji Ogino (Oct. 25, 2012). “Aspirin Use, Tumor PIK3CA Mutation, and Colorectal-Cancer Survival". In: *New England Journal of Medicine* 367.17, pp. 1596–1606.

Lin, Quan, Raymond Lai, Lucian R. Chirieac, Changping Li, Vilmos A. Thomazy, Ioannis Grammatikakis, George Z. Rassidakis, Wei Zhang, Yasushi Fujio, Keita Kunisada, Stanley R. Hamilton, and Hesham M. Amin (Oct. 1, 2005). "Constitutive Activation of JAK3/STAT3 in Colon Carcinoma Tumors and Cell Lines: Inhibition of JAK3/STAT3 Signaling Induces Apoptosis and Cell Cycle Arrest of Colon Carcinoma Cells". In: *The American Journal of Pathology* 167.4, pp. 969–980.

Lin, Zhenlv, Lin Zhang, Junfeng Zhou, and Jiantao Zheng (Oct. 2019). "Silencing Smad4 attenuates sensitivity of colorectal cancer cells to cetuximab by promoting epithelial-mesenchymal transition". In: *Molecular Medicine Reports* 20.4, pp. 3735–3745.

Ling, Shaoping, Zheng Hu, Zuyu Yang, Fang Yang, Yawei Li, Pei Lin, Ke Chen, Lili Dong, Lihua Cao, Yong Tao, Lingtong Hao, Qingjian Chen, Qiang Gong, Dafei Wu, Wenjie Li, Wenming Zhao, Xiuyun Tian, Chunyi Hao, Eric A. Hungate, Daniel V. T. Catenacci, Richard R. Hudson, Wen-Hsiung Li, Xuemei Lu, and Chung-I. Wu (Nov. 24, 2015). "Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution". In: *Proceedings of the National Academy of Sciences* 112.47, E6496–E6505.

Linzer, Daniel I. H. and Arnold J. Levine (May 1, 1979). "Characterization of a 54K Dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells". In: *Cell* 17.1, pp. 43–52.

Liu, Jun S. (Jan. 4, 2008). *Monte Carlo Strategies in Scientific Computing*. 368 pp.

Liu, Yang, Nilay S. Sethi, Toshinori Hinoue, Barbara G. Schneider, Andrew D. Cherniack, Francisco Sanchez-Vega, Jose A. Seoane, Farshad Farshidfar, Reanne Bowlby, Mirazul Islam, Jaegil Kim, Walid Chatila, Rehan Akbani, Rupa S. Kanchi, Charles S. Rabkin, Joseph E. Willis, Kenneth K. Wang, Shannon J. McCall, Lopa Mishra, Akinyemi I. Ojesina, Susan Bullman, Chandra Sekhar Pedamallu, Alexander J. Lazar, Ryo Sakai, Cancer Genome Atlas Research Network, Vésteinn Thorsson, Adam J. Bass, and Peter W. Laird (Apr. 9, 2018). "Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas". In: *Cancer Cell* 33.4, 721–735.e8.

Lohr, Jens G, Viktor A Adalsteinsson, Kristian Cibulskis, Atish D Choudhury, Mara Rosenberg, Peter Cruz-Gordillo, Joshua M Francis, Cheng-Zhong Zhang, Alex K Shalek, Rahul Satija, John J Trombetta, Diana Lu, Naren Tallapragada, Narmin Tahirova, Sora Kim, Brendan Blumenstiel, Carrie Sougnez, Alarice Lowe, Bang Wong, Daniel Auclair, Eliezer M Van Allen, Mari Nakabayashi, Rosina T Lis, Gwo-Shu M Lee, Tiantian Li, Matthew S Chabot, Amy Ly, Mary-Ellen Taplin, Thomas E Clancy, Massimo Loda, Aviv Regev, Matthew Meyerson, William C Hahn, Philip W Kantoff, Todd R Golub, Gad Getz, Jesse S Boehm, and J Christopher Love (May 2014). "Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer". In: *Nature Biotechnology* 32.5, pp. 479–484.

Loo, Peter Van, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles M. Perou, Anne-Lise Børresen-Dale, and Vessela N. Kristensen (Sept. 28, 2010). "Allele-specific copy number analysis of tumors". In: *Proceedings of the National Academy of Sciences* 107.39, pp. 16910–16915.

Lopez-Garcia, Carlos, Allon M. Klein, Benjamin D. Simons, and Douglas J. Winton (Nov. 5, 2010). "Intestinal stem cell replacement follows a pattern of neutral drift". In: *Science (New York, N.Y.)* 330.6005, pp. 822–825.

Lote, H., I. Spiteri, L. Ermini, A. Vatsiou, A. Roy, A. McDonald, N. Maka, M. Balsitis, N. Bose, M. Simbolo, A. Mafficini, A. Lampis, J. C. Hahne, F. Trevisani, Z. Eltahir, G. Mentrasti, C. Findlay, E. a. J. Kalkman, M. Punta, B. Werner, S. Lise, A. Aktipis, C. Maley, M. Greaves, C. Braconi, J. White, M. Fassan, A. Scarpa, A. Sottoriva, and N. Valeri (June 1, 2017). "Carbon dating cancer: defining the chronology of metastatic progression in colorectal cancer". In: *Annals of Oncology* 28.6, pp. 1243–1249.

Love, Michael I., Wolfgang Huber, and Simon Anders (Dec. 5, 2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12, p. 550.

Luo, Jin, Yan-Ni Li, Fei Wang, Wei-Ming Zhang, and Xin Geng (Dec. 6, 2010). "S-adenosylmethionine inhibits the growth of cancer cells by reversing the hypomethylation status of c-myc and H-ras in human gastric cancer and colon cancer". In: *International Journal of Biological Sciences* 6.7, pp. 784–795.

Luo, Qianxin, Dianke Chen, Xinjuan Fan, Xinhui Fu, Tenghui Ma, and Daici Chen (Dec. 1, 2020). "KRAS and PIK3CA bi-mutations predict a poor prognosis in colorectal cancer patients: A single-site report". In: *Translational Oncology* 13.12, p. 100874.

Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M. Opitz, Renata Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A. Haas, Marco Osterwalder, Martin Franke, Bernd Timmermann, Jochen Hecht, Malte Spielmann, Axel Visel, and Stefan Mundlos (May 21, 2015). "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions". In: *Cell* 161.5, pp. 1012–1025.

Luria, S. E. and M. Delbrück (Nov. 20, 1943). "Mutations of Bacteria from Virus Sensitivity to Virus Resistance". In: *Genetics* 28.6, pp. 491–511.

Macdonald, Stuart J. and Anthony D. Long (May 3, 2005). "Prospects for identifying functional variation across the genome". In: *Proceedings of the National Academy of Sciences of the United States of America* 102 (Suppl 1), pp. 6614–6621.

Manly, Bryan F. J. (Dec. 1, 1986). "Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations". In: *Researches on Population Ecology* 28.2, pp. 201–218.

Mantel, N. (Feb. 1967). "The detection of disease clustering and a generalized regression approach". In: *Cancer Research* 27.2, pp. 209–220.

Mardis, Elaine R. (2008). "Next-Generation DNA Sequencing Methods". In: *Annual Review of Genomics and Human Genetics* 9.1, pp. 387–402.

Marie, P (1910). "Contribution a L'etude du developement des tumeurus malignes sur le ulcers de Roentgen". In: *Bull Assoc Francl'eude du Cancer* 3, pp. 404–426.

Marjoram, Paul, John Molitor, Vincent Plagnol, and Simon Tavaré (Dec. 23, 2003). "Markov chain Monte Carlo without likelihoods". In: *Proceedings of the National Academy of Sciences* 100.26, pp. 15324–15328.

Market, Eleonora and F. Nina Papavasiliou (Oct. 13, 2003). "V(D)J Recombination and the Evolution of the Adaptive Immune System". In: *PLOS Biology* 1.1, e16.

Marshall, A. D., C. G. Bailey, K. Champ, M. Vellozzi, P. O'Young, C. Metierre, Y. Feng, A. Thoeng, A. M. Richards, U. Schmitz, M. Biro, R. Jayasinghe, L. Ding, L. Anderson, E. R. Mardis, and J. E. J. Rasko (July 20, 2017). "CTCF genetic alterations in endometrial carcinoma are pro-tumorigenic". In: *Oncogene* 36.29, pp. 4100–4110.

Marshall, C. J., A. Hall, and R. A. Weiss (Sept. 1, 1982). “A transforming gene present in human sarcoma cell lines”. In: *Nature* 299, pp. 171–173.

Martincorena, Inigo (2019). *dndscv: Poisson-based dN/dS models to quantify natural selection in somatic evolution*.

Martincorena, Iñigo and Peter J. Campbell (Sept. 25, 2015). “Somatic mutation in cancer and normal cells”. In: *Science* 349.6255, pp. 1483–1489.

Martincorena, Iñigo, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R. Stratton, and Peter J. Campbell (Nov. 16, 2017). “Universal Patterns of Selection in Cancer and Somatic Tissues”. In: *Cell* 171.5, 1029–1041.e21.

Martinez, Pierre, Diego Mallo, Thomas G. Paulson, Xiaohong Li, Carissa A. Sanchez, Brian J. Reid, Trevor A. Graham, Mary K. Kuhner, and Carlo C. Maley (Feb. 23, 2018). “Evolution of Barrett’s esophagus through space and time at single-crypt and whole-biopsy levels”. In: *Nature Communications* 9.1, p. 794.

Martínez-Jiménez, Francisco, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, Oriol Pich, Jose Bonet, Hanna Kranas, Abel Gonzalez-Perez, and Nuria Lopez-Bigas (Oct. 2020). “A compendium of mutational cancer driver genes”. In: *Nature Reviews Cancer* 20.10, pp. 555–572.

Marušic, Miljenko, Željko Bajzer, Stanimir Vuk-Pavlovic, and James P. Freyer (July 1, 1994). “Tumor growthin vivo and as multicellular spheroids compared by mathematical models”. In: *Bulletin of Mathematical Biology* 56.4, pp. 617–631.

Marušić, M., Ž Bajzer, J. P. Freyer, and S. Vuk-Pavlović (1994). “Analysis of growth of multicellular tumour spheroids by mathematical models”. In: *Cell Proliferation* 27.2, pp. 73–94.

Marutani, Masumi, Hidefumi Tonoki, Mitsuhiro Tada, Masato Takahashi, Haruhiko Kashiwazaki, Yasuhiro Hida, Jun-ich Hamada, Masahiro Asaka, and Tetsuya Moriuchi (Oct. 1, 1999). “Dominant-Negative Mutations of the Tumor Suppressor p53 Relating to Early Onset of Glioblastoma Multiforme”. In: *Cancer Research* 59.19, pp. 4765–4769.

Mason, Penelope A., Elizabeth C. Matheson, Andrew G. Hall, and Robert N. Lightowlers (Feb. 1, 2003). “Mismatch repair activity in mammalian mitochondria”. In: *Nucleic Acids Research* 31.3, pp. 1052–1058.

Masoodi, Tariq Ahmad, Noor Ahmad Shaik, Syed Burhan, Qurratulain Hasan, Gowhar Shafi, and Venkateswara Rao Talluri (June 1, 2019). “Structural prediction, whole exome sequencing and molecular dynamics simulation confirms p.G118D somatic mutation of PIK3CA as functionally important in breast cancer patients”. In: *Computational Biology and Chemistry* 80, pp. 472–479.

Matlashewski, G, P Lamb, D Pim, J Peacock, L Crawford, and S Benchimol (Dec. 20, 1984). “Isolation and characterization of a human p53 cDNA clone: expression of the human p53 gene.” In: *The EMBO Journal* 3.13, pp. 3257–3262.

May, Pierre and Evelyne May (Dec. 1999). “Twenty years of p53 research: structural and functional aspects of the p53 protein”. In: *Oncogene* 18.53, pp. 7621–7636.

Mayneord, W. V. (July 1, 1932). “On a Law of Growth of Jensen’s Rat Sarcoma”. In: *The American Journal of Cancer* 16.4, pp. 841–846.

McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth (May 2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation”. In: *Nucleic Acids Research* 40.10, pp. 4288–4297.

McDonald, Oliver G., Xin Li, Tyler Saunders, Rakel Tryggvadottir, Samantha J. Mentch, Marc O. Warmoes, Anna E. Word, Alessandro Carrer, Tal H. Salz, Sonoko Natsume, Kimberly M. Stauffer, Alvin Makohon-Moore, Yi Zhong, Hao Wu, Kathryn E. Wellen, Jason W. Locasale, Christine A. Iacobuzio-Donahue, and Andrew P. Feinberg (Mar. 2017). “Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis”. In: *Nature Genetics* 49.3, pp. 367–376.

McDonald, Thomas O., Shaon Chakrabarti, and Franziska Michor (Dec. 2018). “Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution”. In: *Nature Genetics* 50.12, pp. 1620–1623.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo (Sept. 2010). “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data”. In: *Genome Research* 20.9, pp. 1297–1303.

McKinley, Trevelyan, Alex R. Cook, and Robert Deardon (July 20, 2009). “Inference in Epidemic Models without Likelihoods”. In: *The International Journal of Biostatistics* 5.1.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham (June 6, 2016). “The Ensembl Variant Effect Predictor”. In: *Genome Biology* 17.1, p. 122.

Mei, Jiandong, Zhilan Xiao, Chenglin Guo, Qiang Pu, Lin Ma, Chengwu Liu, Feng Lin, Hu Liao, Zongbing You, and Lunxu Liu (June 7, 2016). “Prognostic impact of tumor-associated macrophage infiltration in non-small cell lung cancer: A systemic review and meta-analysis”. In: *Oncotarget* 7.23, pp. 34217–34228.

Meng, Shu, Qilin Gu, Xiaojie Yang, Jie Lv, Iris Owusu, Gianfranco Matrone, Kaifu Chen, John P. Cooke, and Longhou Fang (Aug. 28, 2018). “TBX20 Regulates Angiogenesis Through the Prokineticin 2–Prokineticin Receptor 1 Pathway”. In: *Circulation* 138.9, pp. 913–928.

Merlos-Suárez, Anna, Francisco M. Barriga, Peter Jung, Mar Iglesias, María Virtudes Céspedes, David Rossell, Marta Sevillano, Xavier Hernando-Momblona, Victoria da Silva-Diz, Purificación Muñoz, Hans Clevers, Elena Sancho, Ramón Mangues, and Eduard Batlle (May 6, 2011). “The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse”. In: *Cell Stem Cell* 8.5, pp. 511–524.

Meyer, Mona, Jüri Reimand, Xiaoyang Lan, Renee Head, Xueming Zhu, Michelle Kushida, Jane Bayani, Jessica C. Pressey, Anath C. Lionel, Ian D. Clarke, Michael Cusimano, Jeremy A. Squire, Stephen W. Scherer, Mark Bernstein, Melanie A. Woodin, Gary D. Bader, and Peter B. Dirks (Jan. 20, 2015). “Single cell-derived clonal analysis of human glioblastoma links functional and genomic heterogeneity”. In: *Proceedings of the National Academy of Sciences of the United States of America* 112.3, pp. 851–856.

Michor, Franziska, Martin A. Nowak, and Yoh Iwasa (June 21, 2006). “Stochastic dynamics of metastasis formation”. In: *Journal of Theoretical Biology* 240.4, pp. 521–530.

Millane, R. Cathriona, Justyna Kanska, David J. Duffy, Cathal Seoighe, Stephen Cunningham, Günter Plickert, and Uri Frank (June 15, 2011). “Induced stem cell neoplasia in a cnidarian by ectopic expression of a POU domain transcription factor”. In: *Development* 138.12, pp. 2429–2439.

Miller, Christopher A., Brian S. White, Nathan D. Dees, Malachi Griffith, John S. Welch, Obi L. Griffith, Ravi Vij, Michael H. Tomasson, Timothy A. Graubert, Matthew J. Walter, Matthew J. Ellis, William Schierding, John F. DiPersio, Timothy J. Ley, Elaine

R. Mardis, Richard K. Wilson, and Li Ding (Aug. 7, 2014). “SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution”. In: *PLOS Computational Biology* 10.8, e1003665.

Mills, Ryan E., Christopher T. Luttig, Christine E. Larkins, Adam Beauchamp, Circe Tsui, W. Stephen Pittard, and Scott E. Devine (Sept. 2006). “An initial map of insertion and deletion (INDEL) variation in the human genome”. In: *Genome Research* 16.9, pp. 1182–1190.

Mir, Arnau, Francesc Rosselló, and Lucı A. Rotger (Jan. 2013). “A new balance index for phylogenetic trees”. In: *Mathematical Biosciences* 241.1, pp. 125–136.

Mir, Arnau, Lucía Rotger, and Francesc Rosselló (Sept. 25, 2018). “Sound Colless-like balance indices for multifurcating trees”. In: *PLoS ONE* 13.9.

Mishra, Manish and Renu A. Kowluru (Oct. 1, 2014). “Retinal Mitochondrial DNA Mismatch Repair in the Development of Diabetic Retinopathy, and Its Continued Progression After Termination of Hyperglycemia”. In: *Investigative Ophthalmology & Visual Science* 55.10, pp. 6960–6967.

Miyata, Takashi and Teruo Yasunaga (Mar. 1, 1980). “Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application”. In: *Journal of Molecular Evolution* 16.1, pp. 23–36.

Mizuno, Takashi, Jordan M. Cloyd, Diego Vicente, Kiyohiko Omichi, Yun Shin Chun, Scott E. Kopetz, Dipen Maru, Claudius Conrad, Ching-Wei D. Tzeng, Steven H. Wei, Thomas A. Aloia, and Jean-Nicolas Vauthey (May 1, 2018). “SMAD4 gene mutation predicts poor prognosis in patients undergoing resection for colorectal liver metastases”. In: *European Journal of Surgical Oncology* 44.5, pp. 684–692.

Molinari, Francesca and Milo Frattini (2014). “Functions and Regulation of the PTEN Gene in Colorectal Cancer”. In: *Frontiers in Oncology* 3.

Mooers, Arne O. and Stephen B. Heard (1997). “Inferring Evolutionary Process from Phylogenetic Tree Shape”. In: *The Quarterly Review of Biology* 72.1, pp. 31–54.

Moolgavkar, S. H. (1986). “Carcinogenesis modeling: from molecular biology to epidemiology”. In: *Annual Review of Public Health* 7, pp. 151–169.

Mularoni, Loris, Radhakrishnan Sabarinathan, Jordi Deu-Pons, Abel Gonzalez-Perez, and Núria López-Bigas (June 16, 2016). “OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations”. In: *Genome Biology* 17.1, p. 128.

Murakami, Yohei (Aug. 4, 2014). “Bayesian Parameter Inference and Model Selection by Population Annealing in Systems Biology”. In: *PLoS ONE* 9.8.

Murray, Iain and Zoubin Ghahramani (2005). *A note on the evidence and Bayesian Occam’s razor.*

Muse, S. V. and B. S. Gaut (Sept. 1994). “A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome”. In: *Molecular Biology and Evolution* 11.5, pp. 715–724.

Muse, Spencer V (1996). “Estimating synonymous and nonsynonymous substitution rates.” In: *Molecular biology and evolution* 13.1, pp. 105–114.

Muzny, Donna M. et al. (July 2012). “Comprehensive molecular characterization of human colon and rectal cancer”. In: *Nature* 487.7407, pp. 330–337.

Nagtegaal, Iris D., Robert D. Odze, David Klimstra, Valerie Paradis, Massimo Rugge, Peter Schirmacher, Kay M. Washington, Fatima Carneiro, Ian A. Cree, and WHO Classification of Tumours Editorial Board (Jan. 2020). “The 2019 WHO classification of tumours of the digestive system”. In: *Histopathology* 76.2, pp. 182–188.

Narayanan, Deevya L., Rao N. Saladi, and Joshua L. Fox (2010). “Review: Ultraviolet radiation and skin cancer”. In: *International Journal of Dermatology* 49.9, pp. 978–986.

Navin, Nicholas, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W. Richard McCombie, James Hicks, and Michael Wigler (Apr. 7, 2011). “Tumour evolution inferred by single-cell sequencing”. In: *Nature* 472.7341, pp. 90–94.

Nebbioso, Angela, Francesco Paolo Tambaro, Carmela Dell’Aversana, and Lucia Altucci (June 7, 2018). “Cancer epigenetics: Moving forward”. In: *PLOS Genetics* 14.6, e1007362.

Nei, M and T Gojobori (Sept. 1, 1986). “Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.” In: *Molecular Biology and Evolution* 3.5, pp. 418–426.

Neuhauser, Claudia and Stephen M Krone (Feb. 1, 1997). “The Genealogy of Samples in Models With Selection”. In: *Genetics* 145.2, pp. 519–534.

Ng, Jennifer Mun-Kar and Jun Yu (Jan. 22, 2015). “Promoter hypermethylation of tumour suppressor genes as potential biomarkers in colorectal cancer”. In: *International Journal of Molecular Sciences* 16.2, pp. 2472–2496.

Nguyen, Huy, Cristy Loustaunau, Alexander Facista, Lois Ramsey, Nadia Hassounah, Hilary Taylor, Robert Krouse, Claire M. Payne, V. Liana Tsikitis, Steve Goldschmid, Bhaskar Banerjee, Rafael F. Perini, and Carol Bernstein (July 28, 2010). “Deficient Pms2, ERCC1, Ku86, CcOI in Field Defects During Progression to Colon Cancer”. In: *Journal of Visualized Experiments : JoVE* 41.

Nichol, Daniel, Peter Jeavons, Alexander G. Fletcher, Robert A. Bonomo, Philip K. Maini, Jerome L. Paul, Robert A. Gatenby, Alexander R. A. Anderson, and Jacob G. Scott (Sept. 11, 2015). “Steering Evolution with Sequential Therapy to Prevent the Emergence of Bacterial Antibiotic Resistance”. In: *PLOS Computational Biology* 11.9, e1004493.

Nicholson, Michael D. and Tibor Antal (Nov. 2016). “Universal Asymptotic Clone Size Distribution for General Population Growth”. In: *Bulletin of Mathematical Biology* 78.11, pp. 2243–2276.

Nicolas, Pierre, Kyoung-Mee Kim, Darryl Shibata, and Simon Tavaré (Mar. 2, 2007). “The Stem Cell Population of the Human Colon Crypt: Analysis via Methylation Patterns”. In: *PLOS Computational Biology* 3.3, e28.

Nielsen, Rasmus (n.d.). “Statistical tests of selective neutrality in the age of genomics”. In: (), p. 7.

Nik-Zainal, Serena, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, Adam Shlien, Susanna L. Cooke, Jonathan Hinton, Andrew Menzies, Lucy A. Stebbings, Catherine Leroy, Mingming Jia, Richard Rance, Laura J. Mudie, Stephen J. Gamble, Philip J. Stephens, Stuart McLaren, Patrick S. Tarpey, Elli Papaemmanuil, Helen R. Davies, Ignacio Varela, David J. McBride, Graham R. Bignell, Kenric Leung, Adam P. Butler, Jon W. Teague, Sancha Martin, Goran Jönsson, Odette Mariani, Sandrine Boyault, Penelope Miron, Aquila Fatima, Anita Langerød, Samuel A. J. R. Aparicio, Andrew Tutt, Anieta M. Sieuwerts, Åke Borg, Gilles Thomas, Anne Vincent Salomon, Andrea L. Richardson, Anne-Lise Børresen-Dale, P. Andrew Futreal, Michael R. Stratton, and Peter J. Campbell (May 25, 2012a). "The Life History of 21 Breast Cancers". In: *Cell* 149.5, pp. 994–1007.

Nik-Zainal, Serena et al. (May 25, 2012b). "Mutational Processes Molding the Genomes of 21 Breast Cancers". In: *Cell* 149.5, pp. 979–993.

Niknafs, Noushin, Violeta Beleva-Guthrie, Daniel Q. Naiman, and Rachel Karchin (Oct. 5, 2015). "SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing". In: *PLOS Computational Biology* 11.10, e1004416.

Nilsen, Gro, Knut Liestøl, Peter Van Loo, Hans Kristian Moen Vollan, Marianne B. Eide, Oscar M. Rueda, Suet-Feung Chin, Roslin Russell, Lars O. Baumbusch, Carlos Caldas, Anne-Lise Børresen-Dale, and Ole Christian Lingjærde (Nov. 4, 2012). "Copynumber: Efficient algorithms for single- and multi-track copy number segmentation". In: *BMC Genomics* 13.1, p. 591.

Nirenberg, M. W. and J. H. Matthaei (Oct. 15, 1961). "The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides". In: *Proceedings of the National Academy of Sciences of the United States of America* 47, pp. 1588–1602.

Nishisho, I., Y. Nakamura, Y. Miyoshi, Y. Miki, H. Ando, A. Horii, K. Koyama, J. Utsunomiya, S. Baba, and P. Hedge (Aug. 9, 1991). "Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients". In: *Science (New York, N.Y.)* 253.5020, pp. 665–669.

Niu, Beifang, Kai Ye, Qunyuan Zhang, Charles Lu, Mingchao Xie, Michael D. McLellan, Michael C. Wendl, and Li Ding (Apr. 1, 2014). "MSIsensor: microsatellite instability detection using paired tumor-normal sequence data". In: *Bioinformatics (Oxford, England)* 30.7, pp. 1015–1016.

Nixon, Kevin C. (1999). "The Parsimony Ratchet, a New Method for Rapid Parsimony Analysis". In: *Cladistics* 15.4, pp. 407–414.

Noorani, Ayesha, Xiaodun Li, Martin Goddard, Jason Crawte, Ludmil B. Alexandrov, Maria Secrier, Matthew D. Eldridge, Lawrence Bower, Jamie Weaver, Pierre Lao-Sirieix, Inigo Martincorena, Irene Debiram-Beecham, Nicola Grehan, Shona MacRae, Shalini Malhotra, Ahmad Miremadi, Tabitha Thomas, Sarah Galbraith, Lorraine Petersen, Stephen D. Preston, David Gilligan, Andrew Hindmarsh, Richard H. Hardwick, Michael R. Stratton, David C. Wedge, and Rebecca C. Fitzgerald (Jan. 2020). "Genomic evidence supports a clonal diaspora model for metastases of esophageal adenocarcinoma". In: *Nature Genetics* 52.1, pp. 74–83.

Noorbakhsh, Javad and Jeffrey H. Chuang (Sept. 2017). "Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures". In: *Nature Genetics* 49.9, pp. 1288–1289.

Nordling, C. O. (Mar. 1953). "A new theory on cancer-inducing mechanism". In: *British Journal of Cancer* 7.1, pp. 68–72.

Nowell, P. C. (Oct. 1, 1976). "The clonal evolution of tumor cell populations". In: *Science* 194.4260, pp. 23–28.

Nowell, Peter C. and David A. Hungerford (July 1, 1960). "Chromosome Studies on Normal and Leukemic Human Leukocytes". In: *JNCI: Journal of the National Cancer Institute* 25.1, pp. 85–109.

Obenchain, Valerie, Michael Lawrence, Vincent Carey, Stephanie Gogarten, Paul Shannon, and Martin Morgan (2014). “VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants”. In: *Bioinformatics* 30.14, pp. 2076–2078.

Oda, Katsutoshi, Jennifer Okada, Luika Timmerman, Pablo Rodriguez-Viciana, David Stokoe, Keiko Shoji, Yuji Taketani, Hiroyuki Kuramoto, Zachary A. Knight, Kevan M. Shokat, and Frank McCormick (Oct. 1, 2008). “PIK3CA Cooperates with Other Phosphatidylinositol 3’-Kinase Pathway Mutations to Effect Oncogenic Transformation”. In: *Cancer Research* 68.19, pp. 8127–8136.

Oesper, Layla, Gryte Satas, and Benjamin J. Raphael (Dec. 15, 2014). “Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data”. In: *Bioinformatics (Oxford, England)* 30.24, pp. 3532–3540.

Ogino, Shuji, Takako Kawasaki, Katsuhiko Nosho, Mutsuko Ohnishi, Yuko Suemoto, Gregory J. Kirkner, and Charles S. Fuchs (2008). “LINE-1 hypomethylation is inversely associated with microsatellite instability and CpG island methylator phenotype in colorectal cancer”. In: *International Journal of Cancer* 122.12, pp. 2767–2773.

Ohlsson, Rolf, Rainer Renkawitz, and Victor Lobanenkov (Sept. 1, 2001). “CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease”. In: *Trends in Genetics* 17.9, pp. 520–527.

Okugawa, Yoshinaga, William M. Grady, and Ajay Goel (Oct. 1, 2015). “Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers”. In: *Gastroenterology*. Genetics, Genetic Testing, and Biomarkers of Digestive Diseases 149.5, 1204–1225.e12.

Olive, Kenneth P., David A. Tuveson, Zachary C. Ruhe, Bob Yin, Nicholas A. Willis, Roderick T. Bronson, Denise Crowley, and Tyler Jacks (Dec. 17, 2004). “Mutant p53 gain of function in two mouse models of Li-Fraumeni syndrome”. In: *Cell* 119.6, pp. 847–860.

Olshen, Adam B., E. S. Venkatraman, Robert Lucito, and Michael Wigler (Oct. 2004). “Circular binary segmentation for the analysis of array-based DNA copy number data”. In: *Biostatistics (Oxford, England)* 5.4, pp. 557–572.

Ong, Chin-Tong and Victor G. Corces (Apr. 2014). “CTCF: an architectural protein bridging genome topology and function”. In: *Nature Reviews Genetics* 15.4, pp. 234–246.

Opasic, Luka, Da Zhou, Benjamin Werner, David Dingli, and Arne Traulsen (Apr. 29, 2019). “How many samples are needed to infer truly clonal mutations from heterogenous tumours?” In: *BMC cancer* 19.1, p. 403.

Pabinger, Stephan, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke, and Zlatko Trajanoski (Mar. 2014). “A survey of tools for variant analysis of next-generation genome sequencing data”. In: *Briefings in Bioinformatics* 15.2, pp. 256–278.

Pagel, Mark (Oct. 1999). “Inferring the historical patterns of biological evolution”. In: *Nature* 401.6756, pp. 877–884.

Pagès, Hervé (2021). *BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs*.

Pamilo, P and N O Bianchi (Mar. 1, 1993). “Evolution of the Zfx and Zfy genes: rates and interdependence between the genes.” In: *Molecular Biology and Evolution* 10.2, pp. 271–281.

Paradis, Emmanuel and Klaus Schliep (Feb. 1, 2019). “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R”. In: *Bioinformatics* 35.3, pp. 526–528.

Paterson, Chay, Hans Clevers, and Ivana Bozic (Aug. 25, 2020). "Mathematical model of colorectal cancer initiation". In: *Proceedings of the National Academy of Sciences* 117.34, pp. 20681–20688.

Peltomäki, Päivi (Mar. 15, 2003). "Role of DNA Mismatch Repair Defects in the Pathogenesis of Human Cancer". In: *Journal of Clinical Oncology* 21.6, pp. 1174–1179.

Peng, Xudong, Zan Luo, Qingjie Kang, Dawei Deng, Qiang Wang, Hongxia Peng, Shan Wang, and Zhengqiang Wei (2015). "FOXQ1 mediates the crosstalk between TGF-β and Wnt signaling pathways in the progression of colorectal cancer". In: *Cancer Biology & Therapy* 16.7, pp. 1099–1109.

Peters, Antoine H. F. M., Stefan Kubicek, Karl Mechtler, Roderick J. O'Sullivan, Alwin A. H. A. Derijck, Laura Perez-Burgos, Alexander Kohlmaier, Susanne Opravil, Makoto Tachibana, Yoichi Shinkai, Joost H. A. Martens, and Thomas Jenuwein (Dec. 2003). "Partitioning and plasticity of repressive histone methylation states in mammalian chromatin". In: *Molecular Cell* 12.6, pp. 1577–1589.

Petukhov, Viktor, Teun van den Brand, and Evan Biederstedt (2020). *ggrastr: Raster Layers for 'ggplot2'*.

Pfeifer, Gerd P., Mikhail F. Denissenko, Magali Olivier, Natalia Tretyakova, Stephen S. Hecht, and Pierre Hainaut (Oct. 2002). "Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers". In: *Oncogene* 21.48, pp. 7435–7451.

Pfeifer, Gerd P., Young-Hyun You, and Ahmad Besaratinia (Apr. 1, 2005). "Mutations induced by ultraviolet light". In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*. Biological Effects of Ultraviolet Radiation 571.1, pp. 19–31.

Phillips, Jennifer E. and Victor G. Corces (June 26, 2009). "CTCF: master weaver of the genome". In: *Cell* 137.7, pp. 1194–1211.

Phipps, Amanda I., Karen W. Makar, and Polly A. Newcomb (Dec. 2013). "Descriptive profile of PIK3CA-mutated colorectal cancer in postmenopausal women". In: *International Journal of Colorectal Disease* 28.12, pp. 1637–1642.

Picchini, Umberto (Oct. 2, 2014). "Inference for SDE models via Approximate Bayesian Computation". In: *Journal of Computational and Graphical Statistics* 23.4, pp. 1080–1100. arXiv: `1204.5459`.

Pich, Oriol, Ferran Muiños, Radhakrishnan Sabarinathan, Iker Reyes-Salazar, Abel Gonzalez-Perez, and Nuria Lopez-Bigas (Nov. 2018). "Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes". In: *Cell* 175.4, 1074–1087.e18.

Pizzi, C., P. Rastas, and E. Ukkonen (Jan. 2011). "Finding Significant Matches of Position Weight Matrices in Linear Time". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8.1, pp. 69–79.

Platt, Randall J., Sidi Chen, Yang Zhou, Michael J. Yim, Lukasz Swiech, Hannah R. Kempton, James E. Dahlman, Oren Parnas, Thomas M. Eisenhaure, Marko Jovanovic, Daniel B. Graham, Siddharth Jhunjhunwala, Matthias Heidenreich, Ramnik J. Xavier, Robert Langer, Daniel G. Anderson, Nir Hacohen, Aviv Regev, Guoping Feng, Phillip A. Sharp, and Feng Zhang (Oct. 9, 2014). "CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling". In: *Cell* 159.2, pp. 440–455.

Pleasance, Erin D., R. Keira Cheetham, Philip J. Stephens, David J. McBride, Sean J. Humphray, Chris D. Greenman, Ignacio Varela, Meng-Lay Lin, Gonzalo R. Ordóñez, Graham R. Bignell, Kai Ye, Julie Alipaz, Markus J. Bauer, David Beare, Adam Butler, Richard J. Carter, Lina Chen, Anthony J. Cox, Sarah Edkins, Paula I. Kokko-

Gonzales, Niall A. Gormley, Russell J. Grocock, Christian D. Haudenschild, Matthew M. Hims, Terena James, Mingming Jia, Zoya Kingsbury, Catherine Leroy, John Marshall, Andrew Menzies, Laura J. Mudie, Zemin Ning, Tom Royce, Ole B. Schulz-Trieglaff, Anastassia Spiridou, Lucy A. Stebbings, Lukasz Szajkowski, Jon Teague, David Williamson, Lynda Chin, Mark T. Ross, Peter J. Campbell, David R. Bentley, P. Andrew Futreal, and Michael R. Stratton (Jan. 2010a). "A comprehensive catalogue of somatic mutations from a human cancer genome". In: *Nature* 463.7278, pp. 191–196.

Pleasance, Erin D., Philip J. Stephens, Sarah O'Meara, David J. McBride, Alison Meynert, David Jones, Meng-Lay Lin, David Beare, King Wai Lau, Chris Greenman, Ignacio Varela, Serena Nik-Zainal, Helen R. Davies, Gonzalo R. Ordoñez, Laura J. Mudie, Calli Latimer, Sarah Edkins, Lucy Stebbings, Lina Chen, Mingming Jia, Catherine Leroy, John Marshall, Andrew Menzies, Adam Butler, Jon W. Teague, Jonathon Mangion, Yongming A. Sun, Stephen F. McLaughlin, Heather E. Peckham, Eric F. Tsung, Gina L. Costa, Clarence C. Lee, John D. Minna, Adi Gazdar, Ewan Birney, Michael D. Rhodes, Kevin J. McKernan, Michael R. Stratton, P. Andrew Futreal, and Peter J. Campbell (Jan. 14, 2010b). "A small-cell lung cancer genome with complex signatures of tobacco exposure". In: *Nature* 463.7278, pp. 184–190.

Pleguezuelos-Manzano, Cayetano, Jens Puschhof, Axel Rosendahl Huber, Arne van Hoeck, Henry M. Wood, Jason Nomburg, Carino Gurjao, Freek Manders, Guillaume Dalmasso, Paul B. Stege, Fernanda L. Paganelli, Maarten H. Geurts, Joep Beumer, Tomohiro Mizutani, Yi Miao, Reinier van der Linden, Stefan van der Elst, K. Christopher Garcia, Janetta Top, Rob J. L. Willems, Marios Giannakis, Richard Bonnet, Phil Quirke, Matthew Meyerson, Edwin Cuppen, Ruben van Boxtel, and Hans Clevers (Apr. 2020). "Mutational signature in colorectal cancer caused by genotoxic pks + E. coli". In: *Nature* 580.7802, pp. 269–273.

Polak, Paz, Rosa Karlić, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael S. Lawrence, Alex Reynolds, Eric Rynes, Kristian Vlahoviček, John A. Stamatoyannopoulos, and Shamil R. Sunyaev (Feb. 2015). "Cell-of-origin chromatin organization shapes the mutational landscape of cancer". In: *Nature* 518.7539, pp. 360–364.

Popat, S., R. Hubner, and R. S. Houlston (Jan. 20, 2005). "Systematic review of microsatellite instability and colorectal cancer prognosis". In: *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 23.3, pp. 609–618.

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J. Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks (July 24, 2018). "Scaling accurate genetic variant discovery to tens of thousands of samples". In: *bioRxiv*, p. 201178.

Pott, Percival (1775). *Chirurgical Observations relative to the Cataract, the polypus of the nose, the cancer of the scrotum, ... ruptures, and the mortification of the toes, etc.* 222 pp.

Preston, Sean L., Wai-Man Wong, Annie On-On Chan, Richard Poulsom, Rosemary Jeffery, Robert A. Goodlad, Nikki Mandir, George Elia, Marco Novelli, Walter F. Bodmer, Ian P. Tomlinson, and Nicholas A. Wright (July 1, 2003). "Bottom-up Histogenesis of Colorectal Adenomas: Origin in the Monocryptal Adenoma and Initial Expansion by Crypt Fission". In: *Cancer Research* 63.13, pp. 3819–3825.

Price, L. F., C. C. Drovandi, A. Lee, and D. J. Nott (Jan. 2, 2018). "Bayesian Synthetic Likelihood". In: *Journal of Computational and Graphical Statistics* 27.1, pp. 1–11.

Price, Trevor D., Anna Qvarnström, and Darren E. Irwin (July 22, 2003). “The role of phenotypic plasticity in driving genetic evolution”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270.1523, pp. 1433–1440.

Priestley, Peter, Jonathan Baber, Martijn P. Lolkema, Neeltje Steeghs, Ewart de Bruijn, Charles Shale, Korneel Duyvesteyn, Susan Haidari, Arne van Hoeck, Wendy Onstenk, Paul Roepman, Mircea Voda, Haiko J. Bloemendal, Vivianne C. G. Tjan-Heijnen, Carla M. L. van Herpen, Mariette Labots, Petronella O. Witteveen, Egbert F. Smit, Stefan Sleijfer, Emile E. Voest, and Edwin Cuppen (Nov. 2019). “Pan-cancer whole-genome analyses of metastatic solid tumours”. In: *Nature* 575.7781, pp. 210–216.

Prior, Ian A., Paul D. Lewis, and Carla Mattos (May 15, 2012). “A comprehensive survey of Ras mutations in cancer”. In: *Cancer Research* 72.10, pp. 2457–2467.

Pritchard, J K, M T Seielstad, A Perez-Lezaun, and M W Feldman (Dec. 1, 1999). “Population growth of human Y chromosomes: a study of Y chromosome microsatellites.” In: *Molecular Biology and Evolution* 16.12, pp. 1791–1798.

Pybus, Oliver G. and Paul H. Harvey (Nov. 22, 2000). “Testing macro–evolutionary models using incomplete molecular phylogenies”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267.1459, pp. 2267–2272.

Pyke, C., J. Rømer, P. Kallunki, L. R. Lund, E. Ralfkiaer, K. Danø, and K. Tryggvason (Oct. 1994). “The gamma 2 chain of kalinin/laminin 5 is preferentially expressed in invading malignant cells in human cancers.” In: *The American Journal of Pathology* 145.4, pp. 782–791.

Quinlan, Aaron R. and Ira M. Hall (Mar. 15, 2010). “BEDTools: a flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6, pp. 841–842.

Quintás-Cardama, Alfonso and Jorge Cortes (Feb. 19, 2009). “Molecular biology of bcr-abl1–positive chronic myeloid leukemia”. In: *Blood* 113.8, pp. 1619–1630.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Ramírez-Soriano, Anna, Sebastià E. Ramos-Onsins, Julio Rozas, Francesc Calafell, and Arcadi Navarro (May 2008). “Statistical Power Analysis of Neutrality Tests Under Demographic Expansions, Contractions and Bottlenecks With Recombination”. In: *Genetics* 179.1, pp. 555–567.

Ray, Nicolas, Mathias Currat, and Laurent Excoffier (Jan. 2003). “Intra-deme molecular diversity in spatially expanding populations”. In: *Molecular Biology and Evolution* 20.1, pp. 76–86.

Revell, Liam J. (2012). “phytools: An R package for phylogenetic comparative biology (and other things).” In: *Methods in Ecology and Evolution* 3, pp. 217–223.

Reynolds, T. Y., S. Rockwell, and P. M. Glazer (Dec. 15, 1996). “Genetic instability induced by the tumor microenvironment”. In: *Cancer Research* 56.24, pp. 5754–5757.

Rimmer, Andy, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R. F. Twigg, WGS500 Consortium, Andrew O. M. Wilkie, Gil McVean, and Gerton Lunter (Aug. 2014). “Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications”. In: *Nature Genetics* 46.8, pp. 912–918.

Robert, Jacques (Sept. 2010). “Comparative study of tumorigenesis and tumor immunity in invertebrates and nonmammalian vertebrates”. In: *Developmental and Comparative Immunology* 34.9, pp. 915–925.

Roberts, Steven A., Michael S. Lawrence, Leszek J. Klimczak, Sara A. Grimm, David Fargo, Petar Stojanov, Adam Kiezun, Gregory V. Kryukov, Scott L. Carter, Gordon

Saksena, Shawn Harris, Ruchir R. Shah, Michael A. Resnick, Gad Getz, and Dmitry A. Gordenin (Sept. 2013). "An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers". In: *Nature Genetics* 45.9, pp. 970–976.

Roberts, Steven A., Joan Sterling, Cole Thompson, Shawn Harris, Deepak Mav, Ruchir Shah, Leszek J. Klimczak, Gregory V. Kryukov, Ewa Malc, Piotr A. Mieczkowski, Michael A. Resnick, and Dmitry A. Gordenin (May 25, 2012). "Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions". In: *Molecular Cell* 46.4, pp. 424–435.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth (Jan. 1, 2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics (Oxford, England)* 26.1, pp. 139–140.

Robinson, Mark D. and Alicia Oshlack (Mar. 2, 2010). "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biology* 11.3, R25.

Roche-Lestienne, Catherine, Valerie Soenen-Cornu, Nathalie Grardel-Duflos, Jean-Luc Laï, Nathalie Philippe, Thierry Facon, Pierre Fenaux, and Claude Preudhomme (Aug. 1, 2002). "Several types of mutations of the Abl gene can be found in chronic myeloid leukemia patients resistant to STI571, and they can pre-exist to the onset of treatment". In: *Blood* 100.3, pp. 1014–1018.

Rodriguez-Brenes, Ignacio A., Natalia L. Komarova, and Dominik Wodarz (Oct. 1, 2013). "Tumor growth dynamics: insights into evolutionary processes". In: *Trends in Ecology & Evolution* 28.10, pp. 597–604.

Roerink, Sophie F., Nobuo Sasaki, Henry Lee-Six, Matthew D. Young, Ludmil B. Alexandrov, Sam Behjati, Thomas J. Mitchell, Sebastian Grossmann, Howard Lightfoot, David A. Egan, Apollo Pronk, Niels Smakman, Joost van Gorp, Elizabeth Anderson, Stephen J. Gamble, Chris Alder, Marc van de Wetering, Peter J. Campbell, Michael R. Stratton, and Hans Clevers (Apr. 2018). "Intra-tumour diversification in colorectal cancer at the single-cell level". In: *Nature* 556.7702, pp. 457–462.

Rohland, Nadin and David Reich (Jan. 5, 2012). "Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture". In: *Genome Research* 22.5, pp. 939–946.

Rosenthal, Rachel (2016). *deconstructSigs: Identifies Signatures Present in a Tumor Sample*.

Roth, Andrew, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah (Apr. 2014). "PyClone: statistical inference of clonal population structure in cancer". In: *Nature Methods* 11.4, pp. 396–398.

Roufas, Constantinos, Ilias Georgakopoulos-Soares, and Apostolos Zaravinos (Mar. 1, 2021). "Molecular correlates of immune cytolytic subgroups in colorectal cancer by integrated genomics analysis". In: *NAR Cancer* 3 (zcab005).

Rous, Peyton (Sept. 1, 1910). "A transmissible avian neoplasm. (sarcoma of the common fowl.)" In: *The Journal of Experimental Medicine* 12.5, pp. 696–705.

— (Apr. 1, 1911). "A sarcoma of the fowl transmissible by an agent separable from the tumor cells". In: *The Journal of Experimental Medicine* 13.4, pp. 397–411.

— (July 7, 1967). "The Challenge to Man of the Neoplastic Cell". In: *Science* 157.3784, pp. 24–28.

Rowan, A. J., H. Lamlum, M. Ilyas, J. Wheeler, J. Straub, A. Papadopoulou, D. Bicknell, W. F. Bodmer, and I. P. M. Tomlinson (Mar. 28, 2000). "APC mutations in sporadic colorectal tumors: A mutational “hotspot” and interdependence of the “two hits””. In: *Proceedings of the National Academy of Sciences* 97.7, pp. 3352–3357.

Ryland, Georgina L., Sally M. Hunter, Maria A. Doyle, Simone M. Rowley, Michael Christie, Prue E. Allan, David DL Bowtell, Kylie L. Gorringe, and Ian G. Campbell (2013). “RNF43 is a tumour suppressor gene mutated in mucinous tumours of the ovary”. In: *The Journal of Pathology* 229.3, pp. 469–476.

Ryser, Marc D., Byung-Hoon Min, Kimberly D. Siegmund, and Darryl Shibata (May 29, 2018). “Spatial mutation patterns as markers of early colorectal tumor cell mobility”. In: *Proceedings of the National Academy of Sciences* 115.22, pp. 5774–5779.

Sachs, R. K., L. R. Hlatky, and P. Hahnfeldt (June 1, 2001). “Simple ODE models of tumor growth and anti-angiogenic or radiation treatment”. In: *Mathematical and Computer Modelling* 33.12, pp. 1297–1305.

Sackin, M. J. (1972). “”Good” and ”Bad” Phenograms”. In: *Systematic Zoology* 21.2, pp. 225–226.

Salcedo, Adriana, Maxime Tarabichi, Shadrielle Melijah G. Espiritu, Amit G. Deshwar, Matei David, Nathan M. Wilson, Stefan Dentro, Jeff A. Wintersinger, Lydia Y. Liu, Minjeong Ko, Srinivasan Sivanandan, Hongjiu Zhang, Kaiyi Zhu, Tai-Hsien Ou Yang, John M. Chilton, Alex Buchanan, Christopher M. Lalansingh, Christine P’ng, Catalina V. Anghel, Imaad Umar, Bryan Lo, William Zou, Jared T. Simpson, Joshua M. Stuart, Dimitris Anastassiou, Yuanfang Guan, Adam D. Ewing, Kyle Ellrott, David C. Wedge, Quaid Morris, Peter Van Loo, and Paul C. Boutros (Jan. 2020). “A community effort to create standards for evaluating tumor subclonal reconstruction”. In: *Nature Biotechnology* 38.1, pp. 97–107.

Samuels, Yardena, Luis A. Diaz, Oleg Schmidt-Kittler, Jordan M. Cummins, Laura Delong, Ian Cheong, Carlo Rago, David L. Huso, Christoph Lengauer, Kenneth W. Kinzler, Bert Vogelstein, and Victor E. Velculescu (June 2005). “Mutant PIK3CA promotes cell growth and invasion of human cancer cells”. In: *Cancer Cell* 7.6, pp. 561–573.

Samuels, Yardena, Zhenghe Wang, Alberto Bardelli, Natalie Silliman, Janine Ptak, Steve Szabo, Hai Yan, Adi Gazdar, Steven M. Powell, Gregory J. Riggins, James K. V. Willson, Sanford Markowitz, Kenneth W. Kinzler, Bert Vogelstein, and Victor E. Velculescu (Apr. 23, 2004). “High Frequency of Mutations of the PIK3CA Gene in Human Cancers”. In: *Science* 304.5670, pp. 554–554.

Sano, Akinori and Hidenori Tachida (Mar. 2005). “Gene Genealogy and Properties of Test Statistics of Neutrality Under Population Growth”. In: *Genetics* 169.3, pp. 1687–1697.

Santamaría, Rodrigo and Roberto Therón (Aug. 1, 2009). “Treevolution: visual analysis of phylogenetic trees”. In: *Bioinformatics* 25.15, pp. 1970–1971.

Sato, Toshiro, Daniel E. Stange, Marc Ferrante, Robert G. J. Vries, Johan H. Van Es, Stieneke Van den Brink, Winan J. Van Houdt, Apollo Pronk, Joost Van Gorp, Peter D. Siersema, and Hans Clevers (Nov. 2011). “Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett’s epithelium”. In: *Gastroenterology* 141.5, pp. 1762–1772.

Scheinin, Ilari, Daoud Sie, Henrik Bengtsson, Mark A. van de Wiel, Adam B. Olshen, Hinke F. van Thuijl, Hendrik F. van Essen, Paul P. Eijk, François Rustenburg, Gerrit A. Meijer, Jaap C. Reijneveld, Pieter Wesseling, Daniel Pinkel, Donna G. Albertson, and Bauke Ylstra (Dec. 2014). “DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly”. In: *Genome Research* 24.12, pp. 2022–2032.

Schep, Alicia (2020). *motifmatchr: Fast Motif Matching in R*.

Schliep, Klaus Peter (Feb. 15, 2011). “phangorn: phylogenetic analysis in R”. In: *Bioinformatics* 27.4, pp. 592–593.

Schneider, Valerie A., Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, Kim D. Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S. Fulton, Milinn Kremitzki, Vince Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, Will Chow, Joanna Collins, Glenn Harden, Tim Hubbard, Sarah Pelan, Jared T. Simpson, Glen Threadgold, James Torrance, Jonathan Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Heng Li, Chen-Shan Chin, Adam M. Phillippy, Richard Durbin, Richard K. Wilson, Paul Flicek, and Deanna M. Church (Jan. 1, 2016). “Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly”. In: *bioRxiv*, p. 072116.

Schreck, Carl F., Diana Fusco, Yuya Karita, Stephen Martis, Jona Kayser, Marie-Cécilia Duvernoy, and Oskar Hallatschek (Aug. 22, 2019). “Impact of crowding on the diversity of expanding populations”. In: *bioRxiv*, p. 743534.

Schuierer, Sven (2017). “A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples”. In: p. 13.

Schuster-Böckler, Benjamin and Ben Lehner (Aug. 2012). “Chromatin organization is a major influence on regional mutation rates in human cancer cells”. In: *Nature* 488.7412, pp. 504–507.

Schwann, Theodor and M. J. Schleyden (1847). *Microscopical researches into the accordance in the structure and growth of animals and plants*. London.

Schwarz, Roland F., Anne Trinh, Botond Sipos, James D. Brenton, Nick Goldman, and Florian Markowetz (Apr. 17, 2014). “Phylogenetic Quantification of Intra-tumour Heterogeneity”. In: *PLOS Computational Biology* 10.4, e1003535.

Schwarze, Katharina, James Buchanan, Jilles M. Fermont, Helene Dreau, Mark W. Tilley, John M. Taylor, Pavlos Antoniou, Samantha J. L. Knight, Carme Camps, Melissa M. Pentony, Erika M. Kvikstad, Steve Harris, Niko Popitsch, Alistair T. Pagnamenta, Anna Schuh, Jenny C. Taylor, and Sarah Wordsworth (Jan. 2020). “The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom”. In: *Genetics in Medicine* 22.1, pp. 85–94.

Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery (2016). “mclust 5: clustering, classification and density estimation using Gaussian finite mixture models”. In: *The R Journal* 8.1, pp. 289–317.

Seshan, Venkatraman E, Adam B Olshen, et al. (2015). *DNAcopy: a package for analyzing DNA copy data*.

Setlow, R. B. (July 22, 1966). “Cyclobutane-Type Pyrimidine Dimers in Polynucleotides”. In: *Science* 153.3734, pp. 379–386.

Shah, Neil P., John M. Nicoll, Bhushan Nagar, Mercedes E. Gorre, Ronald L. Paquette, John Kuriyan, and Charles L. Sawyers (Aug. 2002). “Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia”. In: *Cancer Cell* 2.2, pp. 117–125.

Shao, Kwang-Tsao (Sept. 1, 1990). “Tree Balance”. In: *Systematic Biology* 39.3, pp. 266–276.

Shen, Lanlan, Yutaka Kondo, Gary L. Rosner, Lianchun Xiao, Natalie Supunpong Hernandez, Jill Vilaythong, P. Scott Houlihan, Robert S. Krouse, Anil R. Prasad, Janine G. Einspahr, Julie Buckmeier, David S. Alberts, Stanley R. Hamilton, and Jean-Pierre J. Issa (Sept. 21, 2005). "MGMT promoter methylation and field defect in sporadic colorectal cancer". In: *Journal of the National Cancer Institute* 97.18, pp. 1330–1338.

Shendure, Jay and Hanlee Ji (Oct. 2008). "Next-generation DNA sequencing". In: *Nature Biotechnology* 26.10, pp. 1135–1145.

Sherry, S. T., M. Ward, and K. Sirotkin (Aug. 1999). "dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation". In: *Genome Research* 9.8, pp. 677–679.

Sherry, S. T., M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin (Jan. 1, 2001). "dbSNP: the NCBI database of genetic variation". In: *Nucleic Acids Research* 29.1, pp. 308–311.

Shibata, D. (2009). "Inferring human stem cell behaviour from epigenetic drift". In: *The Journal of Pathology* 217.2, pp. 199–205.

Shibata, Darryl (Feb. 1, 2011). "Mutation and epigenetic molecular clocks in cancer". In: *Carcinogenesis* 32.2, pp. 123–128.

Shih, Chiaho and Robert A. Weinberg (May 1, 1982). "Isolation of a transforming sequence from a human bladder carcinoma cell line". In: *Cell* 29.1, pp. 161–169.

Shoaib, Muhammad, David Walter, Peter J. Gillespie, Fanny Izard, Birthe Fahrenkrog, David Lleres, Mads Lerdrup, Jens Vilstrup Johansen, Klaus Hansen, Eric Julien, J. Julian Blow, and Claus S. Sørensen (Sept. 12, 2018). "Histone H4K20 methylation mediated chromatin compaction threshold ensures genome integrity by limiting DNA replication licensing". In: *Nature Communications* 9.1, p. 3704.

Sia, E. A., R. J. Kokoska, M. Dominska, P. Greenwell, and T. D. Petes (May 1, 1997). "Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes." In: *Molecular and Cellular Biology* 17.5, pp. 2851–2858.

Siegmund, Kimberly D., Paul Marjoram, Simon Tavaré, and Darryl Shibata (July 15, 2009a). "Many colorectal cancers are "flat" clonal expansions". In: *Cell cycle (Georgetown, Tex.)* 8.14, pp. 2187–2193.

Siegmund, Kimberly D., Paul Marjoram, Yen-Jung Woo, Simon Tavaré, and Darryl Shibata (Mar. 24, 2009b). "Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers". In: *Proceedings of the National Academy of Sciences of the United States of America* 106.12, pp. 4828–4833.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro (Sept. 1995). "Properties of statistical tests of neutrality for DNA polymorphism data". In: *Genetics* 141.1, pp. 413–429.

Sinsheimer, Robert L. (Nov. 1, 1989). "The Santa Cruz Workshop—May 1985". In: *Genomics* 5.4, pp. 954–956.

Sisson, S. A., Y. Fan, and Mark M. Tanaka (Feb. 6, 2007). "Sequential Monte Carlo without likelihoods". In: *Proceedings of the National Academy of Sciences* 104.6, pp. 1760–1765.

Slowikowski, Kamil (2020). *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'.*

Smith, J. Maynard (Sept. 1968). ""Haldane's Dilemma" and the Rate of Evolution". In: *Nature* 219.5159, pp. 1114–1116.

Snippert, Hugo J., Laurens G. van der Flier, Toshiro Sato, Johan H. van Es, Maaike van den Born, Carla Kroon-Veenboer, Nick Barker, Allon M. Klein, Jacco van Rheenen, Benjamin D. Simons, and Hans Clevers (Oct. 1, 2010). "Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells". In: *Cell* 143.1, pp. 134–144.

Sodir, Nicole M., Xuan Chen, Ryan Park, Andrea E. Nickel, Peter S. Conti, Rex Moats, James R. Bading, Darryl Shibata, and Peter W. Laird (Sept. 1, 2006). "Smad3 Deficiency Promotes Tumorigenesis in the Distal Colon of ApcMin/+ Mice". In: *Cancer Research* 66.17, pp. 8430–8438.

Sondka, Zbyslaw, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes (Nov. 2018). "The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers". In: *Nature Reviews. Cancer* 18.11, pp. 696–705.

Song, Min Sup, Leonardo Salmena, and Pier Paolo Pandolfi (May 2012). "The functions and regulation of the PTEN tumour suppressor". In: *Nature Reviews Molecular Cell Biology* 13.5, pp. 283–296.

Sottoriva, Andrea, Chris P Barnes, and Trevor A Graham (Apr. 1, 2017). "Catch my drift? Making sense of genomic intra-tumour heterogeneity". In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*. Evolutionary principles - heterogeneity in cancer? 1867.2, pp. 95–100.

Sottoriva, Andrea, Haeyoun Kang, Zhicheng Ma, Trevor A. Graham, Matthew P. Salomon, Junsong Zhao, Paul Marjoram, Kimberly Siegmund, Michael F. Press, Darryl Shibata, and Christina Curtis (Mar. 2015). "A Big Bang model of human colorectal tumor growth". In: *Nature Genetics* 47.3, pp. 209–216.

Souza-Pinto, Nadja C. de, Penelope A. Mason, Kazunari Hashiguchi, Lior Weissman, Jingyan Tian, David Guay, Michel Lebel, Tinna V. Stevnsner, Lene Juel Rasmussen, and Vilhelm A. Bohr (June 4, 2009). "Novel DNA mismatch-repair activity involving YB-1 in human mitochondria". In: *DNA repair* 8.6, pp. 704–719.

Spratt, JS (1961). "The growth of a colonic adenocarcinoma". In: *Ann Surg* 27, pp. 23–28.

Squires, Donald F. (Apr. 23, 1965). "Neoplasia in a Coral?" In: *Science* 148.3669, pp. 503–505.

Stadler, T., O. G. Pybus, and M. P. H. Stumpf (Jan. 15, 2021). "Phylodynamics for cell biologists". In: *Science* 371.6526.

Stamatoyannopoulos, John A., Ivan Adzhubei, Robert E. Thurman, Gregory V. Kryukov, Sergei M. Mirkin, and Shamil R. Sunyaev (Apr. 2009). "Human mutation rate associated with DNA replication timing". In: *Nature Genetics* 41.4, pp. 393–395.

Steel, G. G. and L. F. Lamerton (Mar. 1966). "The growth rate of human tumours." In: *British Journal of Cancer* 20.1, pp. 74–86.

Stelloo, E., A. M. L. Jansen, E. M. Osse, R. A. Nout, C. L. Creutzberg, D. Ruano, D. N. Church, H. Morreau, V. T. H. B. M. Smit, T. van Wezel, and T. Bosse (Jan. 1, 2017). "Practical guidance for mismatch repair-deficiency testing in endometrial cancer". In: *Annals of Oncology* 28.1, pp. 96–102.

Stintzing, Sebastian and Heinz-Josef Lenz (Dec. 4, 2013). "A Small Cog in a Big Wheel: PIK3CA Mutations in Colorectal Cancer". In: *JNCI: Journal of the National Cancer Institute* 105.23, pp. 1775–1776.

Struhl, Kevin (Jan. 3, 1998). "Histone acetylation and transcriptional regulatory mechanisms". In: *Genes & Development* 12.5, pp. 599–606.

Sun, Ruping, Zheng Hu, Andrea Sottoriva, Trevor A. Graham, Arbel Harpak, Zhicheng Ma, Jared M. Fischer, Darryl Shibata, and Christina Curtis (July 2017). "Between-region genetic divergence reflects the mode and tempo of tumor evolution". In: *Nature Genetics* 49.7, pp. 1015–1024.

Suter, Catherine M., David I. Martin, and Robyn L. Ward (Mar. 2004). "Hypomethylation of L1 retrotransposons in colorectal cancer and adjacent normal tissue". In: *International Journal of Colorectal Disease* 19.2, pp. 95–101.

Suzuki, Miho M., Alastair R. W. Kerr, Dina De Sousa, and Adrian Bird (May 2007). "CpG methylation is targeted to transcription units in an invertebrate genome". In: *Genome Research* 17.5, pp. 625–631.

Suzuki, Yuka, Sarah Boonhsi Ng, Clarinda Chua, Wei Qiang Leow, Jermain Chng, Shi Yang Liu, Kalpana Ramnarayanan, Anna Gan, Dan Liang Ho, Rachel Ten, Yan Su, Alexandar Lezhava, Jiunn Herng Lai, Dennis Koh, Kiat Hon Lim, Patrick Tan, Steven G. Rozen, and Iain Beehuat Tan (2017). "Multiregion ultra-deep sequencing reveals early intermixing and variable levels of intratumoral heterogeneity in colorectal cancer". In: *Molecular Oncology* 11.2, pp. 124–139.

Swofford, David L. and Wayne P. Maddison (Dec. 1, 1987). "Reconstructing ancestral character states under Wagner parsimony". In: *Mathematical Biosciences* 87.2, pp. 199–229.

Szklarczyk, Damian, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T. Doncheva, John H. Morris, Peer Bork, Lars J. Jensen, and Christian von Mering (Jan. 8, 2019). "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets". In: *Nucleic Acids Research* 47 (D1), pp. D607–D613.

Tajima, F. (Nov. 1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism". In: *Genetics* 123.3, pp. 585–595.

Talevich, Eric, A. Hunter Shain, Thomas Botton, and Boris C. Bastian (Apr. 21, 2016). "CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing". In: *PLOS Computational Biology* 12.4, e1004873.

Talkington, Anne and Rick Durrett (Oct. 2015). "Estimating tumor growth rates in vivo". In: *Bulletin of mathematical biology* 77.10, pp. 1934–1954.

Tamura, Tomohiko, Hideyuki Yanai, David Savitsky, and Tadatsugu Taniguchi (2008). "The IRF Family Transcription Factors in Immunity and Oncogenesis". In: *Annual Review of Immunology* 26.1, pp. 535–584.

Tan, S. Y. and J. Brown (July 2006). "Rudolph Virchow (1821-1902): "pope of pathology"". In: *Singapore Medical Journal* 47.7, pp. 567–568.

Tanaka, Mark M., Andrew R. Francis, Fabio Luciani, and S. A. Sisson (July 2006). "Using Approximate Bayesian Computation to Estimate Tuberculosis Transmission Parameters From Genotype Data". In: *Genetics* 173.3, pp. 1511–1520.

Tang, W., M. Dodge, D. Gundapaneni, C. Michnoff, M. Roth, and L. Lum (July 15, 2008). "A genome-wide RNAi screen for Wnt/ -catenin pathway components identifies unexpected roles for TCF transcription factors in cancer". In: *Proceedings of the National Academy of Sciences* 105.28, pp. 9697–9702.

Tarabichi, Maxime, Iñigo Martincorena, Moritz Gerstung, Armand M. Leroi, Florian Markowetz, Paul T. Spellman, Quaid D. Morris, Ole Christian Lingjærde, David C. Wedge, and Peter Van Loo (Dec. 2018). "Neutral tumor evolution?" In: *Nature Genetics* 50.12, pp. 1630–1633.

Tarabichi, Maxime, Adriana Salcedo, Amit G. Deshwar, Máire Ni Leathlobhair, Jeff Wintersinger, David C. Wedge, Peter Van Loo, Quaid D. Morris, and Paul C. Boutros (Jan. 4, 2021). "A practical guide to cancer subclonal reconstruction from DNA sequencing". In: *Nature Methods*, pp. 1–12.

Tate, John G, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes (Jan. 8, 2019). "COSMIC: the Catalogue Of Somatic Mutations In Cancer". In: *Nucleic Acids Research* 47 (D1), pp. D941–D947.

Tavaré, Simon, David J. Balding, R. C. Griffiths, and Peter Donnelly (Feb. 1, 1997). "Inferring Coalescence Times From DNA Sequence Data". In: *Genetics* 145.2, pp. 505–518.

Team, BC and BP Maintainer (2019). *TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s).*

Temin, H. M. and S. Mizutani (June 27, 1970). "RNA-dependent DNA polymerase in virions of Rous sarcoma virus". In: *Nature* 226.5252, pp. 1211–1213.

Temko, Daniel, Inge C. Van Gool, Emily Rayner, Mark Glaire, Seiko Makino, Matthew Brown, Laura Chegwidden, Claire Palles, Jeroen Depreeuw, Andrew Beggs, Chaido Stathopoulou, John Mason, Ann-Marie Baker, Marc Williams, Vincenzo Cerundolo, Margarida Rei, Jenny C. Taylor, Anna Schuh, Ahmed Ahmed, Frédéric Amant, Diether Lambrechts, Vincent THBM Smit, Tjalling Bosse, Trevor A. Graham, David N. Church, and Ian Tomlinson (2018). "Somatic POLE exonuclease domain mutations are early events in sporadic endometrial and colorectal carcinogenesis, determining driver mutational landscape, clonal neoantigen burden and immune response". In: *The Journal of Pathology* 245.3, pp. 283–296.

Thanasopoulou, Angeliki, Alexandra G. Xanthopoulou, Athanasios K. Anagnostopoulos, Eumorphia G. Konstantakou, Lukas H. Margaritis, Isidora S. Papassideri, Dimitrios J. Stravopodis, George T. Tsangaris, and Ema Anastasiadou (Mar. 1, 2012). "Silencing of CCDC6 Reduces the Expression of 14-3-3σ in Colorectal Carcinoma Cells". In: *Anticancer Research* 32.3, pp. 907–913.

Tokheim, Collin, Rohit Bhattacharya, Noushin Niknafs, Derek M. Gygax, Rick Kim, Michael Ryan, David L. Masica, and Rachel Karchin (July 1, 2016). "Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure". In: *Cancer Research* 76.13, pp. 3719–3731.

Tomasetti, Cristian and Doron Levy (Oct. 2010). "An elementary approach to modeling drug resistance in cancer". In: *Mathematical biosciences and engineering: MBE* 7.4, pp. 905–918.

Tomkova, Marketa, Jakub Tomek, Skirmantas Kriaucionis, and Benjamin Schuster-Böckler (Sept. 10, 2018). "Mutational signature distribution varies with DNA replication timing and strand asymmetry". In: *Genome Biology* 19.1, p. 129.

Toni, Tina, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P.H Stumpf (Feb. 6, 2009). "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems". In: *Journal of The Royal Society Interface* 6.31, pp. 187–202.

Tsao, J. L., J. Zhang, R. Salovaara, Z. H. Li, H. J. Järvinen, J. P. Mecklin, L. A. Aaltonen, and D. Shibata (Oct. 1998). "Tracing cell fates in human colorectal tumors from somatic microsatellite mutations: evidence of adenomas with stem cell architecture”. In: *The American Journal of Pathology* 153.4, pp. 1189–1200.

Tsao, Jen-Lan, Yasushi Yatabe, Reijo Salovaara, Heikki J. Järvinen, Jukka-Pekka Mecklin, Lauri A. Aaltonen, Simon Tavaré, and Darryl Shibata (Feb. 1, 2000). “Genetic reconstruction of individual colorectal tumor histories”. In: *Proceedings of the National Academy of Sciences* 97.3, pp. 1236–1241.

Tsao, M. S., J. W. Grisham, and K. G. Nelson (Oct. 1985). “Clonal analysis of tumorigenicity and paratumorigenic phenotypes in rat liver epithelial cells chemically transformed in vitro”. In: *Cancer Research* 45.10, pp. 5139–5144.

Tsuchida, Nobuo, Tom Ryder, and Eiichi Ohtsubo (Sept. 3, 1982). “Nucleotide Sequence of the Oncogene Encoding the p21 Transforming Protein of Kirsten Murine Sarcoma Virus”. In: *Science* 217.4563, pp. 937–939.

Tsukiyama, Tadasuke, Akimasa Fukui, Sayuri Terai, Yoichiro Fujioka, Keisuke Shinada, Hidehisa Takahashi, Terry P. Yamaguchi, Yusuke Ohba, and Shigetsugu Hatakeyama (June 1, 2015). “Molecular Role of RNF43 in Canonical and Noncanonical Wnt Signaling”. In: *Molecular and Cellular Biology* 35.11, pp. 2007–2023.

Turajlic, Samra, Andrea Sottoriva, Trevor Graham, and Charles Swanton (July 2019). “Resolving genetic heterogeneity in cancer”. In: *Nature Reviews Genetics* 20.7, pp. 404–416.

Turcatti, Gerardo, Anthony Romieu, Milan Fedurco, and Ana-Paula Tairi (Mar. 2008). “A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis”. In: *Nucleic Acids Research* 36.4, e25.

Uchi, Ryutaro, Yusuke Takahashi, Atsushi Niida, Teppei Shimamura, Hidenari Hirata, Keishi Sugimachi, Genta Sawada, Takeshi Iwaya, Junji Kurashige, Yoshiaki Shinden, Tomohiro Iguchi, Hidetoshi Eguchi, Kenichi Chiba, Yuichi Shiraishi, Genta Nagae, Kenichi Yoshida, Yasunobu Nagata, Hiroshi Haeno, Hirofumi Yamamoto, Hideshi Ishii, Yuichiro Doki, Hisae Iinuma, Shin Sasaki, Satoshi Nagayama, Kazutaka Yamada, Shinichi Yachida, Mamoru Kato, Tatsuhiro Shibata, Eiji Oki, Hiroshi Saeki, Ken Shirabe, Yoshinao Oda, Yoshihiko Maehara, Shizuo Komune, Masaki Mori, Yutaka Suzuki, Ken Yamamoto, Hiroyuki Aburatani, Seishi Ogawa, Satoru Miyano, and Koshi Mimori (Feb. 18, 2016). “Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution”. In: *PLOS Genetics* 12.2, e1005778.

Uhlen, Mathias, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, Muhammad Arif, Zhengtao Liu, Fredrik Edfors, Kemal Sanli, Kalle von Feilitzen, Per Oksvold, Emma Lundberg, Sophia Hober, Peter Nilsson, Johanna Mattsson, Jochen M. Schwenk, Hans Brunnström, Bengt Glimelius, Tobias Sjöblom, Per-Henrik Edqvist, Dijana Djureinovic, Patrick Micke, Cecilia Lindskog, Adil Mardinoglu, and Fredrik Ponten (Aug. 18, 2017). “A pathology atlas of the human cancer transcriptome”. In: *Science* 357.6352.

Underhill, Peter A., Peidong Shen, Alice A. Lin, Li Jin, Giuseppe Passarino, Wei H. Yang, Erin Kauffman, Batsheva Bonné-Tamir, Jaume Bertranpetit, Paolo Francalacci, Muntaser Ibrahim, Trefor Jenkins, Judith R. Kidd, S. Qasim Mehdi, Mark T. Seielstad, R. Spencer Wells, Alberto Piazza, Ronald W. Davis, Marcus W. Feldman, L. Luca Cavalli-Sforza, and Peter J. Oefner (Nov. 2000). “Y chromosome sequence variation and the history of human populations”. In: *Nature Genetics* 26.3, pp. 358–361.

Van der Auwera, Geraldine A., Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo (2013). “From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline". In: *Current Protocols in Bioinformatics* 43, pp. 11.10.1–11.10.33.

Vander Velde, Robert, Nara Yoon, Viktoriya Marusyk, Arda Durmaz, Andrew Dhawan, Daria Miroshnychenko, Diego Lozano-Peral, Bina Desai, Olena Balynska, Jan Poleszhuk, Liu Kenian, Mingxiang Teng, Mohamed Abazeed, Omar Mian, Aik Choon Tan, Eric Haura, Jacob Scott, and Andriy Marusyk (May 14, 2020). "Resistance to targeted therapies as a multifactorial, gradual adaptation to inhibitor specific selective pressures". In: *Nature Communications* 11.1, p. 2393.

Varmus, Harold E., Robin A. Weiss, Robert R. Friis, Warren Levinson, and J. Michael Bishop (Jan. 1, 1972). "Detection of Avian Tumor Virus-Specific Nucleotide Sequences in Avian Cell DNAs". In: *Proceedings of the National Academy of Sciences* 69.1, pp. 20–24.

Venter, J. Craig et al. (Feb. 16, 2001). "The Sequence of the Human Genome". In: *Science* 291.5507, pp. 1304–1351.

Via, Sara and Russell Lande (1985). "Genotype-Environment Interaction and the Evolution of Phenotypic Plasticity". In: *Evolution* 39.3, pp. 505–522.

Virchow, Rudolf (1860). *Cellular Pathology as based upon physiological and pathological histology. Twenty lectures delivered in... 1858. Translated from the second edition of the original by F. Chance. With notes and numerous emendations principally from MS. notes of the author, and illustrated by... engravings on wood.*

Vogelstein, Bert and Kenneth W. Kinzler (Nov. 11, 2015). *The Path to Cancer — Three Strikes and You're Out*. http://dx.doi.org/10.1056/NEJMp1508811. URL: https://www.nejm.org/doi/10.1056/NEJMp1508811 (visited on 05/22/2021).

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler (Mar. 29, 2013). "Cancer genome landscapes". In: *Science (New York, N.Y.)* 339.6127, pp. 1546–1558.

Vousden, Karen H. and Carol Prives (May 1, 2009). "Blinded by the Light: The Growing Complexity of p53". In: *Cell* 137.3, pp. 413–431.

Waddington, CH (1942). "The epigenotype: Endeavour". In.

Wang, Grace M., Hong Yuen Wong, Hiroyuki Konishi, Brian G. Blair, Abde M. Abukhdeir, John P. Gustin, D. Marc Rosen, Samuel Ray Denmeade, Zeshaan Rasheed, William Matsui, Joseph P. Garay, Morassa Mohseni, Michaela J. Higgins, Justin Cidado, Danijela Jelovac, Sarah Croessmann, Rory L. Cochran, Sivasundaram Karnan, Yuko Konishi, Akinobu Ota, Yoshitaka Hosokawa, Pedram Argani, Josh Lauring, and Ben Ho Park (June 1, 2013a). "Single Copies of Mutant KRAS and Mutant PIK3CA Cooperate in Immortalized Human Epithelial Cells to Induce Tumor Formation". In: *Cancer Research* 73.11, pp. 3248–3261.

Wang, Hurng-Yi, Yuxin Chen, Ding Tong, Shaoping Ling, Zheng Hu, Yong Tao, Xuemei Lu, and Chung-I Wu (Jan. 1, 2018a). "Is the evolution in tumors Darwinian or non-Darwinian?" In: *National Science Review* 5.1, pp. 15–17.

Wang, Kai, Siu Tsan Yuen, Jiangchun Xu, Siu Po Lee, Helen H. N. Yan, Stephanie T. Shi, Hoi Cheong Siu, Shibing Deng, Kent Man Chu, Simon Law, Kok Hoe Chan, Annie S. Y. Chan, Wai Yin Tsui, Siu Lun Ho, Anthony K. W. Chan, Jonathan L. K. Man, Valentina Foglizzo, Man Kin Ng, April S. Chan, Yick Pang Ching, Grace H. W. Cheng, Tao Xie, Julio Fernandez, Vivian S. W. Li, Hans Clevers, Paul A. Rejto, Mao Mao, and Suet Yi Leung (June 2014). "Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer". In: *Nature Genetics* 46.6, pp. 573–582.

Wang, Qiang, Yan-long Shi, Kai Zhou, Li-li Wang, Ze-xuan Yan, Yu-lin Liu, Li-li Xu, Shiwei Zhao, Hui-li Chu, Ting-ting Shi, Qing-hua Ma, and Jingwang Bi (July 3, 2018b). "PIK3CA mutations confer resistance to first-line chemotherapy in colorectal cancer". In: *Cell Death & Disease* 9.7, pp. 1–11.

Wang, Qingguo, Peilin Jia, Fei Li, Haiquan Chen, Hongbin Ji, Donald Hucks, Kimberly Brown Dahlman, William Pao, and Zhongming Zhao (Oct. 11, 2013b). "Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers". In: *Genome Medicine* 5.10, p. 91.

Waris, Gulam and Haseeb Ahsan (May 11, 2006). "Reactive oxygen species: role in the development of cancer and various chronic conditions". In: *Journal of Carcinogenesis* 5, p. 14.

Watson, J. D. and F. H. Crick (Apr. 25, 1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid". In: *Nature* 171.4356, pp. 737–738.

Weghorn, Donate and Shamil Sunyaev (Dec. 2017). "Bayesian inference of negative and positive selection in human cancers". In: *Nature Genetics* 49.12, pp. 1785–1788.

Wegmann, Daniel, Christoph Leuenberger, and Laurent Excoffier (Aug. 1, 2009). "Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood". In: *Genetics* 182.4, pp. 1207–1218.

Weir, B. S. and C. Clark Cockerham (1984). "Estimating F-Statistics for the Analysis of Population Structure". In: *Evolution* 38.6, pp. 1358–1370.

Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G. Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J. M. Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes (Sept. 11, 2014). "Determination and inference of eukaryotic transcription factor sequence specificity". In: *Cell* 158.6, pp. 1431–1443.

Wendt, Kerstin S. and Jan-Michael Peters (Feb. 1, 2009). "How cohesin and CTCF cooperate in regulating gene expression". In: *Chromosome Research* 17.2, pp. 201–214.

Wenzel, Janna, Katja Rose, Elham Bavafaye Haghighi, Constanze Lamprecht, Gilles Rauen, Vivien Freihen, Rebecca Kesselring, Melanie Boerries, and Andreas Hecht (May 2020). "Loss of the nuclear Wnt pathway effector TCF7L2 promotes migration and invasion of human colorectal cancer cells". In: *Oncogene* 39.19, pp. 3893–3909.

Werner, Benjamin, Arne Traulsen, Andrea Sottoriva, and David Dingli (Mar. 27, 2017). "Detecting truly clonal alterations from multi-region profiling of tumours". In: *Scientific Reports* 7.1, p. 44991.

Wetering, Marc van de, Hayley E. Francies, Joshua M. Francis, Gergana Bounova, Francesco Iorio, Apollo Pronk, Winan van Houdt, Joost van Gorp, Amaro Taylor-Weiner, Lennart Kester, Anne McLaren-Douglas, Joyce Blokker, Sridevi Jaksani, Sina Bartfeld, Richard Volckman, Peter van Sluis, Vivian S.W. Li, Sara Seepo, Chandra Sekhar Pedamallu, Kristian Cibulskis, Scott L. Carter, Aaron McKenna, Michael S. Lawrence, Lee Lichtenstein, Chip Stewart, Jan Koster, Rogier Versteeg, Alexander van Oudenaarden, Julio Saez-Rodriguez, Robert G.J. Vries, Gad Getz, Lodewyk Wessels, Michael R. Stratton, Ultan McDermott, Matthew Meyerson, Mathew J. Garnett, and Hans Clevers (May 7, 2015). "Prospective derivation of a Living Organoid Biobank of colorectal cancer patients". In: *Cell* 161.4, pp. 933–945.

Wickham, Hadley (2007). “Reshaping data with the reshape package”. In: *Journal of Statistical Software* 21.12.

— (2016). *ggplot2: Elegant Graphics for Data Analysis*.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller (2020). *dplyr: A Grammar of Data Manipulation*.

Wiles, Elizabeth T. and Eric U. Selker (Apr. 2017). “H3K27 methylation: a promiscuous repressive chromatin mark”. In: *Current opinion in genetics & development* 43, pp. 31–37.

Wilhelm, Stefan and Manjunath B. G (2015). *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*.

WILK, M. B. and R. GNANADESIKAN (Mar. 1, 1968). “Probability plotting methods for the analysis for the analysis of data”. In: *Biometrika* 55.1, pp. 1–17.

Wilke, Claus O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for ’ggplot2’*.

Williams, Marc (2018). *neutralitytestr: Test for a Neutral Evolutionary Model in Cancer Sequencing Data*.

Williams, Marc J (2019). “Methods and practice of detecting selection in human cancers”. PhD thesis. UCL (University College London).

Williams, Marc J., Andrea Sottoriva, and Trevor A. Graham (2019). “Measuring Clonal Evolution in Cancer with Genomics”. In: *Annual Review of Genomics and Human Genetics* 20.1, pp. 309–329.

Williams, Marc J., Benjamin Werner, Chris P. Barnes, Trevor A. Graham, and Andrea Sottoriva (Mar. 2016). “Identification of neutral tumor evolution across cancer types”. In: *Nature Genetics* 48.3, pp. 238–244.

— (Sept. 2017). “Reply: Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures”. In: *Nature Genetics* 49.9, pp. 1289–1291.

Williams, Marc J., Benjamin Werner, Timon Heide, Chris P. Barnes, Trevor A. Graham, and Andrea Sottoriva (Dec. 2018a). “Reply to ‘Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data’”. In: *Nature Genetics* 50.12, pp. 1628–1630.

Williams, Marc J., Benjamin Werner, Timon Heide, Christina Curtis, Chris P. Barnes, Andrea Sottoriva, and Trevor A. Graham (June 2018b). “Quantification of subclonal selection in cancer from bulk sequencing data”. In: *Nature Genetics* 50.6, pp. 895–903.

Willis, Amy, Eun Joo Jung, Therese Wakefield, and Xinbin Chen (Mar. 2004). “Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes”. In: *Oncogene* 23.13, pp. 2330–2338.

Woese, C. R. (May 22, 1964). “UNIVERSALITY IN THE GENETIC CODE”. In: *Science (New York, N.Y.)* 144.3621, pp. 1030–1031.

Wong, W.-M., N. Mandir, R. A. Goodlad, B. C. Y. Wong, S. B. Garcia, S.-K. Lam, and N. A. Wright (Feb. 1, 2002). “Histogenesis of human colorectal adenomas and hyperplastic polyps: the role of cell proliferation and crypt fission”. In: *Gut* 50.2, pp. 212–217.

Wong-Staal, F., R. Dalla-Favera, G. Franchini, E. P. Gelmann, and R. C. Gallo (July 10, 1981). “Three distinct genes in human DNA related to the transforming genes of mammalian sarcoma retroviruses”. In: *Science* 213.4504, pp. 226–228.

Wood, Simon N. (Aug. 2010). “Statistical inference for noisy nonlinear ecological dynamic systems”. In: *Nature* 466.7310, pp. 1102–1104.

Woods, Mae L. and Chris P. Barnes (Oct. 14, 2016). “Mechanistic Modelling and Bayesian Inference Elucidates the Variable Dynamics of Double-Strand Break Repair”. In: *PLOS Computational Biology* 12.10, e1005131.

Wright, Nicholas A, Malcolm Alison, et al. (1984). *The biology of epithelial cell populations*. Vol. 1.

Wright, Nicholas A. and Richard Poulsom (2012). “Omnis cellula e cellula revisited: cell biology as the foundation of pathology”. In: *The Journal of Pathology* 226.2, pp. 145–147.

Wright, Sewall (1931). “Evolution in Mendelian populations”. In: *Genetics* 16.2, p. 97.

Wu, Chung-I, Hurng-Yi Wang, Shaoping Ling, and Xuemei Lu (Nov. 23, 2016). “The Ecology and Evolution of Cancer: The Ultra-Microevolutionary Process”. In: *Annual Review of Genetics* 50.1, pp. 347–369.

Xia, Li C., Paul Van Hummelen, Matthew Kubit, HoJoon Lee, John M. Bell, Susan M. Grimes, Christina Wood-Bouwens, Stephanie U. Greer, Tyler Barker, Derrick S. Haslem, James M. Ford, Gail Fulde, Hanlee P. Ji, and Lincoln D. Nadauld (Mar. 19, 2020). “Whole genome analysis identifies the association of TP53 genomic deletions with lower survival in Stage III colorectal cancer”. In: *Scientific Reports* 10.1, p. 5009.

Xie, Wen, David L. Rimm, Yong Lin, Weichung J. Shih, and Michael Reiss (Aug. 2003). “Loss of Smad Signaling in Human Colorectal Cancer Is Associated with Advanced Disease and Poor Prognosis”. In: *The Cancer Journal* 9.4, pp. 302–312.

Xing, Dong, Longzhi Tan, Chi-Han Chang, Heng Li, and X. Sunney Xie (Feb. 23, 2021). “Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands”. In: *Proceedings of the National Academy of Sciences of the United States of America* 118.8, e2013106118.

Xu, Chang (Jan. 1, 2018). “A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data”. In: *Computational and Structural Biotechnology Journal* 16, pp. 15–24.

Xu, Huilei, John DiCarlo, Ravi Vijaya Satya, Quan Peng, and Yexun Wang (Mar. 28, 2014). “Comparison of somatic mutation calling methods in amplicon and whole exome sequence data”. In: *BMC Genomics* 15.1, p. 244.

Xu, X., S. G. Brodie, X. Yang, Y. H. Im, W. T. Parks, L. Chen, Y. X. Zhou, M. Weinstein, S. J. Kim, and C. X. Deng (Apr. 6, 2000). “Haploid loss of the tumor suppressor Smad4/Dpc4 initiates gastric polyposis and cancer in mice”. In: *Oncogene* 19.15, pp. 1868–1874.

Xu, Xun, Yong Hou, Xuyang Yin, Li Bao, Aifa Tang, Luting Song, Fuqiang Li, Shirley Tsang, Kui Wu, Hanjie Wu, Weiming He, Liang Zeng, Manjie Xing, Renhua Wu, Hui Jiang, Xiao Liu, Dandan Cao, Guangwu Guo, Xueda Hu, Yaoting Gui, Zesong Li, Wenyue Xie, Xiaojuan Sun, Min Shi, Zhiming Cai, Bin Wang, Meiming Zhong, Jingxiang Li, Zuhong Lu, Ning Gu, Xiuqing Zhang, Laurie Goodman, Lars Bolund, Jian Wang, Huanming Yang, Karsten Kristiansen, Michael Dean, Yingrui Li, and Jun Wang (Mar. 2, 2012). “Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor”. In: *Cell* 148.5, pp. 886–895.

Xue, BingKan and Stanislas Leibler (Dec. 11, 2018). “Benefits of phenotypic plasticity for population growth in varying environments”. In: *Proceedings of the National Academy of Sciences* 115.50, pp. 12745–12750.

Yachida, Shinichi, Siân Jones, Ivana Bozic, Tibor Antal, Rebecca Leary, Baojin Fu, Mihoko Kamiyama, Ralph H. Hruban, James R. Eshleman, Martin A. Nowak, Victor E. Velculescu, Kenneth W. Kinzler, Bert Vogelstein, and Christine A. Iacobuzio-Donahue

(Oct. 2010). “Distant metastasis occurs late during the genetic evolution of pancreatic cancer”. In: *Nature* 467.7319, pp. 1114–1117.

Yamagiwa, Katsusaburo (1915). “Experimentelle studie uber die pathogenese der epithelialgeschwulste”. In: *Mitt. Med. Fad. Tokio* 15, pp. 295–344.

Yanai, Hideyuki, Hideo Negishi, and Tadatsugu Taniguchi (Nov. 1, 2012). “The IRF family of transcription factors: Inception, impact and implications in oncogenesis”. In: *Oncoimmunology* 1.8, pp. 1376–1386.

Yang, Ziheng and Joseph P. Bielawski (Dec. 1, 2000). “Statistical methods for detecting molecular adaptation”. In: *Trends in Ecology & Evolution* 15.12, pp. 496–503.

Yang, Ziheng and Rasmus Nielsen (Jan. 1, 2000). “Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models”. In: *Molecular Biology and Evolution* 17.1, pp. 32–43.

Yang, Ziheng, Rasmus Nielsen, Nick Goldman, and Anne-Mette Krabbe Pedersen (May 1, 2000). “Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites”. In: *Genetics* 155.1, pp. 431–449.

Yatabe, Yasushi, Simon Tavaré, and Darryl Shibata (Sept. 11, 2001). “Investigating stem cells in human colon by using methylation patterns”. In: *Proceedings of the National Academy of Sciences* 98.19, pp. 10839–10844.

Yates, Andrew D et al. (Jan. 8, 2020). “Ensembl 2020”. In: *Nucleic Acids Research* 48 (D1), pp. D682–D688.

Yates, Lucy R., Moritz Gerstung, Stian Knappskog, Christine Desmedt, Gunes Gundem, Peter Van Loo, Turid Aas, Ludmil B. Alexandrov, Denis Larsimont, Helen Davies, Yilong Li, Young Seok Ju, Manasa Ramakrishna, Hans Kristian Haugland, Peer Kaare Lilleng, Serena Nik-Zainal, Stuart McLaren, Adam Butler, Sancha Martin, Dominic Glodzik, Andrew Menzies, Keiran Raine, Jonathan Hinton, David Jones, Laura J. Mudie, Bing Jiang, Delphine Vincent, April Greene-Colozzi, Pierre-Yves Adnet, Aquila Fatima, Marion Maetens, Michail Ignatiadis, Michael R. Stratton, Christos Sotiriou, Andrea L. Richardson, Per Eystein Lønning, David C. Wedge, and Peter J. Campbell (July 2015). “Subclonal diversification of primary breast cancer revealed by multiregion sequencing”. In: *Nature Medicine* 21.7, pp. 751–759.

Yates, Lucy R., Stian Knappskog, David Wedge, James H. R. Farmery, Santiago Gonzalez, Inigo Martincorena, Ludmil B. Alexandrov, Peter Van Loo, Hans Kristian Haugland, Peer Kaare Lilleng, Gunes Gundem, Moritz Gerstung, Elli Pappaemmanuil, Patrycja Gazinska, Shriram G. Bhosle, David Jones, Keiran Raine, Laura Mudie, Calli Latimer, Elinor Sawyer, Christine Desmedt, Christos Sotiriou, Michael R. Stratton, Anieta M. Sieuwerts, Andy G. Lynch, John W. Martens, Andrea L. Richardson, Andrew Tutt, Per Eystein Lønning, and Peter J. Campbell (Aug. 14, 2017). “Genomic Evolution of Breast Cancer Metastasis and Relapse”. In: *Cancer Cell* 32.2, 169–184.e7.

Yeh, Chien-Hung, Marcia Bellon, and Christophe Nicot (Aug. 7, 2018). “FBXW7: a critical tumor suppressor of human cancers”. In: *Molecular Cancer* 17.1, p. 115.

Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam (2017). “ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data”. In: *Methods in Ecology and Evolution* 8.1, pp. 28–36.

Yu, Guangchuang, Li-Gen Wang, and Qing-Yu He (July 15, 2015). “ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization”. In: *Bioinformatics* 31.14, pp. 2382–2383.

Yu, Jia, Permeen A. Mohamed Yusoff, Daniëlle T. J. Woutersen, Pamela Goh, Nathan Harmston, Ron Smits, David M. Epstein, David M. Virshup, and Babita Madan (Dec. 15, 2020). "The Functional Landscape of Patient-Derived RNF43 Mutations Predicts Sensitivity to Wnt Inhibition". In: *Cancer Research* 80.24, pp. 5619–5632.

Yue, Xuetian, Yuhan Zhao, Yang Xu, Min Zheng, Zhaohui Feng, and Wenwei Hu (June 2, 2017). "Mutant p53 in cancer: accumulation, gain-of-function and therapy". In: *Journal of molecular biology* 429.11, pp. 1595–1606.

Zaccaria, Simone and Benjamin J. Raphael (Feb. 2021). "Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL". In: *Nature Biotechnology* 39.2, pp. 207–214.

Zack, Travis I., Steven E. Schumacher, Scott L. Carter, Andrew D. Cherniack, Gordon Saksena, Barbara Tabak, Michael S. Lawrence, Cheng-Zhong Zhang, Jeremiah Wala, Craig H. Mermel, Carrie Sougnez, Stacey B. Gabriel, Bryan Hernandez, Hui Shen, Peter W. Laird, Gad Getz, Matthew Meyerson, and Rameen Beroukhim (Oct. 2013). "Pan-cancer patterns of somatic copy number alteration". In: *Nature Genetics* 45.10, pp. 1134–1140.

Zaidi, Syed H. et al. (July 20, 2020). "Landscape of somatic single nucleotide variants and indels in colorectal cancer and impact on survival". In: *Nature Communications* 11.1, p. 3644.

Zakut-Houri, R, B Bienz-Tadmor, D Givol, and M Oren (May 1985). "Human p53 cellular tumor antigen: cDNA sequence and expression in COS cells." In: *The EMBO Journal* 4.5, pp. 1251–1255.

Zapata, Luis, Oriol Pich, Luis Serrano, Fyodor A. Kondrashov, Stephan Ossowski, and Martin H. Schaefer (May 31, 2018). "Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome". In: *Genome Biology* 19.1, p. 67.

Zeira, Ron and Benjamin J Raphael (July 1, 2020). "Copy number evolution with weighted aberrations in cancer". In: *Bioinformatics* 36 (Supplement_1), pp. i344–i352.

Zhang, B., B. Zhang, X. Chen, S. Bae, K. Singh, M. K. Washington, and P. K. Datta (Feb. 18, 2014a). "Loss of Smad4 in colorectal cancer induces resistance to 5-fluorouracil through activating Akt pathway". In: *British Journal of Cancer* 110.4, pp. 946–957.

Zhang, Bixiang, Sunil K. Halder, Nilesh D. Kashikar, Yong–Jig Cho, Arunima Datta, D. Lee Gorden, and Pran K. Datta (Mar. 1, 2010). "Antimetastatic Role of Smad4 Signaling in Colorectal Cancer". In: *Gastroenterology* 138.3, 969–980.e3.

Zhang, Jianjun, Junya Fujimoto, Jianhua Zhang, David C. Wedge, Xingzhi Song, Jiexin Zhang, Sahil Seth, Chi-Wan Chow, Yu Cao, Curtis Gumbs, Kathryn A. Gold, Neda Kalhor, Latasha Little, Harshad Mahadeshwar, Cesar Moran, Alexei Protopopov, Huandong Sun, Jiabin Tang, Xifeng Wu, Yuanqing Ye, William N. William, J. Jack Lee, John V. Heymach, Waun Ki Hong, Stephen Swisher, Ignacio I. Wistuba, and P. Andrew Futreal (Oct. 10, 2014b). "Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing". In: *Science (New York, N.Y.)* 346.6206, pp. 256–259.

Zhang, Jingsong, Jessica J. Cunningham, Joel S. Brown, and Robert A. Gatenby (Nov. 28, 2017). "Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer". In: *Nature Communications* 8.1, p. 1816.

Zhang, Junjun, Joachim Baran, A. Cros, Jonathan M. Guberman, Syed Haider, Jack Hsu, Yong Liang, Elena Rivkin, Jianxin Wang, Brett Whitty, Marie Wong-Erasmus, Long Yao, and Arek Kasprzyk (2011). "International Cancer Genome Consortium Data

Portal–a one-stop shop for cancer genomics data”. In: *Database: The Journal of Biological Databases and Curation* 2011, bar026.

Zhang, Yiqun, Lixing Yang, Melanie Kucherlapati, Fengju Chen, Angela Hadjipanayis, Angeliki Pantazi, Christopher A. Bristow, Eunjung A. Lee, Harshad S. Mahadeshwar, Jiabin Tang, Jianhua Zhang, Sahil Seth, Semin Lee, Xiaojia Ren, Xingzhi Song, Huandong Sun, Jonathan Seidman, Lovelace J. Luquette, Ruibin Xi, Lynda Chin, Alexei Protopopov, Wei Li, Peter J. Park, Raju Kucherlapati, and Chad J. Creighton (July 10, 2018). “A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases”. In: *Cell Reports* 24.2, pp. 515–527.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu (Sept. 17, 2008). “Model-based Analysis of ChIP-Seq (MACS)”. In: *Genome Biology* 9.9, R137.

Zou, Yang, Feng Wang, Fa-Ying Liu, Mei-Zhen Huang, Wei Li, Xiao-Qun Yuan, Ou-Ping Huang, and Ming He (Nov. 15, 2013). “RNF43 mutations are recurrent in Chinese patients with mucinous ovarian carcinoma but absent in other subtypes of ovarian cancer”. In: *Gene* 531.1, pp. 112–116.

Zuckerkandl, E. and L. Pauling (Mar. 1965). “Molecules as documents of evolutionary history”. In: *Journal of Theoretical Biology* 8.2, pp. 357–366.