Exploring tumour evolution through single-cell sequencing

Haixi Yan

Institute of Cancer Research

and

The Francis Crick Institute

PhD Supervisor: Peter Van Loo

A thesis submitted for the degree of

Doctor of Philosophy

Institute of Cancer Research

April 2023

Declaration

I, Haixi Yan 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 thesis.

Abstract

Tumours are composed of heterogeneous populations of cells which, under the pressure of the host and external treatments, evolve across time and space. Recent advances in next-generation sequencing technologies have allowed the profiling of cells across different modalities. These techniques have revealed insights into tumourigenesis and progression with previously unachievable degrees of resolution, spatiality, and throughput.

In this thesis, I perform in-depth profiling of a malignant peripheral nerve sheath tumour with a multi-omics approach. These different sequencing methods are then integrated together, enabling characterisation of the tumour through different lenses and mitigating the limitations of individual techniques. A detailed evolutionary history of this tumour is inferred down to the single-cell level, revealing that intra-tumour heterogeneity is predominantly driven by chromosomal instability. Using this extensive CNA heterogeneity, lineage tracing of tumour subclones was performed across space and used to reconstruct intricate growth paths.

In addition, I attempt to detect rare single disseminated tumour cells by applying single-cell sequencing techniques across different cancer stages. Multiple normal tissues across different cancer types were collected through research autopsies from patients with metastatic disease. Although a disseminated tumour cell belonging to a micrometastasis was profiled, single-cell sequencing could not be performed at scale due to poor tissue quality and the lack of a specific tumour marker. In the limited disease setting, bone marrow aspirates were collected from patients with clear cell renal cell carcinoma undergoing surgery. Several cells with an abnormal chromosome complement were detected, although they did not appear to be obvious disseminated tumour cells.

Overall, these results demonstrate compelling use cases for single-cell sequencing and the power of integrating multi-omics to reconstruct the development of a tumour in a spatio-temporal manner. These detailed evolutionary histories will be critical for understanding the mechanisms of tumourigenesis and developing more effective patient-specific anti-cancer therapies.

Acknowledgement

First, I would like to thank all the patients and families who contributed to this work. In particular, I am incredibly grateful to the seven patients who bravely donated their bone marrow for this project at what must be a very difficult time. I am also indebted to Cancer Research UK and all their fundraisers for their support which made this work possible.

Thank you to my supervisor Peter Van Loo for showing faith in me and giving me the opportunity to work in such a fantastic lab. Jonas Demeulemeester and Maxime Tarabichi have been incredible mentors during my time in the lab and have continuously inspired me to challenge myself. The learning curve was very steep, but I hope I have made you proud. Thank you also to Annelien Verfaillie, Cristina Cotobol Martin and Christie English who were instrumental in the wet lab. Toby, Carla and Nana, thank you for encouragement and making every day in the lab so enjoyable. I must also thank Sara for her incredible cakes which made all the difference during the tough times.

I am incredibly grateful to my secondary supervisor Samra Turajlic and for the help her incredible team have given me. The TRACERx Renal and PEACE studies are huge team efforts, and I would like to acknowledge everyone involved who have made these projects possible. Thank you also my thesis committee - Francesca Ciccarelli and Nicholas Turner- for your guidance.

To Sammy, thank you for your flow cytometry expertise and baking. Thank you also to my friends Jonny, Emily, Anthony, Matthew and Razwana for all for your support. Whether on court or over chat, you have all made this time so memorable.

Finally, I would like to thank my parents and my sister for supporting me all these years. Haihui, I am also looking forward to reading your thesis. To all of you, I dedicate this thesis.

Table of Contents

List of figures

List of tables

Abbreviations

ASCAT	Allele-Specific Copy number Analysis of Tumours
BAF	B-Allele Frequency
BMA	Bone Marrow Aspirate
CAIX	Carbonic Anhydrase 9
CCF	Cancer Cell Fraction
ccRCC	Clear Cell Renal Cell Cancer
CGH	Comparative Genome Hybridization
CIN	Chromosomal Instability
CNA	Copy Number Aberrations
CTC	Circulating Tumour Cell
DLP	Direct Library Preparation
DTC	Disseminated Tumour Cell
EpCAM	Epithelial Cell Adhesion Molecule
FACS	Fluorescence-Activated Cell Sorting
FISH	Fluorescent In Situ Hybridization
GRN	Gene Regulatory Network
G&T-seq	Genome and Transcriptome Sequencing
ICC	Immunocytochemistry
IHC	Immunohistochemistry
ITH	Intra-Tumour Heterogeneity
KNN	K-Nearest Neighbour
LCM	Laser Capture Microdissection
LOH	Loss Of Heterozygosity
MCSP	Melanoma-associated Chondroitin Sulphate Proteoglycan
MPNST	Malignant Peripheral Nerve Sheath Tumour
MRCA	Most Recent Common Ancestor
NF1	Neurofibromatosis type 1
NGS	Next-Generation Sequencing
NMF	Non-negative Matrix Factorisation
PAX8	Paired Box Gene 8
PCF	Piecewise Constant Fitting
PCR	Polymerase Chain Reaction

PEACE	Posthumous Evaluation of Advanced Cancer Environment
PRC2	Polycomb Repressive Complex 2
RT	Room Temperature
scDNA-seq	single-cell DNA sequencing
scRNA-seq	single-cell RNA sequencing
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
STP	Science Technology Platform
t-SNE	t-distributed Stochastic Neighbour Embedding
TME	Tumour Microenvironment
TRACERx Renal	TRAcking Renal Cell Cancer Evolution Through Therapy (Rx)
UMAP	Uniform Manifold Approximation And Projection
UMI	Unique Molecular Index
VAF	Variant Allele Frequency
WGA	Whole-Genome Amplification
WGD	Whole-Genome Doubling
WGS	Whole-Genome Sequencing

Chapter 1. Introduction

1.1 Carcinogenesis

Cancers arise from the uncontrolled clonal proliferation of abnormal cells. Normal cells must acquire certain key traits to transform into cancer cells. Ten such features have been described in the Hallmarks of Cancer by Hanahan and Weinberg with cancers acquiring many but not necessarily all of these traits (Hanahan & Weinberg, 2011). This is predominantly achieved through genetic changes that disrupt the normal function of genes resulting in cancer.

1.1.1 Cancer as a genetic disease

The earliest theory that cancer is a genetic disease came from David von Hansemann and Theodor Boveri over a hundred years ago (Hansemann, 1890). Hansemann first reported an association between an abnormal number of chromosomal (aneuploidy) and cancer and suggested this as a mechanism of tumour formation although he did not consider aetiologies which would cause aneuploidy. Subsequently, Boveri formalised this hypothesis that cancerous cells result from the scrambling of chromosomes from his studies of sea urchins, in which he found that all chromosomes are necessary for proper embryonic development (Boveri, 1914, 2008). As microscopy techniques and our knowledge of chromosomes improved, aberrant chromosomes were identified in specific subtypes of leukaemia such as the Philadelphia chromosome (t9;22) in chronic myeloid leukaemia (Rowley, 1973). The chromosomal translocation results in the fusion of the *BCR* and *ABL1* genes, resulting in a constitutively active tyrosine kinase. This results in uncontrolled cell growth and division, fulfilling one of the hallmarks of cancer.

Further characterisation of chromosomal aberrations became possible with the development of fluorescent in situ hybridization (FISH) and microarray-based comparative genome hybridization (CGH). Array CGH involves labelling tumour and normal DNA from an individual patient with different fluorophores. The mixed, labelled DNAs are then hybridized onto microarrays laid with small fragments of DNA belonging to different loci followed by laser-based scanning. By comparing signal intensities, chromosomal gains, losses, and other structural changes can be identified at much higher resolutions.

Following the discovery of DNA, it was shown that chemicals which alter DNA also cause cancer, thus, postulating a molecular genetic basis of cancer (Loeb & Harris, 2008). The ground-breaking discovery of *HRAS* as the first naturally occurring somatic (occurring after fertilisation) mutation causing cancer has led to a search for additional cancer genes (Reddy *et al.*, 1982; Tabin *et al.*, 1982). Techniques such as polymerase chain reaction (PCR) and Sanger sequencing were initially used to identify mutations in suspected cancer genes (Davies *et al.*, 2002; Samuels *et al.*, 2004). Some early candidate cancer genes were often chosen based on linkage studies from cancer-prone families. Discovery of mutations in the germline provided further evidence for the genetic basis of cancer. These mutations are inherited from parent to child resulting in multiple family members being affected by cancer, often at an early age (Brown *et al.*, 2020). An example of one such inherited cancer syndrome is Li-Fraumeni syndrome where germline loss of the tumour suppressor *TP53* commonly leads to soft tissue sarcomas, breast cancer, central nervous system tumours, and adrenocortical carcinoma (McBride *et al.*, 2014).

As the list of known cancer genes has expanded, so too have the mechanisms by which these mutations cause cancer been illuminated. Normal genes which are involved in cell growth and division are termed “proto-oncogenes”, which when mutated or upregulated are referred to as “oncogenes” due to their contribution to carcinogenesis. For example, the *MYC* oncogene is activated in most cancers and confers a multitude of hallmarks of cancer e.g. uncontrolled proliferation, genomic instability, angiogenesis and immune evasion (Dhanasekaran *et al.*, 2022). In addition, another class of cancer genes was identified – “tumour suppressors”. Loss of these genes, whether physical or functional due to inactivation, resulted in a lack of inhibition of cell division or other critical inhibitors of tumour development. Unlike oncogenes, which mostly act in a dominant fashion, both copies of tumour suppressors must be lost or inactivated. Knudson's studies on retinoblastoma led to the discovery of the tumour suppressor *RB1* and the concept of loss of heterozygosity (LOH) where one of two alleles is lost. This led to the two hit-hypothesis which proposed that two inactivating events are required, with the first occurring in the inherited germline allele and the second due to a somatic deletion (Knudson, 1971).

Furthermore, as most cancer types tend to contain multiple mutated genes, studies were undertaken to ascertain whether there is a conserved ordering of events. Vogelstein's multi-step model posited that cancer was driven by sequential mutations in oncogenes and tumour suppressor genes (Vogelstein *et al.*, 1988). Based on frequencies of shared cancer genes across a cohort of colorectal cancer at different stages, a model with ordered events was proposed. Inactivation of *APC* occurs as the first event, followed by oncogenic *KRAS* mutations in the adenomatous stage, deletion of chromosome 18q, and eventually, inactivation of *TP53* on chromosome 17p during the transition to malignancy. Following the Human Genome Project, studies could now be performed to systematically identify all cancer genes. Large-scale discovery studies today allow the cataloguing of virtually every mutation across thousands of individual tumours. Although certain genes are recurrently mutated in certain tumour types such as *RB1* in retinoblastoma, most common cancers exhibit combinations of recurrent but low-prevalence cancer genes (The Cancer Genome Atlas Research Network, 2011; Stephens *et al.*, 2012; The Cancer Genome Atlas Network, 2012).

1.1.2 Classes of mutations

The application of next-generation sequencing (NGS) to large cohort cancer genome studies has transformed our understanding of cancer development and progression (Stratton *et al.*, 2009; Vogelstein *et al.*, 2013). This has allowed in-depth characterisation of DNA mutations, copy number changes, and structural variations (Alexandrov *et al.*, 2020; Li *et al.*, 2020; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Cancer Consortium, 2020). The results revealed a complex landscape of somatic genomic changes, showing that each cancer type has a distinct genetic signature and that different subtypes of a cancer type can have different genomic changes. These variants can be divided into the following types:

Single nucleotide variants (SNVs) are mutations in a single nucleotide. These mutations can result in a change in amino acid (non-synonymous) or not (synonymous). It can also result in a stop codon generating a non-functional truncated protein (nonsense). *Insertions or deletions* (indels) are the addition or deletion of one or more nucleotides to the DNA sequence. If the length of the indel is not a multiple of three, this can result in a frameshift, dramatically changing the translated protein and typically introducing an early stop codon.

Larger rearrangements of the genome (typically larger than 50 base pairs) or *structural variants* (SVs) include insertions, deletions, translocations, inversions, duplications, or a combination of these. Aneuploidy or an abnormal complement of chromosomes is a key feature of cancer. These include gains or losses of whole or parts of chromosomes or *copy number aberrations* (CNAs). Furthermore, when the maternal and paternal haplotypes are known, cases of copy-number neutral LOH can be identified where there is a loss of one allele but without a corresponding change in the overall copy number of that region. The two parental haplotypes typically can be identified through phasing (identification of alleles co-located on the same parental chromosome) of heterozygous single nucleotide polymorphisms (SNPs) using haplotypes derived from population studies.

Furthermore, the plethora of mutations identified through NGS studies have allowed the identification of different patterns of SNVs. The most studied are SNVs and their trinucleotide sequence context which have been linked to specific mutational processes (Nik-Zainal *et al.*, 2012a). Large pan-cancer cohorts have enabled the extraction of mutational signatures using non-negative matrix factorisation (NMF) (Alexandrov *et al.*, 2013, 2020; Degasperi *et al.*, 2022). Individual mutational signatures can be present in many cancer types whilst others are restricted to just one, revealing potential endogenous and environmental sources of mutagenesis (Kucab *et al.*, 2019). It is also possible to profile which mutational signatures (and mutagenic processes driving tumour evolution) were active in that period by identifying SNVs which occurred between two time points in a tumour's development, e.g. processes developed after primary treatment and active following relapse (Yates *et al.*, 2017).

Finally, epigenetic modifications are reversible alterations in the DNA or its associated proteins that influence gene expression without altering the underlying genetic sequence. Studies have shown modifications in every aspect of epigenetic mechanisms in cancer including DNA methylation, histone modifications, and noncoding RNA transcription and microRNA expression (Sharma *et al.*, 2010). For example, global DNA hypomethylation removes silencing of transposons and promotes genomic instability (another hallmark of cancer) whereas local hypermethylation can silence tumour suppressor genes (Jones & Baylin, 2002; Howard *et al.*, 2008). The chromatin accessibility machinery can also be hijacked by cancer to silence tumour suppressors through histone modifications such as H3K27me3 by the histone methyltransferase EZH2 (Das & Taube, 2020). These abnormal changes in the epigenome result in the dysregulation of transcriptional programmes, allowing cancer cells to adopt different phenotypes from the same genotype. Flexible switching between phenotypes enables cellular plasticity and this adaptability is important in the context of cancer evolution.

1.2 Cancer evolution

1.2.1 Intra-tumour heterogeneity and selection

A critical feature of cancer making eradication difficult is its capacity to continuously evolve (Nowell, 1976; Yates & Campbell, 2012). As with other evolutionary processes, cancer progression requires variation and selection.

Peter Nowell's seminal paper proposed that when a mutation in a single cell causes it to proliferate, this cell can accumulate further mutations resulting in successive clonal expansions and progressively more malignant cells (Nowell, 1976). Occasionally, a selective sweep occurs and clones which are outcompeted disappear. If this sweep is complete, then all surviving tumour cells have descended from this one cell which is the most recent common ancestor (MRCA). Later clonal expansions or genetic drift can then generate descendant subclones which can be identified by the mutations that they bear. Recent large-scale studies have confirmed that this variation is common across cancer types and that tumours are composed of a patchwork of heterogeneous subclones (Dentro *et al.*, 2021).

This remarkable degree of intra-tumour heterogeneity (ITH) has been revealed through an array of alterations including SNVs, CNAs and transcriptomic changes which differentiate subclones (McGranahan & Swanton, 2017). Chromosomal instability (CIN) is a key mechanism by which tumours continuously generate genetically distinct subclones through changes to the number and structure of chromosomes. CIN allows for large-scale changes to the genome and has been linked to poor prognosis across tumour types (Carter *et al.*, 2006; Birkbak *et al.*, 2011; Jamal-Hanjani *et al.*, 2017). Further epigenetic changes can result in additional diversification with cancer cells from the same subclone able to exist in different cell states. The importance of ITH has been demonstrated in key cancer processes such as immune evasion, metastasis and acquiring resistance against treatments (Caswell & Swanton, 2017).

Natural selection can then act upon this repertoire of subclones, resulting in clones with a selective advantage expanding more. Clones which are outcompeted will disappear and as the tumour microenvironment (TME) changes, so too will the selective criteria. Mutations that confer a selective advantage on a cell are defined as driver mutations with over 500 reported (Bailey *et al.*, 2018; Sondka *et al.*, 2018; Martínez-Jiménez *et al.*, 2020). The remaining mutations, which make up the majority, are considered passengers which do not provide a selective advantage. Treatment can also introduce an additional source of “artificial” selection in the form of drugs or radiotherapy. In addition, some chemotherapeutic agents can induce additional mutations in tumour cells which generates further diversity and sources of resistance.

1.2.2 Tumour phylogenetic reconstruction

ITH can be resolved by identifying subclones. One key feature of mutations is that all descendants of a cell which develops a mutation will inherit that mutation except in cases where there is a deletion of that locus, and the mutation is lost. Therefore, whilst passenger mutations do not benefit the tumour, these scars in the genome can be used to identify individual subclones and infer their origin and relative relationships. Thus, mutations that occurred before the MRCA are carried by all tumour cells and can be used as markers of the clonal population. Mutations occurring after the MRCA will not be present in all cells, therefore, identify specific subclones.

The method of sampling deployed by the study has important implications for the phylogenetic reconstruction of tumours (Nam *et al.*, 2021). Single-region bulk sampling is the most commonly used method due to its simplicity. However, the offers limited resolution and may give the illusion of clonality when all cells from a small sample are from a subclone but the much larger tumour contains additional subclones. Multi-region sampling or multiple samples from the same tumour can increase the sensitivity of subclone detection and help resolve some cases of ambiguity by providing the frequencies of subclones shared between different regions (Gerlinger *et al.*, 2012; Yates *et al.*, 2015). Despite this, the tissue undergoing analysis is still a tiny fraction of the overall tumour which methods such as representative sequencing that samples the entire tumour aims to resolve (Litchfield *et al.*, 2020). Ultimately, for full resolution of the tumour phylogenetic tree, single-cell sequencing is required which will be discussed later.

A variety of methods exist for inferring tumour phylogenies from bulk samples, with most relying on SNV data (Schwartz & Schäffer, 2017). Subclonal reconstruction involves identifying the somatic mutations, relative proportion, and ancestral relationships of major populations of cells in a tumour (Tarabichi *et al.*, 2021). Most methods use the variant allele frequency (VAF) of SNVs to infer the fraction of all cells with the SNV present or cellular prevalence. As tumour samples often contain admixed TME cells, the sample purity can be used to estimate the cancer cell fraction (CCF) or fraction of *tumour* cells harbouring the SNV. Subclones can then be identified by clustering together SNVs based on their CCFs, as SNVs belonging to the same subclone will have a similar CCF. It is important to note that CNAs can also dramatically affect VAFs, therefore the number of DNA copies the SNV is present on, or the multiplicity must be corrected for. CNAs can be inferred using the relative read depth between tumour and normal samples (logR). The ratio of alleles at heterozygous SNP positions or B-allele frequency (BAF) can then inform the allelic composition of the number of copies at that locus. Methods such as Allele-Specific Copy number Analysis of Tumours (ASCAT) jointly estimate purity and ploidy using the logR and BAF to derive allele-specific CNAs (Van Loo *et al.*, 2010). By reviewing the CCFs of subclones, evolutionary relationships between subclonal populations can be inferred (Dentro *et al.*, 2017). This involves application of the “pigeonhole principle”, where if the sum CCF of two subclones exceeds that of their parent clone, then the smaller subclone must be descended from the larger subclone.

Alternatively, CNAs can also be used to infer the phylogenies of subclones. This relies on defining each copy number change of a chromosomal segment as an event and creates a tree which minimises the total number of events between subclones (Letouzé *et al.*, 2010; Kaufmann *et al.*, 2022). These methods are useful where there are multiple samples for a patient or where SNV information is limited such as in single-cell studies.

1.2.3 Models of tumour evolution

Different models of tumour evolution have been hypothesised to explain how ITH arises during tumour progression (Davis *et al.*, 2017; Turajlic *et al.*, 2019). The four main models of tumour evolution include linear, branched, neutral and punctuated evolution (Figure 1.1) (Vendramin *et al.*, 2021). With the development of methods to infer the evolutionary history of tumours at the subclonal level, different models of tumour evolution can now be distinguished in individual tumours.

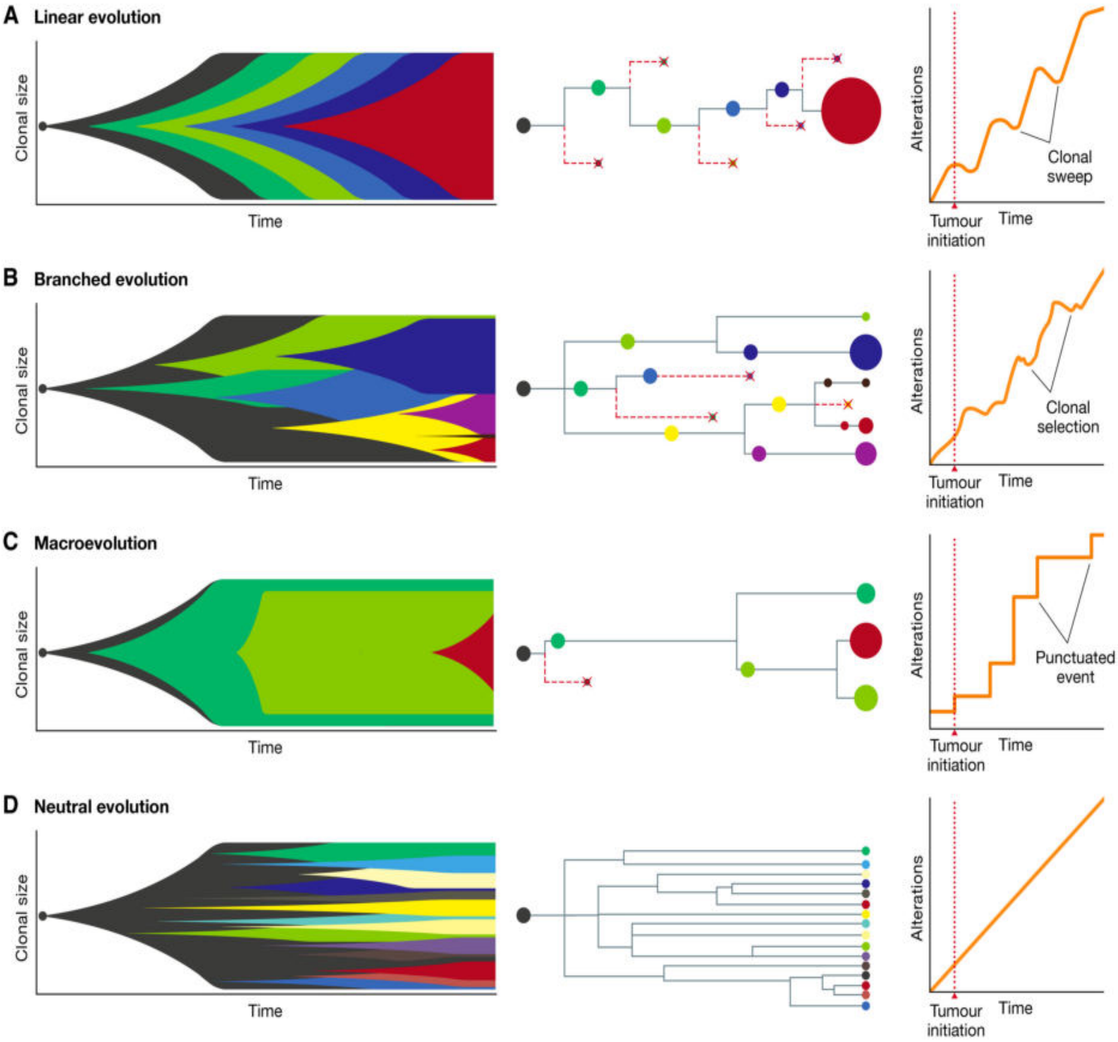


Figure 1.1 Models of tumour evolution. Models of linear evolution (A), branched evolution (B), macroevolution or punctuated evolution (C) and neutral evolution (D) showing changes over time in the size of clones (left), phylogenetic trees of clones (centre) and the number of alterations (right). Figure reproduced from Vendramin *et al*., 2021, licensed under CC BY 4.0.

Linear evolution proposes that each new driver provides that cell such a selective advantage that it completely outcompetes its ancestor resulting in a clonal sweep and only one surviving clone. However, inadequate sampling or lack of resolution can potentially confound observations of linear evolution. Branching or divergent evolution refers to the emergence of two or more distinct subclones which have evolved in parallel from the MRCA. These subclones have individual selective advantages or are influenced by local spatial features resulting in infrequent complete sweeps. A plethora of studies have described branching evolution in human cancers with variations in the trunk length and the number of branches. Neutral evolution occurs in the absence of selection pressures resulting in many different subclones with equal fitness and is powered by genetic drift. This implies that although ITH is a feature of tumours, it has no beneficial impact in driving tumour growth. The degree to which neutral evolution occurs has been a contentious topic (Williams *et al.*, 2016; Tarabichi *et al.*, 2018). Recently, evidence for punctuated evolution has also emerged. This involves significant changes to the genome in a short period of time resulting in bursts of adaptation rather than gradual evolution. Examples include “firestorms” or multiple gains closely located on one chromosome arm and *chromothripsis*, which refers to a cluster of complex rearrangements and oscillating copy number changes located on a single chromosome or chromosome arm (Hicks *et al.*, 2006; Korbel & Campbell, 2013). In addition, a “Big Bang” model has been proposed, in which a burst of CNAs and SVs occurring rapidly produces many subclones. Following selection, a few of the surviving subclones then expand in a stable fashion. This typically involves extreme CIN and has been described in breast and colorectal cancer (Sottoriva *et al.*, 2015; Gao *et al.*, 2016).

Finally, spatial or physical constraints can also influence the mode of tumour evolution and may explain why the variation in patterns between different tumour classes (Noble *et al.*, 2021). For example, haematological malignancies can grow without spatial constraints due to mixing of tumour cells in the bloodstream, whereas colorectal cancer grows through a fission process of colonic crypts (Clapp & Levy, 2015; Sun *et al.*, 2017). Meanwhile, boundary-driven or surface growth refers to when only cells close to the border of the tumour can proliferate but those in the core of the tumour lack sufficient space to do so (Rodriguez-Brenes *et al.*, 2013; Chkhaidze *et al.*, 2019). These growth environments can all determine the pattern of cancer evolution in a tissue.

1.3 ‘Omics’ approaches to study tumour evolution

Multiple features of tumours must be profiled to identify alterations acquired during cancer evolution and fully understand their functional effects. These “Omics” approaches have traditionally focused on comprehensively characterising the genomes, transcriptomes, and epigenomes of cells. However, the importance of the proteome, metabolome, and microbiome has also been shown and are now increasingly profiled. The different layers of information from these techniques must then be integrated to reveal causative changes in cancer. The following sections will discuss single-cell and spatial methods, focusing primarily on genome and transcriptome profiling, and strategies for data integration.

1.3.1 Single-cell transcriptome profiling

Measuring gene expression profiles allows us to detail phenotypic changes in a cell resulting from mutations in cancer. Conventionally, RNA sequencing has been used to derive bulk expression profiles. This relies on reverse transcription of RNA into cDNA which allows for quantification of each gene transcript and identification of differentially expressed genes between samples or conditions. However, the variation in the cell types comprising a tumour and their prevalence can confound analysis. Although methods exist to deconvolve bulk expression, these are not always reliable and cannot fully uncover the complexity of all cell types and their cell states (Newman *et al.*, 2019; Menden *et al.*, 2020; Chu *et al.*, 2022). Therefore, single-cell technologies have been developed which enable the study of interactions between the tumour and the TME or between individual cancer cells at high resolution (Potter, 2018).

Following the first manual single-cell RNA sequencing (scRNA-seq) study, a variety of methods to isolate and profile individual cells have come to fruition (Tang *et al.*, 2009). Plate-based methods like Cel-seq and SMART-seq2 utilise microwells for each cell to increase the throughput of cells profiled (Hashimshony *et al.*, 2012; Picelli *et al.*, 2013). Microdroplet-based methods such as Drop-seq and Chromium can also profile thousands or even millions of cells by using microfluidics to generate thousands of drops (Macosko *et al.*, 2015). Each drop contains a single bead with a unique barcode which is incorporated into the cDNA after a cell is introduced into the drop. This technology was commercialised by 10X Genomics and has seen widespread adoption. Meanwhile, *in situ* barcoding methods such as SPLiT-seq use the cells themselves as vessels in which to perform combinatorial barcoding through successive rounds of splitting and ligation (Rosenberg *et al.*, 2018). This method has major cost and logistical advantages as it requires samples to be permeabilised prior to barcoding to allow the barcode to enter the cell, therefore samples can be fixed and stored separately before undergoing barcoding together. The coverage between methods also differs as Chromium only captures the 3' or 5' ends of RNA molecules whereas SMART-seq2 provides full-length transcripts, allowing detection of isoforms and mutations throughout the coding region of a gene. However, Chromium makes use of unique molecular indices (UMIs) which enable detection of duplicates, giving a direct readout of gene expression (Islam *et al.*, 2014).

As the popularity of single-cell studies increased, whole ecosystems of analysis methods have been developed to analyse these vast datasets. Although some existing methods for analysing RNA-seq such as gene set enrichment analysis have been extended to scRNA-seq, dedicated scRNA-seq methods have been required (Borcherding *et al.*, 2021; Wu *et al.*, 2021b). scRNA-seq data is typically sparse due to measuring all genes in every single cell. Unsupervised graph-based clustering methods such as k-nearest neighbour (KNN) followed by community detection with techniques such as the Louvain algorithm have been used to group similar cells together and have been shown to outperform *k*-means clustering (Stuart *et al.*, 2019; Duò *et al.*, 2020). This enables reliable measurement of each cell type by pooling information from similar cells together and mitigates against dropout (expressed genes whose transcripts are not captured). Dimension reduction techniques such as t-distributed stochastic neighbour embedding (t-SNE) and uniform manifold approximation and projection (UMAP) can then be applied to visualise cells and their clusters in 2D or 3D (Luecken & Theis, 2019). Clusters can then be annotated manually based on known marker genes or using automated annotation algorithms (Abdelaal *et al.*, 2019; Pliner *et al.*, 2019; Zhang *et al.*, 2019). Atlas projects have profiled all cells from most tissues to greatly increase our collection of cell type markers and can also be used as a reference for querying cell types in new datasets (Regev *et al.*, 2017; Jones *et al.*, 2022). In addition, computational methods can mitigate some of the shortfalls of droplet-based technologies. For example, contaminant ambient cell-free RNA captured in droplets can be removed in silico using SoupX and doublets can be identified using methods such as demuxlet and DoubletFinder (Kang *et al.*, 2018; McGinnis *et al.*, 2019).

Several studies have identified cell states exhibited by cancer cells in different tumour types such as glioblastoma and breast cancer (Neftel *et al.*, 2019; Wu *et al.*, 2021a). These can include states related to hypoxia, stress response or dormancy, thereby contributing to ITH and increasing the resilience of tumour cells. These states are coordinated by transcription factors that activate a network of genes. Whilst differences in gene expression between clusters of cells have been relatively simple to detect, gene regulatory networks are more complex and require more complex algorithms to profile (Aibar *et al.*, 2017). Recently, NMF has been used to extract recurrent cancer cell states across tumour types and depicts the cellular mechanisms at work in cancer cells (Barkley *et al.*, 2022). Further profiling of these cell states in response to treatment may give insights into how tumour cells acquire resistance mechanisms.

Additional biological features have also been inferred using computational algorithms. Trajectory analysis methods have contributed to capturing transitions between cell types; revealing developmental paths of subpopulations of cells in pseudotime (a proxy for developmental time) (Trapnell *et al.*, 2014; Saelens *et al.*, 2019). Meanwhile, CNAs can also be inferred from transcriptomic data by exploiting the rough correlation between the number of copies of a DNA locus and the degree to which genes on this locus are expressed (Tickle *et al.*, 2019; Gao *et al.*, 2021). This has been used to differentiate malignant cells from normal cells and has recently been extended to allele-specific CNAs using population-based phasing (Gao *et al.*, 2023).

Adaptations for scRNA-seq have also been developed with V(D)J T/B cell receptor sequencing for T or B cells or barcoded antibodies with cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) for concurrent epitope detection with antibody-oligo conjugates (Kim *et al.*, 2012; Stoeckius *et al.*, 2017). Furthermore, epigenetic single-cell profiling techniques have also emerged. Chromatin accessibility can be measured using single-cell assays for transposase-accessible chromatin (scATAC-seq), histone modifications can be profiled using single-cell chromatin immunoprecipitation (ChIP-seq) and Cleavage Under Targets and Tagmentation (CUT&Tag), and DNA methylation can be detected using single-cell reduced representation bisulfite sequencing (scRRBS-seq) (Guo *et al.*, 2013; Buenrostro *et al.*, 2015; Rotem *et al.*, 2015; Kaya-Okur *et al.*, 2019).

1.3.2 Single-cell genome profiling

When applied to genomics, single-cell techniques enable the study of cancer genomes at the cellular level – the atomic unit of cancer evolution (Gawad *et al.*, 2016; Bowes *et al.*, 2022). Just as in scRNA-seq, single cells can be isolated with microwell plate or microdroplet-based methods. Where a particular cell type is of interest, samples can be enriched with antibodies against specific markers, followed by fluorescence-activated cell sorting (FACS). Alternative methods for isolating cells of interest include manual micromanipulation with capillaries if the required number of cells is low. In addition, techniques such as laser capture microdissection (LCM) where individual cells are cut out and dropped or catapulted into collection tubes have been developed (Casasent *et al.*, 2018). LCM may be desirable if the tissue type or cell size is incompatible with other methods.

A key challenge in singe-cell DNA sequencing (scDNA-seq) is that genetic information must be derived from the minuscule 6pg of genomic DNA in a cell. Therefore, this single copy of the genome is typically first amplified before library preparation resulting in technical artefacts such as bias in coverage, allelic imbalance (an allele being preferentially amplified) and allelic dropouts (loss of one allele during amplification). Several whole-genome amplification (WGA) methods have been developed to achieve this. Degenerate oligonucleotide primer (DOP) PCR is based on PCR amplification of randomly fragmented genomic DNA with universal primers. However, it suffers from higher dropout and amplification errors due to the use of a thermostable polymerase. Multiple displacement amplification (MDA) uses random priming of the high-fidelity Φ29 DNA polymerase to perform non-PCR amplification but suffers from non-uniform coverage affecting detection of CNAs (Ning *et al.*, 2015). These shortcomings have been improved with hybrid methods such as multiple annealing and looping-based amplification cycles (MALBAC) or Picoplex (Zong *et al.*, 2012; Voet *et al.*, 2013).

Recently, a direct library preparation (DLP) method was developed which utilises a Tn5 transposase to simultaneously fragment DNA and ligate adapter oligonucleotides before PCR incorporation of cell barcodes and sequencing adapters (Zahn *et al.*, 2017). Due to each sequencing insert being unique, duplicates introduced during PCR can be computationally removed, removing biases in coverage. When applied in nanowell arrays and combined with robotics, thousands of cells can be sequenced with matched cell images (Laks *et al.*, 2019). Acoustic cell tagmentation (ACT) is a comparable Tn5 tagmentation technology which utilises robotic acoustic liquid transfer to dispense unique barcodes and minimise reagent volumes (Minussi *et al.*, 2021). The development of these technologies has enabled profiling of hundreds or thousands of cells from individual tumours.

Single-cell genomics allows two key genetic alterations to be detected: SNVs and CNAs. SNVs calling can be difficult due to the sparse coverage and errors introduced during WGA. Although this has led to the development of SNV callers for scDNA-seq data, they have not seen widespread adoption due to higher rates of false positives (Wang *et al.*, 2014; Dong *et al.*, 2017; Singer *et al.*, 2018). To mitigate this, mutations confidently identified in paired bulk data can be genotyped in scDNA-seq data. In contrast, CNA profiling of single cells is more mature. This is because the principles of bulk copy-number calling can be applied to single-cells with sequencing reads counted for each genomic “bin” used for logR calculation. Therefore, some bulk methods such as HMMcopy have been applied to single-cell data (Ha *et al.*, 2012). As the field has grown, bespoke single-cell methods are now available with CHISEL facilitating allele-specific calling (Mallory *et al.*, 2020; Zaccaria & Raphael, 2021). Where this is sufficient variation in CNAs between cells, clustering approaches can be applied to cluster cells with similar CNA profiles together to identify subclones as shown in the first scDNA-seq study by Navin et al. (Navin *et al.*, 2011). Since then, a larger-scale study has shown that breast cancer subclones – identified through single-cell CNAs – maintain their diversity, highlighting ongoing copy-number evolution (Minussi *et al.*, 2021). Furthermore, allele-specific CNA-based evolutionary inference methods have been extended to create single-cell resolution phylogenetic trees for these cells (Kaufmann *et al.*, 2022). As further scDNA-seq studies are carried out, more detailed evolutionary histories of tumours will be revealed.

1.3.3 Spatial profiling technologies

Whilst single-cell methods provide exquisite resolution, by dissociating cells, their spatial context and the subcellular location of molecules within them are inherently lost. This spatial information may provide critical information about tumour TME interactions and their local neighbourhood. LCM was one of the first spatial profiling technologies developed and has used to isolate small but pure populations of cells of interest based on histology (Espina *et al.*, 2006). LCM-captured cells can then be used for a variety of downstream applications such as whole-genome sequencing (WGS), single-cell DNA or RNA sequencing demonstrating its flexibility. Nevertheless, LCM can be technically challenging, labour-intensive, and susceptible to bias due to its low throughput.

Recently, a multitude of spatial technologies has been developed to systematically profile tissues, whilst retaining spatial information(Moffitt *et al.*, 2022). Although this field is developing rapidly, major limitations still exist and there are trade-offs between resolution, scale, detection efficiency, sample compatibility and ability to multiplex. Image-based profiling techniques of the transcriptome or proteome using fluorescence such as multiplexed error-robust fluorescence *in situ* hybridization (MERFISH) or imaging mass cytometry (IMC) have been developed and applied to cancer samples (Giesen *et al.*, 2014; Moffitt *et al.*, 2018; Ali *et al.*, 2020; Schürch *et al.*, 2020).

Spatial indexing methods coupled with sequencing have also rapidly advanced with single-cell and even subcellular resolutions now possible (Vickovic *et al.*, 2019). Spatial transcriptomics utilises spots with unique barcode sequences to map captured RNA back to spatial locations. This technology has been subsequently developed into the commercial Visium platform by 10X Genomics. When applied to cancer datasets, spatial transcriptomics has enabled inference of CNAs based on gene expression of specific regions in prostate cancer and histologically normal tissues (Erickson *et al.*, 2022). Other methods such as Slide-seq (which uses beads instead of spots) and deterministic barcoding in tissue for spatial omics sequencing (DBiT-seq) have also been developed, with the advantage of greater resolution or compatibility with fixed tissue (Rodriques *et al.*, 2019; Liu *et al.*, 2020).

One key difference between single-cell and spatial indexing methods is that spots or beads will not perfectly align with individual cells, resulting in a mix of multiple cells or part of a single-cell, depending on the resolution. This leads to mixing of transcripts from cells, causing difficulties in cell-type identification, and limits spatial resolution. To overcome this limitation, scRNA-seq data can be integrated with spatial data to assign cells to spatial locations (Stuart *et al.*, 2019; Biancalani *et al.*, 2021).

Furthermore, methods like RCTD and Cell2location have enabled the deconvolution of individual spots to estimate the relative abundance of each cell type (Cable *et al.*, 2022; Kleshchevnikov *et al.*, 2022).

As with single-cell technologies, other omics modalities can be profiled, such as the profiling of chromatin accessibility with Spatial-ATAC-seq, and epigenomic profiling of histone modifications using Spatial-CUT&Tag (Deng *et al.*, 2022). Finally, spatial DNA profiling methods such as Slide-DNA-seq and base-specific *in situ* sequencing (BaSISS) have given us an enticing glimpse at spatial relationships between specific subclones and detailed growth patterns (Lomakin *et al.*, 2022; Zhao *et al.*, 2022). Indeed, these methods may prove the most useful for studying how tumours grow, given they can directly detect SNVs and CNAs.

1.3.4 Multi-omic data integration

Multimodal technologies (measuring different molecular features in the same cell in parallel) such as genome and transcriptome sequencing (G&T-seq), single-cell nucleosome, methylation and transcription sequencing (scNMT-seq), scTrio-seq and others have been recently developed (Macaulay *et al.*, 2015; Hou *et al.*, 2016; Clark *et al.*, 2018; Nam *et al.*, 2019; Yu *et al.*, 2023). However, most studies to date only profile one dimension or independently profile data from more than one dimension due to considerations regarding cost or throughput. Therefore, a key challenge is to develop methods which can integrate data from across different 'omics' layers (Argelaguet *et al.*, 2021). This allows for a more comprehensive understanding of the biology of individual cells by offering orthogonal validation, as no single technology can fully portray the complexity of cellular mechanisms.

A key concept in integration involves the use of anchors to connect data between modalities and placing cells in a shared representation space based on these different modalities. Single-cell data integration methods can be classified based on whether anchors are used and how they are used. These strategies for integration include horizontal, vertical, and diagonal integration (Figure 1.2).

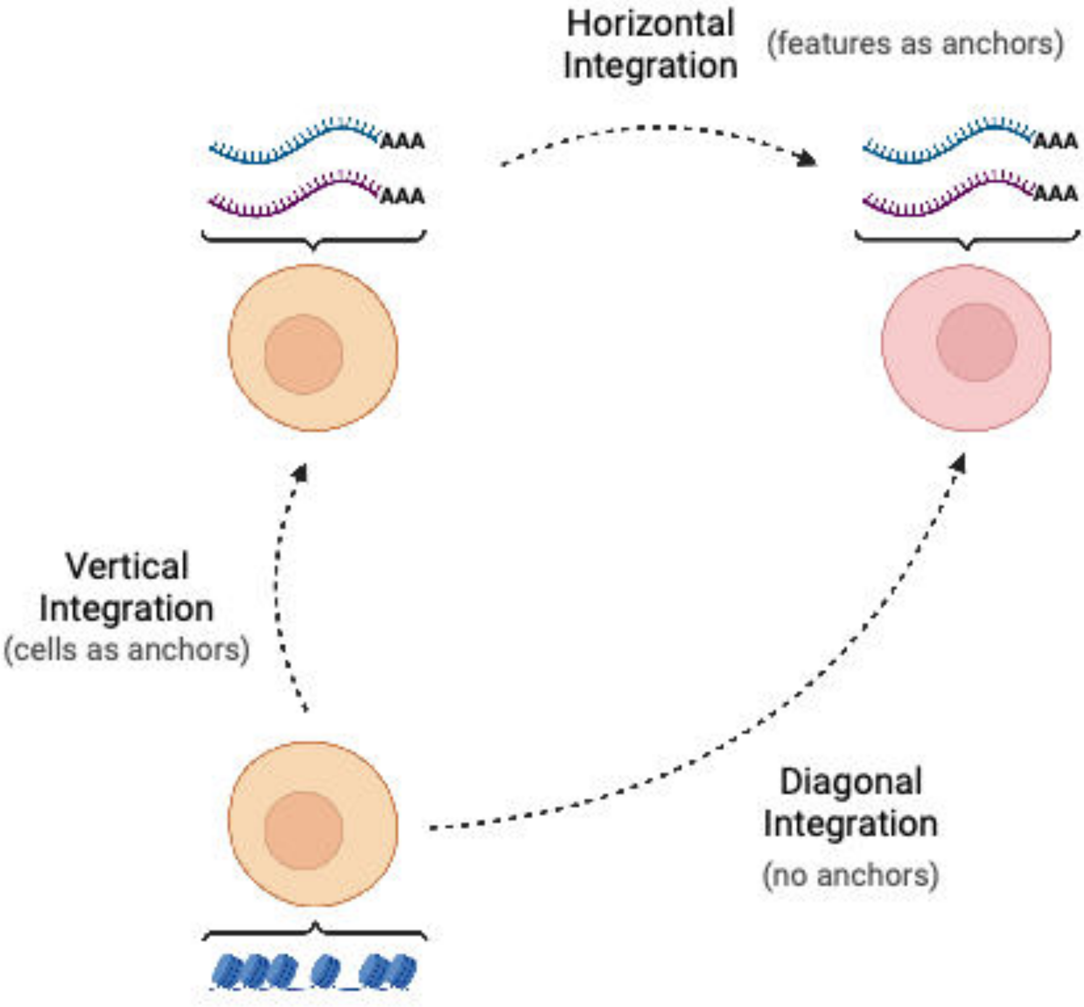


Figure 1.2 Strategies for data integration. Example of horizontal, vertical, and diagonal integration between gene expression and chromatin accessibility in singe cells. Created with Biorender.com.

Horizontal integration refers to the use of genomic features as anchors when the same feature is profiled across datasets, such as in studies where the same technology is used on different samples. Techniques based on mutual nearest neighbours (MNN) such as Seurat v3 have been used to integrate or batch correct these single-cell datasets (Haghverdi *et al.*, 2018; Stuart *et al.*, 2019). MNN identifies pairs of cells from each modality which are amongst each other's set of nearest neighbours in a shared low-dimensional space (thus using the same genomic features) and uses these anchors to identify matching cell types. Batch correction vectors can then be calculated and applied to integrate the datasets together, although a key challenge is distinguishing batch effects from true biological differences.

Meanwhile, vertical integration uses shared cells as anchors which have undergone profiling across multiple data modalities. This involves profiling the same cell using multi-omic techniques, thereby ensuring there is no ambiguity in the origin of each multimodal profile. Nearest neighbour methods have been adapted to multimodal settings such as the weighted nearest neighbour (WNN) method used in Seurat v4 which has been used to integrate joint scRNA-seq and scATAC-seq data (Hao *et al.*, 2021).

Finally, for diagonal integration, anchors do not exist in high-dimensional space. This is the case when different 'omics' techniques are performed in independent groups of cells. Naturally, this is the most challenging situation in which integration is required. However, real-world datasets frequently require this form of integration due to the maturity, availability, and scalability of single-omics technologies. Diagonal integration relies on a hidden or latent space that is retained between different data modalities applied to cells from the same sample (Argelaguet *et al.*, 2021). However, the extent to which biological assumptions that underpin the link between different measurements can be relied upon is unclear. In summary, as the number and modalities of datasets increases, integration methods have been developed and will become increasingly important for analysis and interpretation.

1.4 Metastasis, dormancy, and progression

1.4.1 Disseminated tumour cells and metastasis

As cancers evolve, they typically acquire a deadly hallmark which is responsible for the vast majority of cancer deaths – metastasis (Chaffer & Weinberg, 2011). Routes of metastatic spread include haematogenous, lymphatic, and transcoelomic (across a body cavity) spread. Tumour cells which undergo haematogenous spread do so by intravasating, travelling in the blood stream as circulating tumour cells (CTCs), and extravasating into distant organs such as the bone marrow to become disseminated tumour cells (DTCs) (Pantel *et al.*, 2009). DTCs in the bone marrow can enter dormancy for years, acting as a reservoir of tumour cells. These dormant DTCs can be refractory to adjuvant therapy before reactivating and becoming overt metastases (Risson *et al.*, 2020). This is thought to explain why in most cancer types, despite complete removal of the primary tumour, patients still have a significant risk of relapse for several years.

Despite their importance, the molecular nature of DTCs remains unclear, as well as the timing of dissemination and precise source of tumour cells generating metastasis. Two models of metastasis have been proposed (Klein, 2009). The linear progression model suggests cancer cells acquire clonal mutations in the primary site, and dissemination of tumour cells genetically similar to the primary occurs late in molecular time. In cases where there is linear progression, targeting early clonal mutations may be effective in treating the primary tumour but also in preventing relapse. In contrast, the parallel progression model proposes early dissemination of tumour cells which independently acquire mutations and evolve at their target site. This results in cells that are genetically dissimilar from the primary tumour and each other, therefore representing a greater challenge to target.

How DTCs behave in different models of tumour evolution is unknown. In clear cell renal cell cancer (ccRCC), different patterns of tumour evolution in the primary tumour have been shown to be associated with different spatio-temporal patterns of metastatic spread (Turajlic *et al.*, 2018a). Patients with tumours that were less heterogeneous as a result of punctuated evolution rapidly progressed with widespread dissemination resulting in poor prognosis, whereas polyclonal tumours generated by branching evolution progressed slowly, often with solitary metastases developing sequentially. Therefore, primary tumours with branching evolution might be releasing DTCs continuously from different subclones and would be expected to be more heterogeneous than those with punctuated evolution (Figure 1.3).

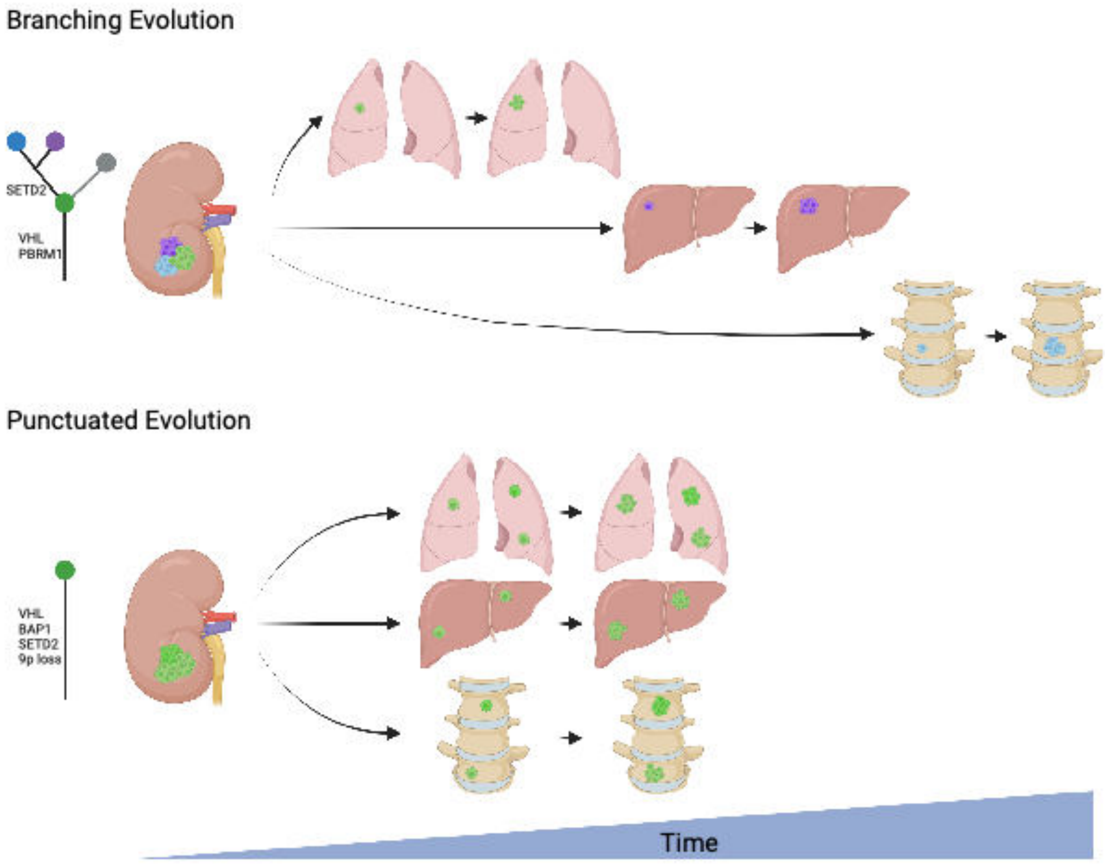


Figure 1.3 DTC dissemination under models of ccRCC tumour evolution.
In tumours with branching evolution, only certain clones are capable of metastasis and result in solitary or oligometastases over time. In contrast, tumours with punctuated evolution rapidly disseminate to multiple organs. Typical evolutionary trees are shown next to the primary tumour with DTCs shown as single cells and metastases shown as clusters of cells. Created with Biorender.com.

Determining the genetic profiles of DTCs will give further insights into the timing and source of dissemination. This may lead to opportunities for intervention prior to the development of metastasis, resulting in improved clinical management (Naxerova & Jain, 2015).

1.4.2 Clinical studies of DTCs

DTCs can be sampled through bone marrow aspirates from patients without overt metastatic disease and have been identified using epithelial or tissue-specific markers in several tumour types (Braun *et al.*, 2000; Weckermann *et al.*, 2009; Eide *et al.*, 2013). The majority of studies on DTCs have been carried out in breast cancer, as relapse can occur decades after seemingly curative treatment of the primary tumour leading to a need to understand and prevent this phenomenon. Several studies in this setting have shown that the presence of DTCs in the bone marrow is a prognostic marker for relapse and poor survival (Braun *et al.*, 2005; Janni *et al.*, 2011). Accordingly, a further study has shown that successful eradication of DTCs with secondary adjuvant therapy removes this excess risk (Naume *et al.*, 2014). However, as there was no control group for patients receiving secondary adjuvant therapy, it is unclear whether these DTCs were killed by the extra chemotherapy or if they would have perished regardless.

Historically, DTCs have primarily been studied in the bone marrow due to the relative ease of sample collection. However, DTCs are also expected to seed other organs where tumours relapse and have been found there in mouse models (Piranlioglu *et al.*, 2019). Furthermore, DTCs have been detected in liver samples acquired from autopsies of pancreatic cancer patients, although no further molecular or genomic characterisation of these cells was undertaken in this study (Pommier *et al.*, 2018). This demonstrates the value of sampling normal tissues. The factors which govern organotropism (cancers metastasising to particular organs) and which tissues are permissive for DTC reactivation are currently unknown (Vanharanta & Massagué, 2013; Chen *et al.*, 2018). Differences in interactions between DTCs and tissues which permit future metastases and those that do not may reveal the mechanisms underlying organ specificity.

1.4.3 Genomic profiling of DTCs

Studies carrying out molecular profiling of DTCs have been very limited in scope and scale. Previous genomic studies utilising array CGH have found aberrant CNA profiles in DTCs which were seemingly unrelated to the primary tumour, supporting parallel progression (Schmidt-Kittler *et al.*, 2003; Gangnus *et al.*, 2004). In addition, individual DTCs isolated from the same patient displayed diverse sets of CNAs which were non-recurrent (Klein *et al.*, 2002). In contrast, CNAs in DTCs from patients with metastatic disease were similar to those found in the primary tumour and lymph node metastases, thereby supporting a model of linear progression (Mathiesen *et al.*, 2012; Czyż *et al.*, 2014).

However, recent work by Demeulemeester et al. using scDNA-seq on bone marrow DTCs in primary breast cancer patients has allowed our group to map DTCs onto tumour phylogenetic trees showing they disseminate relatively late (Demeulemeester *et al.*, 2016). In addition, a population of cells with CNAs which did not match the primary tumour was also identified which were termed "aberrant cells of unknown origin". Similar "pseudodiploid" cells have also been reported in earlier studies profiling normal tissues with a prevalence of 1-8% depending on the tissue (Knouse *et al.*, 2014; Gao *et al.*, 2016). When genotyped, Demeulemeester et al. found that these cells did not share any of the tumour specific SNVs belonging to the primary, suggesting they are not true DTCs. Furthermore, the prevalence of these cells appeared to be age-dependent and may be related to clonal mosaicism. Previous studies would have likely incorrectly considered these aberrant cells as DTCs and grouped them with true DTCs. This could explain the diverse CNA profiles which do not appear to be related to the primary tumour observed in studies of patients with early-stage disease which were taken as evidence of parallel progression.

1.4.4 Detection of DTCs

Traditionally, minimal residual disease has been detected with reverse transcriptase PCR (RT-PCR). This has since been superseded by antigen dependent methods for the detection of DTCs. As many solid tumours are of epithelial origin, unlike native bone marrow cells, DTCs from these tumours can be detected using antibodies against epithelial markers such as epithelial cell adhesion molecule (EpCAM) in the case of breast cancer (Braun *et al.*, 2000). Cytokeratins have also been used for detection of DTCs from oesophageal cancer, pancreatic cancer, and prostate cancer (Thorban *et al.*, 1999; Gužvić *et al.*, 2014; Schumacher *et al.*, 2017).

DTCs are estimated to have a prevalence of 1 in 10^7–10^8 in the bone marrow and various DTC detection methods have been used in the literature (Banys *et al.*, 2014; Gilje *et al.*, 2014). To profile these rare cells, they must first be isolated as single cells which can be performed with several different techniques. Flow cytometry has also been successfully used to detect CTCs and DTCs in peripheral blood and bone marrow (Merugu *et al.*, 2020). In addition, micromanipulation has been successfully applied for isolation in previous studies after detection of DTCs using immunocytochemistry or immunomagnetic beads (Morgan *et al.*, 2009; Mathiesen *et al.*, 2012). Finally, antigen-independent size-based microfluidic techniques have also been applied for isolation DTCs or CTCs (Hvichia *et al.*, 2016; Lin *et al.*, 2017).

1.4.5 Research autopsy studies

Most studies of cancer are carried out on primary tumours as these are accessible after surgical removal. Once a known primary cancer metastasises, there is little value to be gained from a further biopsy of metastatic lesions. This is because the additional risks of the procedure are not outweighed by improbable changes in management resulting from a further sample. Indeed, a recent pan-cancer study found minimal change in clinically actionable targets after repeat metastatic biopsies (van de Haar *et al.*, 2021). Thus, the acquisition of metastatic samples (preferably paired with their primary) is a key challenge for studies in the advanced cancer setting.

One way to overcome this difficulty is through the practice of research autopsy studies of cancer patients (Iacobuzio-Donahue *et al.*, 2019). Whilst these studies are logistically challenging and require the selfless donation of cancer patients, early research studies have revealed novel insights about advanced cancer. Research autopsy programs extend the number of regions spatially and temporally that can be sampled allowing the reconstruction of more detailed evolutionary trees that highlight the characteristics of lethal subclones (Makohon-Moore *et al.*, 2017). They have unveiled the mechanisms behind treatment resistance with parallel evolution of *PTEN* inactivating mutations noted in a metastatic breast cancer patient (Juric *et al.*, 2015). Finally, tumour samples from rapid or “warm” autopsies have enabled the creation of cell lines and patient derived xenograft models (Rubin *et al.*, 2000). Therefore, research autopsies have the potential to enable the detection and profiling DTCs in various normal tissues.

1.5 Clinical impact of cancer evolution

Knowledge gained from the study of cancer evolution has important clinical implications with biological results beginning to be translated into the clinic. This impact can be broadly divided into the areas of early detection, selection, and therapeutic monitoring.

1.5.1 Early detection

The earliest genetic or epigenetic drivers that contribute to cancer formation can be determined by studying cancer evolution. These changes can then form the basis of early detection strategies to prevent the formation of new tumours or remove cells with malignant potential. Early changes can be measured directly through premalignant lesions producing pre-cancer atlases (Srivastava *et al.*, 2018). Alternatively, evolutionary timelines inferred from WGS can also inform the timing of mutations during tumour development (Jolly & Van Loo, 2018). Recently the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium reconstructed evolutionary histories across cancer types and showed driver events occur many years before diagnosis giving a window of opportunity for intervention (Gerstung *et al.*, 2020).

However, early detection will be more complex than simply detecting driver mutations as these are not exclusive to cancer (Acha-Sagredo *et al.*, 2022). Advances in sequencing technology have allowed us to interrogate mutations in macroscopically normal tissues. This has revealed surprising findings of driver mutations frequently being found in histologically normal tissues showing selection is not exclusive to cancer (Martincorena *et al.*, 2015). This raises questions about why these subclones do not develop into cancer and whether healthy competition between subclones prevents the development of cancer. Furthermore, recent studies have shown that clonal haematopoiesis of indeterminate potential (CHIP) – a condition where driver mutations result in clonally expanded haematological cells – is associated with atherosclerosis (Jaiswal *et al.*, 2017). Patients with *TP53* mutant clones are particularly at risk, indicating that somatic mutations may also be involved in non-cancer diseases (Zekavat *et al.*, 2023).

1.5.2 Evolution-guided therapy

Since the development of the first targeted therapies, identification of recurrent mutations in subtypes of cancer has been a cornerstone for rational cancer therapeutics. The addition of tumour evolution can give a more detailed perspective on potential targets (Fittall & Van Loo, 2019). For example, mutations that are identified as clonal are attractive targets as they will be present in every cell, unlike subclonal mutations. This is likely the underlying basis of effectiveness of targeting oncogene addiction, as tumours rely on the selective advantage provided by these drivers. Furthermore, targeting combinations of clonal mutations may create more durable responses, as this necessitates the development of multiple resistance mechanisms simultaneously.

Characterisation of a tumour's evolutionary trajectory can also be informative for predicting a patient's outcome. This is particularly important in the setting of minimal residual disease (MRD) where the patient's risk of relapse can vary greatly. For example, the TRACERx Renal (TRAcking Renal Cell Carcinoma Evolution Through Therapy (Rx)) evolutionary subtypes mentioned earlier relate to clinical phenotypes with different patterns and risks of relapse. Furthermore, information gained from tumour evolution studies can also be used to identify biomarkers predictive of poor prognosis (Biswas *et al.*, 2019). Together, these biomarkers can potentially stratify patients and guide identification of patients who require intense treatment as they are more likely to relapse.

Recently, trials utilising cancer genomics to guide therapy have begun to read out. Andre et al. reported improved progression-free survival in a trial for metastatic breast cancer using clinically actionable genetic alterations to guide therapy (Andre *et al.*, 2022). Mutational signatures have also demonstrated their clinical impact. For example, the single base substitution signature 3 is associated with homologous recombination deficiency (HRD), often due to inactivating mutations in *BRCA1* or *BRCA2*. A panel of mutational signatures has been used clinically to detect functional *BRCA1/BRCA2* deficiency. This has clinical implications as poly(ADP-ribose) polymerase (PARP) inhibitors have been shown to be effective in HRD and mutational signature-based HRD prediction has been incorporated into clinical trials (Chopra *et al.*, 2020).

1.5.3 Therapeutic monitoring and treatment resistance

In addition to imaging, disease burden during treatment can be tracked non-invasively through circulating tumour DNA (ctDNA) produced by dying cancer cells.

By quantifying the level of cancer mutations in the blood, ctDNA can be used as a surrogate marker of disease burden and detect relapse earlier than imaging, or track growth of individual subclones (Abbosh *et al.*, 2017). However, many of these methods are personalised or reliant on mutations informed by the primary tumour, therefore will not track mutations developed after the primary tumour was removed. Furthermore, if a subclone is not cycling quickly or resulting in cell death, it will not contribute significantly to the pool of ctDNA.

Therapeutic resistance can develop as the result of mutations in targets or evasion of the immune system. Identifying these key subclones and characterising the phenotypic properties which allow them to succeed, is critical for the prevention of resistance. Therefore, understanding common evolutionary escape paths and pre-empting them may be an effective strategy. For example, responses to EGFR inhibitors in non-small cell lung cancer (NSCLC) are often short-lived due to mutations that block EGFR inhibitor binding. As resistance mechanisms are conserved, third-generation inhibitors have been designed to pre-empt this by covalently bonding to the active site of EGFR (Shi *et al.*, 2022).

The emergence of immunotherapies has been a landmark event in the treatment of many cancers. These therapies generate an anti-tumour response through release of immune suppression by immune checkpoint inhibitors or directly trigger an anti-tumour response through cellular therapies or cancer vaccines. However, cancerous cells can evolve, with those that lose their ability to present neoantigens effectively able to survive and proliferate (McGranahan *et al.*, 2017). As a result, potentially more aggressive subclones will be selected for, which are then able to deploy additional immune evasion strategies, further complicating their treatment. Therefore, a better understanding of the interplay between subclonal evolution and immune evasion is required for maximising the efficacy of these new treatments.

Overall, the knowledge gained from our study of tumour evolution has the potential to radically improve cancer treatment and bring about changes in clinical practice which benefit patients.

1.6 Summary

After decades of progress in the field of cancer genomics, tumour heterogeneity and cancer evolution can now be studied down to each individual cell. This thesis applies single-cell and spatial sequencing technologies in three settings: characterising a rare soft tissue sarcoma, profiling disseminated tumour cells in limited stage disease, and correspondingly in the metastatic setting.

Although the original intention of the PhD project was to primarily study disseminated tumours cells in detail, the emergence of the COVID-19 pandemic during the PhD necessitated a shift to analysis of data which had already been generated. Therefore, much of this thesis is based on the multi-omic high-resolution profiling of a rare malignant nerve sheath tumour. This study utilised a variety of single-cell and spatial techniques to perform high resolution profiling and spatial phylogenetic reconstruction. Chapter 3 reports the extensive genomic analysis performed on this tumour and the significant copy number heterogeneity detected. The impact on gene expression resulting from this chromosomal instability and integration of different data modalities is then described in Chapter 4.

Single-cell sequencing was also deployed for two projects aiming to genomically profile disseminated tumour cells in early and late stage disease. Unfortunately, these projects required collection of new samples from patients in a clinical setting and were significantly impacted by the COVID-19 pandemic. Therefore, the scale and scope of these projects was limited. The detection and profiling of DTCs in the metastatic setting through research autopsy where sample collection was paused for a prolonged period is described in Chapter 5. Finally, Chapter 6, presents work on detection of DTCs in bone marrow aspirates from patients with early-stage clear cell renal cell carcinoma where approval for sampling was delayed until the final year of the overall project.

Together, this work demonstrates our current capabilities to examine the evolutionary history of tumour and highlights valuable use-cases where single-cell technologies are required.

Chapter 2. Materials & Methods

2.1 Overview of methods

The samples studied in this thesis came from 3 sources. Firstly, the primary tumour and 5 recurrence regions were collected from a patient with a malignant peripheral nerve sheath tumour (MPNST) and underwent multi-omic analysis. Secondly, samples were collected through research autopsy from a cohort of patients with advanced metastatic clear cell RCC and melanoma. Finally, bone marrow aspirate samples were collected from a cohort of patients undergoing resection of clear cell RCC at the time of surgery.

2.2 Data analysis and figure generation

Analysis was mostly performed in R (3.6.0 and 4.0.0) with bespoke scripts. Some packages requiring Python were run in conda environments with Anaconda3 (2022.05) and Singularity (v3.6.4) containers were also used where specific images of software was required. Figures were generated with R, Illustrator and Biorender.com. Flow cytometry analysis for identifying populations of cells or ploidy analysis was performed using FlowJo (v10.8.1).

2.3 Multi-omics analysis of an MPNST

The primary MPNST was collected from theatre at the Royal National Orthopaedic Hospital, Stanmore, underwent processing in pathology and was snap frozen in liquid nitrogen. Five local recurrence region samples were subsequently collected from amputation tissue and snap-frozen in liquid nitrogen or collected in PBS, manually minced, and dissociated into single-cell suspensions using Collagenase II. All samples were then transferred to −80°C for long-term storage. Clinical annotation was collected from electronic health records from the Royal National Orthopaedic Hospital.

Experiments and library preparation for bulk, 10X scDNA-seq, 10X scRNA-seq and 10X Visium spatial transcriptomics and flow cytometry for ploidy analysis was performed by Annelien Verfaillie. Single cells or nuclei were sorting into 96-well PCR plates for the genome and transcriptome sequencing experiment (also performed by Annelien Verfaillie). These plates of single cells were transferred to the Wellcome Sanger Institute, Cambridge and G&T-seq was kindly performed by members of the Voet group. Cryosectioning, imaging, and LCM of tissue sections and subsequent library preparation of DNA amplified from LCM spots was performed by Cristina Cotobal Martin. Cryosectioning of tissues and Slide-seq V2 was performed by Christie English with the help of Laura Cubitt from the Rodriques lab. Sequencing was performed by the Advanced Sequencing Facility science technology platform (STP) at the Crick Institute. Unless otherwise stated, bioinformatics analysis was performed by myself under the supervision of Jonas Demeulemeester, Maxime Tarabichi and Peter Van Loo.

A methods overview for Chapters 3 and 4 is shown in Figure 2.1.

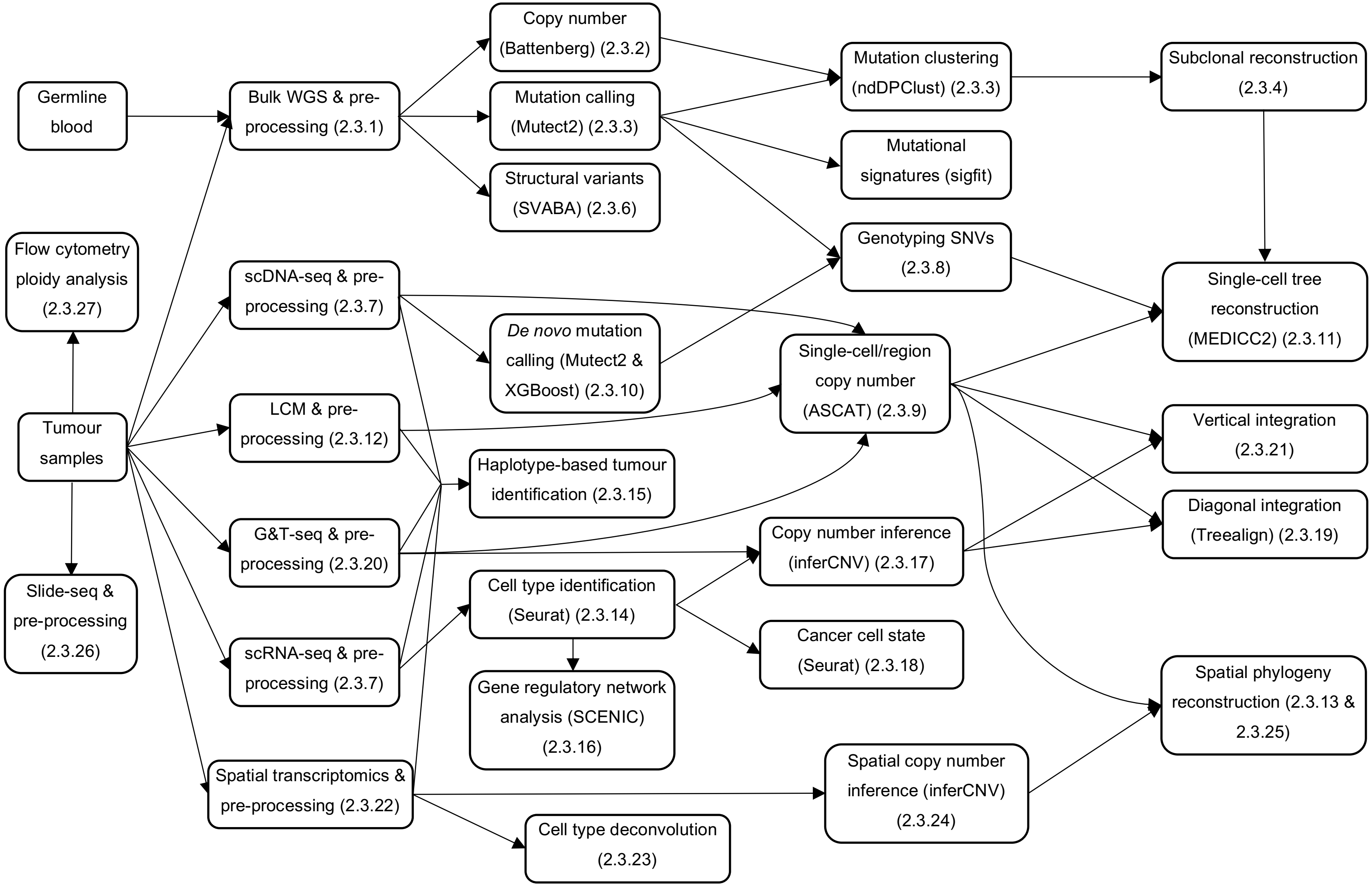

Germline blood
Bulk WGS & pre-processing (2.3.1)
Copy number (Battenberg) (2.3.2)
Mutation calling (Mutect2) (2.3.3)
Structural variants (SVABA) (2.3.6)
Mutation clustering (ndDPClust) (2.3.3)
Subclonal reconstruction (2.3.4)
Mutational signatures (sigfit)
Genotyping SNVs (2.3.8)
Flow cytometry ploidy analysis (2.3.27)
scDNA-seq & pre-processing (2.3.7)
De novo mutation calling (Mutect2 & XGBoost) (2.3.10)
Single-cell/region copy number (ASCAT) (2.3.9)
Single-cell tree reconstruction (MEDICC2) (2.3.11)
Tumour samples
LCM & pre-processing (2.3.12)
Haplotype-based tumour identification (2.3.15)
Vertical integration (2.3.21)
Diagonal integration (Treealign) (2.3.19)
G&T-seq & pre-processing (2.3.20)
Copy number inference (inferCNV) (2.3.17)
Slide-seq & pre-processing (2.3.26)
Cell type identification (Seurat) (2.3.14)
Cancer cell state (Seurat) (2.3.18)
scRNA-seq & pre-processing (2.3.7)
Gene regulatory network analysis (SCENIC) (2.3.16)
Spatial phylogeny reconstruction (2.3.13 & 2.3.25)
Spatial transcriptomics & pre-processing (2.3.22)
Spatial copy number inference (inferCNV) (2.3.24)
Cell type deconvolution (2.3.23)

Figure 2.1 Flowchart of methods used for Chapters 3 and 4.

2.3.1 Bulk WGS and pre-processing

DNA was extracted from germline blood and tissue sections cut from tumour samples. Libraries were made from native DNA using the NEBNext Ultra II FS DNA Library Prep kit (New England BioLabs) and 100 base paired-end sequencing was performed on the HiSeq 4000 system according to Illumina protocols. Paired-end reads were aligned to the reference human genome (GRCh38) with BWA-MEM (Li, 2013). Duplicates were marked using MarkDuplicates (Picard v2.19.1) and base quality scores were recalibrated with Genome Analysis Toolkit (GATK v4.1.2.0) BaseRecalibrator (Broad Institute, 2019; Van der Auwera *et al.*, 2020).

2.3.2 Copy number calling in bulk data

The Battenberg pipeline (v2.2.9) was used to obtain subclonal copy-number profiles for the primary tumour and each region of the recurrence (Nik-Zainal *et al.*, 2012b). This pipeline first counts the number of reads for each allele at SNP positions from the 1000 Genomes Project in the tumour and normal BAMs. The counts from SNPs are then used to derive a logR (log2 of the normalised ratio between tumour and normal reads) and BAF (ratio between the paternal and maternal alleles). Battenberg then uses population-based phasing for haplotype reconstruction with Beagle5 (Browning *et al.*, 2021). Segmentation by Piecewise Constant Fitting (PCF) is then performed on the logR and BAF values of heterozygous SNPs to determine genomic segments with the same copy-number state (Nilsen *et al.*, 2012). It then co-estimates purity and clonal copy number states using logR and BAF values and the following pair of equations, as in ASCAT:

$$logR = log_2\left(\frac{\rho(n_A + n_B) + (1-\rho)2}{\psi}\right)$$

$$BAF = \frac{\rho n_B + (1-\rho)1}{\rho(n_A+n_B) + (1-\rho)2}$$

Where ρ is the sample purity, n_A and n_B are the major and minor allele integer copy numbers and ψ is the sample ploidy which is derived from $\psi = \rho\psi_T + (1-\rho)2$ using the tumour ploidy ψ_T. Finally, a T-test is performed to determine if there is evidence of subclonal copy number event for a genomic segment.

As several samples were available, multi-sample mode was used with external phasing information for haplotypes also incorporated from linked-read sequencing performed on the same tumour samples (provided by Jonas Demeulemeester) to improve the phased haplotype blocks and segmentation. Most tumour samples were highly pure, resulting in a BAF very close to 0 or 1 and difficulty in detection of subtle shifts in BAF. Therefore, to artificially reduce the purity, I used an *in-silico* spike-in of reads from the matched normal, which resulted in improved segmentation and copy-number calling accuracy. The modified Battenberg plot to displayed copy number profiles of all regions was created by overlaying the CNA profiles of all samples and colouring the copy number value of the major allele by region.

2.3.3 Mutation calling and clustering

Somatic variants were called using MuTect2 (GATK v4.1.8.0) in a Singularity container (v3.6.4) for each chromosome to decrease run time. Variant Call Format (VCF) files were then merged and indexed and MergeMutectStats from GATK was used to combine the stats files for each chromosome. The LearnReadOrientationModel, GetPileupSummaries and CalculateContamination steps in the GATK pipeline were then run to allow SNV filtering using FilterMutectCalls and those which passed were annotated with Funcotator.

Multidimensional Bayesian Dirichlet Process-based mutation clustering (ndDPClust) (v2.2.8) was run in multi-sample mode on all SNVs present in all regions (Nik-Zainal *et al.*, 2012b; Bolli *et al.*, 2014). This algorithm clusters mutations based on the CCF of each SNV based on the following equation:

$$CCF = \frac{f}{m\rho}(\rho N_T + 2(1 - \rho))$$

Where *f* is VAF, *m* is the number of copies of DNA bearing the mutation (multiplicity), *ρ* is the sample purity and N_T is the local copy number. The R package dpclust3p (v1.0.8) was used to prepare inputs for ndDPClust. Only SNVs on chr1-22 and X were retained. Due to memory constraints, all combinations of 4 from 6 samples were run for 12,000 iterations (2,000 burn-in). ClusterID-based consensus clustering (CICC) was used to generate 40 consensus clusters across the 15 runs and allowed calculation of CCFs of each cluster (Tarabichi, 2018; Dentro *et al.*, 2021). This revealed clonal clusters for all regions and a subclonal cluster for R4. Subclonal clusters of SNVs for other regions were identified from a large cluster of SNVs with low CCF (~0.05) by extracting SNVs with CCF of 0 in all other regions. Remaining clusters which violated the Pigeonhole principle or crossing rule such as those with constant CCFs across samples were removed due to likely being artefactual clusters. Based on these criteria 2,223 out of the total 13,957 SNVs were filtered out. The remaining clusters were taken forward for further analysis.

2.3.4 Bulk Phylogenetic tree reconstruction

The phylogenetic tree was reconstructed manually by assessing the CCFs of each subclone in each region. The "pigeon-hole" principle was then applied to mutational clusters only present in one region to determine their relationship (i.e., branching *vs.* linear). This principle states that if the CCFs of two mutational clusters sum up to more than that of their shared ancestral cluster, they must be collinear (Tarabichi *et al.*, 2021).

2.3.5 Mutational signature profiling

The R package Sigfit (v2.2.0) was used to fit SNVs from each mutational cluster against a catalogue of mutational signatures found in a cohort of MPNSTs from Genomics England (GEL) provided by Steven Hargreaves (Gori & Baez-Ortega, 2020). These included signatures SBS1, SBS2, SBS5, SBS8, SBS9, SBS13, SBS17a, SBS17b, SBS18, SBS28, SBS30, SBS35, SBS36 and SBS39 (COSMIC v3.2) (Tate *et al.*, 2019). Sigfit was run for 10,000 iterations with a burn-in of 5,000 iterations with each mutational cluster as a sample, thereby deriving the mutational signatures for each clonal/subclonal cluster. Mutational signatures with a 95% highest posterior density interval lower limit of greater than 1% of SNVs in any cluster were retained (signatures SBS13, SBS17a, SBS17b, SBS28 and SBS36 were removed) and mutations were refitted to the remaining signatures.

2.3.6 Structural variant calling

SVABA (v1.1.0) was used to perform structural variant calling on the primary tumour and each recurrence region (Wala *et al.*, 2018). Circos plots showing SVs were made with the R package circlize (Gu *et al.*, 2014). Types of structural variants were annotated as inter-chromosomal translocation, inversion, insertion, deletion, or duplication.

2.3.7 10X single-cell sequencing and pre-processing

Nuclei were isolated from all tumour samples using EZ Prep (Sigma-Aldrich) and single cell sequencing was performed on single-nuclei suspensions using the 10X Genomics Chromium platform. For scDNA-seq, the Chromium Single Cell CNV kit (10X Genomics) was used and for scRNA-seq, the Chromium Single Cell 3’ v3 kit (10X Genomics) was used. 10X Genomics-generated DNA/cDNA libraries were sequenced on an Illumina HiSeq 4000 using 150 base paired-end and 100 base single-end sequencing, respectively. Reads for scDNA-seq were aligned to GRCh38 using BWA-MEM. Cellranger-DNA (v1.1.0) and Cellranger (v3.0.1) were used to demultiplex and align reads to GRCh38 for scDNA-seq and scRNA-seq respectively (Zheng *et al.*, 2017). 6,795 scDNA-seq cells and 37,716 scRNA-seq cells in total passed the UMI noise filter.

2.3.8 Genotyping SNVs in single cells

Clusters of mutations identified in bulk WGS through DPClust and CICC were genotyped in the scDNA-seq data using alleleCount (v4.0.0). A binary co-occurrence matrix was generated where co-occurrence was considered to be present if one or more reads with the ALT allele were present for both SNVs in the same cell (CASM/Cancer IT Wellcome Sanger Institute, 2020).

2.3.9 Allele-specific copy number calling in single cells

ASCAT.sc (single-cell or shallow coverage) (v0.1) was developed alongside this project by Maxime Tarabichi who inferred copy number profiles for scDNA-seq data (Tarabichi, 2020). First, pre-computed variable-size bins for GRCh38 included in ASCAT.sc were loaded. Read counts were then derived in each bin with MAPping Quality (MAPQ)>=30 and not counting duplicates. As GC-content, the proportion of G and C nucleotides in a genomic bin, affects read coverage, this was computed for each 500kb bin along the genome. The log read counts were then smoothed by applying a loess fit against GC-content to obtain the corrected logR.

To identify purity and ploidy in a similar fashion to ASCAT, a grid search was then performed on all ploidy values between 1.7 and 5 by steps of 0.01 and a purity of 0.5 or 1 to fit copy number profiles from the logR track. Noisy cells were filtered out based on the average standard deviation of the logR across segments and number of reads per bin per tumour chromosomal copy (NRPCC) resulting in 4,408 cells with satisfactory copy number profiles.

Fitted integer copy number profiles revealed populations of cells present in all regions which appeared to have undergone further whole-genome doubling (WGD) (Appendix 8.1.6). Whilst the existence of these sub-populations was confirmed by using FACS ploidy analysis, it was not possible to accurately assign which cells had undergone WGD as a copy number profile can always be fitted to a doubled version of itself (Tarabichi *et al.*, 2021).

Using these profiles, I defined copy-number based subclones by first removing normal and WGD cells by applying the ward.D2 method of hierarchical clustering on Manhattan distances of copy number profiles. A K-means approach with K set to 20 clusters was then applied to the remaining cells. A cluster with fewer than 5 cells and the cluster of normal cells were removed. As only 82 cells from R3 were sequenced, these were manually split from their shared cluster with cells from R2 through hierarchical clustering. This yielded 19 tumour subclones which were used in subsequent analyses.

Allele specific copy-number profiles were also derived using ASCAT.sc leveraging BAF values calculated from the phased haplotypes derived from the bulk Battenberg runs. As single-cell data is particularly sparse, logR and BAF are clustered across cells to improve their estimation and thus, allele-specific copy number profile fitting. A similar strategy has been deployed in other single-cell copy number calling methods such as CHISEL (Zaccaria & Raphael, 2021). Multi-PCF (mPCF) was then used for segmentation of shared genomic segments with different copy number states across all cells to produce profiles with the same set of breakpoints (Nilsen *et al.*, 2012).

UMAP dimension reduction of all single-cell copy number profiles using Manhattan distances was performed with the R package umap (v0.2.7). Copy number profiles of bulk samples were then transformed or projected into this learned space. Similarly, copy number profiles inferred from Visium spatial transcriptomics were also projected onto this shared UMAP (see section 2.3.24).

ASCAT.sc was also used to perform integer and allele-specific copy number calling on shallow coverage WGS data from G&T-seq cells and LCM single spots (see sections 2.3.13 and 2.3.20). This was done by applying the breakpoints identified using mPCF and applying them to the logR and BAF tracks derived from G&T-seq or LCM spots to create recurrent breakpoints across sequencing methods. This enabled integration of copy number profiles from different experiments for downstream analysis.

Copy number profiles were plotted by taking the integer value for each genomic bin per cell and plotting them as a heatmap using a custom wrapper function around the ComplexHeatmap package which annotates the chromosomes (Gu *et al.*, 2016). For allele specific copy number profiles, the states of the two alleles were converted to a single value to allow plotting as a heatmap in a similar fashion.

2.3.10 *De novo* SNV calling and subclonal mutation clustering

I sought to identify additional SNVs specific to copy-number based subclones by performing somatic variant calling in the scDNA-seq data. First, mutations were called using Mutect2 against the bulk normal and also pooled scDNA-seq normal cells as matched controls. scDNA-seq cells were identified as normal by having diploid copy number profiles and a 1:1 ratio of the two haplotype counts (see section 2.3.15). Only shared SNVs detected against both bulk and pooled normal were retained.

Next, a XGBoost machine learning classifier from the R package xgboost (v1.3.2.1) was trained to filter out false SNV calls (Chen & Guestrin, 2016). The following features of SNVs were used for statistical modelling and as input for the XGBoost algorithm: number of reads with the reference and alternative allele in the bulk tumour, number of reads with the reference and alternative allele in pooled scDNA-seq tumour cells, number of reads with the reference and alternative allele in the bulk normal, number of reads with the reference and alternative allele in pooled scDNA-seq normal cells and whether the SNV is within 1Mb of the centromere. A manually curated subset of 200 SNVs was used to train the classifier reaching 96% accuracy on a hold-out set of 100 SNVs after 100 iterations.

Clusters of SNVs specific to subclones were determined by pooling cells for each subclone together and genotyping SNVs. SNVs were considered present in a subclone if one read with the alternative allele was found in at least 3 cells from the subclone. Hierarchical clustering was performed on SNVs using the ward.D2 method after removal of SNVs assigned to be truncal by ndDPClust and SNVs present on 1000 Genomes Project SNP positions. *De novo* SNVs were almost exclusively found to be subclone specific as expected due to their lower VAF or complete absence in the bulk resulting in them not being identified. Among the subclones from R1, *de novo* SNVs were almost predominantly in the R1_4 and R1_5 subclones which exhibited a loss of 10p and segment of 10q that was not detected in the bulk copy-number profile (Figure 3.11 and Appendix 8.1.3). This finding supports the hypothesis that these subclones were only sampled in the single-cell and not the bulk experiment.

Two conflicting clusters of SNVs shared between the subclones of R5_1 and R5_3 and those of R5_2 and R5_3 were observed. Due to the sparsity of data in single-cell sequencing and variability in size of subclones, SNVs can be missed when genotyped. This leads to subclonal specific clusters of SNVs being observed by chance even if a cluster of SNVs is shared by 2 subclones. By modelling probabilities of observed clusters of SNVs using a binomial distribution, I determined whether a subclone specific cluster of SNVs was real or simply observed due to low coverage. This revealed that the shared cluster between R5_1 and R5_3 was observed due to chance as these SNVs are likely also present in R5_2 but were not seen due to shallow coverage (Appendix 8.1.8). This allowed me to resolve this conflict and determine the relationship between these three subclones to create a SNV-based phylogenetic tree of subclones which formed the main stems of the fully resolved single-cell tree reconstructed in section 2.3.11.

2.3.11 Single-cell tree reconstruction

Single cells from each subclone were used to generate a minimum-event distance tree with MEDICC2 (v0.5b) with WGD events enabled and minimum segment length of 1.5Mb (Kaufmann *et al.*, 2022). Briefly, MEDICC2 infers phylogenetic relationships between tumour samples based on the minimum-event distance between their copy-number profiles (the minimum number of gains and losses required to transform one copy number profile into another). It then infers the tree topology from pairwise minimum-event distances between all samples using neighbour joining before reconstructing ancestral copy number profiles to minimise the total number of events in the tree.

During this study, I noted that uneven numbers of cells from different regions can influence the structure of the tree at the level of large branches. This is because events present in more cells will be placed higher in the tree to minimise the total number of events, resulting in incorrect ordering of high-level events. For example, if a certain gain is only present in one region (and therefore occurred later in evolution) but this region massively dominates the number of cells profiled, then MEDICC2 will place this gain as an early event in the tree, and other regions will feature gain then loss of this segment to minimise the overall number of events. To overcome this, the subclone tree was used for the relationship between subclones and copy-number based single-cell trees were then grafted onto these subclone tips to create a phylogenetic tree refined down to the single-cell level.

2.3.12 Laser capture microdissection and pre-processing

Tumour tissues from the primary tumour, R1 and R4 were embedded in optimal cutting temperature (OCT) compound and stored at −80 °C. 16 µm sections were cut from the front, side and back of tissue blocks, mounted on PEN-Membrane Slides (Leica) and stained with H&E. LCM of 16 to 24 spots per section was performed according to the manufacturer’s protocol on the Leica LMD7000 system. Whole genome amplification was performed on extracted DNA using the Genome Plex kit (Sigma-Aldrich). Libraries were prepared from 134 successful amplifications using the NEBNext Ultra II FS DNA Library Prep kit (New England BioLabs) and 100 base paired-end sequencing was performed on an Illumina HiSeq 4000. Reads were aligned to GRCh38 using BWA-MEM (v0.7.17).

2.3.13 Phylogenetic analysis of LCM spots

Allele-specific copy-number profiles were derived using ASCAT.sc as described in section 2.3.9. The R package Shiny (v1.6.0) was used to create an interactive Shiny app with various features. These include selection of LCM spots to view its allele-specific copy number profile and ratio of haplotype counts (see 2.3.15), and selection of a genomic segment to view the copy-number state in spots for all sides belonging to that region. It also allows the relationship between LCM spots and single cells to be observed by displaying the fully resolved single-cell phylogenetic tree with LCM spots. This was done by running MEDICC2 on the combined scDNA-seq and LCM copy number profiles to generate an event-distance matrix and joining each LCM spot to the closest single cell based on event distance. This phylogenetic tree was then pruned for the regions studied with LCM (primary tumour, R1 and R4) to show the location of LCM spots on phylogenetic trees for these specific regions.

MEDICC2 was used to infer phylogenetic relationships between spots from each side of the tumour block with normal spots excluded. These phylogenetic trees were then represented on histological images by placing successive ancestral nodes equidistant between descendants until the root (MRCA) of the tree was placed. These intermediate node positions were then iteratively adjusted to move 30% of the distance between its original position and its immediate ancestor. Subclone specific SNVs derived from single cells were genotyped in LCM spots from each region in a similar fashion to section 2.3.8 to derive the VAF for each cluster of SNVs. These VAFs were then plotted onto the histological image of each slide to show the spatial distribution of SNVs.

2.3.14 Cell type identification

Gene count matrixes were loaded into the R package Seurat (v3.2.1) with removal of cells with more than 5% mitochondrial reads. Counts were then transformed and normalised with the SCTransform function using the top 3000 variable genes. Sample batch correction was not performed as we noted that batch correction was inappropriately removing biological signal. For example, when batch correction was performed using Seurat's FindIntegrationAnchors and IntegrateData functions, copy-number driven differences between regions were removed with cells from all regions placed together on the UMAP (Appendix 8.1.15). PCA and UMAP dimensionality reduction was performed with 30 principal components used for UMAP calculations. Cycling cells were identified using the function CellCycleScoring with a list of cell cycle markers provided in Seurat.

To identify cell types, I performed clustering of cells using the function FindClusters with a resolution of 0.5. Marker genes for the resulting 17 cell clusters were identified using the function FindAllMarkers. They were then annotated to known cell types based on canonical markers: macrophages (*MRC1*, *F13A1*, *CD163*), T cells (*PTPRC*, *THEMIS*), endothelial cells (*VWF*, *PTPRB*) and skeletal muscle cells (*TTN*, *MYOG*). Further sub-clustering of cells belonging to the macrophage, T cell and endothelial clusters was performed to identify additional sub-populations based on the following markers: fibroblasts (*FGF7*, *LAMA2*), B cells (*CD19*, *CD79A*, *MS4A1*), regulatory T cells (*FOXP3*, *IL2RA*), and pericytes (*MYO1B*, *PDGFRB*).

2.3.15 Validation of tumour cell identification through genotyping haplotype phased SNPs

The remaining cells were annotated as tumour cells and confirmed to be malignant using a genotyping approach by generating BAF values at 1000 Genomes Project heterozygous SNP positions using alleleCount (v4.0.0) and assigning them to the two haplotypes determined through linked reads of the bulk tumour. Cells annotated as tumour featured significantly fewer reads from the lost alleles compared to TME cells without allelic imbalance. This approach also revealed an additional population of cells with intermediate BAF but clustering with non-malignant cells on the UMAP (Appendix 8.1.17). The BAF of these cells was between the BAF of tumour and normal cells, and exactly matched the expected BAF of simulated tumour–normal doublets which was performed by merging the counts of tumour cells and normal cells (Appendix 8.1.18).

This genotyping process was also performed to identify cells or spots as normal *vs.* tumour in scDNA-seq, G&T-seq, LCM, spatial transcriptomics and Slide-seq data.

2.3.16 Gene regulatory network analysis

Gene regulatory networks were identified using the tool SCENIC by running SCENICprotocol (v0.25.0) in a Singularity container (Aibar *et al.*, 2017; Van de Sande *et al.*, 2020). The outputs were then analysed in R with functions from the SCENIC package (v1.1.2). Dimension reduction of the regulons was carried out using the python package umap-learn (v0.5.1). The top transcription factors for each cluster were identified using the calcRSS function from SCENIC.

2.3.17 Copy number inference from single cell transcriptomes

CNAs in scRNA cells were detected using the R package inferCNV (v1.0.2) which uses gene expression intensity across a sliding window of 101 genes to infer copy-number changes. Doublets and cells annotated as normal but with a haplotype count ratio that of tumour cells were excluded. Cells annotated as TME were used as reference normal cells with a cut-off for the minimum average number of reads per gene for reference cells of 0.1, analysis_mode set to “samples” and cluster_by_groups set to true. Chromosome X was also included by setting chr_exclude to “c(“chrY”,”chrM”)”. 24 clusters were identified using hierarchical clustering of correlation distances between inferCNV intensity values using Ward’s minimum variance method. A heatmap showing scRNA-seq inferCNV profiles was created using a similar function to that used to plot total copy number profiles in section 2.3.9. These clusters were visualised back onto the Seurat umap by adding the cluster information to the Seurat metadata and using the Dimplot function to annotate these cells by cluster.

2.3.18 Cancer cell state scoring

Modules from Barkley et al. were used to score cancer cell states in individual cells (Barkley *et al.*, 2022). The AddModuleScore from Seurat was used to calculate the module score for each gene module for 5,000 downsampled cells. Briefly, this function places all genes into 24 bins in order of average expression across all cells. It then randomly selects 100 control genes from bins in which the genes in the module being tested were placed. These control genes are then used to calculate the mean expression in each cell from which the expression of genes in the module being tested are subtracted to give a module score reflecting expression of those genes compared to control.

2.3.19 Diagonal integration of scRNA-seq and scDNA-seq data

TreeAlign (v.10) was used to perform integration in a conda environment (Shi *et al.*, 2023). Briefly, TreeAlign models gene dosage effects from subclonal copy number changes to assigns scRNA-seq cells to subclones. Consensus copy number profiles of each scDNA-seq subclone defined in section 2.3.9 along with the raw expression matrix from scRNA-seq were provided and scRNA-seq cells were assigned using the “Total CN model” with the CloneAlignClone function.

2.3.20 Joint genome and transcriptome sequencing and analysis

Single cells dissociated at time of tissue collection were sorted into 96-well PCR plates for the recurrence regions. Nuclei were prepared from the primary tumour and also sorted into a 96-well PCR plate. Plates were then sent to the Wellcome Sanger Institute where G&T-seq was performed. Unfortunately, scDNA-seq for the R1 region failed. In total, 216 single-cell genomes and 300 single-cell transcriptomes were successfully sequenced resulting in 158 paired genomes and transcriptomes from the same cell.

Genomes were aligned to GRCh38 BWA-MEM (v0.7.17) and duplicates were marked with SAMtools markdup (v1.3.1). Transcriptomes were aligned to GRCh38 and gene counts generated with STARsolo (v 2.7.1a).

Total and allele-specific copy number calling was performed as described earlier in section 2.3.9, producing copy number profiles with the same breakpoints as those in the scDNA-seq and LCM data. MEDICC2 was also run to derive a minimum-event distance matrix to scDNA-seq cells. This was then used to graft G&T-seq cells to the fully resolved phylogenetic tree in the same way as LCM spots, by placing them next to the closest scDNA-seq cell based on event-distance. Identification of tumour and normal cells was performed by genotyping haplotype counts as described previously (see section 2.3.15).

For G&T-seq transcriptomic analysis, a Seurat object was created with genes present in at least 3 cells and with cells expressing at least 200 genes. Cells with more than 10% of counts coming from mitochondrial genes were removed. Cell cycle scoring was performed as described in section 2.3.14. G&T-seq cells were integrated with scRNA-seq to identify cell types using the SelectIntegrationFeatures for 3000 variable features, PrepSCTIntegration, FindIntegrationAnchors and IntegrateData functions from Seurat. Integration appeared to be largely accurate with G&T-seq tumour cells annotated to clusters which matched the region they were derived from (Appendix 8.1.22).

InferCNV was run as described previously in section 2.3.17 using cells integrated into non-tumour clusters as a reference and with a cut-off of 1 due to G&T-seq transcriptomes being SMART-seq as opposed to 10X. As copy number profiles and inferCNV intensities are derived from the same cell, gene dosage effects could be explored for every segment for that cell. In addition, inferCNV was run with G&T-seq data combined with scRNA-seq for the purposes of vertical integration using a cut-off of 0.1 (see section 2.3.21).

2.3.21 Vertical integration of scRNA-seq and scDNA-seq using G&T-seq

Batch correction methods which identify and utilise cells in both datasets as anchors have been used for data integration. As the origin of genomes and transcriptomes from G&T-seq are unambiguous, they were used as anchors for batch correction in a similar manner to Seurat v3 (Stuart *et al.*, 2019). Batch correction was performed to convert inferCNV signal intensities inferred from transcriptomes into total copy number values derived from genomes. Let $Y_{g,c}$ be a matrix of inferCNV intensities from the genomic bins g_1, g_2, …, g_n by cells c_1, c_2, …, c_m and $X_{g,d}$ be a matrix of total copy number values of the same genomic bins g_1, g_2, …, g_n by cells d_1, d_2, …, c_m.

First, a distance matrix *D* for each pair of cell, *c*, and anchor, *a*, was created by taking the Euclidean distance of a UMAP dimension reduction of their inferCNV signal intensities derived from the transcriptomes:

$$D_{c,a} = dist(c, a)$$

Normalisation of distances with feature scaling was performed with:

$$D' = \frac{D - D_{min}}{D_{max} - D_{min}}$$

Next, the distance for each anchor was weighted by identifying the nearest *k* neighbours from the distance matrix:

$$\widetilde{D}_{c,a} = 1 - \frac{D'_{c,a}}{D'_{c,k}}$$

The weight matrix, *W,* was then constructed with:

$$W_{c,i} = \frac{\widetilde{D}_{c,a}}{\sum_{1}^{j=k} \widetilde{D}_{c,j}}$$

Which produces weights for each anchor as a proportion of the total distance to the cell of all *k* anchors. After the weight matrix was constructed, a transformation matrix, *C,* was calculated with:

$$C = BW^T$$

Where *B* is the difference in between the original inferCNV signal matrix, *Y*, and the total copy number matrix, *X*, using $B = Y[,i] - X[,i]$. This allows the transformation matrix, *C*, to be subtracted to from the original transcriptomic matrix, *Y*, to give the batch corrected transcriptome-derived total copy number matrix $\hat{Y}$:

$$\hat{Y} = Y - C$$

To assess the performance of my vertical integration method, half of G&T-seq cells were used as anchors for batch correction and the other half used for validation with *k* set to 4. This method was then applied to all cells profiled with 10X scRNA-seq with all G&T-seq cells used as anchors.

2.3.22 10X Visium spatial transcriptomics

All tumour tissues were embedded in OCT and stored at −80 °C. 10 µm sections cut for each tissue and processed using the Visium Spatial Gene Expression Kit (10X Genomics). Tissues were permeabilised for 3 minutes after optimisation of permeabilisation conditions. Sections were stained with H&E and imaged using a

VS120 Slide Scanner (Olympus Life Science) under 20x magnification. Visium libraries were constructed according to 10X protocols and sequenced on an Illumina HiSeq 4000. Spaceranger (v1.1.0) was used to demultiplex reads and align to GRCh38 after manual alignment of fiducial markers and marking of tissue boundaries for each slide using Loupe Browser (v4.1.0).

Gene/cell count matrices were loaded into Seurat with spatial information incorporated into the metadata. Transformation and normalisation was performed with SCTransform followed by PCA and UMAP dimensionality reduction. Clustering on the 30 principal components was performed with a resolution of 0.8 to identify clusters of spots. These were visualised spatially with the SpatialDimPlot function. Identification of tumour and normal spots was attempted by genotyping haplotype counts as described previously (see section 2.3.15).

2.3.23 Spatial transcriptomic deconvolution

As Visium spots are 55μm in diameter, multiple cells can overlie the same spot and their transcriptomes will be combined. Deconvolution of the spot was performed to determine the cell types present in each spot. scRNA-seq data can be used to provide the expression profiles of cell types which have already been annotated as a reference for deconvolution. The package RCTD was run in doublet mode with a minimum number of 25 cells for each type of cell in the reference to derive deconvolved cell types of spatial transcriptomics data (Cable *et al.*, 2022). Deconvolved cell types were then plotted on Visium slides using the Spatialplot function in Seurat.

2.3.24 Inferred spatial transcriptomic copy number profile

InferCNV was used to derive relative copy number difference intensities for each slide with each cluster of spots defined by Seurat as a sample. The slide from R1 was used as a reference and the same settings as described in section 2.3.17 were used. Note the reference R1 spot did not show any subclonal copy number changes relative to each other, confirming this slide indeed displayed uniform copy-number profiles.

The scDNA-seq cells with a matching copy number profile for spots from the R1 slide were identified as a subset of cells from the R1_1 clone. This was done by looking at the relative changes of segments which differ between R1 subclones but are clonal in other regions and deducing the identity of this subclone through a process of elimination. For example, a relative gain was not detected in other regions in chr10p, therefore this slide is not part of the R1_4 or R1_5 subclone (Figure 4.16 and Appendix 8.1.29). In addition, relative losses in chr1p were seen in the R2, R3, and R4 slides but not Primary_A, Primary_B, R5_A, or R5_B, suggesting the slide was not from R1_2. Finally, relative losses in chr5 were not present in Primary_A or Primary_B but a relative gain in chr5 was seen for R2 and R3 confirming that this R1 subclone had a copy number state of 2 for chr5 and belonged to R1_1. The consensus copy number profile was generated by cutting the hierarchical clustering tree of R1_1 subclone generated with the Ward.D2 method into two, selecting the cluster without the loss of part of chr1q (as this segment likely had a copy number state of 2 given no relative gains were seen in other samples) and taking the mode of the copy number state for each genomic bin (Appendix 8.1.30).

The mean relative inferCNV intensities for each cluster of spots was calculated and centred around zero for each segment, as defined by the shared breakpoints determined from scDNA-seq (see section 2.3.9). This mean was then normalised by the number of copies of R1 for that segment. This is required as a loss of 1 copy from 2 results in a 50% decrease in expression intensity, whereas the same loss from 3 copies would only be a 33% reduction, assuming a linear gene dosage effect. The normalised expression was then discretized into five different states: gain of one copy, gain of two copies, no difference, loss of one copy and loss of two copies. As these slides came from the same tumour and subclones only differ in copy number to a small degree, these five states would cover most of the relative differences from the R1_1 subclone. Inferred total copy number profiles were then calculated by adding the relative discretized gains or losses to the reference R1_1 subclone copy number profile (Appendix 8.1.31).

Total copy number profiles calculated for Visium consensus cluster profiles were then projected onto the scDNA-seq UMAP (see section2.3.9). Similarly, copy number profiles calculated for individual Visium spots from a slide were also projected to detect rare individual spots which are not represented when taking the consensus for a cluster of spots.

2.3.25 Spatial phylogenetic reconstruction from spatial transcriptomics

MEDICC2 was used to derive phylogenetic trees for each slide using the total copy number profiles inferred from Visium clusters. The phylogenetic tree was then overlaid on the slide as described in section 2.3.13 with the centroid of each cluster of spots used as its position for plotting.

2.3.26 Slide-seq spatial transcriptomics

Cryosections of tumour tissues from the R2, R3, and R5 were prepared in a similar fashion to those for LCM (see section 2.3.12). Tissue was placed on Slide-seq V2 sequenced pucks provided by the Rodriques lab and library preparation was performed as per the Slide-seq V2 protocol (Stickels *et al.*, 2021). Initial 100 base paired-end sequencing was performed on an Illumina Novaseq 6000 for two pucks per region, followed by deeper sequencing of a puck from R2 and R5.

The Snakemake pipeline Spacemake (0.4.3) was used to perform preprocessing and alignment of Slide-seq sequencing data to GRCh38 (Sztanka-Toth *et al.*, 2022). The cell barcode was read from bases 1 to 11 and the UMI from bases 15 to 23. The spot diameter was set to 10µm and the puck width to 3,000µm. Gene/cell count matrices were loaded in Seurat and underwent the same analyses as previously described (see section 2.3.22). Identification of tumour and normal spots was again attempted as described previously (see section 2.3.15).

2.3.27 Flow cytometry ploidy analysis

Small fragments of tumour tissue were placed in Pepsin-HCl solution (5 mg/ml pepsin in 10 mM HCl solution). A Bioruptor sonication system (Diagenode) was used to dissociate the tissue and incubated at 37˚C for 30mins. The solution was then spun down and washed with 1X PBS. 1 ml of 70% ethanol was added drop-wise whilst vortexing, then incubated at room temperature (RT) for 1 hour before a further spin and wash with 1X PBS. The sample was then incubated with propidium iodide (1 µg/ml) overnight at 37˚C.

Flow cytometry was performed on a LSR II Flow Cytometer (BD Biosciences). Nuclei were gated based on the FSC *vs*. SSC plot, then single cells were gated on the area *vs*. height plot for the yellow-green laser with 610/20 bandpass filter. Histograms were plotted for each region, revealing subpopulations with different ploidies. The ploidy for each peak was calculated by taking its median signal intensity value and dividing by the median of the diploid peak then multiplying by 2.

2.4 Detection of DTCs in late-stage disease

2.4.1 Patient samples

Patients were recruited to the Posthumous Evaluation of Advanced Cancer Environment (PEACE) study (NCT03004755) and provided their written informed consent for research autopsy and post-mortem sampling of tumour and normal tissues. This study was approved by the National Health Service Research Ethics Committee (13/LO/0972/AM05). Patients were introduced to the trial by their clinical team and their written informed consent was obtained. In rare cases where consent could not be taken from the patient, consent was given posthumously by the next-of-kin. Upon the patient's death, the study team are contacted by the family and the body is transferred for autopsy. In addition, two melanoma metastases collected from surgical resections for the TRACERx Melanoma trial were used for flow cytometry.

For this project, autopsies were attended by me and processing of tissues into single-cell suspensions was also performed by myself. Flow cytometry was performed by Cristina Cotobal Martin and Christie English, and single cells were sorted with the assistance of the Flow Cytometry STP. Library preparation was also performed by Cristina Cotobal Martin and Christie English. Immunohistochemistry (IHC) of tissue sections for carbonic anhydrase 9 (CAIX) and paired box gene 8 (PAX8) was performed by Anne-Laure Cattin and Sucheta Mahapatra. Fixation of samples for scRNA-seq was performed by me and sorting was performed with the assistance of the Flow Cytometry STP. Combinatorial indexing and library preparation for scRNA-seq was performed by Fiona Byrne. Sequencing was performed by the Advanced Sequencing Facility STP. All bioinformatics analysis was conducted by me alone.

2.4.2 Collection of tissues during research autopsy

Autopsies were performed at St Thomas's Hospital Mortuary, London by pathologists Dr. Ula Mahadeva and Dr. Anna Green. Prior to the autopsy, a medical report of the patient including their clinical history, treatment, and recent imaging scans, was prepared. This report guided the autopsy, with particular attention paid to known sites of metastasis informed by imaging. Tumour tissues were identified by the pathologist and dissected. For larger tumour tissues, multi-region sampling was carried out with images of the location of each sample taken. In addition to known metastases, all major organs, including the brain, heart, lungs, digestive tract, liver, kidneys, adrenals, pancreas, pituitary, thyroid, skin, muscle, fat, and lymphatic systems were examined. Each autopsy yielded up to (or occasionally more than) 100 samples depending on disease burden. A sample collection proforma is shown in Appendix 8.2.5.

As part of the PEACE sampling team, I aided in systematic sampling of tumour tissues and macroscopically normal tissues from all organs examined for over 50 autopsies. For each sample (normal or tumour), a piece of tissue approximately 1cm cubed was collected and bisected. Half was placed in a cryovial then snap-frozen in liquid nitrogen and the other half was fixed in formalin and inverted several times. Thus, fresh frozen and formalin fixed paraffin embedded (FFPE) samples can be traced back to the exact same piece of tissue. For larger pieces of tissue where there was value in keeping the tissue as a complete piece, "megablocks" were taken and fixed in formalin in larger containers. In addition, fluids such as pleural fluid, ascitic fluid and cerebrospinal fluid was also collected. In cases where germline blood could not be taken prior to death, blood was collected from the cardiac ventricles.

In addition to standard PEACE sampling, I collected a number of tumour or normal tissues in 1X PBS which were processed into single-cell suspensions immediately. For sampling of bone, a piece was taken from the vertebral strip performed during autopsy to detect spinal metastases. Furthermore, I also collected bone marrow aspirates (BMA) using a 13G Jamshidi needle (Argon Medical Devices) into a lithium heparin tube. After testing different sites for aspiration, I found the area of the ileum adjacent to the sacrum (accessed from inside as the donor could not be turned during autopsy) to be the most reliable and providing the largest volume of aspirate.

2.4.3 Dissociation of autopsy samples

Fresh samples were transported to the Francis Crick Institute on ice. Samples were removed from PBS and manually cut into approximately 2 mm cubed fragments. Tissue fragments were then placed into 50 ml Falcon tubes with 25 ml of digestion media (24.5 ml HBSS with no calcium, no magnesium (Gibco), 400 µL Collagenase Type II (25 U/µL, Gibco) and 300 µL 1 M Calcium Chloride (Thermo Fisher Scientific)). Tissue fragments were digested on a horizontal shaker at 200 RPM 37°C for 45 minutes. Samples filtered through a 100 µM filter and spun down at 500G for 10 minutes at 4°C. The supernatant was removed, and 2 ml of red blood cell (RBC) lysis buffer (Thermo Fisher Scientific) was added for 10 minutes and incubated at room temperature. The sample was then quenched with 10 ml of wash buffer (1X PBS with 2%FBS and 2 mM EDTA) and spun down. After two further washes with wash buffer, cells were stained with trypan blue, and the number of cells and viability was recorded. Cells were then resuspended in 1 ml of Recovery Cell Culture Freezing Medium (Thermo Fisher Scientific) and frozen at −80°C. For lung and bone samples, significant cell clumping was noted after digestion, resulting in loss of most of the samples. Addition of 1 mg/ml DNase I (Sigma Aldrich) to the digestion media prevented this.

For three patients, pieces of normal pancreas were also collected, and single-cell suspensions were generated with the same protocol. However, due the high degree of autolysis, these samples featured almost no viable cells so were not utilised and further normal pancreas samples were not collected.

2.4.4 Processing of bone marrow aspirate samples

BMA samples were diluted 1:2 in 1X PBS and passed through a 100 µM filter. SepMate PBMC Isolation Tubes (Stemcell technologies) were loaded with 3.5 ml of Ficoll-Paque Plus (Sigma Aldrich) and the diluted BMA was then slowly loaded into the tube. Tubes were spun at 1200G for 40 minutes at room temperature with the brake on. The supernatant was removed, and the layer of mononuclear cells was collected. Collected cells were washed with wash buffer (1X PBS + 2% FBS) by spinning at 500G for 5 minutes at 4°C and discarding the supernatant. The pellet was resuspended in 5 ml of RBC lysis buffer (Thermo Fisher Scientific) for 10 minutes at room temperature. Following another two washes, cells were resuspended in Recovery Cell Culture Freezing Medium (Thermo Fisher Scientific) and frozen at −80°C.

2.4.5 Flow cytometry detection and sorting of marker-positive cells

Single-cell suspensions were thawed at 37°C with 1 ml of 1X PBS added dropwise on ice with swirling. A further 3 ml of 1X PBS was added and the suspension was centrifuged at 500G for 5 minutes at 4°C. After removal of the supernatant, the cell pellet was resuspended in 5 mL of wash buffer (1x PBS + 2% FCS + 2mM EDTA) and centrifuged again at 500G for 5 minutes at 4°C. Following this, the cells were resuspended in 1 mL of wash buffer filtered into FACS tubes. The phycoerythrin (PE) conjugated anti-CAIX primary antibody (REA658, Miltenyi Biotec) and the Brilliant Violet 421 (BV421) conjugated anti-EpCAM antibody (324220, BioLegend) were added to the FACS tubes and incubated for 30 minutes in the dark in at 4°C. 1 mL of wash buffer was added to the suspension before spinning for 5 minutes at 500G and 4°C. The cells were then washed with 500 µL of 1X PBS and spun again. Finally, the cells were resuspended in 500 µL of wash buffer and filtered into FACS tubes, where DRAQ7 was added for live/dead staining where required.

Samples were analysed on a LSRFortessa Flow Cytometer (BD Biosciences). PE was detected by the yellow laser with a 582/15 bandpass filter, BV421 was detected by the violet laser with a 450/50 bandpass filter and DRAQ7 was detected by the red laser with a 780/60 bandpass filter.

Sorting of cells was performed on a FACSAria Fusion Cell Sorter (BD Biosciences) or a FACSAria III Cell Sorter (BD Biosciences) with a 100 µm nozzle depending on availability. Cells were sorted into 96-well PCR plates with each well filled with 1 µL of 1X PBS. Plates were spun down and stored at −80°C. Flow cytometry against melanoma-associated chondroitin sulphate proteoglycan (MCSP) (PE conjugated, 562415, BD Biosciences) and CD146 (fluorescein isothiocyanate (FITC) conjugated AB_1210462, Thermo Fisher Scientific) was performed using the same protocol as above with the corresponding primary antibodies.

For flow cytometry ploidy analysis, nuclei were prepared from samples and cells lines, following a recently published protocol (Slyper *et al.*, 2020). Frozen melanoma metastasis samples were homogenized in 1 mL of 1X Tween with salts and Tris (TST) buffer on ice, followed by filtration through a 40 µm cell strainer and further wash with TST into a 50 mL tube. The suspension was then centrifuged for 5 minutes at 500G and 4°C, and the resulting pellet was resuspended in 500 µL of 1X salt-Tris buffer. Following further centrifugation for 5 mins at 500G and 4°C, the pellet was resuspended in 500 µL of 1X PBS + 0.5% BSA, followed by filtration through a 40 µm Flowmi tip filter. Hoechst stain (a non-intercalating DNA binding dye) was added to the suspension and stained for 30 mins at 4°C. After centrifugation for 5 mins at 500g and 4°C, the pellet was resuspended in 500 µL of PBS +2%BSA filtered and transferred to FACS tubes.

Flow cytometry of melanoma metastases and cell line nuclei was performed using a LSRFortessa Flow Cytometer (BD Biosciences). Ploidy analysis was performed as described in section 2.3.27, but using the UV laser and a 450/50 bandpass filter.

2.4.6 Whole-genome amplification and library preparation of single cells

The first pilot cohort of cells were amplified using the Ampli1 WGA kit (Menarini Silicon Biosystems) following the suggested protocol. However, although amplification of most cells produced DNA, these were subsequently found to be derived from *E. coli*. This is likely because enzymes in this kit are made in *E. coli* and *E. coli* DNA present in the reagents is being amplified. Therefore, subsequent WGA of single cells was performed with Genomeplex WGA4 kit (Sigma-Aldrich).

Amplification was performed on cells sorted into 96-well PCR plates with 1 µl of PBS per well. Additional UltraPure DNase and RNase free water (Invitrogen) was added to the single-cell sample for a final volume of 9 µL. A working Lysis and Fragmentation Buffer Solution was prepared by adding 2 µL of Proteinase K Solution into 32 µL of the 10X Single Cell Lysis & Fragmentation Buffer. 1 µL of the working Lysis and Fragmentation Buffer Solution was then added to the single-cell sample. This DNA mix was incubated at 50°C for 1 hour then heated to 99°C for four minutes in a MiniAmp thermal cycler (Thermo Fisher Scientific).

Library preparation was performed by adding 2 µL of 1X Single Cell Library Preparation Buffer and 1 µL of Library Stabilization Solution to each sample. Samples were then brought to 95°C for 2 minutes. 1 µL of Library Preparation

Enzyme was then added to the sample, mixed thoroughly, and incubated in the thermal cycler at 16 C for 20°minutes, 24°C for 20 minutes, 37°C for 20 minutes, 75°C for 5 minutes then held at 4°C.

Amplification was performed by adding 7.5 µL of 10X Amplification Master Mix, 48.5 µL of Water, Molecular Biology Reagent and 5.0 µL of WGA DNA Polymerase to the entire 14 µL reaction then mixing thoroughly. The sample was then denatured at 95°C for 3 minutes, then underwent 25 cycles of denaturation at 94°C for 30 seconds and annealing/extension at 65°C for 5 minutes in the thermal cycler.

A 1.5% agarose gel was then used to assess the products of WGA by loading 7 µL of the final reaction alongside a 1kb DNA ladder. For cells successfully amplified, DNA purification was performed with the GenElute PCR Clean-Up kit (Sigma-Aldrich).

Library preparation was performed using NEBNext Ultra II FS DNA Library Prep kit (New England BioLabs) according to the protocol and 100 base paired-end sequencing was performed on a HiSeq 4000.

2.4.7 Genomic profiling of single cells

A newer version of ASCAT.sc (v1.0) compared to the MPNST project was used to call total copy number for single cells, but largely features the same functionality (see section 2.3.9). As matched haplotype information was not available, allele-specific calling was not performed. However, 1000 Genomes Project SNP positions could be genotyped in each cell using alleleCount (v4.0.0) and SNP positions with both alleles detected could be identified, suggesting heterozygosity.

2.4.8 Dual IHC staining of CAIX and PAX8

CAIX-PAX8 double staining was performed on FFPE slides. FFPE slides were heated in an oven at 60°C for 1 hour then deparaffinized using a Sakura Tissue-Tek Prisma 6130 machine. Slides were then placed in distilled water before performing the antigen retrieval step by adding 1X EDTA buffer and heating in a microwave at maximum power (1000W) for 23 minutes. After 10 minutes at room temperature to cool down, slides were placed under running water for a further 10 minutes.

A hydrophobic (ImmEdge) pen was then used to mark around the sample for each slide, and TBS (50 mM Tris HCL pH 7.5, 150 mM NaCl) was added immediately to keep it hydrated. Following marking, the TBS was drained, and slides were placed in a humidified chamber. Slides were incubated with 1X BSA solution diluted in TBS with 0.3% triton for 30 minutes and 150-200 µl of the primary antibodies were added. The antibodies for CAIX (ab128883, abcam) and PAX8 (ab124445, abcam) were diluted 1:6000 and 1:20 respectively. The slides were incubated overnight at 4°C.

On the following day, slides were washed twice in TBS for 5 minutes, incubated with 3% H_2O_2 (diluted 30% H_2O_2, 1:10) for 15 minutes, and then washed twice in TBS for 5 minutes. Slides were then washed with diluted 1X BSA for 5 minutes. The secondary antibody for PAX8 (HRP anti-mouse, MP-7422, Vector Laboratories) was added and incubated for 1 hour. Slides were then washed three times in TBS for 5 minutes before the DAB (3, 3'-diaminobenzidine) reaction was performed by adding the DAB solution. After 15 minutes, slides were rinsed in distilled water then washed in TBS for 5 minutes.

For CAIX staining, the AP anti-rabbit secondary antibody (MP-5401, Vector Laboratories) was added, and slides were incubated for 1 hour. After a further three washed in TBS for 5 minutes, the alkaline phosphatase reaction was performed by adding the AP solution to the slides for 12 minutes. Slides were rinsed in distilled water then taken to a Sakura Tissue-Tek Prisma 6130 machine for haematoxylin counterstaining. Some slides underwent single staining with CAIX using the above protocol but without the steps for staining PAX8.

2.4.9 Fixation and sorting of CAIX$^+$ cells

For scRNA-seq, single-cell suspensions were thawed at 37°C with 1 ml of 1X PBS added dropwise on ice with swirling. In order to enrich for viable cells, the Dead Cell Removal kit (Miltenyi Biotec) and LS Columns (Miltenyi Biotec) were used. Fixation of the viable fraction of cells was performed according to protocol for the V2 Cell Fixation kit (Parse Biosciences). After fixation, cells were stained overnight at 4°C in the dark with the same anti-CAIX antibody as in previous flow cytometry experiments (REA658, Miltenyi Biotec). Cells were washed and sorting into Eppendorf tubes containing the included Cell Buffer on a FACSAria Fusion Cell Sorter (BD

Biosciences) or an Influx Cell Sorter (BD Biosciences) with a 100 µm nozzle depending on availability. Barcoding of samples, amplification and library preparation was performed accord to the Evercode Whole Transcriptome Mini kit protocol (Parse Biosciences). 100 base paired-end sequencing of one sub-library containing half of the barcoded cells was performed on a HiSeq 4000.

2.4.10 scRNA-seq of autopsy samples

The Parse Biosciences pipeline (v1.0.3p) was used in a conda environment to identify barcodes for single cells, pre-process reads, and align them to GRCh38. The gene/cell count matrix was loaded into Seurat (v4.2.1) with cells with less than 300 genes and more than 30% mitochondrial reads removed. Counts were then normalised, and the top 2000 variable genes were used to perform PCA. The first 30 principal components were used for UMAP dimensionality reduction and clustering of cells was performed with a resolution of 0.3. Differentially expressed genes for each of the 5 clusters was performed and these were used along with canonical markers in the literature to identify cell types.

De-multiplexing of samples was also attempted with souporcell (v2.0) in a singularity container using 1000 Genomes Projects SNP positions with allele frequency greater than 2% in the population. However, the number of cells per sample was too low to achieve this with most of the cells unassigned to donors.

2.5 Detection of DTCs in early stage ccRCC

2.5.1 Patient samples

Patients were recruited to the bone marrow sub-study in addition to the main TRACERx Renal study (NCT03226886) and provided their written informed consent for bone marrow sampling during surgery. This study was approved by the National Health Service Research Ethics Committee with the bone marrow sub-study introduced by me approved in amendment 10 (11/LO/1996/AM10). Upon being listed for nephrectomy, or rarely metastatectomy, patients were contacted for the TRACERx Renal study. If the patient consented to the main study, they were also given the option to enrol for the bone marrow sub-study. Patients were provided with a patient information sheet written by myself, given time to consider the sub-study and signed the consent form for the sub-study (Appendix 8.3.5 and 8.3.6).

In addition, single-cell suspensions collected for TRACERx Renal from normal kidney, kidney primary tumour and pancreatic metastases were also kindly provided by Geoffrey Feng and Daqi Deng. These samples were used as positive and negative controls to optimise detection and profiling of cells.

For this project, bone marrow aspirations were performed by myself in theatre and isolation of mononuclear cells was also performed by myself. Flow cytometry was performed by Cristina Cotobal Martin and myself and single cells were sorted with the assistance of the Flow Cytometry STP. Immunocytochemistry (ICC), micromanipulation and WGA of single cells was performed by me. Library preparation and sequencing was performed by the Advanced Sequencing Facility STP. All bioinformatics analysis was conducted by myself.

2.5.2 Peri-operative collection of bone marrow aspirates

Following theatre briefing and induction of anaesthesia, the patient was placed on their left or right lateral decubitus position. Up to 40 ml of BMA was taken from the posterior iliac crests using a 15G Jamshidi needle (TIN3015, BD) and transferred quickly into blood bottles containing lithium heparin and inverted to prevent clotting of the sample. The guidance document implemented for bone marrow aspirate collection is shown in Appendix 8.3.7. BMAs were transported on ice to the Crick and mononuclear cells were isolated as described in section 2.4.4.

2.5.3 Flow cytometry detection and sorting of marker positive cells

Single-cell suspensions were stained and analysed with flow cytometry for CAIX with the same protocol as described in section 2.4.5. For PAX8, permeabilization was required for staining and was carried out with the BD Intrasure Kit (BD Biosciences) and the Intracellular Fixation & Permeabilization Buffer Set (Thermo Fisher Scientific), according to their protocols. Despite staining with the AlexaFluor 350 conjugated primary antibody against PAX8 (NBP3-08274F350, Novus Biologicals) or unconjugated mouse primary antibody (ab53490, abcam) and the AlexaFluor 405 conjugated goat anti-mouse secondary antibody (ab175660, abcam), no stained cells were detected for normal kidney or kidney tumour positive controls.

2.5.4 Immunocytochemistry detection of PAX8-positive cells

Cytospins of single-cell suspensions were prepared to enable IHC staining of cells. Briefly, 20,000 – 40,000 cells were deposited into Cytospin funnels held against SuperFrost Plus slides (Fisher Scientific). The Shandon CytoSpin III was then run on setting 1 and spun at 800 rpm for 7 minutes. The area of cells was marked with a hydrophobic pen and cells were fixed with 80% methanol at −20°C for 15 minutes, followed by two washes in 1X PBS. The slide was kept in a humidified box from this point onwards until staining was complete. The primary antibody against PAX8 (ab53490, abcam) was then added and incubated at 4°C overnight. Following another two washes in 1X PBS, the horse anti-mouse horseradish peroxidase (HRP) secondary antibody (MP-7422, Vector Laboratories) was then added for 30 minutes. After another two washes in 1X PBS, staining was performed with AEC (3-amino-9-ethylcarbazole) solution for 10 minutes, followed by a final wash in 1X PBS. AEC was chosen as the peroxidase substrate as DAB has been shown to interfere with DNA polymerase (Dölle *et al.*, 2018). RPTEC cells displayed strongly positive staining with MCF-7 cells not staining as expected (Appendix 8.3.2).

As cells on Cytospin slides could not be easily dislodged and collected for downstream analysis, ICC of cells in solution was also performed in a similar fashion to staining in preparation for flow cytometry. After thawing, cells were washed in 5 ml of 1X PBS and spun at 500G for 5 minutes at 4°C. BLOXALL Blocking Solution (Vector Laboratories) was used to prevent endogenous peroxidase activity which was seen in BMA samples. Following a wash with 1 ml of 1X PBS, staining was performed in FACS tubes with the same antibodies and incubation times used as above. A further two washes were performed before AEC solution was added for colour development followed by a final wash. Positive staining RPTEC cells were once again seen, although the staining was distributed throughout the cell in contrast to the nuclear localisation seen on Cytospins (Appendix 8.3.2).

2.5.5 Isolation of cells through micromanipulation

PAX8$^+$ cells were transferred from FACS tubes to Superforst microscope slides (Fisher Scientific) and visualised under a light microscope. A TransferMan NK 2 (Eppendorf) micromanipulator was used to isolate single cells with Biopsy Tip I microcapillaries (Eppendorf), which have an internal diameter of 19 µm as described in the literature (Mathiesen *et al.*, 2012; Sanchez-Luque *et al.*, 2017). After a single positive cell was extracted, the cell was deposited in 1 µl of 1X PBS on a clean slide. The PBS containing the cell was then aspirated with a 1 µl pipette and transferred into a PCR tube, which was spun at 500G for 1 minute and stored at −20°C until WGA was performed.

2.5.6 Genomic profiling of isolated cells.

WGA and library preparation of isolated cells was performed as described in section 2.4.6. Copy number calling was performed using ASCAT.sc (v1.0) as described in section 2.4.7.

Chapter 3. Exploring chromosomal instability in a malignant peripheral nerve sheath tumour

3.1 Introduction

MPNSTs or malignant schwannomas, are sarcomas of neural origin. Although rare, these tumours are highly aggressive, and management represents a huge challenge.

3.1.1 Clinical features of MPNST

MPNSTs are a rare type of soft tissue sarcoma comprising 2% of all sarcomas. They are believed to be derived from the Schwann cells or neural crest progenitors of peripheral nerves (Sun *et al.*, 2021). Around half of all cases occur in the context of the condition neurofibromatosis type 1 (NF1). NF1 is a common genetic disorder caused by inherited or spontaneous loss of the *NF1* gene on chromosome 17q (Gutmann *et al.*, 2017). In addition to cutaneous hyperpigmentation and numerous cutaneous neurofibromas, patients can develop more troublesome plexiform neurofibromas. These larger lesions can cause pain or tissue damage due to physical compression and have a lifetime risk of 9-13% of transformation into an MPNST (Evans *et al.*, 2012).

Histologically, MPNSTs are composed of uniform spindle-shaped cells with enlarged hyperchromatic nuclei and high mitotic counts. They display alternating hyper- and hypocellular regions and focal areas of necrosis (Farid *et al.*, 2014). Although there are no pathognomonic IHC markers, due to their neural crest origin, S-100 and SOX10 have been used. In addition, a subset of MPNSTs can undergo heterologous rhabdomyoblastic differentiation (when cartilage, bone or skeletal muscle cells are present) and are known as malignant triton tumours.

Treatment options for MPNSTs are limited with surgical resection of the primary tumour as the mainstay of treatment in localised disease (Dunn *et al.*, 2013). However, due to their size and proximity to important structures, adequate resection margins can be difficult to achieve, and recurrence rates are high. Furthermore, MPNSTs display high metastatic potential and commonly spread to the lungs and bones. Whilst adjuvant radiotherapy is used to prevent recurrence for larger high-grade lesions (>5cm) or those with positive resection margins, there is no evidence that it improves overall survival in these cases (Kahn *et al.*, 2014). Chemotherapy with agents typically used in other soft-tissue sarcomas such as doxorubicin and ifosfamide, is largely reserved for the metastatic setting but again, provides minimal benefit (Farid *et al.*, 2014). Disappointingly, small molecule inhibitors have not shown any activity although immunotherapy has recently begun to be explored for the treatment of MPNST.

Overall, the limited effectiveness of treatment options results in a poor prognosis, with a dismal 5-year overall survival rate of 26-39% (Dunn *et al.*, 2013). Therefore, there is a need to study these rare tumours to inform the development of new therapies and improve treatment outcomes.

3.1.2 Genomics of MPNST

Genomic profiling of MPNSTs, although limited in scale in date, has shown that these tumours are genomically complex (Pemov *et al.*, 2020). Candidate gene studies have identified several genes and signalling pathways which are frequently mutated or lost in MPNSTs, such as *NF1*, the p53 pathway (*TP53* and *MDM2*) and *CDKN2A/B*. As our knowledge of MPNSTs has progressed, the ordering of events in MPNST pathogenesis has also been illuminated through next-generation sequencing studies. In patients with NF1, loss of the remaining wild-type *NF1* allele is widely accepted as the key contributing event for the formation of benign plexiform neurofibromas. Subsequent deletions of *CDKN2A/B* result in premalignant atypical neurofibromas and are typically followed by other mutations driving malignant transformation. These driver mutations have been found in the p53, RB and epidermal growth factor (EGF) pathways with biallelic loss in either *TP53* or *CDKN2A* present in nearly all MPNSTs.

In contrast to the relatively modest number of point mutations observed, array CGH studies have found high levels of CNAs with every chromosome found to display numerical or structural abnormalities across MPNSTs. A higher frequency of gains compared to losses has been observed with MPNSTs typically exhibiting a high ploidy or even in rare cases, near-tetraploidy. However, there does not appear to be any chromosomal gains or losses which are specific to, or are recurrently found in MPNSTs, except for patients with NF1 who are more likely to display CNAs on chr17.

Epigenetic modifications have also been identified in MPNST with loss of function mutations in the polycomb repressive complex 2 (PRC2) components *EED* and *SUZ12* shown to be critical drivers in MPNST pathogenesis (Lee *et al.*, 2014). The PRC2 complex primarily methylates H3K27me3, therefore, its inactivation results in the loss of H3K27me3. The loss of this repressive epigenetic mark leads to aberrant transcriptional activation of several developmental pathways and is associated with poor prognosis (Lyskjær *et al.*, 2020). This is in contrast to most other tumour types where H3K27me3 is intact and is exploited by cancers cells in transcriptional repression of tumour suppressors.

Recently, a large in-depth genomics study of 90 MPNSTs was performed describing their genomic landscape and evolutionary history (Cortes-Ciriano *et al.*, 2023). 95 samples from these 90 tumours underwent WGS, RNA-seq and DNA methylation array profiling. This confirmed landmark events found in previous studies such as biallelic inactivation of *NF1* in 88% percent of MPNSTs arising on a background of NF1 through loss of the remaining wild type allele. Other drivers previously described, such as *TP53*, *CDKN2A* and the PRC complex genes, were also found to be frequently mutated. Mutational signature analysis found the clock-like signature SBS5 to be the predominant signature. WGD was found in 73% and 59% of tumours with or without H3K27me3 loss, respectively. Furthermore, mutation timing analysis confirmed that *CDKN2A* is an early event, whereas WGD is consistently a late event.

Although there was a high burden of CNAs across the cohort, recurrent CNAs were identified specifically in a subset of cases with H3K27me3 loss. Alterations included high levels of chromosome 8 amplification and LOH of chromosomes 1, 10, 11, 16, 17 and 22. These CNA patterns were also detectable in cfDNA and could form the basis of a prognostic biomarker by informing H3K27me3 status. Finally, a model of MPNST evolution was proposed for the two subtypes of MPNSTs based on H3K27me3 status (Figure 3.1). Following inactivation of the PRC2 complex, H3K27me3 deficient tumours evolve by acquiring different degrees of genome-wide LOH and gains of chromosome 8 with WGD occurring late in tumour evolution. In contrast, MPNSTs retaining H3K27me3 do not develop recurrent CNAs and evolve through ongoing chromosomal instability.

Overall, MPNSTs are tumours with high levels of genome instability and display unique aberrations in their genomes. An understanding of their evolution, mechanism of growth and phenotypic effects of these aberrations will be critical for developing new treatments and improving survival.

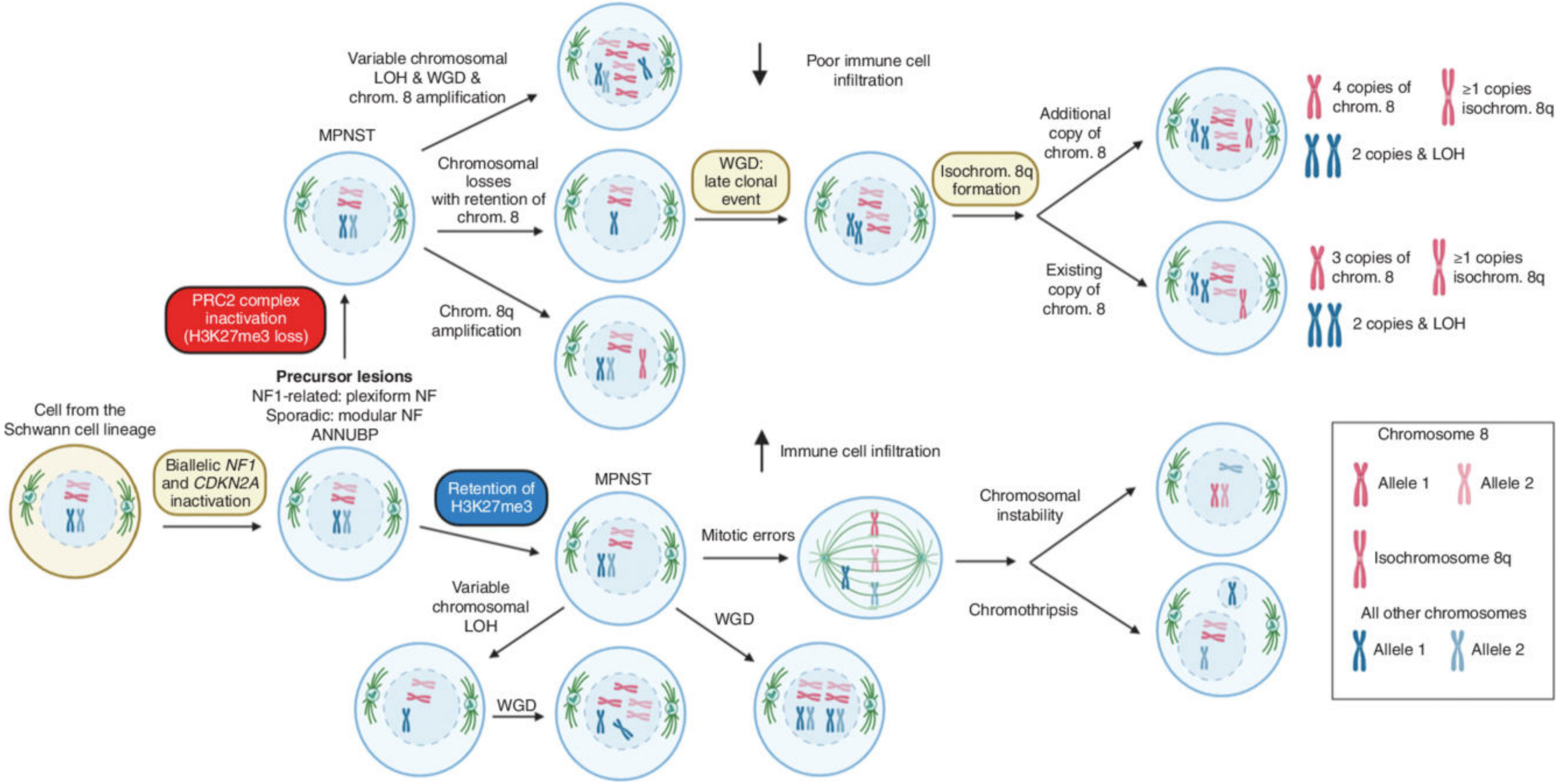


Figure 3.1 Evolutionary pathways of MPNST. Schematic showing the sequence and timing of key events which reoccur in MPNST evolution. Figure reproduced from Cortes-Ciriano *et al*., 2023, licensed under CC BY-NC-ND 4.0.

3.2 Aims

Given the extensive CIN and copy number changes seen in MPNSTs, single-cell and spatial genomic methods which can detect CNAs were applied to profile intra-tumour heterogeneity in ultra-fine detail. Specifically, our aims were to:

- Study tumour evolution and intra-tumour heterogeneity down to single-cell resolution.
- Explore the limit of resolution of different genomic methods for phylogenetic reconstruction.
- Localise specific tumour subclones and retrace evolution in a spatial manner both through genomics and transcriptomics.

3.3 Clinical history

The index case was a 36 year old male who exhibited an NF1 phenotype clinically and subsequently developed an MPNST on the right forearm. This tumour was found to be a triton tumour given the presence of rhabdomyoplastic differentiation at histological review. Immunohistochemical staining revealed loss of H3K27me3. Despite surgical resection of the primary tumour and receiving 6 cycles of adjuvant doxorubicin and ifosfamide, the patient developed a multifocal local recurrence and underwent palliative resection involving amputation of the right arm.

The 5 recurrence regions were sampled immediately at surgery and a fresh frozen sample of primary tumour was retrieved from tissue archives. All samples underwent multi-region bulk and single-cell and spatial multi-omics analysis (Figure 3.2). An overview table of the different techniques applied to all samples and the number of cells or spots profiled is shown in Table 3.1.

Results from genomic methods (bulk WGS, scDNA-seq and LCM) are described in this chapter with results from methods profiling the transcriptome (scRNA-seq, G&T-seq, Visium spatial transcriptomics and Slide-seq) and integration of data modalities described in Chapter 4.

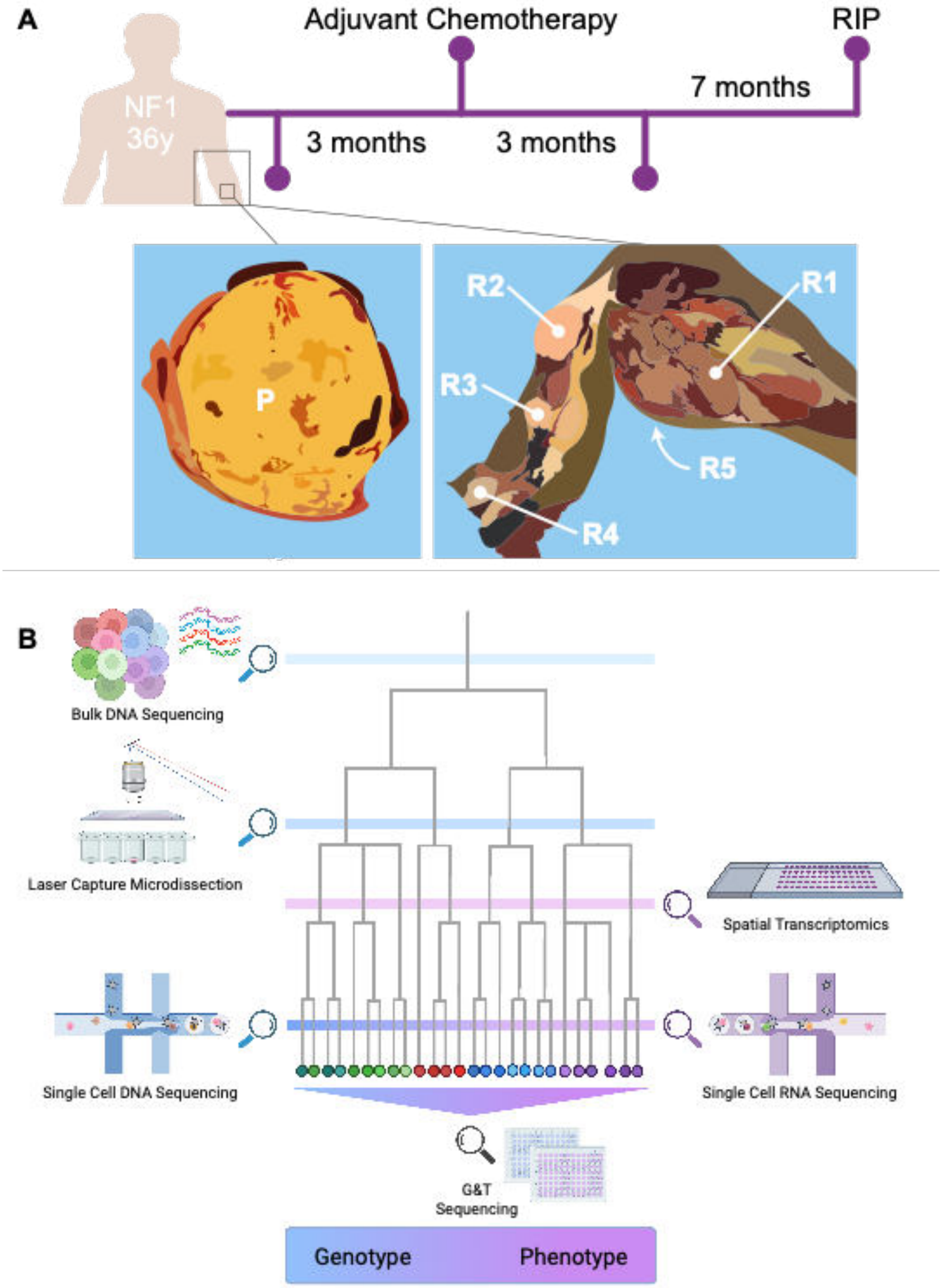


Figure 3.2 Study design. A. Clinical history of patient with MPNST with adapted images showing the location of the primary tumour and those of the recurrence regions sampled from the amputated arm. B. Techniques applied to samples with a range of resolution and features measured. Panel B created with Biorender.com.

Technique Region	Bulk WGS	10X scDNA-seq	LCM			10X scRNA-seq	G&T-seq		10X Visium	Slide-seq
			Front	Side	Back		scDNA-seq	scRNA-seq		
Primary	✓	612	14	12	11	4019	38	17	✓ ✓	-
R1	✓	1421	13	13	12	5049	0	73	✓	-
R2	✓	996	-	-	-	6758	34	46	✓	✓
R3	✓	82	-	-	-	5153	35	51	✓	✓
R4	✓	265	10	7	9	8608	56	54	✓	-
R5	✓	1032	-	-	-	8129	53	59	✓ ✓	-
Germline Blood	✓	-	-	-	-	-	-	-	-	-

Table 3.1 Overview of techniques and samples utilised in this study. For single cell experiments and laser capture microdissection (LCM), the number of cells/spots passing QC is shown. For Visium and Slide-seq, the number of ticks refers to the number of slides/pucks profiled.

3.4 Phylogenetic reconstruction using bulk WGS

All samples underwent bulk whole-genome sequencing to a median depth of 54X (Range 46 to 67) and a matched germline blood normal was sequenced to a median depth of 63X.

3.4.1 Subclonal allele-specific copy number calling

For bulk phylogenetic reconstruction, the purity, ploidy, and copy number profiles of the sampled regions must first be determined. Ploidies estimated with Battenberg ranged from 2.50 to 2.86 across the different regions (Table 3.2). These ploidy values were confirmed with flow cytometry analysis of ploidy experiments (Appendix 8.1.1). All recurrence regions were highly pure in contrast to the primary which had a purity of ~60%.

Region	Battenberg Purity	Battenberg Ploidy	Flow Cytometry Ploidy
Primary	58.1%	2.50	2.65
R1	76.5%	2.60	2.45
R2	96.2%	2.62	2.60
R3	91.5%	2.82	2.62
R4	84.5%	2.79	2.52
R5	81.6%	2.86	2.57

Table 3.2 Purity and ploidy values from Battenberg and flow cytometry.

Allele-specific subclonal CNA profiles were derived for each region from bulk WGS (Figure 3.3 and Appendix 8.1.2). Of note, there was near-genome-wide LOH (>91.7%) with high levels of chromosome 8 amplifications (up to 10 copies in total). Subclonal copy number heterogeneity was detected in all regions and there was significant inter-regional heterogeneity. Finally, there was evidence of multiple breakpoints and changes in copy number in chromosome 7p, highlighting possible chromothripsis.

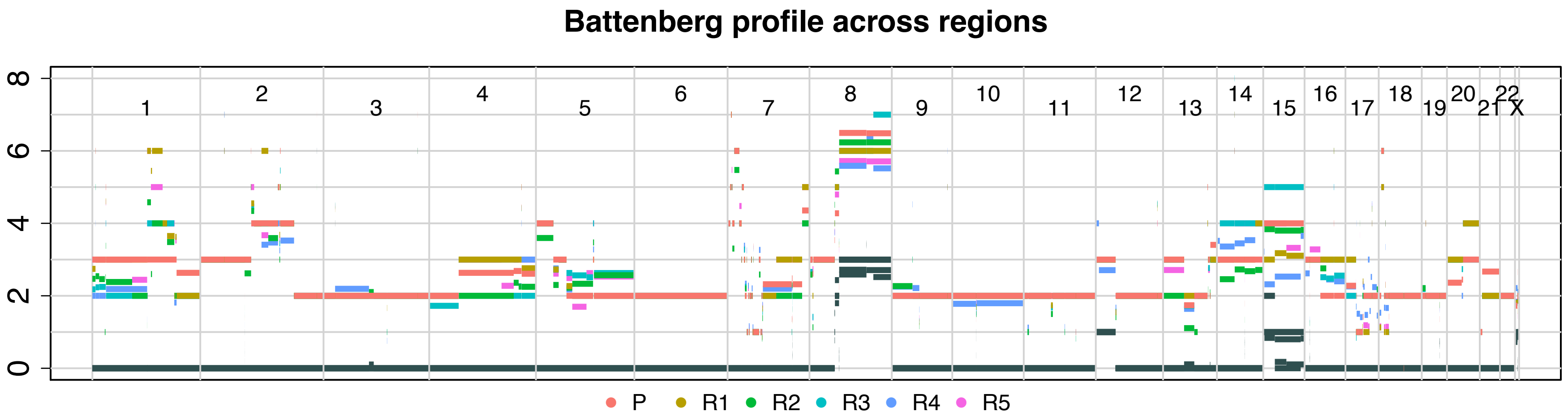


Figure 3.3 Subclonal copy number profiles of all regions. Copy number values for all chromosomes with major alleles coloured by region and the minor alleles in teal. Non-integer copy number values indicate the presence of a subclone with the size of each subclone proportional to the distance between integer states.

3.4.2 SNV-based subclone detection

16,330 SNVs were identified in bulk WGS across all regions with significantly more detected in R4 than in other recurrence regions. The computed CCFs of each SNV was used to identify clusters of SNVs using ndDPClust. SNVs with a similar CCF across samples are clustered together and represent specific clones or subclones. Truncal (those belonging to the MRCA for all tumour samples profiled), exclusive clonal (belonging to clone present in just one region) and subclonal mutation clusters were detected in all regions (Figure 3.4 and Table 3.3). In addition, shared clonal clusters were identified for R1 and R5; and R2, R3 and R4 suggesting these clones shared a common ancestor.

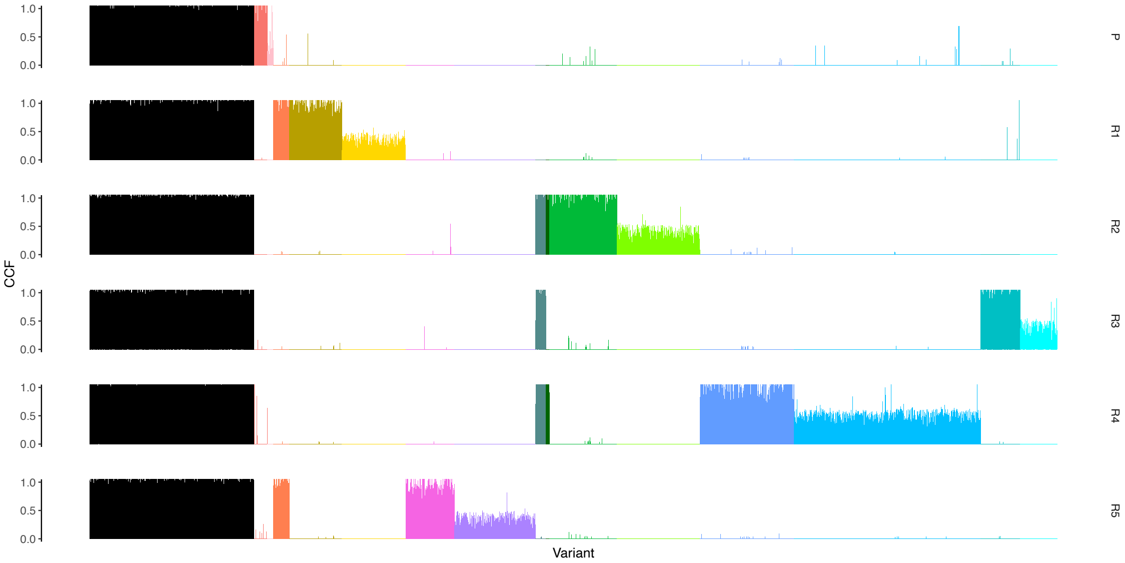


Figure 3.4 Truncal, clonal and subclonal mutation clusters. CCFs of each cluster of SNVs across regions. Truncal mutations with CCF of 1 in all samples shown in black. Shared clonal clusters for R1 and R5; and R2, R3 and R4 are shown in orange and dark green respectively.

3.4.3 Bulk tree reconstruction

The phylogeny of this tumour at the bulk level was reconstructed manually. First, each cluster of SNVs identified by ndDPClust was assigned a name based on the CCF of its SNVs across regions (Table 3.3). By assessing the CCF of each of these clusters across samples and applying the “pigeon-hole principle”, I manually reconstructed an evolutionary tree for this tumour to the resolution of a single subclone for each sample (Figure 3.5). For example, the subclonal 300000 cluster (CCF 0.80) identified in R4 must be a daughter of the X00000 cluster of R1 (CCF 0.25) as their CCFs sum to greater than that of its direct ancestor X000X0 (CCF 0.87) and so cannot be siblings.

The shared clonal clusters revealed that whilst one tumour lineage seeded the recurrences of regions R1 and R5, a sibling lineage gave rise to R2, R3 and R4, representing two major branches of the phylogenetic tree. Interestingly, the phylogenetic relation of these sibling clones recapitulated their spatial proximity on the arm as R1 and R5 were sampled from the forearm and R2, R3 and R4 were sampled from the upper arm. In addition, reconstruction with a copy number-based maximum parsimony method also resulted in a sample tree which mostly agreed with the SNV-based subclone tree (Appendix 8.1.4).

Cluster	R1	R2	R3	R4	R5	P	No. SNVs
XXXXXX	1.02	1.08	1.05	1.12	1.01	1.04	1993
00000X	0.00	0.00	0.00	0.02	0.01	0.83	161
000003	0.00	0.00	0.00	0.00	0.00	0.26	74
X000X0	0.87	0.00	0.00	0.00	0.86	0.00	194
X00000	0.80	0.00	0.00	0.00	0.00	0.00	638
300000	0.25	0.00	0.00	0.00	0.00	0.00	772
0000X0	0.00	0.00	0.00	0.00	0.77	0.00	593
000030	0.00	0.00	0.00	0.00	0.25	0.00	983
0XXX00	0.00	0.95	0.94	1.01	0.00	0.00	125
0X0X00	0.00	1.10	0.00	1.13	0.00	0.00	41
0X0000	0.00	0.90	0.00	0.00	0.00	0.00	822
030000	0.00	0.28	0.00	0.00	0.00	0.00	1005
000X00	0.00	0.00	0.00	0.89	0.00	0.00	1145
000400	0.00	0.00	0.00	0.38	0.00	0.00	2260
00X000	0.01	0.00	0.90	0.00	0.00	0.00	481
003000	0.00	0.00	0.29	0.00	0.00	0.00	447

Table 3.3 CCFs and number of SNVs in each cluster across regions. Clusters are named by their CCF in each region. For example, XXXXXX denotes the truncal SNV cluster present in 100% of tumour cells in all regions and 000003 denotes a subclonal SNV cluster present in 30% of cells in the primary.

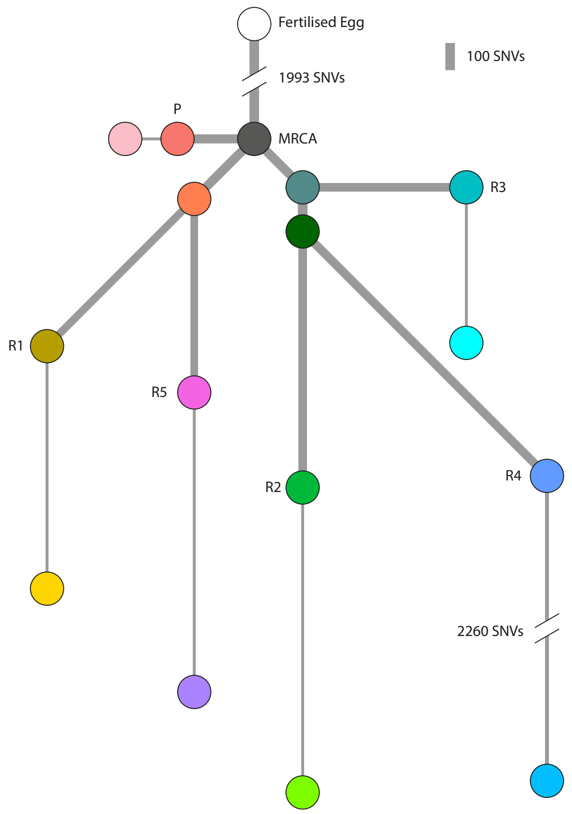


Figure 3.5 Phylogenetic tree from bulk WGS. Length and width of branches is proportional to the number of SNVs and CCF of subclones respectively.

3.4.4 Mutational signatures of subclones

To examine mutational processes active during this tumour's development, I performed fitting of SNVs of each SNV cluster to mutational signatures detected in an external cohort of MPNSTs. This revealed the clock-like signature of SBS1, reactive oxygen species associated SBS18, and large contributions by signatures with unknown mechanism such as SBS5, 8, and 39 (Figure 3.6). There did not appear to be any obvious patterns of signatures exposure between subclones. In addition, SBS35, which is associated with platinum therapy was detected. Although the patient was not known to be treated with platinum agents, this signature was not active in the truncal cluster but present in all descendant clusters. Therefore, the exposure timing could be in keeping with doxorubicin or ifosfamide which may result in this mutational signature through a similar mechanism of action. Alternatively, this could be explained by incomplete clinical annotation although this is unlikely.

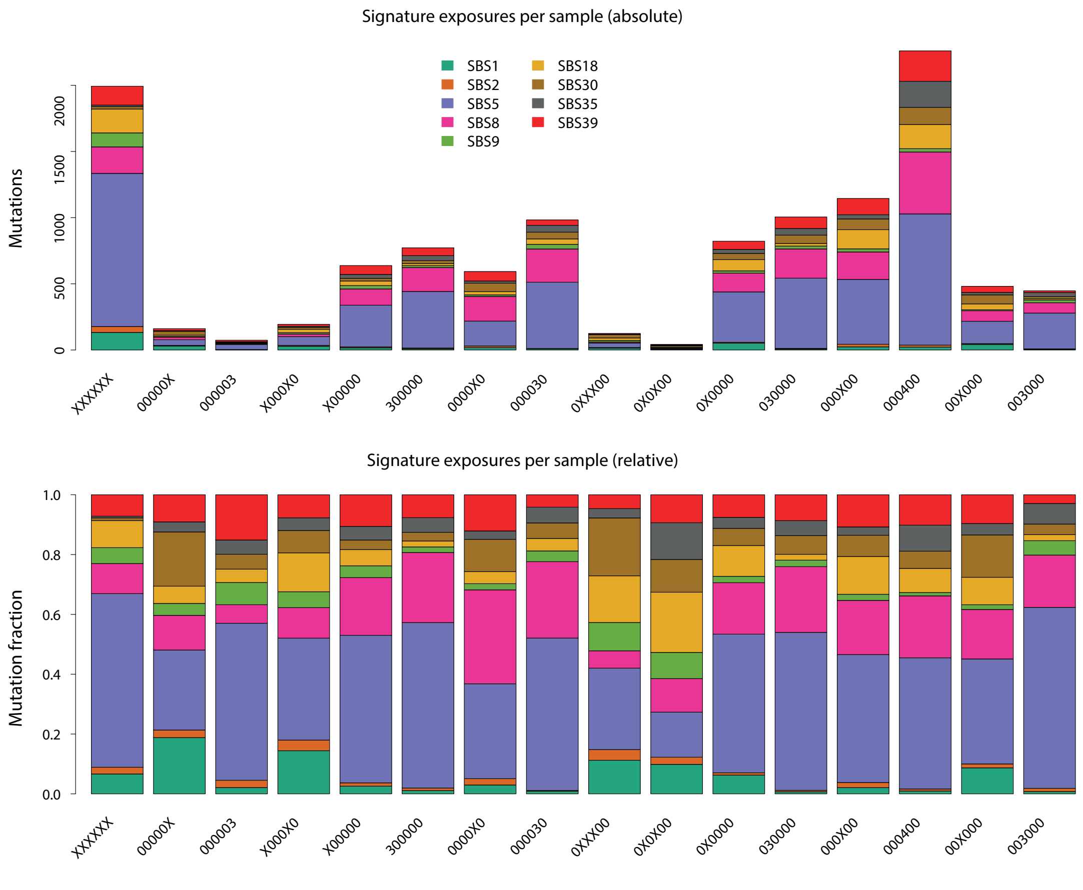


Figure 3.6 Mutational signature activity of subclones. Absolutive and relative exposures of COSMIC mutational signatures in each mutational cluster.

3.4.5 Structural variants

Structural variants were identified in every chromosome with a chr2:chr7 inter-chromosomal translocation recurrently found across all samples (Figure 3.7 and Appendix 8.1.5). Chromosome 7p displayed many inversions and duplications. Together with the oscillating copy number changes of small segments, this provides evidence of likely chromothripsis in chromosome 7p. In addition, as with SNVs, R4 displayed significantly more structural variants and possible chromothripsis in 17q compared to other regions suggesting there was ongoing and significantly higher genomic instability in this region.

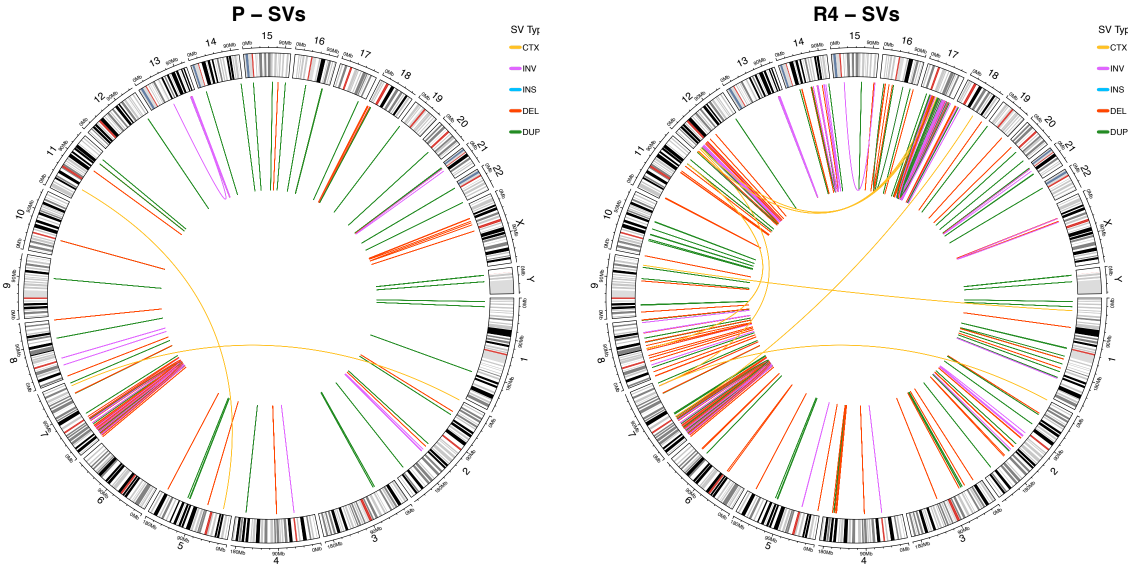


Figure 3.7 Structural variants detected through bulk WGS. Types of structural variants identified in the primary and R4 region: chromosomal translocations (CTX), inversion (INV), insertion (INS), deletion (DEL) and duplication (DUP).

3.5 scDNA-seq enables full resolution inference of the phylogenetic tree

In order to further explore the heterogeneity of this tumour, 10X Genomics scDNA-seq was performed on nuclei from the same samples as the bulk at a median depth of 0.034X. In total, 4,408 single cells from all regions passed quality control with a breakdown of the number of cells per region shown in Table 3.1.

3.5.1 scDNA-seq validates bulk phylogenetic tree

To validate the mutation clusters inferred using the subclonal reconstruction approach from the bulk data, I genotyped these variants in the single cells and assessed their co-occurrence within the same single cell. Indeed, co-occurrence of mutations was only seen for pairs of SNVs either in the same mutation cluster or clusters that were part of the same lineage, validating our bulk phylogenetic tree (Figure 3.8). For example, subclonal SNVs from R4 (000400) co-occur in the same cell with the truncal (XXXXXX) and shared clonal clusters (0XXX00) but not with SNVs from the primary, R1 or R5 clusters.

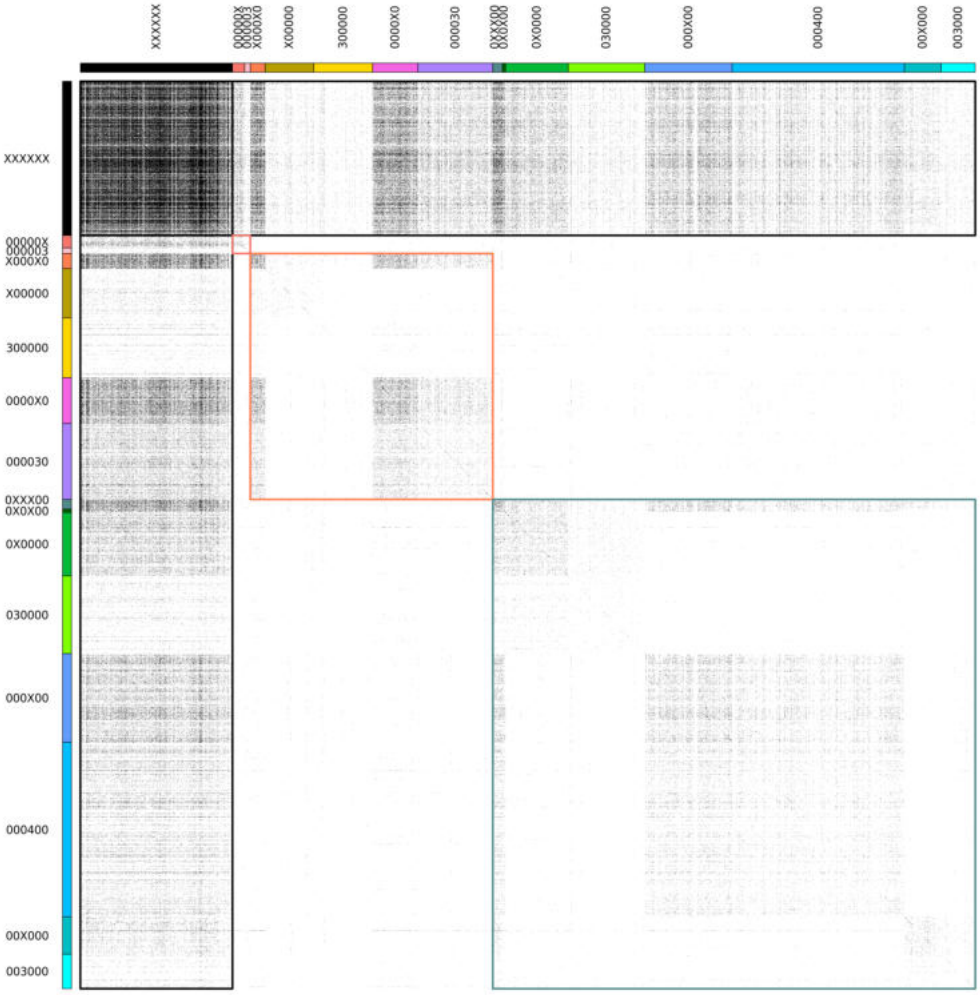


Figure 3.8 Co-occurrence matrix of bulk-detected SNVs in scDNA-seq. Boxes in black, demonstrate truncal SNVs which co-occur with all other clusters of SNVs. Orange and green boxes highlight SNVs which belong to the R1 and R5, and R2, R3, and R4 branches of the tree respectively. Finally, the small pink box highlights the cluster of SNVs from the primary. SNV clusters are annotated as previous.

3.5.2 Single-cell total and allele-specific copy number calling

Integer copy number calling at the single-cell level revealed further complex intra-regional heterogeneity. Multiple subclones distinguishable by their CNAs were found within all regions except R3 (where only a small number of cells were profiled), demonstrating the additional resolution afforded by scDNA-seq over bulk sequencing (Figure 3.9). Using a K-means clustering approach on single-cell copy number profiles, 19 distinct subclones were defined. Dimensionality reduction of all CNA profiles confirmed that the CNA profiles of single cells generally clustered closely with those of their corresponding bulk regions, providing confidence for these single-cell copy number profiles (Appendix 8.1.7).

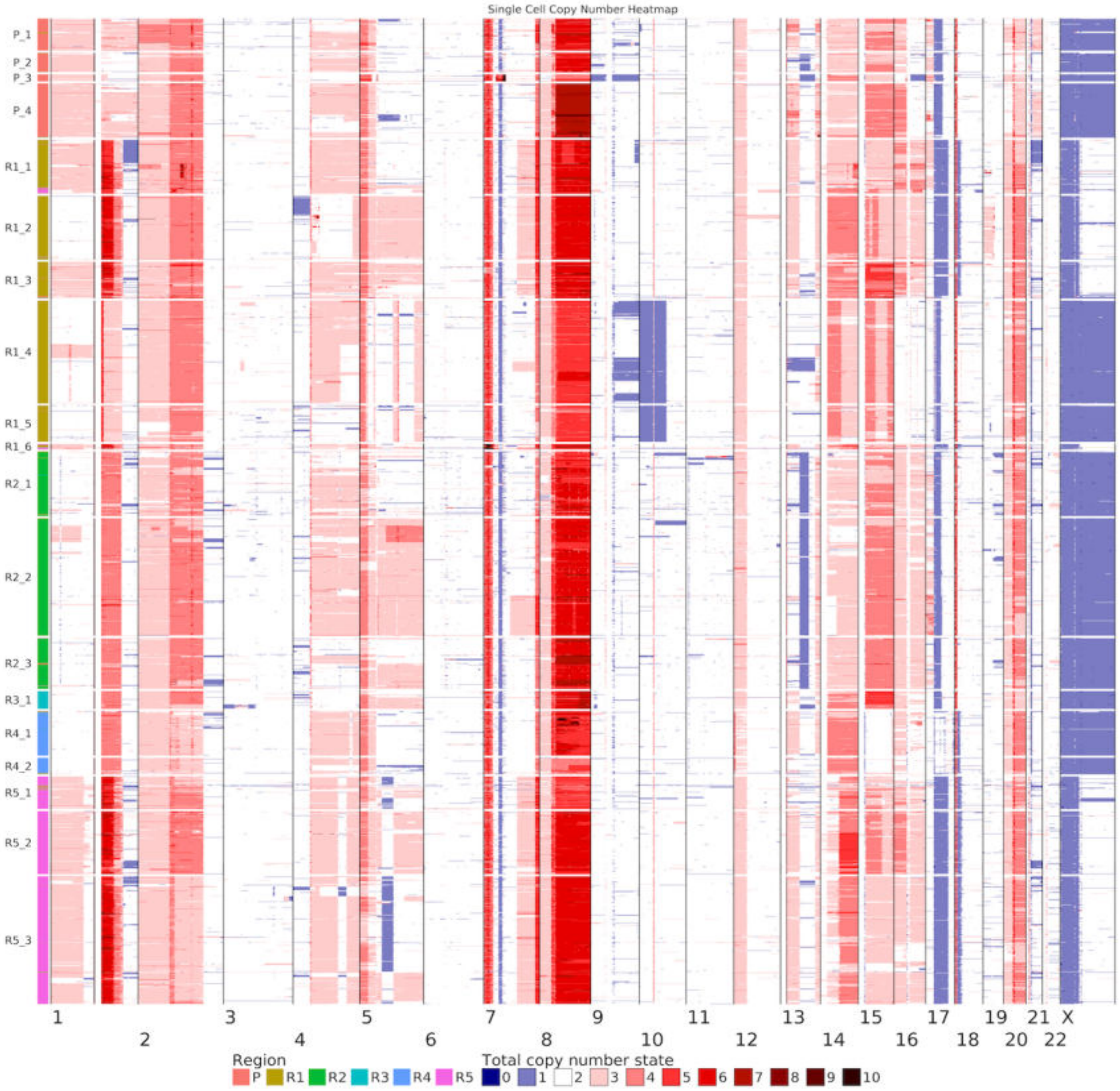
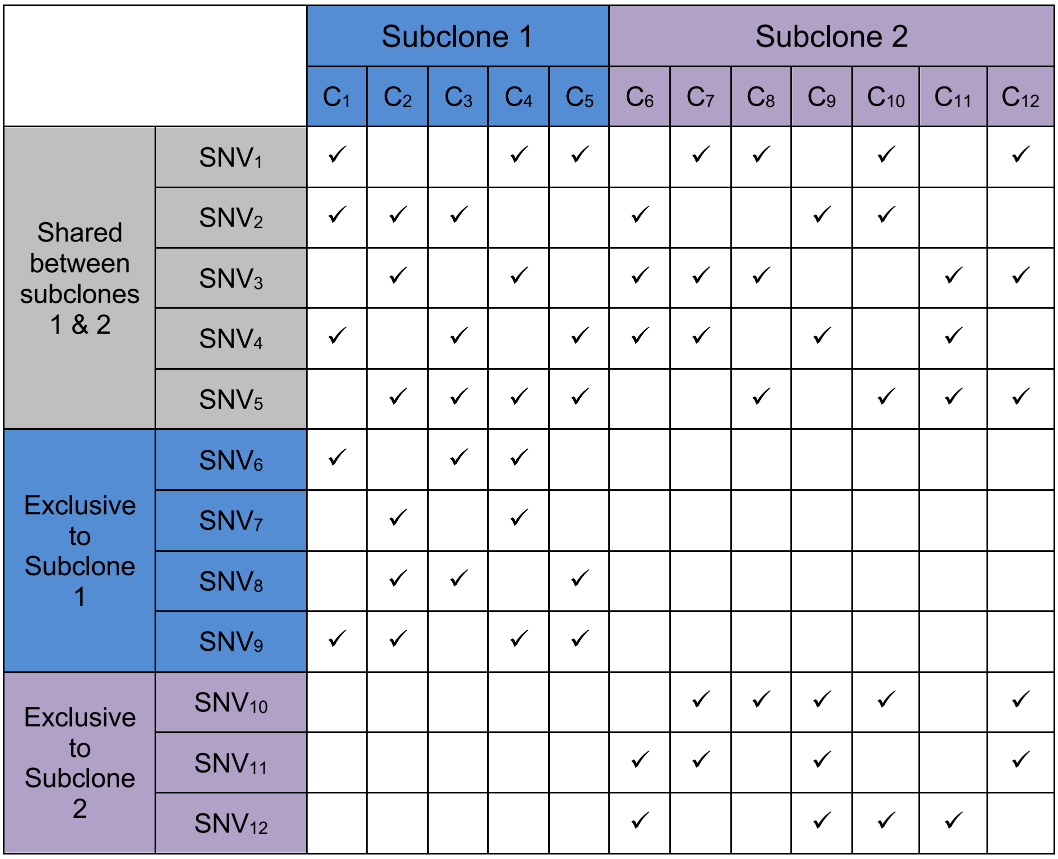


Figure 3.9 Total integer copy number profiles of single cells. Cells are split and annotated by the subclone they belong to and coloured by their region of origin.

3.5.3 *De-novo* SNV calling in single cells identifies subclone-specific mutation clusters

Although specific SNVs may not be genotyped in individual cells due to the low coverage in scDNA-seq, the presence of an SNV in a subclone can be detected by pooling reads from cells from the same subclone. Clustering of these SNVs can then reveal those which are shared between subclones and those that are specific to a subclone (Figure 3.10).



		Subclone 1					Subclone 2						
		C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}
Shared between subclones 1 & 2	SNV_1	✓			✓	✓		✓	✓		✓		✓
	SNV_2	✓	✓	✓			✓			✓	✓		
	SNV_3		✓		✓		✓	✓	✓			✓	✓
	SNV_4	✓		✓		✓	✓	✓		✓		✓	
	SNV_5		✓	✓	✓	✓			✓		✓	✓	✓
Exclusive to Subclone 1	SNV_6	✓		✓	✓								
	SNV_7		✓		✓								
	SNV_8		✓	✓		✓							
	SNV_9	✓	✓		✓	✓							
Exclusive to Subclone 2	SNV_{10}							✓	✓	✓	✓		✓
	SNV_{11}						✓	✓		✓			✓
	SNV_{12}						✓			✓	✓	✓	

Figure 3.10 Schematic showing SNV detection in subclones. Detection of the ALT allele of SNVs (SNV_1…SNV_{12}) in each cell (C_1…C_{12}) shown with tick marks. SNVs detected in cells belonging to both subclones shown in grey and those exclusive to Subclone 1 and 2 shown in blue and purple respectively.

To identify clusters of subclone-specific SNVs, I first performed *de novo* somatic variant calling in the scDNA-seq data for SNVs missed by bulk calling due to their low VAF. Due to the higher levels of technical noise in these data, I trained a machine-learning classifier on a manually annotated subset of candidate variants to filter and identify a further 1107 SNVs (see section 2.3.10).

By genotyping cells in the same subclone for SNVs as described in Figure 3.10 and performing hierarchical clustering, I was able to augment the SNV clusters derived from bulk WGS with these additional *de novo* calls and expand upon the phylogenetic tree to the level of several subclones for each region (Figure 3.11). As expected, the *de novo* SNV calls were almost all exclusive to a particular subclone whereas clusters of SNVs shared between subclones were identified from SNVs called in the bulk data.

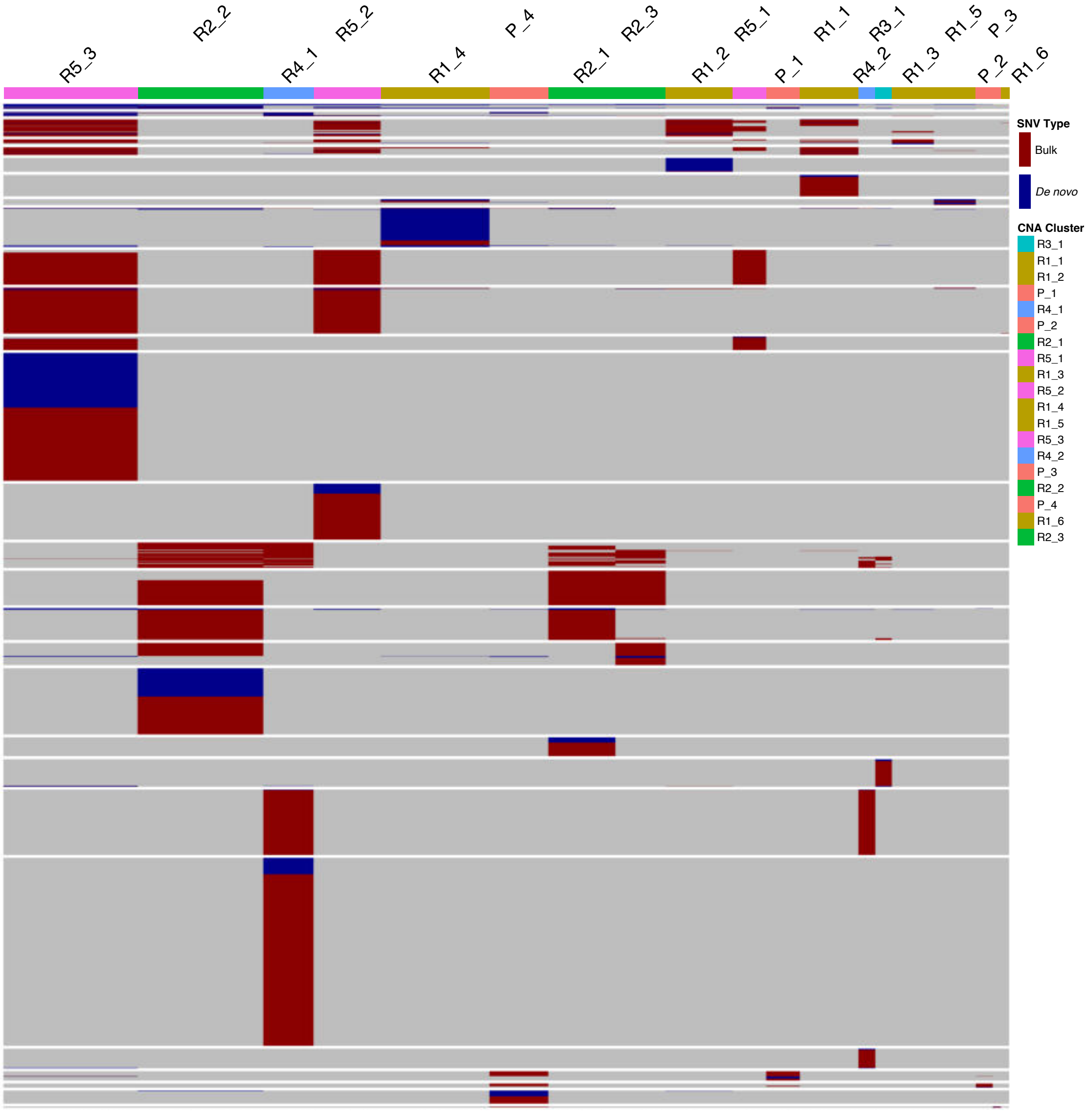


Figure 3.11 Presence of SNVs in each copy number subclone. Subclonal SNVs called in the bulk and *de novo* calls from scDNA-seq are shown in red and blue respectively. Column width is proportional to the total read depth of each subclone.

3.5.4 Single-cell resolution phylogenetic tree reconstruction

Given the sparse nature of scDNA-seq data, single-cell resolution phylogenies cannot be reconstructed using SNV data. Recently, an allele-specific CNA-based minimum-event distance approach for reconstructing phylogenies (MEDICC2) was developed and could be extended to the single-cell level (Kaufmann *et al.*, 2022). Therefore, I fitted allele-specific copy number profiles from LogR and BAF values calculated for each genomic bin for single cells by using the phased haplotypes generated from bulk data (Appendix 8.1.9). These single-cell allele-specific copy number profiles allowed me to fully refine the phylogenetic tree of this tumour down to the single-cell level (Figure 3.12).

Taken together, this patient’s tumour displayed extreme levels of chromosomal instability: at the bulk level the entire genome was observed to exhibit either copy number aberrations or copy neutral LOH. Even more strikingly, at the single cell level 99.5% of the genome had subclonal events supported by at least 10 cells for each copy number state while the median number of different allele-specific copy number states per segment was 3.

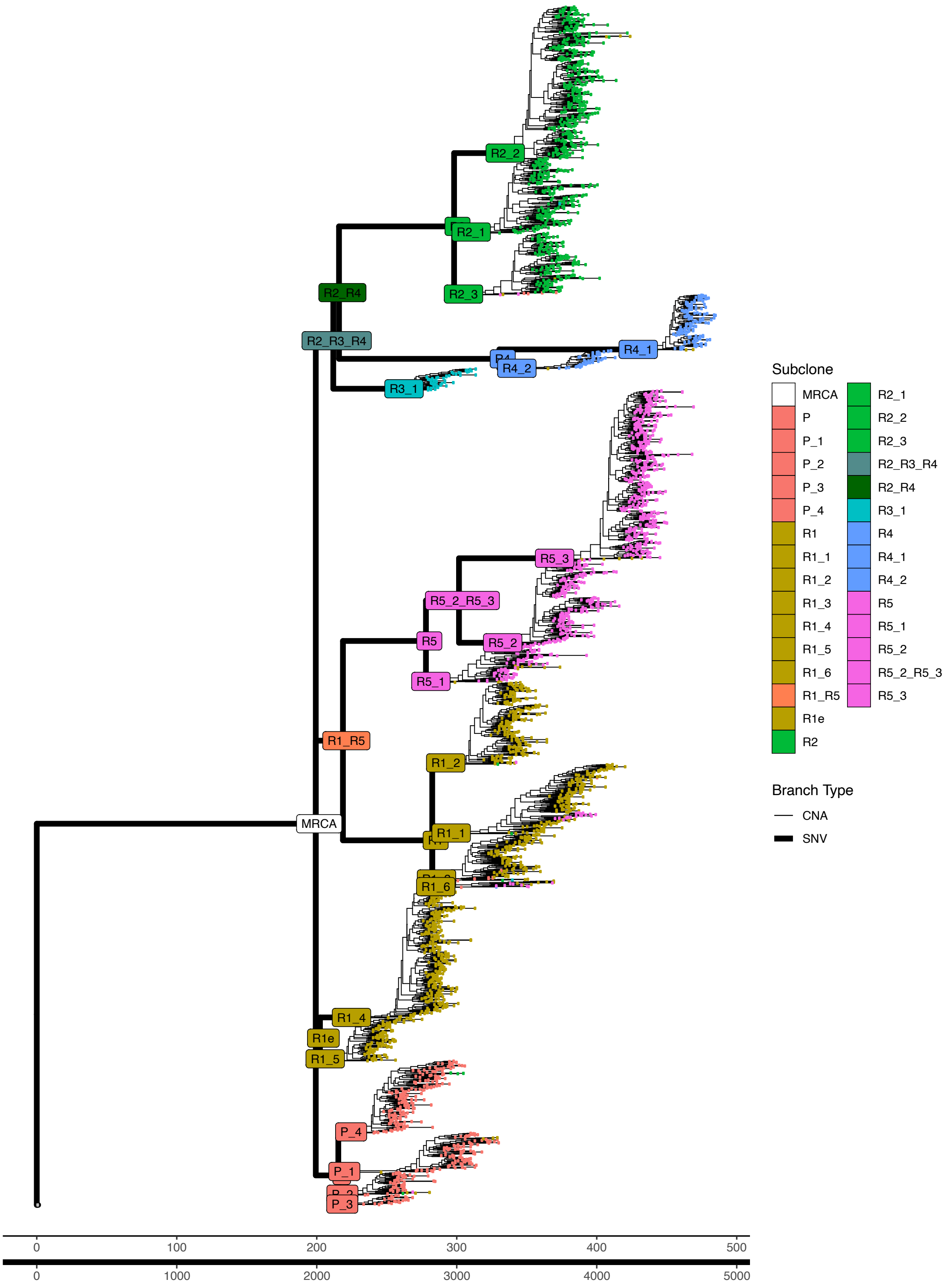


Figure 3.12 Single-cell resolution phylogenetic tree. SNV-based subclone tree and CNA-based phylogeny of single cells in each subclone are shown with thick and thin branches respectively. Cells or subclones are coloured by their region or SNV cluster and the MRCA is shown in white.

3.6 Spatial lineage tracing through laser capture microdissection

One consequence of this high degree of heterogeneity is that even neighbouring samples from the same tumour region may comprise different subclones, complicating analyses. This was demonstrated by the presence of two significant subclones (R1_4 and R1_5) displaying a loss of chr10p (Figure 3.9). While these clones together comprised ~50% of cells in the single-cell data, they were absent or contributed little to the bulk sequencing data (Appendix 8.1.3). Note also that no subclone carrying a reciprocal gain was identified, which could have provided an alternative explanation for the absence of a bulk signal. As the bulk and scDNA-seq samples were taken from adjacent but not identical tissue curls, this pointed to the presence of a focal microscopic subclonal expansion. Therefore, to determine the spatial distribution of subclones in close proximity to each other and determine their phylogenetic relationships, LCM was performed on three sections lining the front, side and back of tumour blocks (Figure 3.13). The primary tumour, and regions R1 and R4 of the recurrence were chosen for this experiment due to the higher intra-tumour heterogeneity seen in their scDNA-seq copy number profiles.

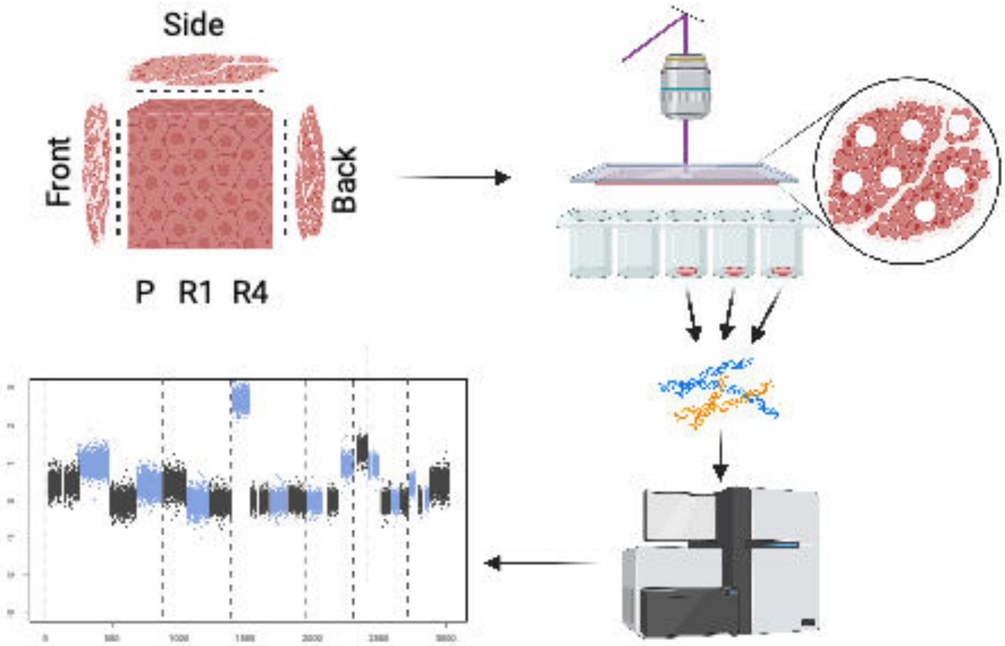


Figure 3.13 Experimental design of the LCM experiment. Sections from the front, side and back of each tissue block underwent LCM. The capture areas were amplified and underwent shallow WGS. Created with Biorender.com

3.6.1 Mapping LCM regions to the single-cell tree

101 spots were captured and underwent successful DNA amplification (Table 3.1). Mini-bulk whole-genome sequencing was performed with a median coverage per spot of 1.35X. Allele-specific CNA profiles were derived from these mini-bulk LCM samples and demonstrated a correlation between physical and genetic distances between spots for 3 of the 9 sections (Appendix 8.1.10). To encourage reuse of the data and aid exploration of the spatial relationships between subclones, I developed an interactive R Shiny app (Appendix 8.1.11). This enabled the identification of genomic segments with gradients of copy number gains across a tissue section revealing spatial patterns of copy number evolution (Figure 3.14).

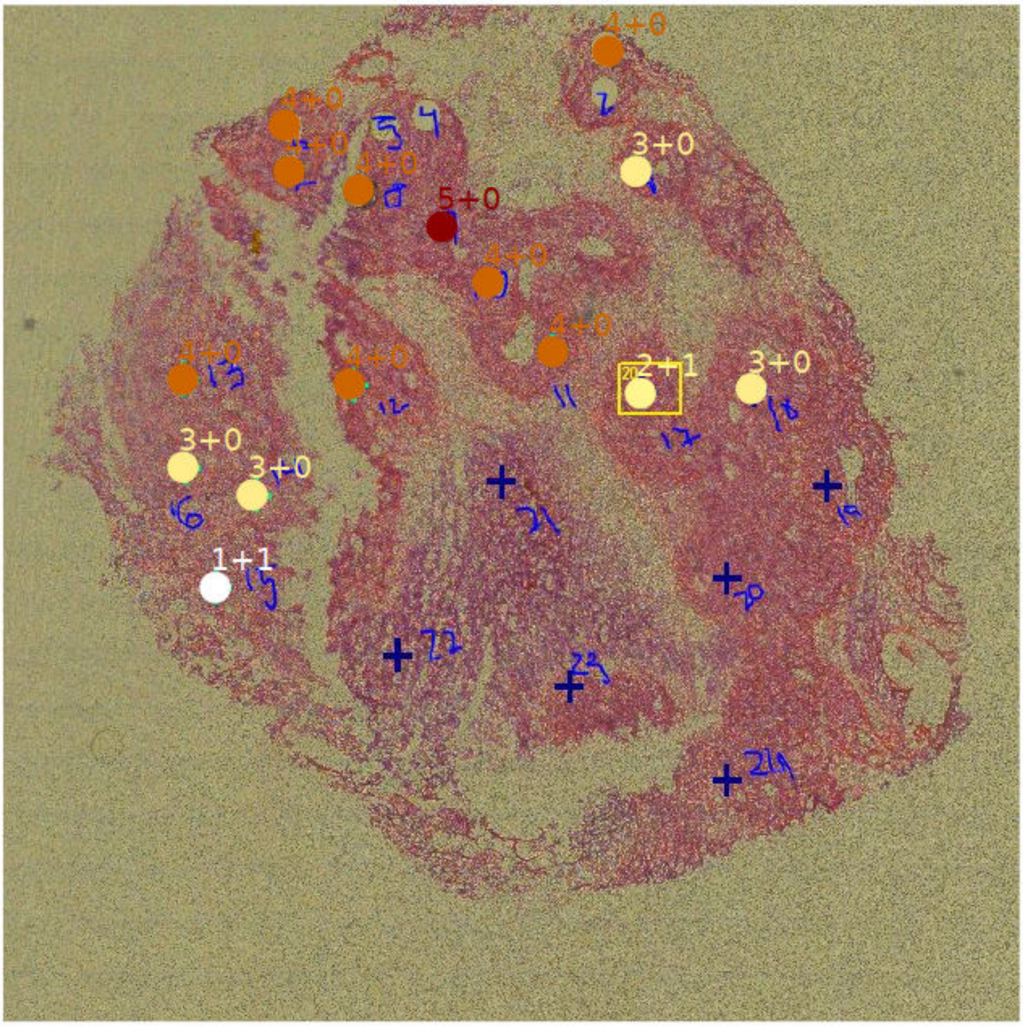


Figure 3.14 Spatial gradients of copy number gains. Allele-specific copy number states of the Chr2:36.7 – 88.7Mb segment from LCM captured spots in the Primary front section. Copy number states are shown as major allele + minor allele. Spot locations marked and annotated during LCM shown in blue.

I then utilised the mini-bulk allele-specific copy number profiles to place spots onto the fully refined phylogenetic tree based on the closest single-cell by event-distance. This revealed several attributes of the physical expansion in this tumour. Firstly, as expected, mini-bulk spots were placed exclusively with subclones from their corresponding region (Figure 3.15). However, they did not always cluster by the section they were captured from, which is in line with our observation of highly localised clonal expansions which are present on part of more than one section (Appendix 8.1.12). Furthermore, spots from the primary were representative of multiple subclones present in the scDNA-seq sample whereas those captured from R1 and R4 only derived from one specific subclone, suggesting expansions of larger sizes in the recurrences compared to the primary (Figure 3.15).

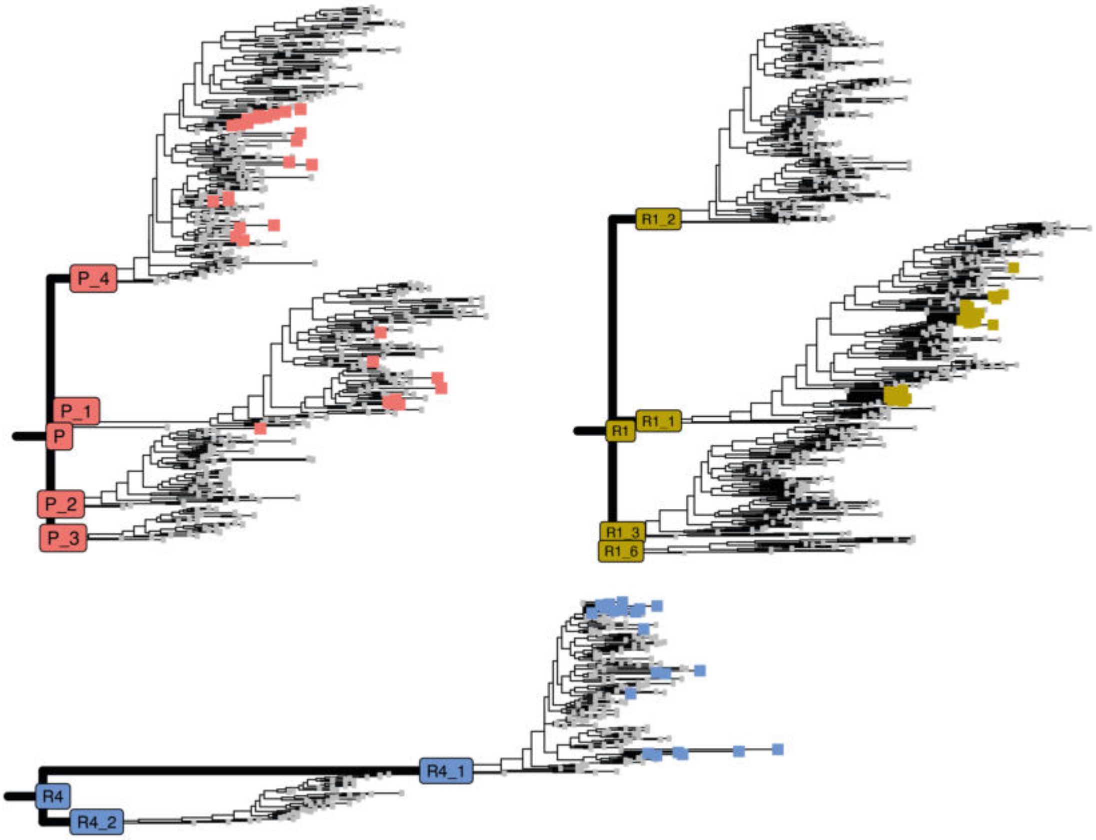


Figure 3.15 Phylogenetic relationships of LCM subclones. Phylogenetic trees of cells from the primary, R1 and R4 recurrence regions pruned from the full resolution phylogenetic tree (Figure 3.12). scDNA-seq cells comprising of each subclone are shown in grey and LCM spots are highlighted in colour by their region of origin and located based on the most similar single cell.

3.6.2 Spatial CNA tree inference

Next, I created phylogenetic trees for each tissue slide, again using allele-specific copy number profiles. These copy number based phylogenetic relationships can then be overlaid onto the tissue sections where the location of spots has been marked (Figure 3.16). For some sections, this revealed intricate growth paths of the tumour at microscopic resolution that recapitulate the infiltration of the tumour. For others, a more complex and improbable mixing of paths was observed (Appendix 8.1.13). This may be explained by the cryo-sectioning being performed in a different plane from that which the tumour was growing in. To my knowledge, this is the first time that growth trajectories have been inferred through spatial lineage tracing for multiple subclones from a human tumour.

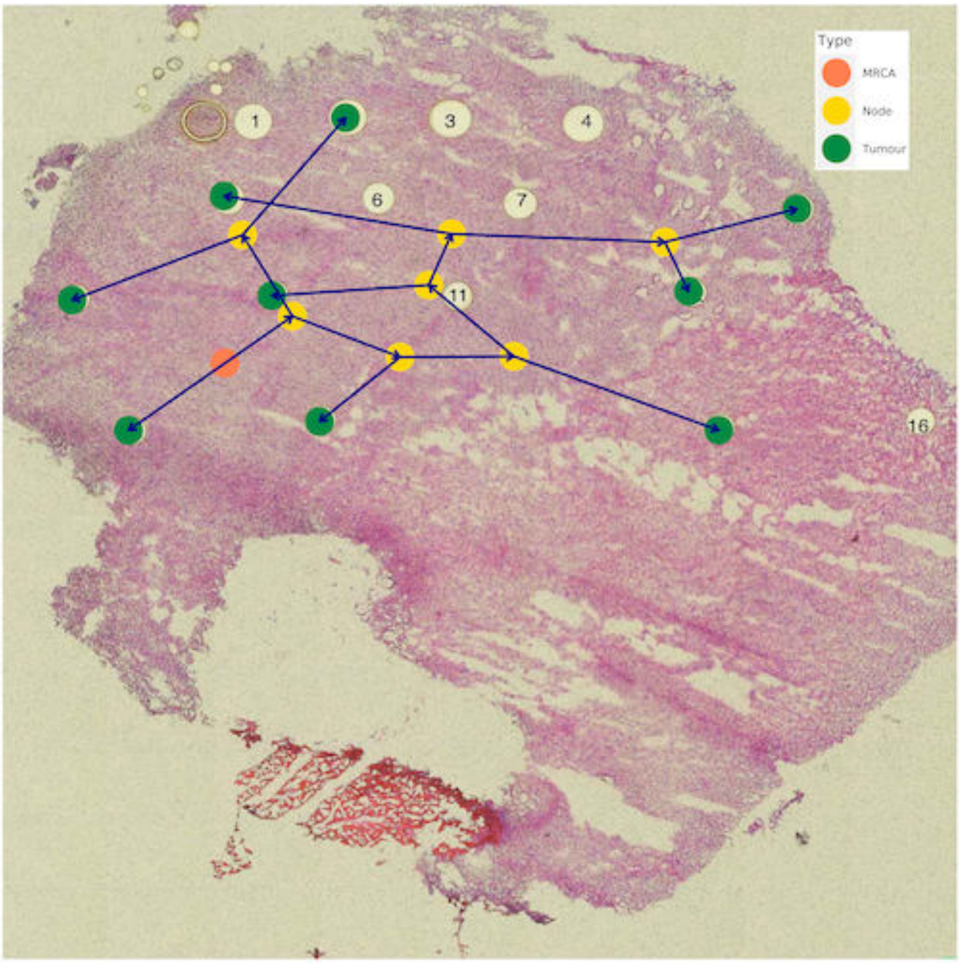


Figure 3.16 LCM derived spatial phylogeny. The "Back" section from R4 with CNA-based phylogenies superimposed. Tumour spots, intermediate nodes and MRCA shown in green, yellow, and orange respectively.

3.6.3 Genotyping subclone-specific SNVs

As LCM typically captures small areas of tissue, I next sought to determine the spatial distribution of SNVs in these tumour sections. The cluster of shared R1 and R5 clonal SNVs was present in all spots captured from R1 as expected (Figure 3.17). Reassuringly, the two spots featuring a diploid copy number profile from the “Side” section of R1 did not display any SNVs from this shared clonal cluster (Appendix 8.1.14). This provides confidence in our copy number calling and confirms that the tissue from these spots was truly adjacent normal tissue.

LCM spots from R1 were only mapped to the R1_1 subclone based on copy number on the full-resolution tree and not to other subclones such as R1_2 or R1_4 (Figure 3.15). Therefore, only the SNVs which identified the R1_1 subclone should be detected in the LCM spots. Indeed, genotyping subclonal clusters of SNVs in LCM spots confirmed that only SNVs specific to R1_1 (and not those specific to R1_2 or R1_4) were present in the LCM sections from R1 (Figure 3.17). However, I did not observe further hyper-localisation of this cluster of SNVs to specific regions of the tissue sections suggesting that this subclone encapsulated a greater area than that sampled in our LCM study. This further provides orthogonal validation of our single-cell copy number based subclone-specific SNV detection method and confirms the localisation of specific subclones in these tissue sections. Further increases in resolution in subclone detection may be achieved by *de novo* SNV calling in the LCM data by pooling LCM spots from each tissue section. However, this would require a library preparation method without PCR amplification to reduce artefacts.

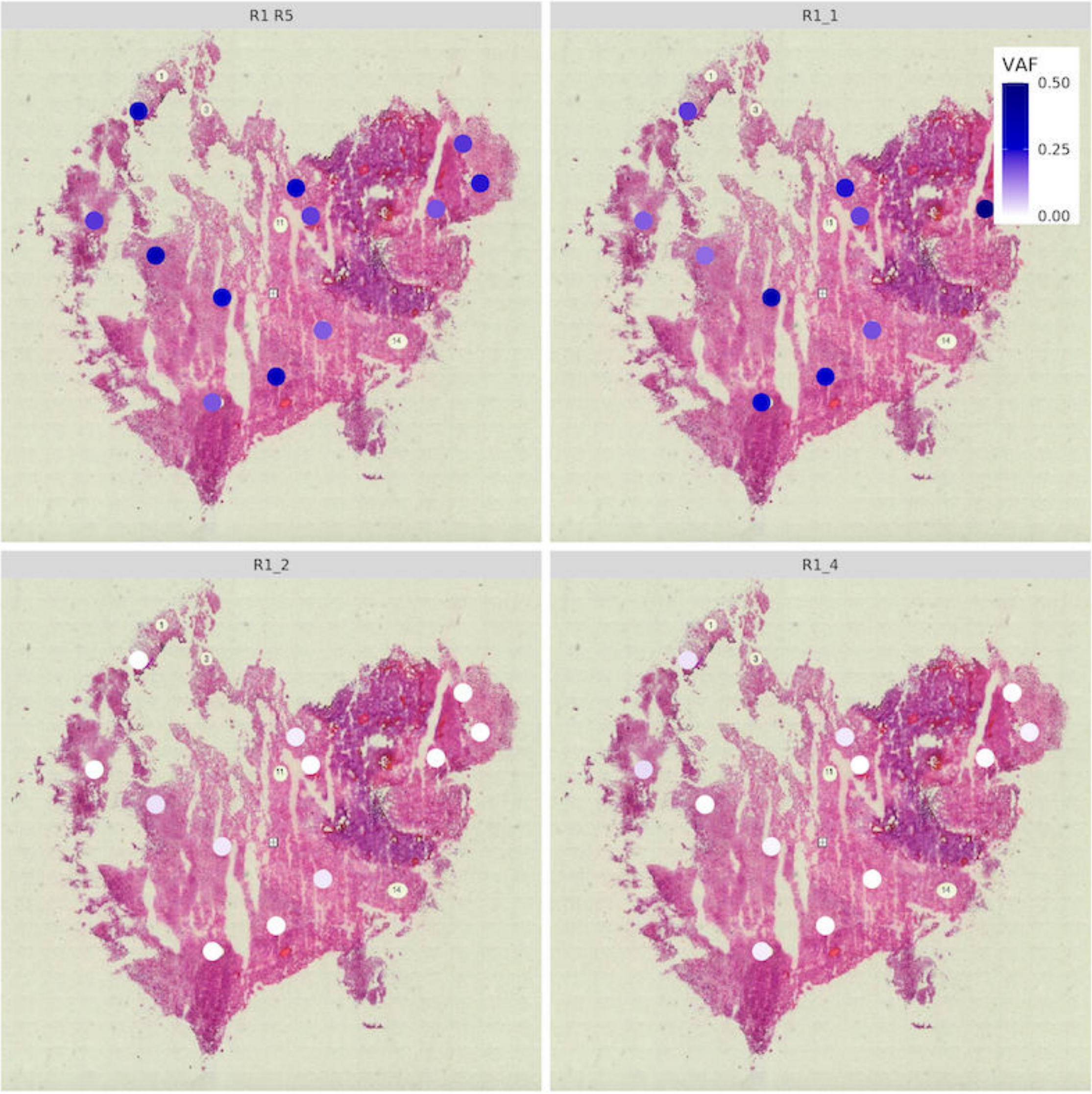


Figure 3.17 Genotyping clusters of SNVs in LCM spots. VAF shown for subclonal clusters of SNVs from R1 for each spot from the R1_Back section. Shared R1 and R5 clonal cluster shown in top left.

3.7 Discussion

In the first part of this study, three genomic profiling techniques with increasing resolution was applied to MPNST primary tumour and 5 local recurrence regions. Bulk WGS enabled phylogenetic reconstruction to the level of a single subclone for each sample. scDNA-seq was used to validate bulk reconstruction and enabled full reconstruction down to the single-cell level. LCM further added spatial information and demonstrated the growth pattern of this tumour.

3.7.1 MPNSTs are primarily driven by chromosomal instability

The primary tumour was found to be genomically unstable and strikingly, more than 90% of the genome was in a state of copy number neutral LOH. This near-haploidisation is likely an extreme case of the LOH of multiple chromosomes that was identified as a recurrent feature in MPNSTs with H3K27me3 loss (Cortes-Ciriano *et al.*, 2023). Following LOH of most chromosomes, this MPNST then likely underwent a WGD to arrive at a ploidy of over 2, in keeping with the observation of Cortes-Ciriano et al. that WGD is a late event in MPNSTs. Therefore, this MPNST must at one time have existed in a near haploid state, although we did not identify these cells through our single-cell profiling. Furthermore, whether this large-scale haploidy occurred sequentially or as a punctuated event is unclear as losses in chromosomes cannot be timed. This is because the genetic material bearing the mutations used in timing analysis are lost. These observations raise questions as to how these haploid tumour cells can survive and further divide, and whether WGD is necessary to rescue the presumably low fitness of these haploid cells. Haploid tumour genomes are extremely uncommon across tumour types but have been reported in sarcomas and other rare tumour types such as giant cell glioblastoma (Steele *et al.*, 2019; Baker *et al.*, 2020). However, the underlying mechanism, how this haploidy is tolerated by tumour cells, as well as its clinical significance is unclear. Understanding this unique feature of MPNSTs and exploiting any therapeutic vulnerabilities that arise from it may be a key for improving clinical outcomes.

The pattern of evolution resulting from this high degree of CIN was also revealed with multifurcation events observed at several levels. These multifurcations were present at the level of bulk regions, and of subclones within individual regions, resulting in a progressively branching phylogenetic tree. The initial punctuated event could be explained by polyphyletic seeding of tumour cells prior to resection of the primary tumour (as the primary harboured exclusive mutations). However, these cells were not eradicated by adjuvant chemotherapy, potentially due to dormancy, and rapidly proliferated resulting in the local recurrence. Subsequent branching was likely generated by ongoing chromosomal instability with multiple subclones identified via their unique CNA profiles. Given the plethora of copy number events displayed by the vast number of subclones, it remains unclear if there is selection of clones with certain copy number events. Tools for detecting copy number selection have not yet been developed but may identify specific combinations of copy number gains or losses which are advantageous. This may reveal synergistic or mutually exclusive interactions between copy number events which can be exploited for therapy.

In conclusion, this tumour exhibited extreme chromosomal instability generating a vast number of subclones which required single-cell methods to fully characterise. Further discussion of the overall study including its limitations and unique features of this tumour type which were exploited is included in Chapter 4.

Chapter 4. Transcriptional consequences of intra-tumour heterogeneity in a MPNST

4.1 Introduction

The transcriptome of a tumour can be profiled using RNA-sequencing and is a key method for characterising the phenotype of individual tumours. This has been performed at scale across tumour types using bulk samples and is increasingly carried out using single-cell methods.

In MPNST, RNA-seq of bulk tumours has been performed with the aim of identifying a specific gene expression signature this tumour type. The studies have identified altered expression in genes involved in processes such as neural crest (e.g., *SOX9* and *TWIST1*) or Schwann cell differentiation (e.g., *MBP*, *S100B* and *SOX10*) with the transcriptomes of these tumours resembling immature Schwann cells (Pemov *et al.*, 2020). However, contrasting findings on whether they are upregulated or downregulated in specific genes have been reported in these studies. This could be explained by the use of bulk samples in these studies which are a mixture of tumour and tumour microenvironment cells, or the use of cell lines which may not accurately recapitulate the transcriptomes of primary tumour cells. As a result, a consistent gene expression signature for MPNSTs has not been generated and how genetic inter- and intra-tumour heterogeneity impacts on the transcriptome of tumour cells remains unclear.

Recent developments in technologies to profile tumours have not yet been applied to MPNSTs. For example, single-cell studies of primary tumours hold promise for unravelling these transcriptomic changes as they can determine the gene expression of individual cell types. Matched recurrence or metastasis samples when analysed in combination with the primary tumour can also contribute by further characterising cell states before and after treatment.

Linking genetic changes to transcriptional activity in tumour cells is critically important in understanding the biology of MPNSTs. Advances in technologies and methodolgies will allow us to interrogate the relationship between tumour genotypes and phenotypes. G&T-seq can be used to directly link genotype to phenotype, however, its throughput is limited, so has not been widely adopted. Alternatively, scDNA-seq and scRNA-seq can be integrated together by detecting gene dosage effects in both types of sequencing. This is where number of copies of a gene is known to affect levels of gene expression (Stranger *et al.*, 2007). Finally, recently developed spatial transcriptomics methods have the potential to characterise the phenotype of spatially constrained subclones and may shed further insights into the pathogenesis of MPNSTs (Hunter *et al.*, 2021).

4.2 Aims

Having detected significant chromosomal instability in this MPNST, I sought to determine the downstream effects on the transcriptome resulting from this extreme level of intra-tumour heterogeneity.

- Identify the cellular composition of this tumour and study transcriptional intra-tumour heterogeneity down to single-cell resolution.
- Determine the relationship between genotype and phenotype in individual subclones.
- Infer the location of microscopic tumour subclones and their phylogenetic relationships through spatial transcriptomics.

4.3 Phenotypic profiling of subclones through scRNA-seq

scRNA-seq was performed on the same samples described in Chapter 3 to profile the cellular composition and assess the impact of this high level of copy number heterogeneity on cell transcriptomes. A total of 37,716 cells were profiled and passed quality control. I then performed dimensionality reduction, unsupervised clustering, and visualised clusters using UMAP.

4.3.1 Identifying cell types

Marker-based annotation of the differentially expressed genes between clusters of cells showed that the TME was composed mostly of macrophages with small numbers of T cells, endothelial cells, and skeletal muscle cells present (Figure 4.1).

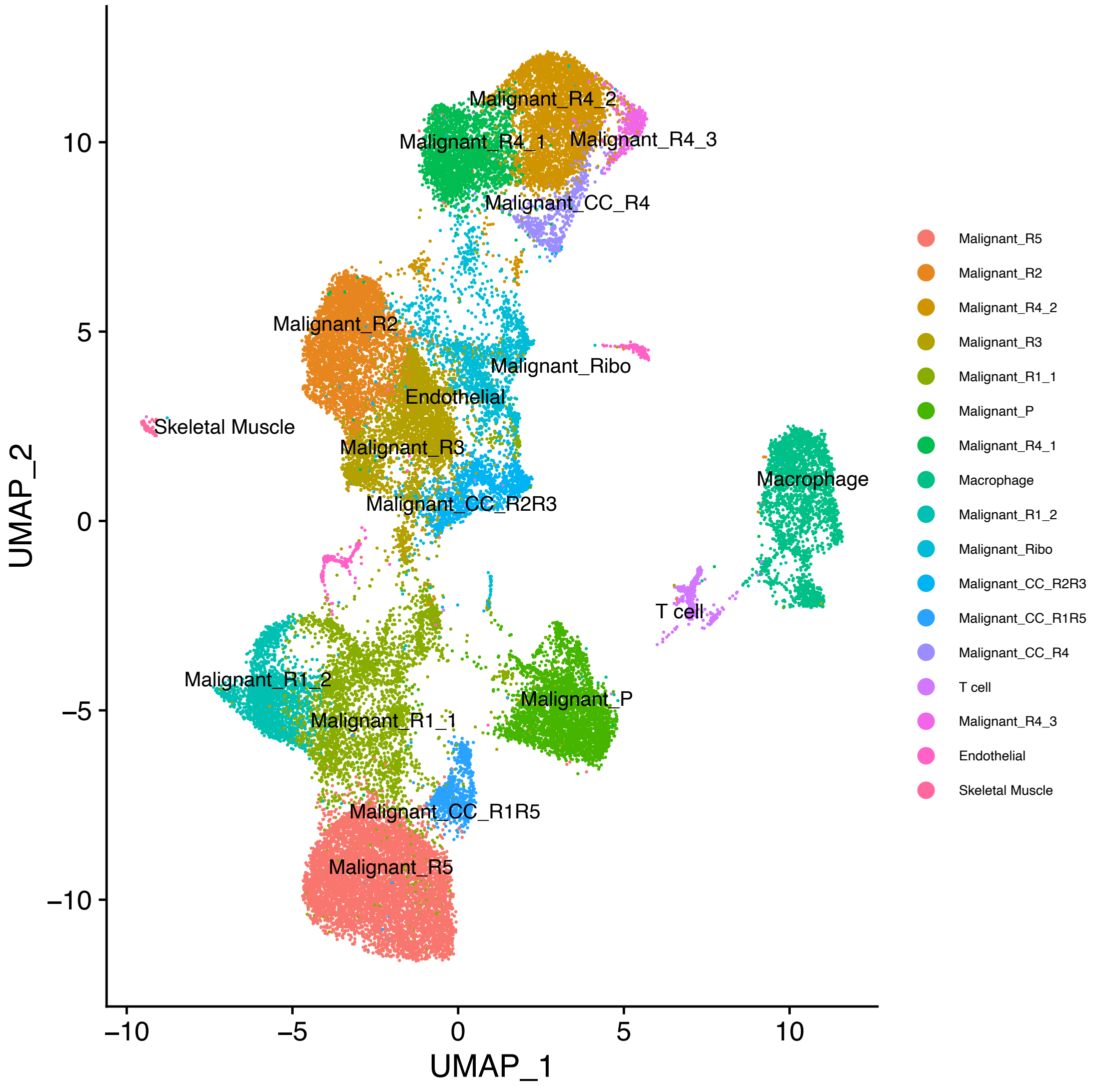


Figure 4.1 UMAP projection of all scRNA-seq cells. Tumour cells clusters are labelled by their region of origin. Cycling cells and ribosomal clusters are labelled “CC” and “Ribo” respectively.

Further sub-clustering and fine annotation of all cells initially classified as macrophage, T cell or endothelial cells identified additional cell types including regulatory T cells, pericytes and fibroblasts (Appendix 8.1.16). Detailed cell types identified and their frequency in each region are shown in Table 4.1. Interestingly, B cells were nearly exclusively found in the primary tumour, suggesting they were not recruited or were unable to infiltrate the recurrence regions. The remaining non-TME cells clustered together by their region of origin and as there are no reliable expression markers for MPNST, were identified as tumour cells in the first instance (Figure 4.1). Their high proportion matched the purity estimates from bulk WGS.

Region	P	R1	R2	R3	R4	R5
Cycling Tumour	2	59	753	562	570	723
Cycling TME	15	0	11	28	3	3
Macrophage	346	112	414	603	404	247
Neuronal	0	2	87	41	128	93
Endothelial	54	11	33	26	24	28
Fibroblast	41	0	1	2	93	7
Pericyte	15	2	9	25	10	5
B cell	108	1	1	5	8	5
T cell	40	18	12	19	99	52
Treg cell	4	0	0	2	19	11
Skeletal Muscle	34	0	2	0	86	1
Tumour	3360	4844	5435	3840	7164	6954
Total	4019	5049	6758	5153	8608	8129

Table 4.1 Cell types identified through scRNA-seq.

4.3.2 Confirming cell type annotation through haplotype counts

Given the lack of canonical markers for MPNST cells, I utilised an orthogonal approach to confirm labelling of tumour cells, by leveraging the near genome-wide LOH in this tumour. Using phased haplotypes, I genotyped single cells at heterozygous SNPs across the genome and generated BAF values representing the minor alleles. As expected, cells identified as malignant cells had significantly fewer reads reporting the lost alleles than TME cells (Figure 4.2) showing this to be a marker-independent cell type annotation method. This has the added benefit of allowing identification of tumour-normal doublets and is sensitive enough to classify cells based on SNPs from only one chromosome which has undergone LOH (Appendix 8.1.19).

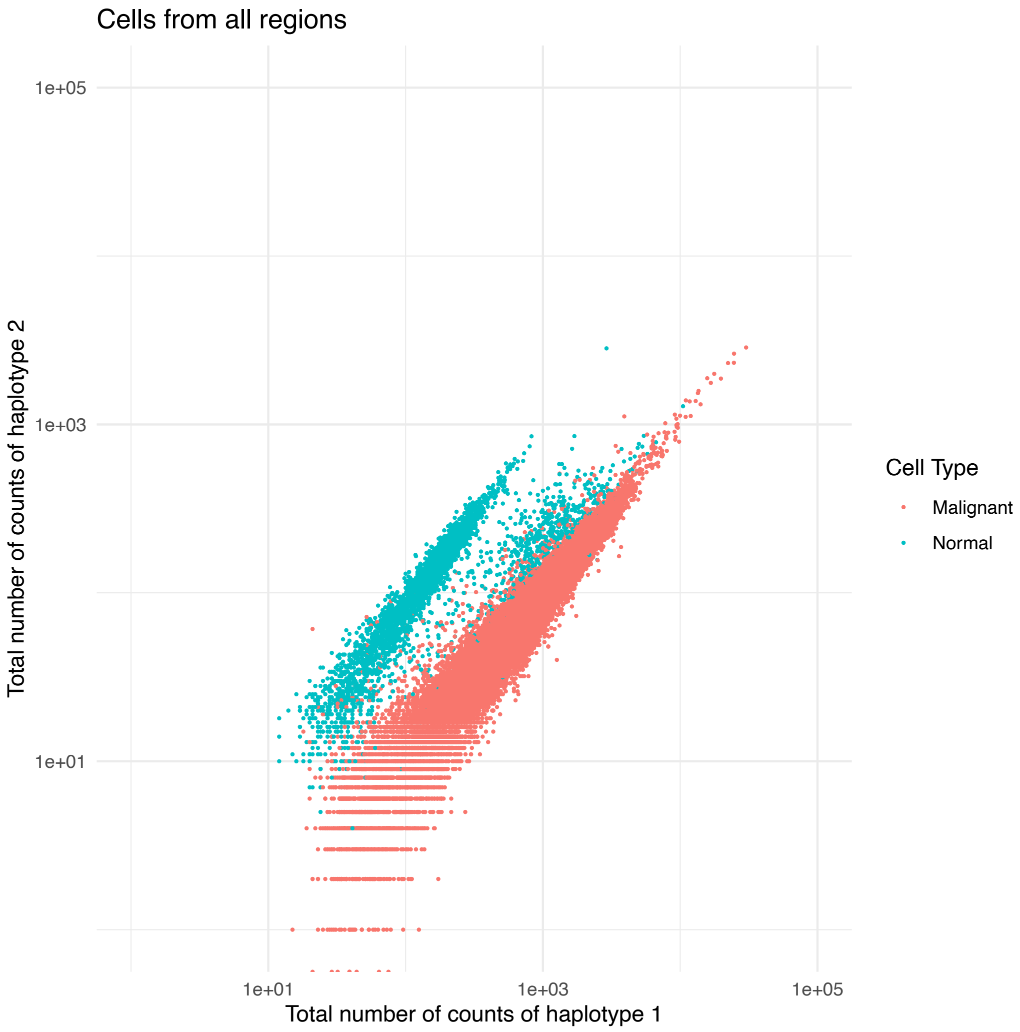


Figure 4.2 Haplotype counts of scRNA-seq cells. Total allele counts of the retained (1) and lost haplotype (2) of cells annotated by expression-based identity.

4.3.3 Gene regulatory network analysis

To identify transcription factors controlling sets of genes, I utilised SCENIC to identify these gene regulatory networks (GRN) in each cell. Once again, cells clustered based on GRN activity colocalised by their region, and regions which were physically adjacent were also closer on the UMAP (Figure 4.3).

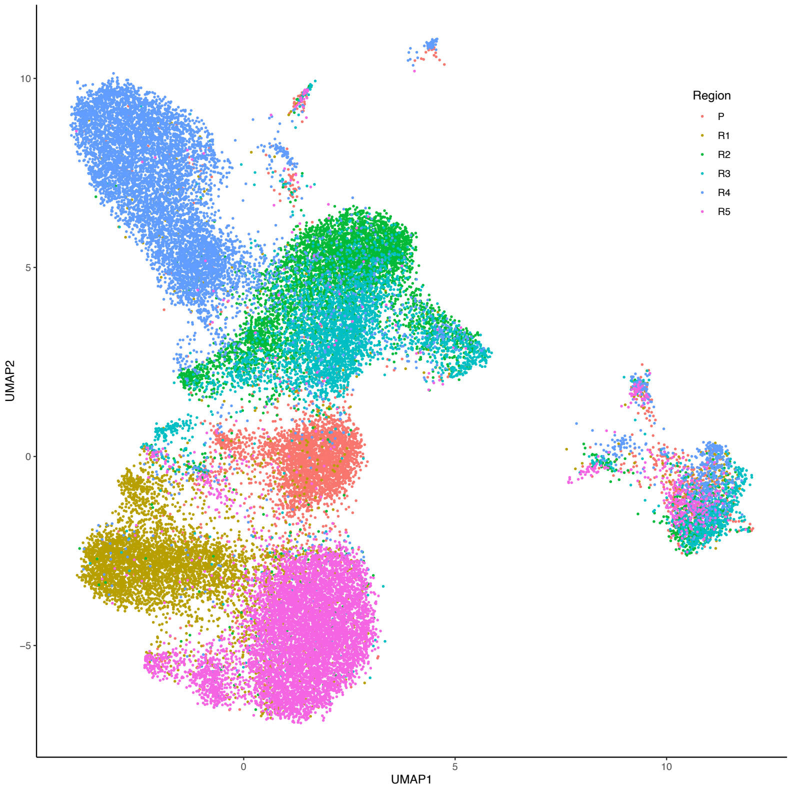


Figure 4.3 GRN-based UMAP projection of scRNA-seq cells. Cells are annotated by their region of origin. The ribosomal cluster of cells can be seen at UMAP1:UMAP2 coordinates (5,3).

In addition, GRN-based clustering identified a cluster of cells from R2 and R3 which clustered separately from other cells from these regions (Figure 4.3). This population of cells was also seen in the scRNA-seq data and expressed a high number of ribosomal protein genes (labelled “Malignant_Ribo” in Figure 4.1). Myc Associated Zinc Finger (MAZ) was the most active GRN for this group of cells and higher activity was also seen for MYC (Figure 4.4). Interestingly, expression of MAZ itself did not appear significantly not higher in these cells, suggesting additional mechanisms enhancing activity of this transcription factor and increasing transcription of downstream genes (Figure 4.5). However, the exact nature of this subset of cells cannot be reliably determined from just one tumour and further investigation in larger studies is required to determine their significance.

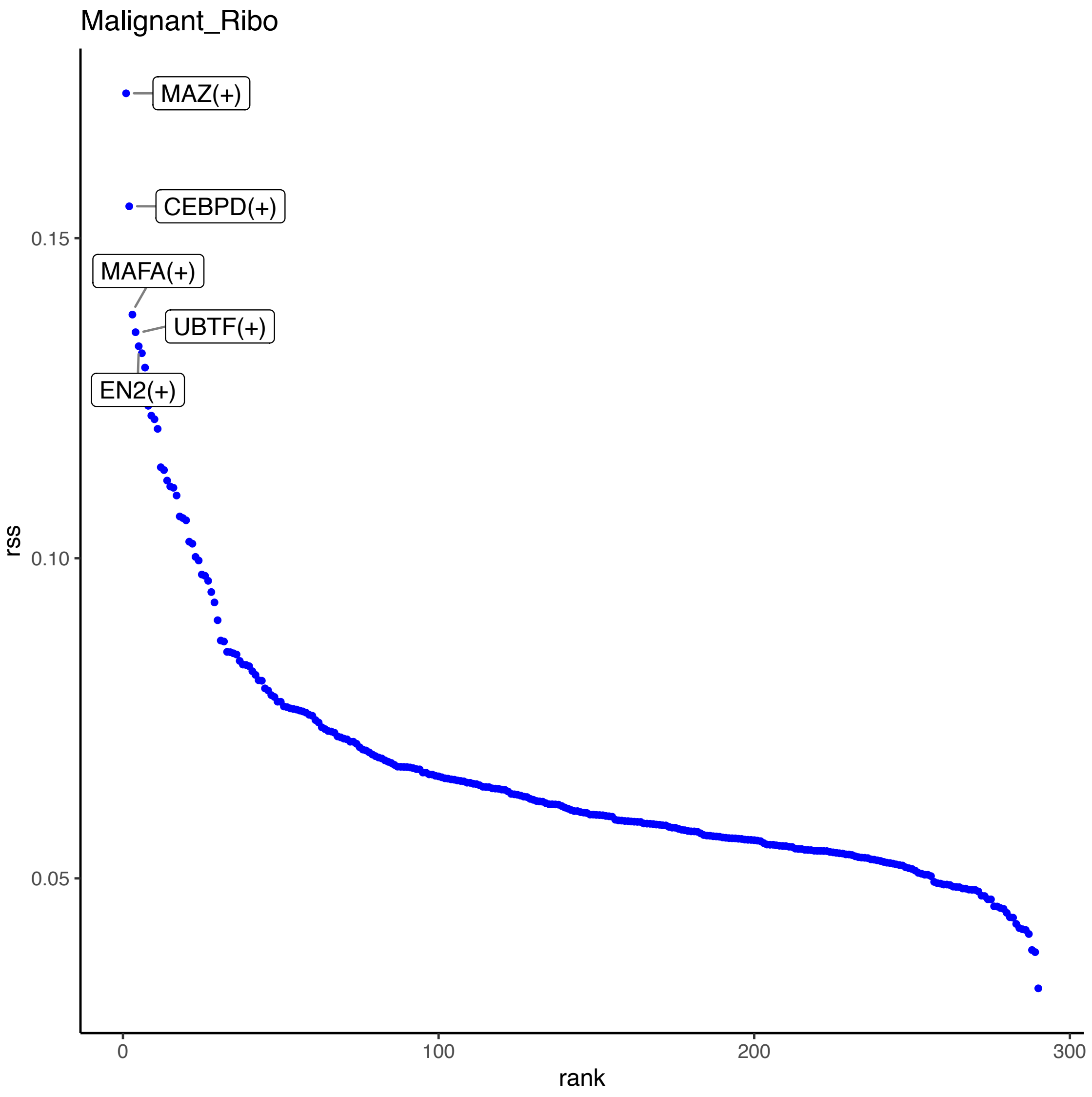


Figure 4.4 Active GRNs in the Malignant_Ribo cluster. Regulon specificity score (RSS) plot of all transcription factors for cells in the Malignant_Ribo cluster with those most active shown on the left of the plot.

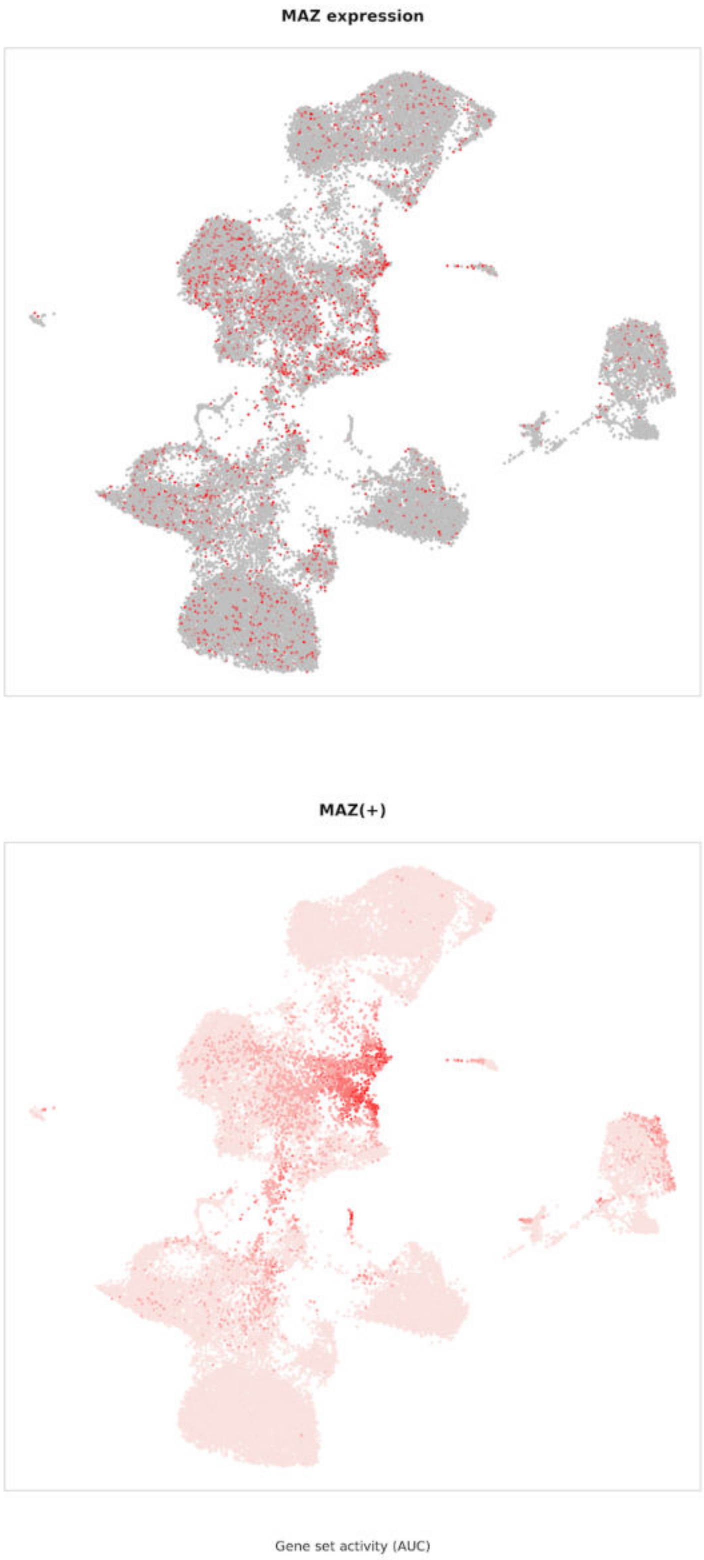


Figure 4.5 Gene expression and GRN activity of MAZ. Cells expression MAZ are highlighted in red in the upper panel and gene set activity is shown in the lower panel.

4.3.4 Inferring CNAs from the transcriptome

To explore the CNA heterogeneity within the scRNA-seq cells, I inferred CNA profiles using inferCNV (Figure 4.6). Inferred copy number profiles differed between regions and cells with similar CNA profiles within a region clustered together on the UMAP, suggesting the majority of this clustering was driven by effects from differences in copy number (Figure 4.7). In addition, cells clustered together on a projection based on GRN activity by their region of origin and by their inferred CNAs (Appendix 8.1.20). This further suggest the mechanism by which CNAs drive differences in gene expression is largely through effects mediated via GRNs.

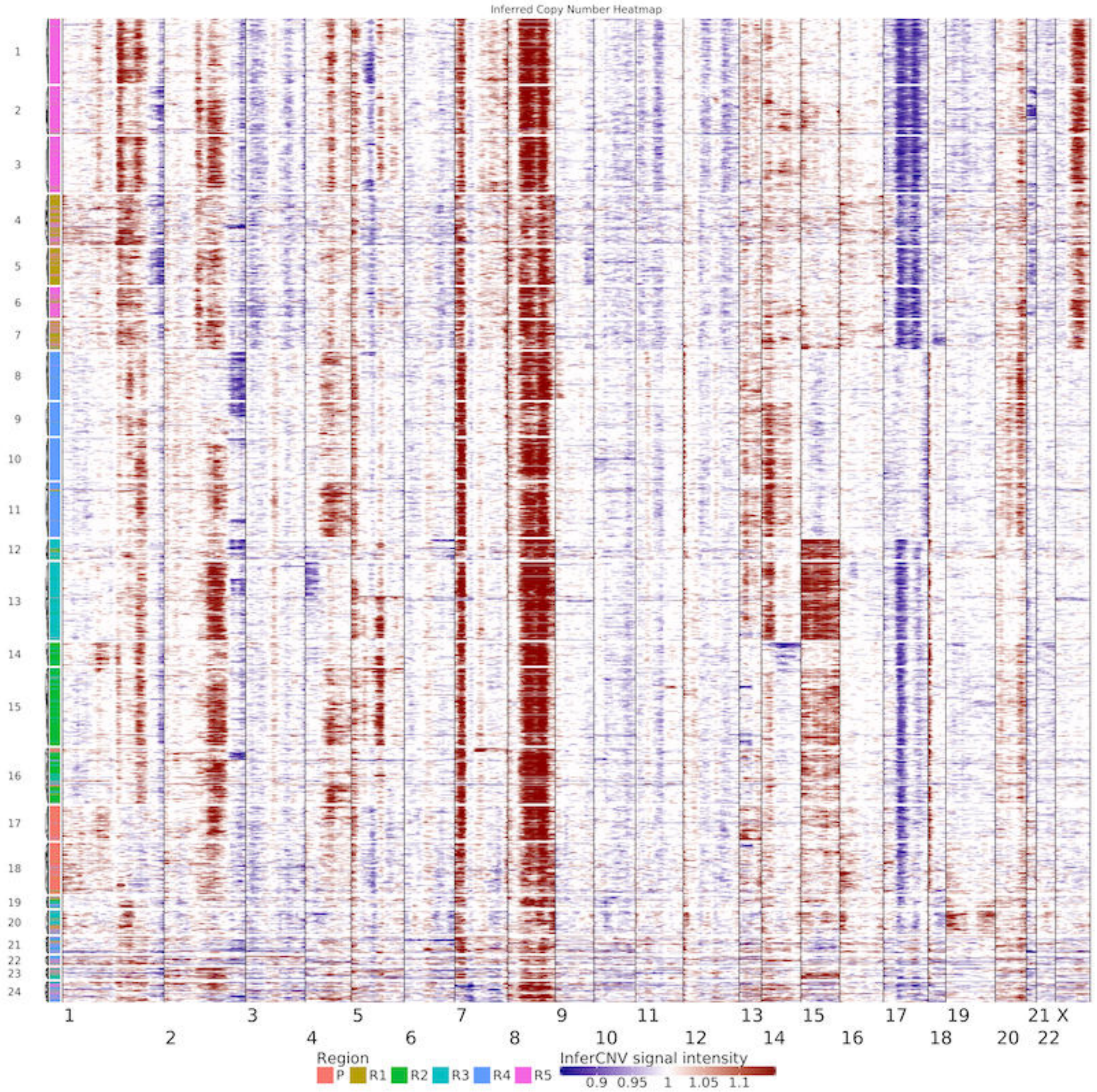


Figure 4.6 InferCNV profiles. Gains and losses relative to normal TME cells are shown in red and blue respectively. Cells are annotated by their region of origin.

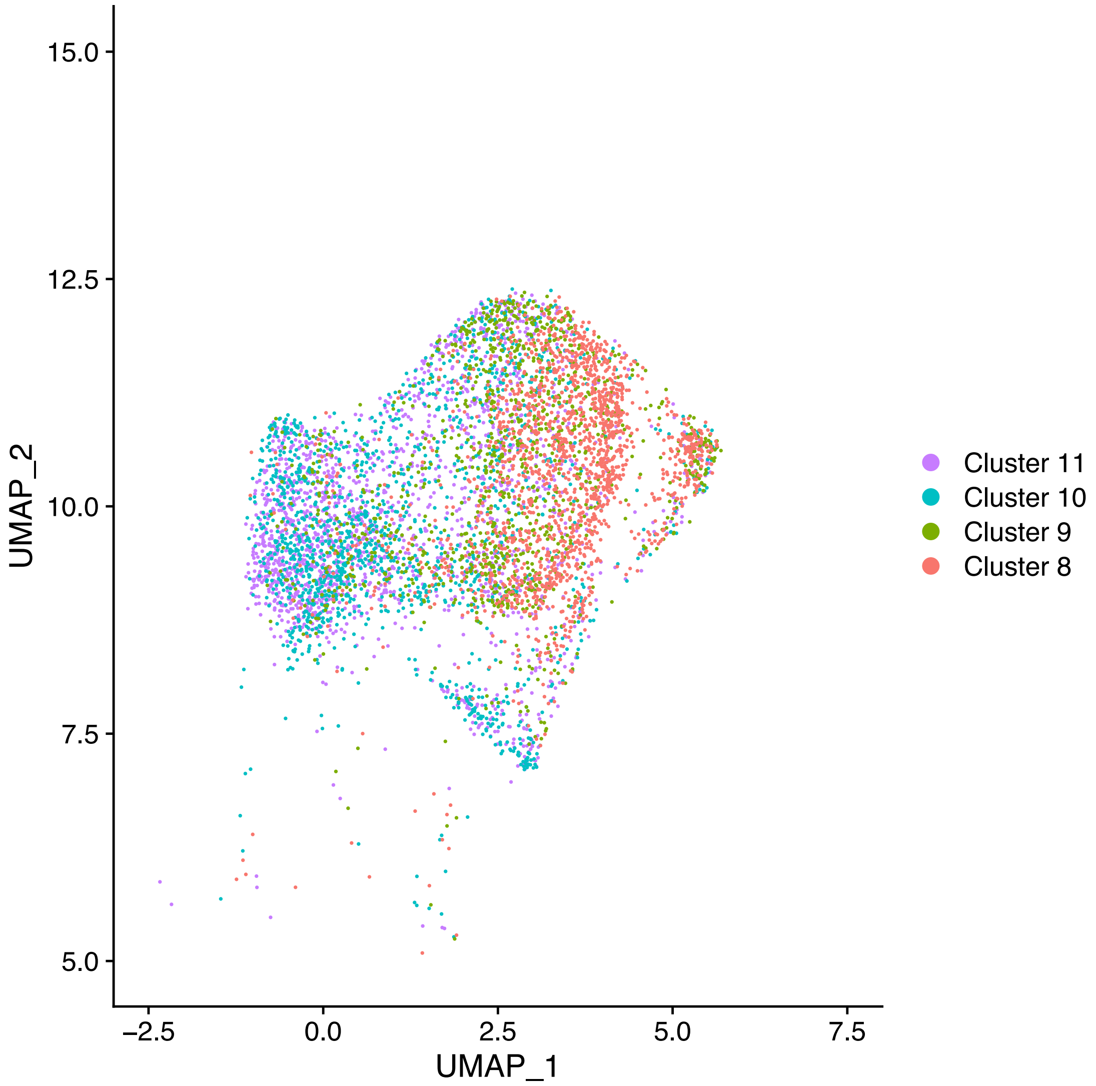


Figure 4.7 UMAP localisation of copy number-based clusters. Four subclones from R4 with unique gains or losses occupy different areas on the expression based UMAP.

4.3.5 Cancer cell state extraction

Identifying cancer cell states through scRNA-seq studies can reveal how tumour-stroma interactions enable immune evasion and resistance. Clusters of cells displayed differences between and within regions for some specific cell states but not others (Figure 4.8 and Appendix 8.1.21). As expected, cycling cells scored highly for the "Cycle" state. Cells from all 5 recurrence regions had a higher "Metal" cell state than those from the untreated primary in keeping with adaptation following systemic anti-cancer therapy. Furthermore, subclone-specific states of “Stress” for the Malignant_R1_2 and Malignant R4_2 clusters were revealed. These findings highlight the ability to characterise functional phenotypes of specific subclones.

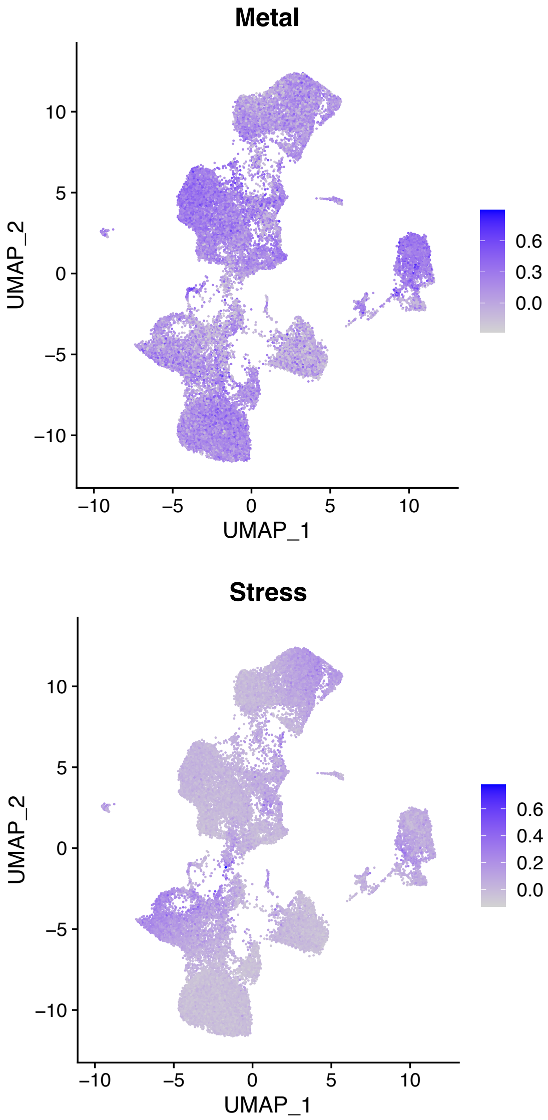


Figure 4.8 Region and subclone specific cell states. “Metal” and “Stress” cell state scores for each cell on UMAP projection.

4.4 CNA-based integration of genotypes and phenotypes

In order to explore phenotypic effects by specific genetic events, integration of single-cell genomic and transcriptomic is required. I attempted different strategies and utilised additional G&T-seq data as part of my own integration method.

4.4.1 Diagonal integration of scRNA-seq cells to scDNA-seq subclones

Given that copy number profiles can be directly determined in the genome and inferred through the transcriptome, I attempted to perform integration across the scDNA-seq and scRNA-seq datasets.

Using Treealign, which models gene dosage effects of CNAs for a subset of genes, scRNA-seq cells were assigned to the copy number subclones established through scDNA-seq. Cells from the primary and R4 were predominantly correctly assigned (Figure 4.9). However, cells from R2 and R3 were assigned to the primary, those from R1 and R5 were shuffled. More concerningly, around one third of cells were unassigned despite seeming to have similar inferred CNAs to those which did undergo assignment.

Whilst this method has been implemented successfully on a small number of datasets, the number of subclones assigned in each case was small (5 to 6) and featured large distinct copy number events from each other. In contrast, many of the subclones of the MPSNT have smaller copy number differences to each other and the subclonal complexity of this MPNST may be too great resulting in incorrect assignments and unreliable performance.

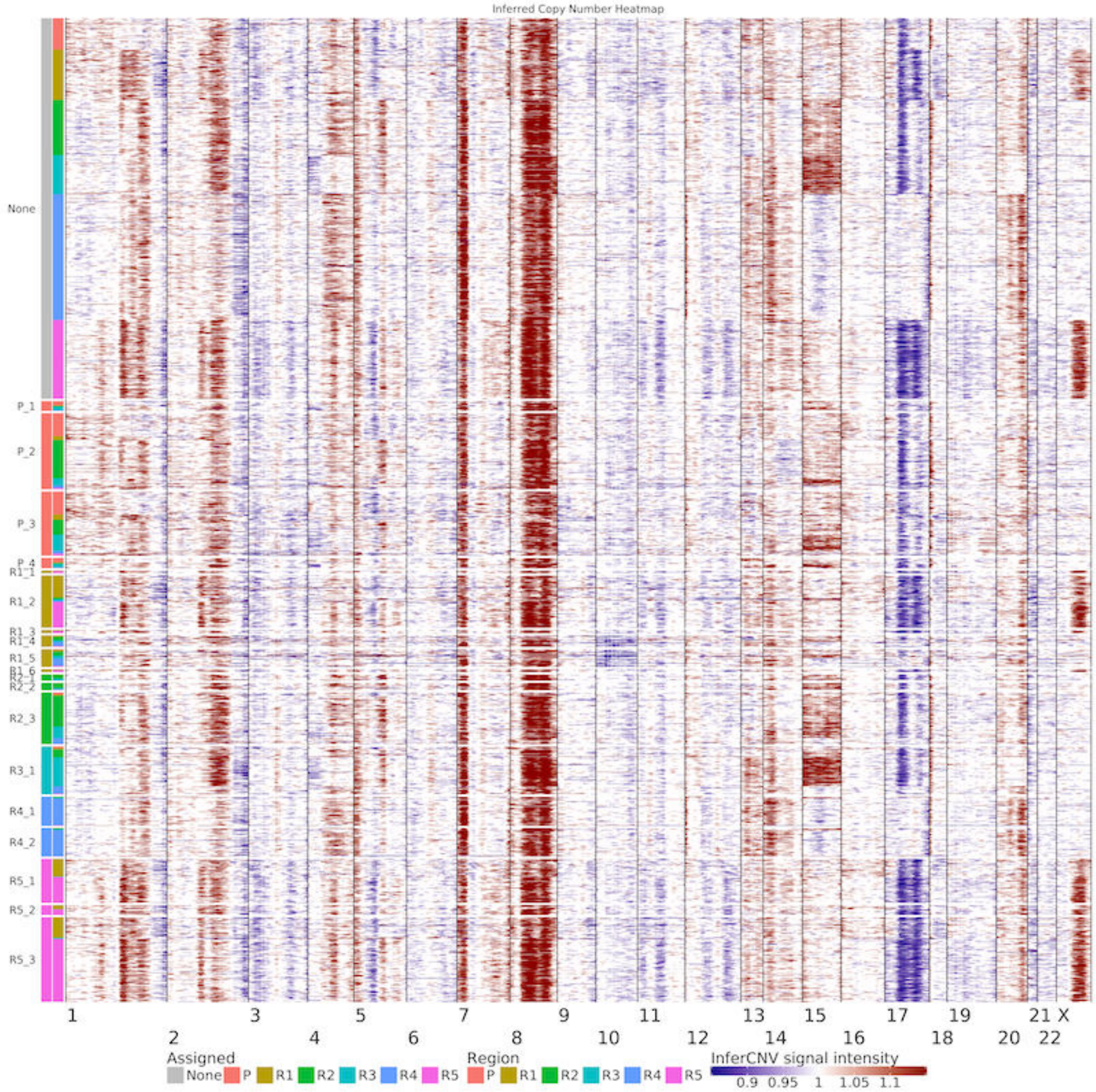


Figure 4.9 Treealign assigned scRNA-seq cells. Cells are annotated by the region of the assigned scDNA-seq subclone on the left column and the region of origin of the cell is shown on the right column. Unassigned cells are shown in grey.

4.4.2 Vertical integration with G&T-seq

Given the inadequacies displayed by Treealign and other packages, I attempted to utilise alternative integration methods. G&T-seq is plate-based technique that involves separation of DNA and full-length mRNA, thereby allowing sequencing of the genome and transcriptome from the same cell. G&T-seq was performed on 96 nuclei sorted from the primary tumour and 96 cells from each recurrence region.

Whilst G&T-seq directly informs which genomes and transcriptomes are derived from the same cell and removes the need for data integration, this methodology is low-throughput. Therefore, I attempted to develop a bespoke integration method by utilising the small number of G&T-seq cells as anchors between the much larger scDNA-seq and scRNA-seq datasets to convert RNA-based inferCNV profiles into integer total copy number profiles.

First, allele-specific copy number profiles were derived for 216 cells across 5 samples and were broadly similar to those from scDNA-seq (Table 3.1 and Figure 4.10). Unfortunately, DNA profiling of the plate of cells from R1 was unsuccessful highlighting the inconsistent success rate of this method. In addition, the presence of an additional clone not seen in the bulk from R4 with a lower level of chr8 amplification and loss of chr10p was revealed. Next, inferCNV profiles were inferred from the matching transcriptomes of these cells (Appendix 8.1.23). Interestingly, the profiles of the primary tumour looked significantly different to others, likely due to those from the primary being derived from nuclei versus cells for the recurrence regions.

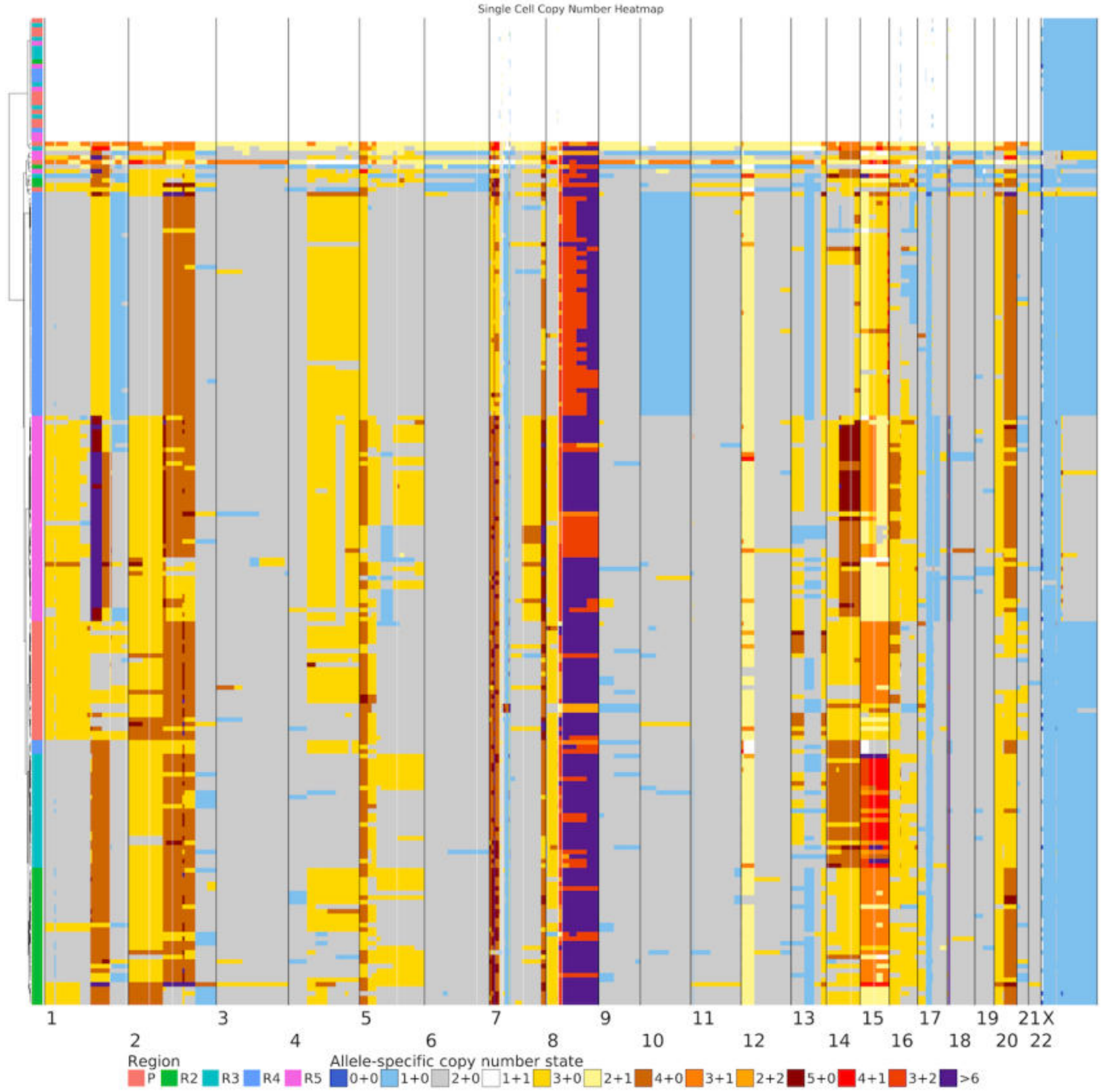


Figure 4.10 Allele-specific G&T-seq copy number profiles.

Inspired by existing batch correction algorithms, I utilised a subset of cells as anchors to derive total copy number profiles from inferCNV values. G&T cells can act as ideal anchors as there is no ambiguity between their genomic and transcriptomic identity. A training set of cells were used to calculate a weight matrix for batch correcting inferCNV profiles. Using a K-nearest neighbours approach, a transformation matrix was then applied to a test set of cells, resulting in copy number profiles which were similar profiles to the ground truth copy number profiles derived from their genomes (Appendix 8.1.24). Furthermore, these corrected copy number profiles showed cells localising mostly by their region of origin on a UMAP projection (Figure 4.11).

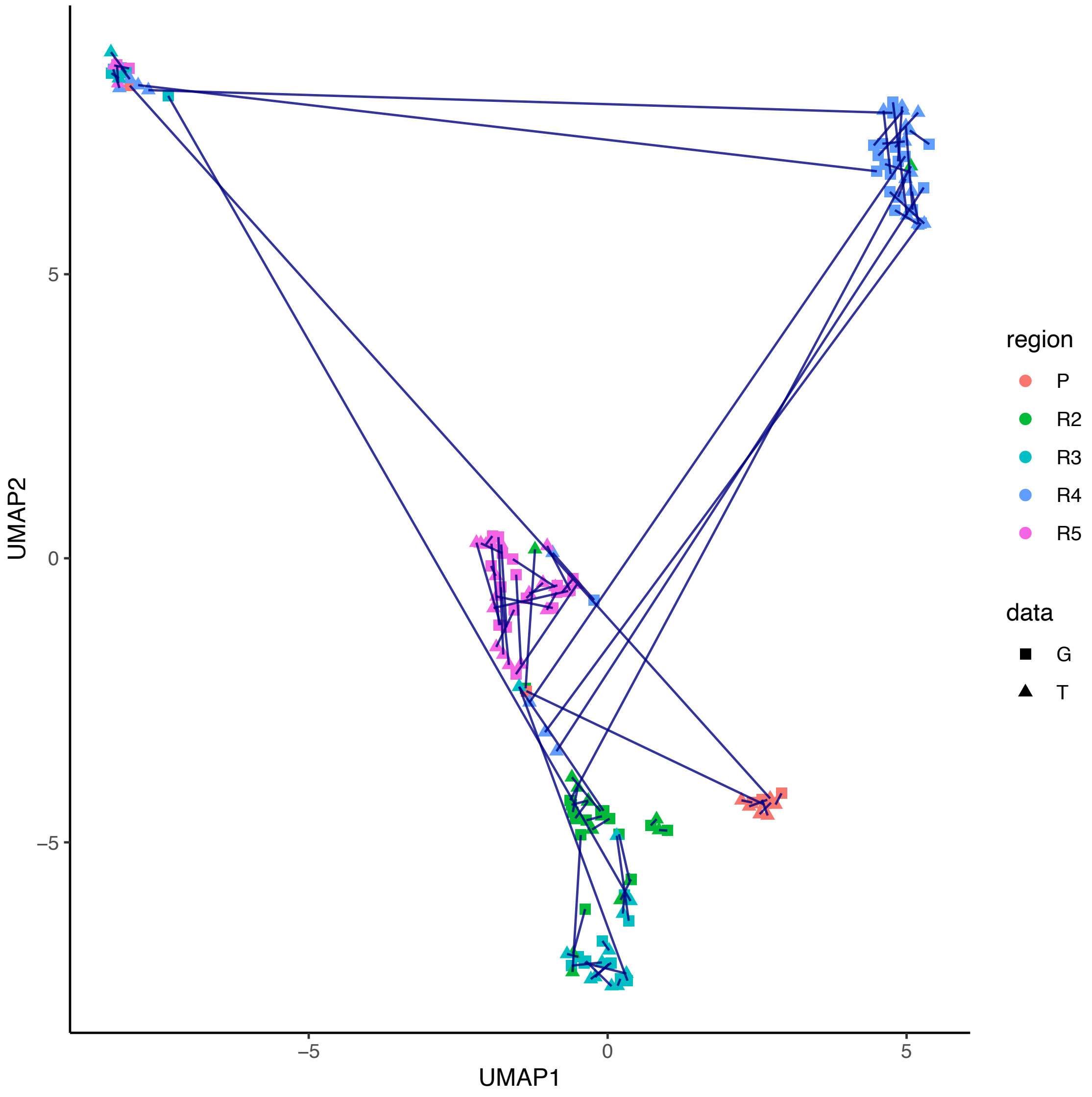


Figure 4.11 UMAP projection of integrated G&T-seq copy number profiles. Test cells are annotated by their genome (G), or transcriptome (T) derived profiles and region of origin. Normal diploid cells from all regions are in the top left.

When applied to scRNA-seq cells, corrected copy number profiles displayed events that were uniquely characteristic for that region. For example, cells from R3 displayed 5 copies of chr15, cells from the Primary, R1, and R5 shared a gain in chr1p and larger loss of chr17 (Figure 4.12). However, a subset of cells was incorrectly corrected to be diploid, likely due to absence of anchors other than those which were diploid. Furthermore, the subtle changes seen in the raw inferCNV intensities such as the subclonal loss of chr2q in R4 was not detected, as their corresponding anchors were not present in the G&T-seq cells due to differences in sampling (Appendix 8.1.25). Therefore, uniform and wide-ranging representation of copy number information by anchors in both datasets is critical for anchor-based batch correction as if anchors are not present then the batch correction becomes incorrect. Overall, whilst this integration has shown it can correctly batch correct copy number profiles for cells with appropriate anchors, it is not yet sufficiently reliable for downstream analysis of assigned scRNA-seq cells for all regions and subclones.

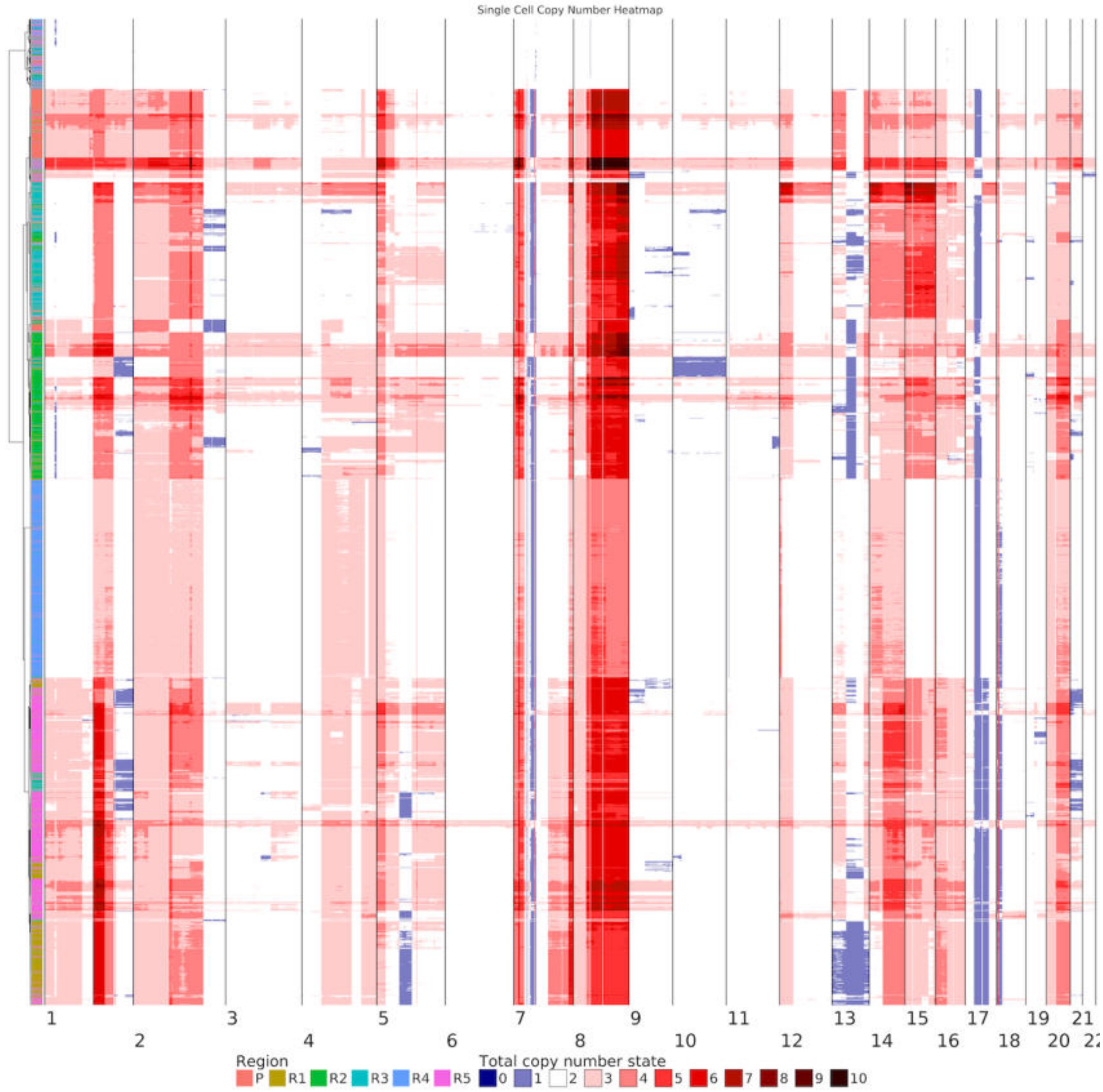


Figure 4.12 Batch corrected total copy number profiles of scRNA-seq cells.

4.4.3 Exploring gene dosage effects

With G&T, gene dosage effects can be explored directly – albeit in a small number of cells. There was a positive correlation between number of copies and expression for most chromosomes (Figure 4.13). However, this correlation gradually diminishes at very high copy number states as shown by chr1, 7, and 8. This implies there may be negative feedback mechanisms which prevent excessive expression. However, not all chromosomes display this correlation with some chromosomes, such as chr19, 20, and 22, seemingly stably expressed despite copy number changes. This may explain the inconsistent performance of diagonal integration which relies upon this relationship.

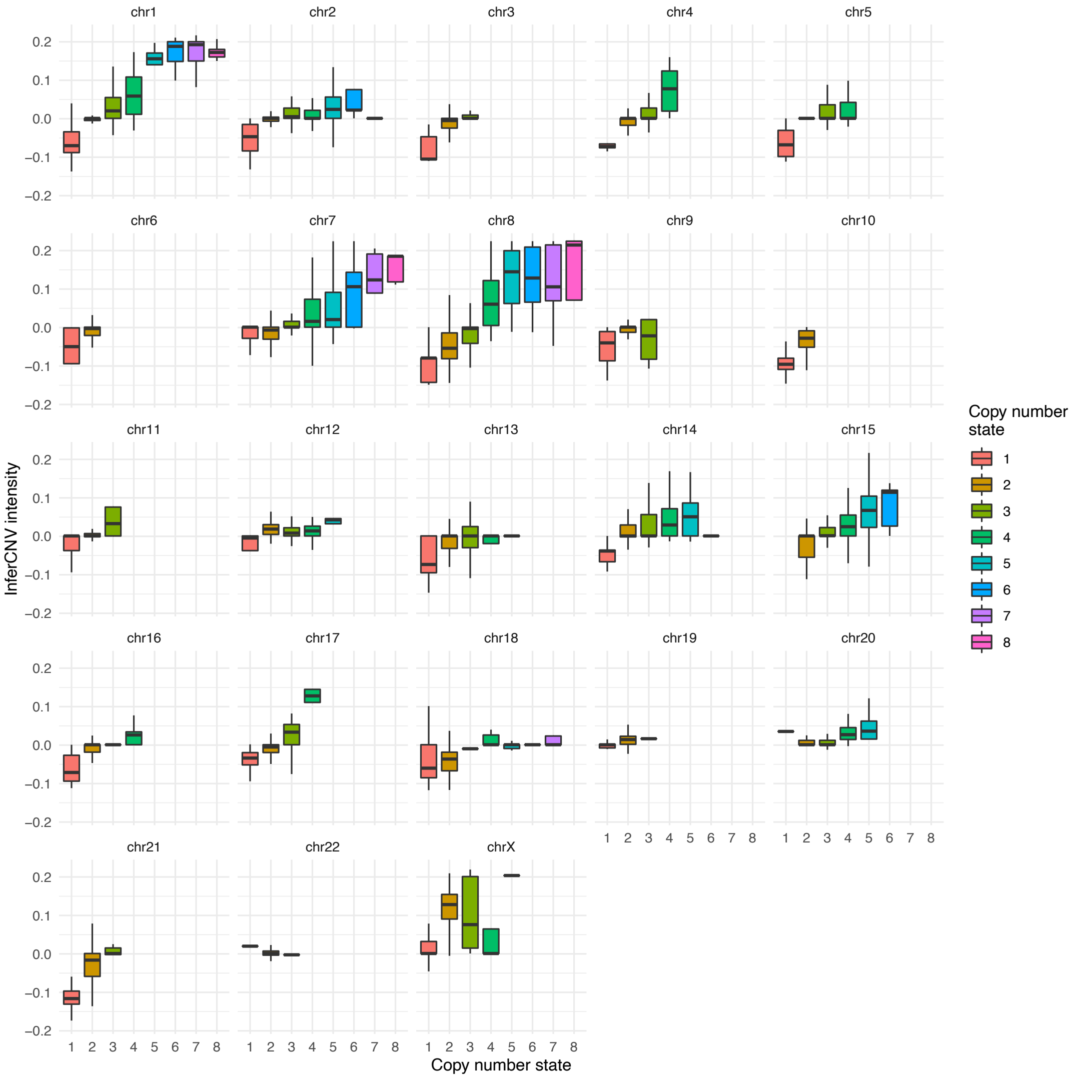


Figure 4.13 Gene dosage effects in G&T-seq cells. InferCNV intensity for each total copy number state for each chromosome.

4.5 Spatial profiling of the tumour microenvironment

Finally, to explore and map transcriptomic effects in a spatial manner and profile the structure of the primary and recurrence regions, 10X Visium spatial transcriptomics was performed on 8 slides with two sections of tissue from the primary tumour (Primary_A and Primary_B) and R5 (R5_A and R5_B), and one section of tissue each from R1, R2, R3, and R4. Briefly, the 10X Visium platform utilises a slide with 4,992 spots which are 55µm in diameter. Each spot is uniquely barcoded and captures mRNAs once the overlaid sectioned tissue is imaged and permeabilised. In addition, single-cell resolution spatial transcriptomics was also attempted with an in-house adaption of Slide-seqV2.

4.5.1 Spot gene expression correlates with morphologically distinct areas

There was morphological variation between slides taken across regions and also between slides from the same region. Most slides, such as the slides from the primary tumour, featured expression-based clusters of spatial transcriptomic spots occupying specific areas which correlated with structures visible on H&E staining suggesting that these areas were composed of different cell types (Figure 4.14 and Appendix 8.1.26). However, the slide from R1 was extremely homogenous and clusters identified based on gene-expression did not appear to be as distinct as those of other tissue sections (Figure 4.14). Furthermore, spots from these clusters were dispersed throughout the slide rather than in specific territories, suggesting there were few local differences in gene expression. This homogeneity was also evident upon histological examination of the section with a uniform sheet of cells present and no distinct structures or regions seen.

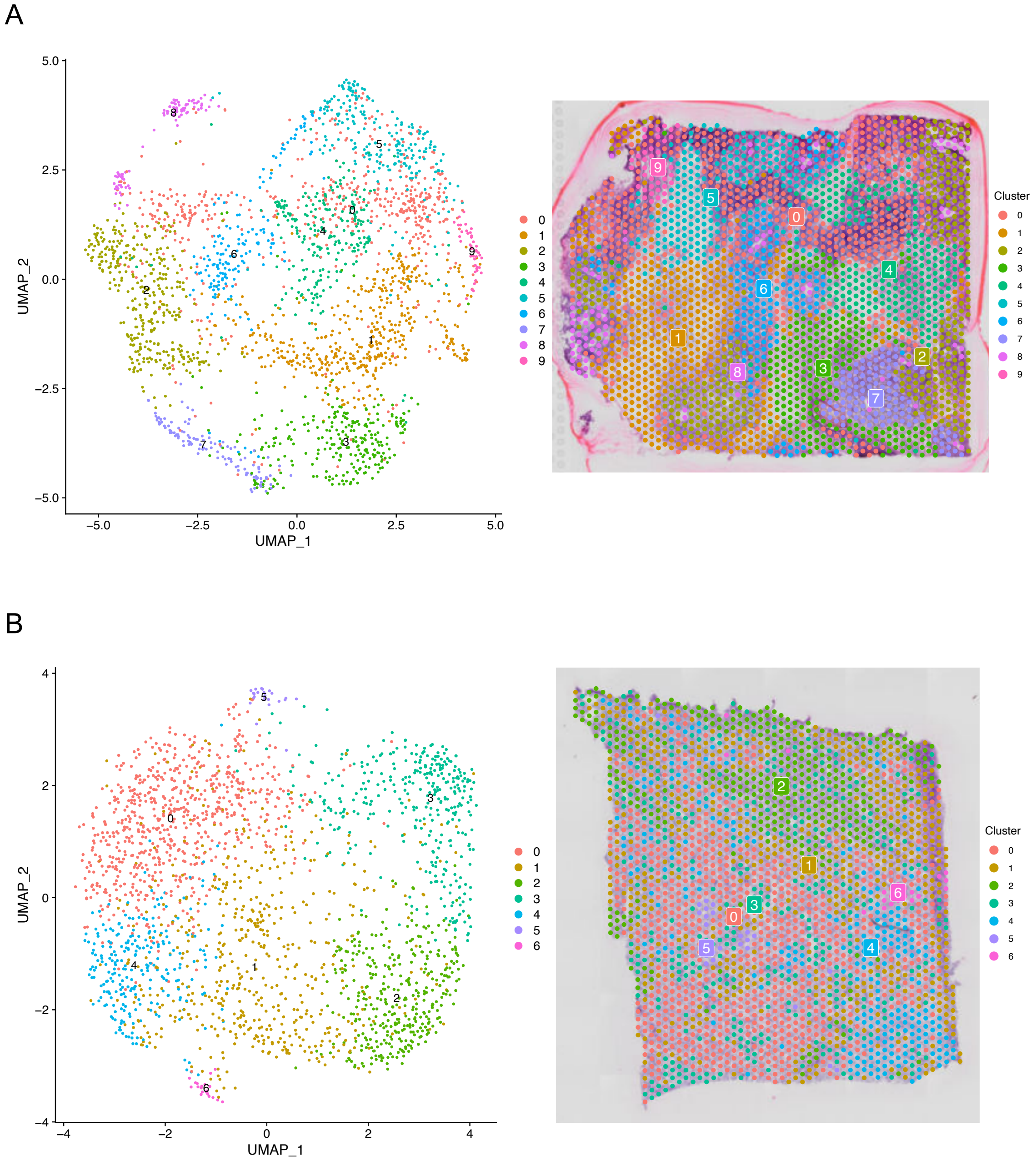


Figure 4.14 Spatial transcriptomics clusters. UMAP projection of gene expression-based spot clusters are shown on the left and cluster annotations overlaid on the tissue are shown on the right for Primary_A (A) and R1 (B) slides.

4.5.2 Deconvolution of spatial spots

Although differential gene expression can be derived for clusters of spatial transcriptomic spots, identifying cell types of spots can be challenging due to mixtures of cell types within a spot. I genotyped heterozygous SNPs for each spot with the aim of determining malignant and non-malignant spots. However, virtually all spots had a constant ratio of counts of each haplotype which was near-identical to those seen for tumour cells in the scRNA-seq data (Appendix 8.1.27). As spots typically cover several cells, the transcriptomes of each spot were likely composed of mostly tumour cells and the rare TME cell. Note, this is in contrast to the LCM sections where several spots of tissue were confirmed to be normal through both copy number profiling and SNV detection (see section 3.6.3).

Therefore, I performed deconvolution of the spots with RCTD, a method to deconvolve cell types in spatial transcriptomic data, by using cell-type profiles learned from scRNA-seq as a reference to extract the expression contribution of different cell types for each spot (Cable *et al.*, 2022). This revealed that most spots were, in fact, predominantly a mixture of tumour and macrophage cells (Figure 4.15 and Appendix 8.1.28). This implies that TME and tumour cells are closely admixed in most regions rather than in certain areas of the profiled tissue section resulting in a high frequency of spots with mRNA captured from a TME cell. In addition, the cell types present in each spot was not easily discernible through histological review of H&E-stained slides as MPNST tumour cells often appear similar to macrophages. The proportions of each cell types were in line with those found in scRNA-seq (Table 4.1) with macrophages being the most abundant TME cell type and rare endothelial and T cells occasionally seen. Furthermore, for the Primary_A section, spots with a higher or pure macrophage presence and spots with endothelial cell presence correlated with a fibrotic area and blood vessels respectively, revealing the cellular composition of certain structures within this tumour (Figure 4.15).

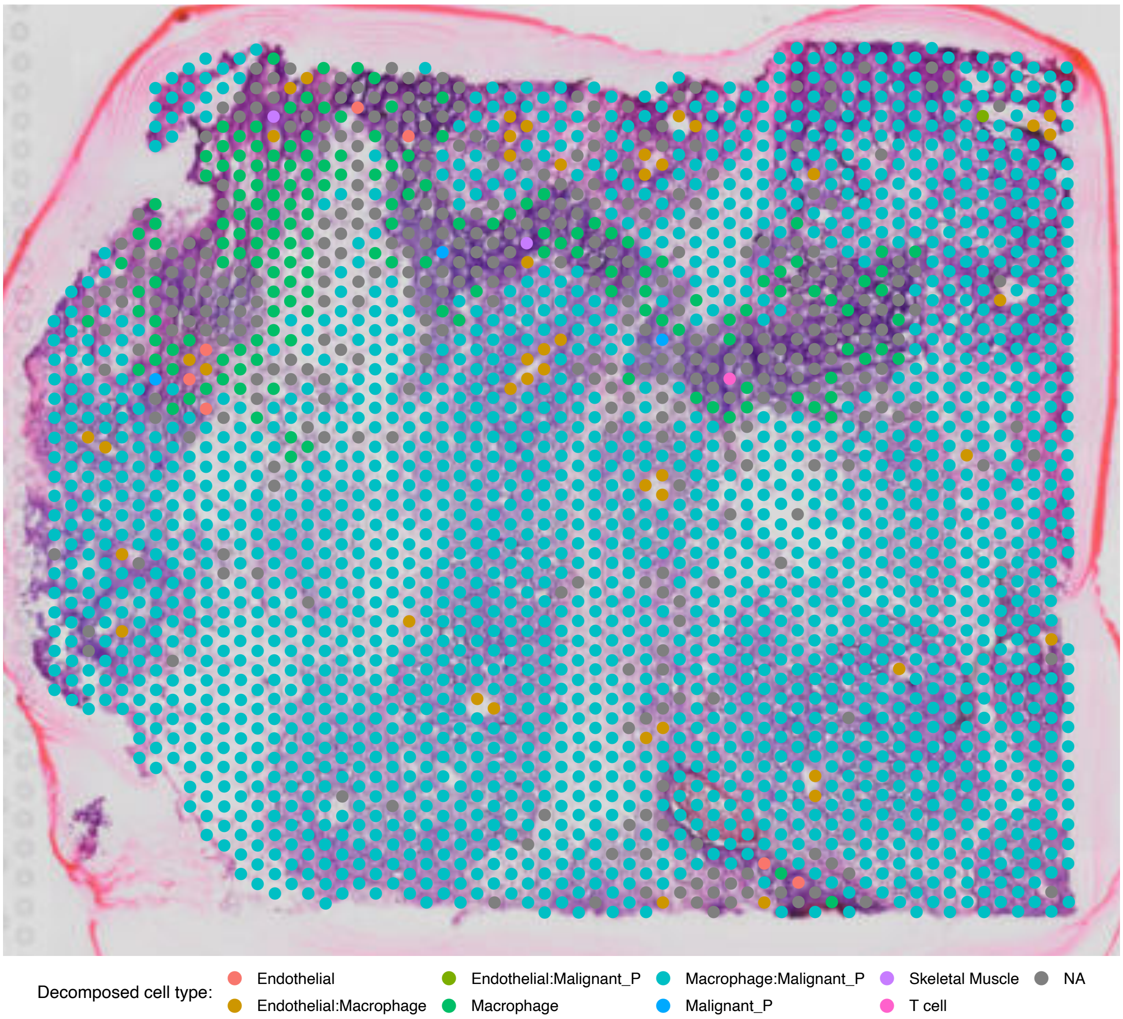


Figure 4.15 Deconvolved cell types in spatial transcriptomics. Spots are annotated by the scRNA-seq cell type present within the spot or the two cell types present if spots are comprised of a mixture of two cells. Spots for which it was not possible to deconvolve cellular composition are shown in grey.

4.5.3 Inference of CNA profiles and spatial tree reconstruction

Whilst the expression profiles of cells from deconvolved spots can be extracted, these do not provide sufficient gene counts for inferring CNAs and are biased due to the reference scRNA-seq expression profiles used to derive them. Therefore, relative copy number differences for each slide were inferred against spots from R1. R1 was selected as a reference as it was the most homogenous and provided a consistent baseline copy number. Relative copy number profiles revealed subclonal heterogeneity in most sections with also different degrees of losses for the same genomic loci present within a slide (Figure 4.16 and Appendix 8.1.29). Copy number differences once again explained most differences in gene-expression as many of the gene-expression based clusters displayed unique relative copy number differences. These subclones featuring specific copy number changes occupied distinct areas of the tissue section reflecting the observations made in the LCM data but at a higher resolution.

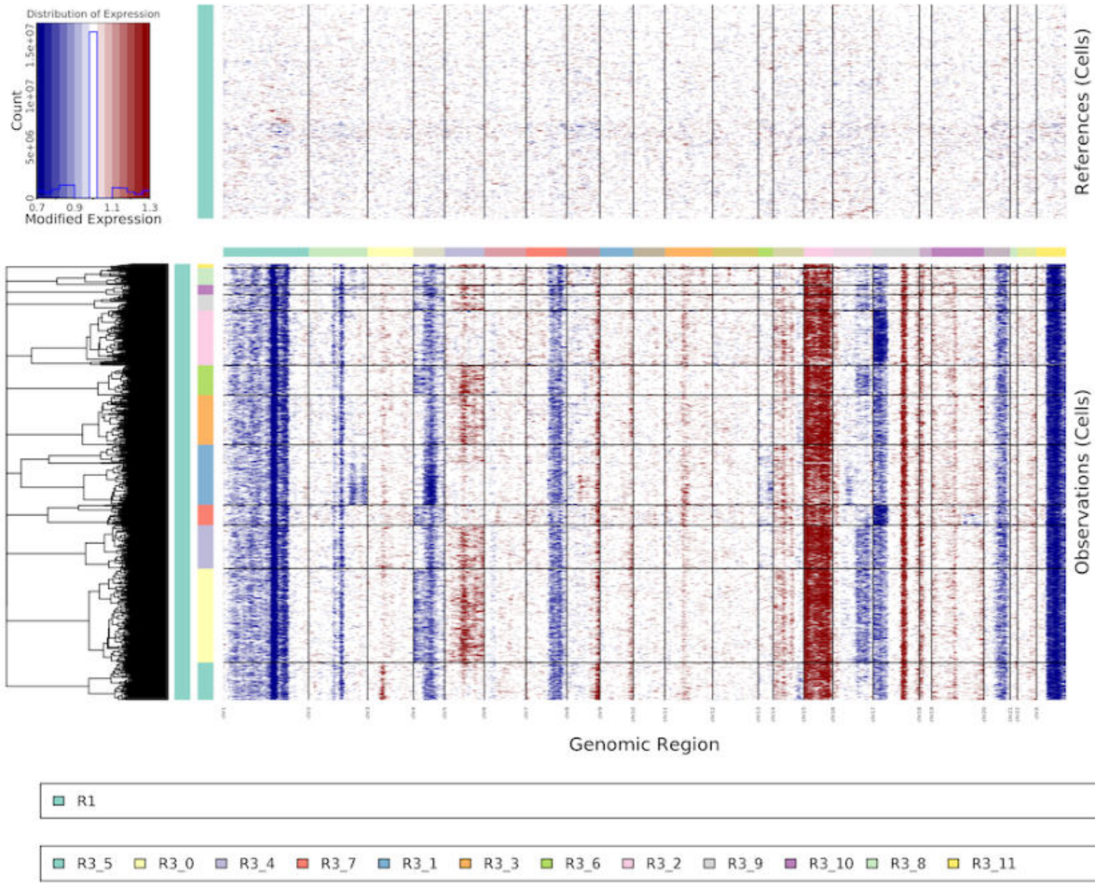


Figure 4.16 Spatial transcriptomic copy number changes. Gains and losses are shown in red and blue respectively. Top panel showing copy number changes in the reference R1 section and bottom panel showing the same for spots from each expression-based cluster relative to spots from R1.

To derive total copy number from relative copy number changes, I utilised the consensus copy number profile of an R1 subclone derived from scDNA-seq data as the baseline state for the R1 reference slide (Appendix 8.1.30). Gains and losses of specific segments relative to this baseline copy number state were then used to calculate integer copy number profiles for each cluster. The total copy number profiles of nearly all clusters matched the scDNA-seq profiles from their region of origin except for four clusters from the Primary_B slides, which were placed closely to single-cells from R1 (Figure 4.17). Furthermore, rare spots in the primary were revealed with profiles similar to those from the R2, R3, and R4 recurrence regions, hinting at the presence of micro-subclones which would eventually proliferate and develop into the recurrence regions (Appendix 8.1.32).

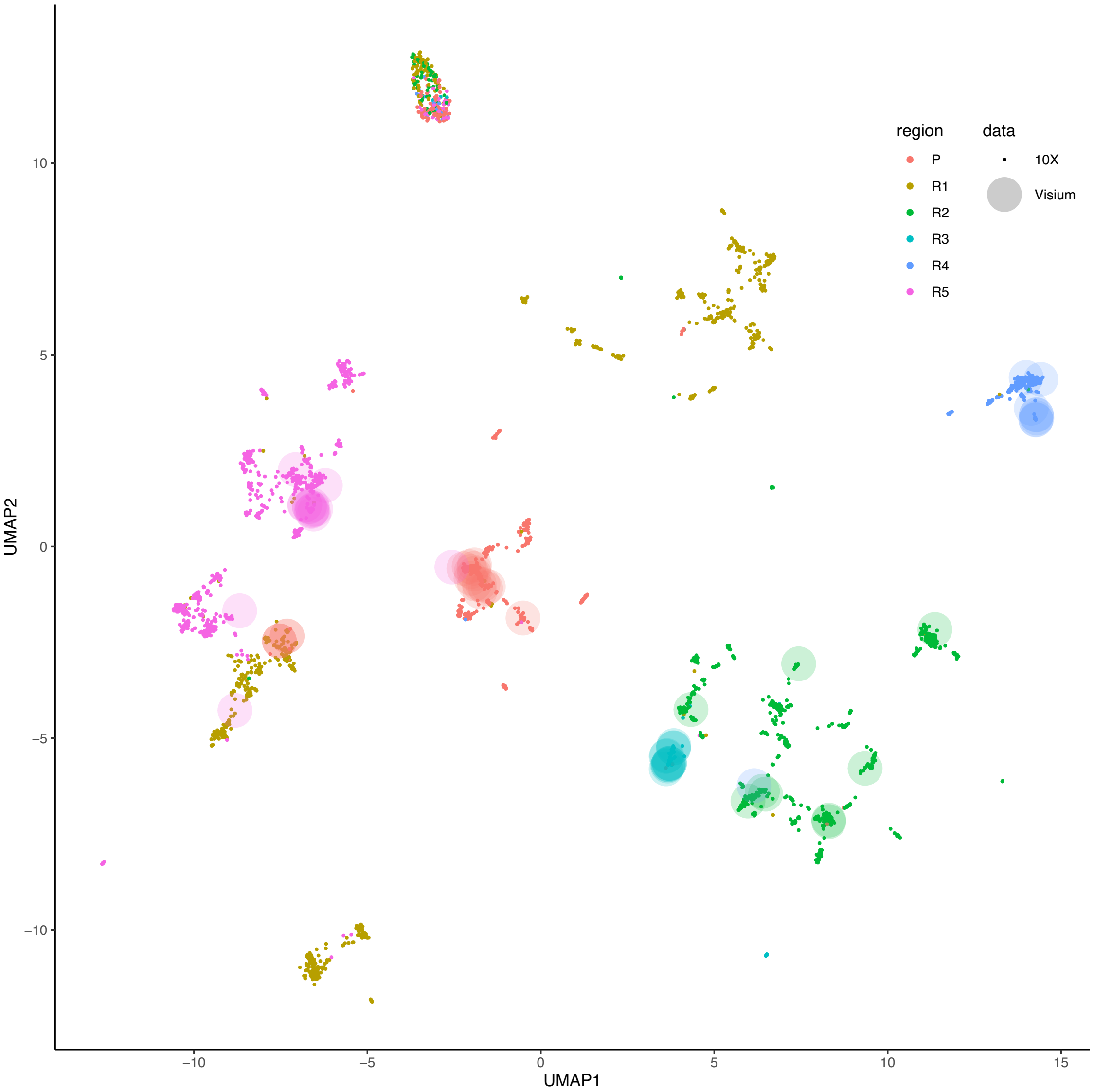


Figure 4.17 UMAP projection of CNAs from scDNA-seq cells and spatial transcriptomic clusters. Cells from scDNA-seq are shown as points and Visium clusters are shown as larger circles. Both are annotated by their region of origin.

Finally, I utilised these clusters as transcriptomic subclones and used their consensus total copy number profiles to reconstruct spatial phylogenetic trees for the centroids of each subclone for each slide (Figure 4.18). Again, detailed subclonal relationships and growth paths for each slide were revealed for some slides. Similar to the LCM phylogenetic trees, some slides displayed unlikely crossing paths which could be explained by sectioning in a perpendicular plane to the growth direction of the tumour. Alternatively, some spots belonging to the same subclone but are assigned as separate clusters due to differences in gene expression or high resolution of cluster detection will result in additional unnecessary branches of the tree (Appendix 8.1.33). Whilst phylogenetic relationships between spatial subclones could be determined, an analysis of their phenotype is unlikely to be informative given the mixture of TME cells.

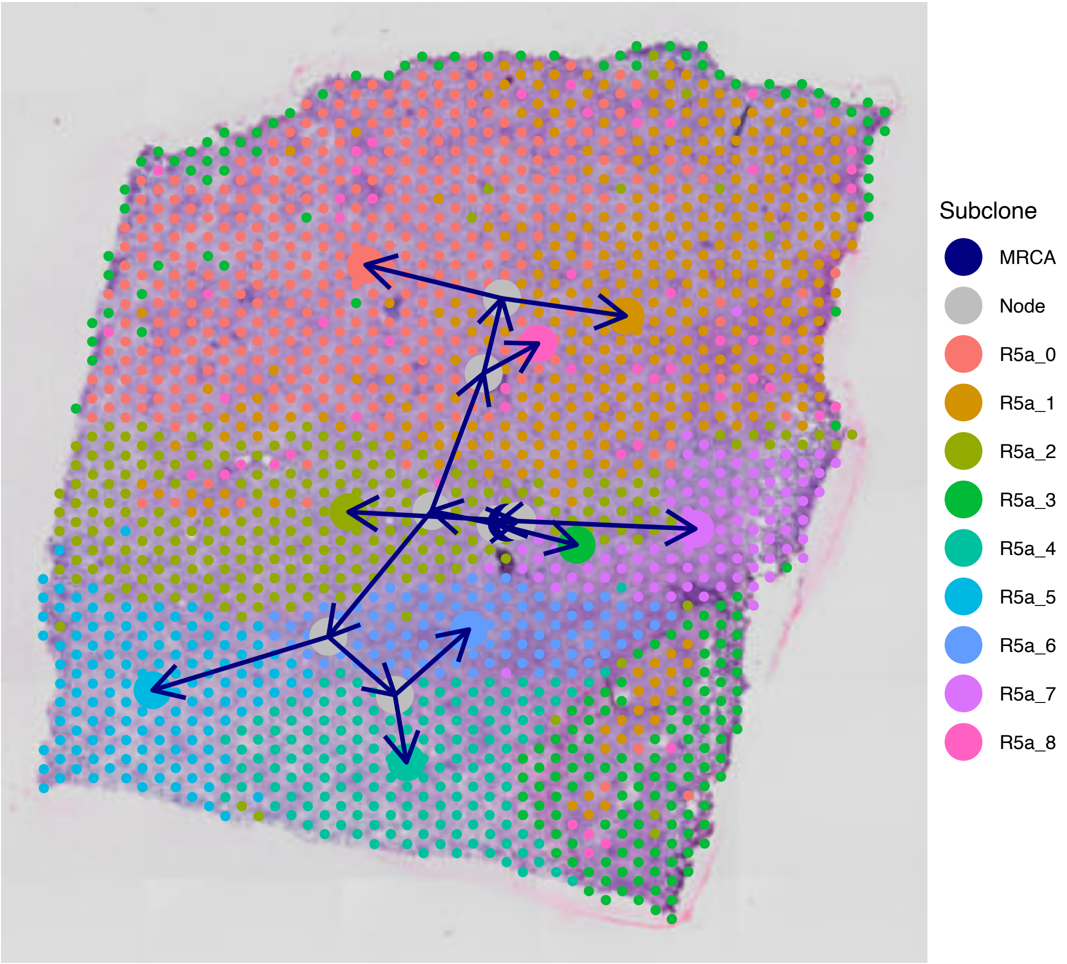


Figure 4.18 Spatial transcriptomic phylogeny of R5_A. Centroids are shown with larger spots and annotated by expression-based cluster. MRCA and intermediate nodes of the tree are shown in navy and grey respectively.

4.5.4 Near-cellular resolution spatial profiling through Slide-seq

Given the limited resolution of the 10X Visium platform, we also sought to spatially profile the transcriptome of this tumour at the single-cell level. Briefly, slide-seqV2 utilises approximately 100,000 uniquely DNA-barcoded 10µm beads deposited as a monolayer on each coverslip or “puck” of 3mm diameter. The barcode and location for each bead can be determined by sequencing by oligonucleotide ligation and detection (SOLiD) which has high accuracy. These beads can capture mRNA released from cryosectioned tissues and undergo reverse transcription to generate a cDNA library which is sequenced to achieve transcriptome-wide spatial profiling of tissues at high resolution.

Slide-seqV2 was performed on two tissue sections each from R2, R3, and R5 as a pilot experiment. However, of the ~5 million reads per puck sequenced, only around 30% (range 12.7% to 35.8%) of reads had a barcode which matched those determined by SOLiD. This was likely due to difficulty in determining barcodes with SOLiD as segmentation of individual beads in the acquired images was challenging and inconsistent.

Two pucks from R2 and R5 were chosen for deeper sequencing with 400 and 447 million reads generated respectively. Subsequently, 56% and 44% of reads were mapped to the transcriptome and matched the barcodes of their puck. This resulted in a relatively low median number of UMIs per bead for R2 and R5 of 823 and 368 respectively.

Whilst the puck from R5 demonstrated some spatial patterns in terms of UMIs per bead and expression-based clusters, it was not possible to identify cell types in either puck (Figure 4.19). Attempts, to integrate the pucks with scRNA-seq to match cell types to beads was also unsuccessful with a negligible number of beads assigned as a cell type. InferCNV profiles also did not detect any relative copy number events. Furthermore, genotyping the pucks for heterozygous SNPs did not identify subsets of tumour or normal beads like in other single-cell data (Appendix 8.1.34). Overall, our in-house implementation of high-resolution spatial transcriptomic profiling through Slide-seqV2 is not yet robust enough for application to tumour tissues and the identification of cell types or tumour subclones.

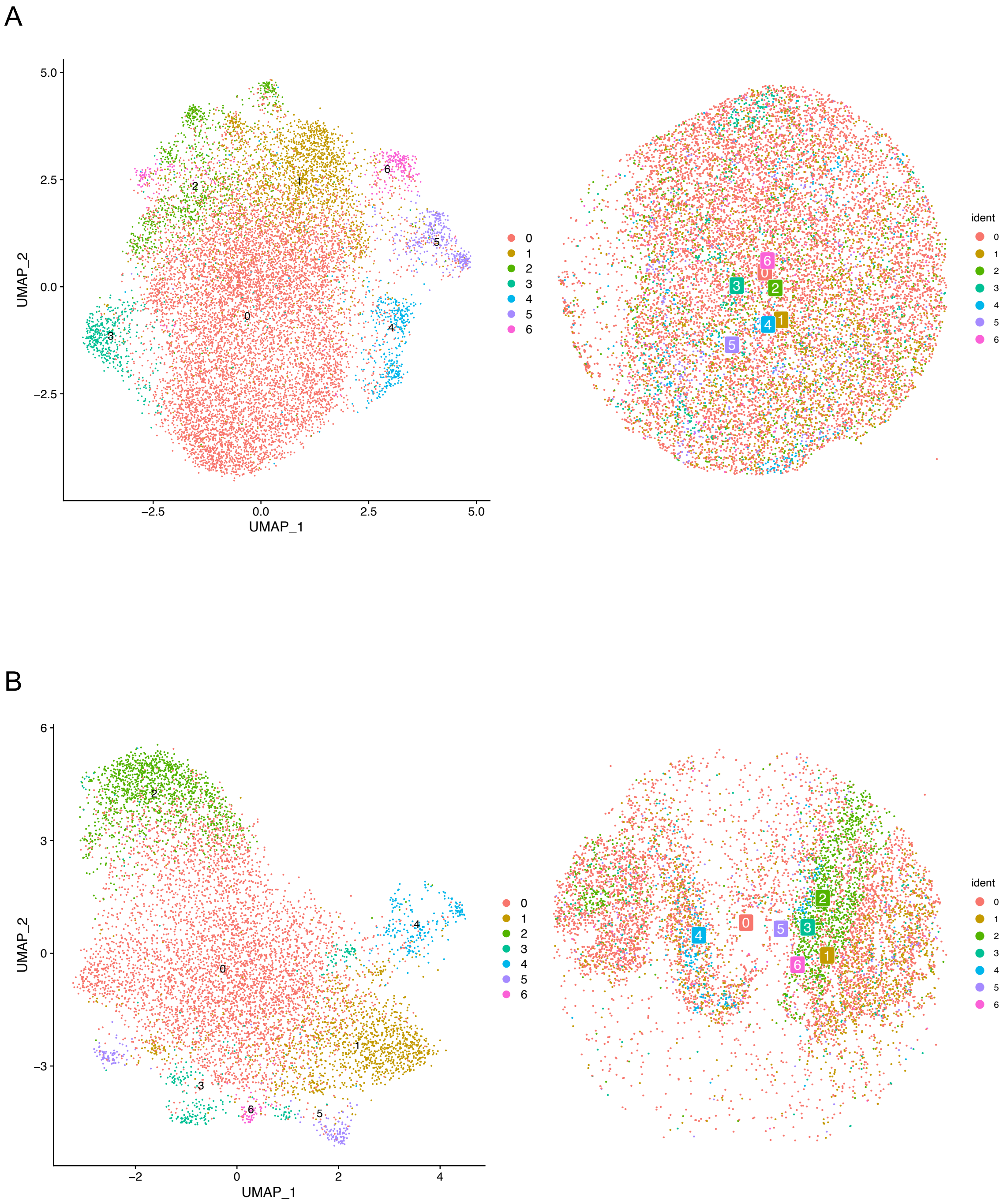


Figure 4.19 Slide-seq pucks and clusters. UMAP projection of gene expression-based clusters of beads are shown on the left and cluster annotations overlaid on the pucks are shown on the right for pucks from R2 (A) and R5 (B).

4.6 Discussion

In this study, by applying state-of-the-art multi-omic bulk, single-cell, and spatial sequencing technologies, I characterised the spatial and temporal evolution of a rare MPNST in fine detail.

4.6.1 Phenotypic diversity is determined by chromosomal instability

Chapter 3 extensively described the extreme diversity generated by CIN in this tumour. This CIN also appeared to have signification implications for the expression profile and phenotypes of tumour cells. Most differences in gene expression were driven by copy number changes at the level of bulk regions but also down to the level of individual subclones. Furthermore, these findings were also translated spatially with gene-expression defined clusters overlapping with tumour areas displaying unique copy number events which were highly localised. This suggests that cells belonging to the same subclone have similar expression and are in close proximity to each other. Therefore, it is likely that ongoing CIN, enables this tumour to generate a vast and diverse set of subclones with differing phenotypes which expand locally, further fuelling cancer evolution.

Phenotypic changes of CIN are likely mediated by gene-dosage effects where more copies of a gene results in more transcripts. These gene dosage effects have been shown to be significant in cancer with gene dosage sensitivity detected in 99% of highly expressed genes (Fehrmann *et al.*, 2015). Gene dosage effects were directly quantified in this tumour through G&T-seq with a correlation between copy number and transcriptional activity observed in most chromosomes. Diagonal integration was also attempted but current methods did not prove sufficiently robust for downstream analysis. This remains an area where methodological development is required to enable higher-throughput studies of these effects. This would allow genes which are under the influence of gene dosage and those where there is dosage compensation to be separated. This dosage compensation may reveal genes whose expression must be finely controlled in tumour cells to allow for survival and may represent potential therapeutic targets.

4.6.2 Unique properties of this MPNST allow detailed evolutionary history reconstruction

This study was made possible by utilising a wide suite of existing packages for a diverse set of data modalities. Established methods for bulk WGS were successfully applied to this tumour with the results of bulk mutational clustering validated in the scDNA-seq data. Whilst mature methods for clustering and inferring copy number changes could be directly applied to scRNA-seq data, newly developed packages for scDNA-seq data, such as ASCAT.sc and MEDICC2 required adaptation and extension to lineage trace and fully resolve the phylogenetic tree using allele-specific CNAs in single cells. Furthermore, given several unique features of the tumour, I also optimised existing packages and developed strategies to maximise inference from the multiple cutting-edge data modalities.

Firstly, by taking advantage of the extreme CNA heterogeneity, we were able to identify single cells with similar CNA profiles as members of a common subclone. I leveraged this native barcoding system for identification of clusters of cells belonging to the same subclones. This enabled me to identify specific SNVs harboured by these subclones to reconstruct the phylogenetic tree to a higher resolution. Furthermore, by applying this principle to scRNA-seq cells, I detected a “Stress” state of cancer cells from specific subclones and a higher “Metal” state in those from the recurrence regions treated with chemotherapy compared to the untreated primary tumour. The “Stress” state may be used to respond to environmental stresses, evade apoptosis and seed metastasis (Baron *et al.*, 2020). Similarly, the “Metal” state increases expression of metallothioneins which may result in chemotherapy and radiation resistance through antioxidant actions and sequestering chemotherapeutic agents (Pedersen *et al.*, 2009). These cells states may be the mechanisms by which the same tumours can display such pleotropic properties such as dormancy, immune evasion, and therapy resistance. Induction of different states in different genetic subclones could suggest that these phenotypes are controlled by genetic mechanisms whereas those generated by the same subclone may be under the influence of epigenetic changes. Overall, this tumour has offered us a glimpse into the degree of detail on tumour progression we can extract by utilising native barcoding and highlights the need for further lineage tracing methods both at the genotypic and phenotypic level.

Despite recent studies investigating how different clones compete and expand, our knowledge of how tumours grow in 3D remains limited, due to difficulties in determining the phylogenetic relationship between cells. By applying copy number-based lineage tracing to the LCM data, I demonstrated the consequences of CIN extended down to small local expansions and that lineage tracing of multiple subclones in a human tumour could be performed. Recently, different computational models of tumour growth have to been used to predict phylogenetic trees and growth processes (Lewinsohn *et al.*, 2023). The phylogenetic trees and geospatial patterns of tumour growth revealed by our LCM data could be used to assess these models and determine the mode of tumour growth. I was also able to extend this lineage tracing into spatial transcriptomics, highlighting the potential for profiling spatial subclones and characterising their interaction with the TME.

In addition, the rare near-haploidisation of the genome enabled me to reliably annotate cells as tumour, normal or tumour-normal doublet using BAF values calculated from chromosome-scale phased haplotypes. Compared to other annotation methods commonly used for scRNA-seq datasets, this method has the advantage of being expression-marker independent. This method was also easily applied to the other data modalities and provided a reliable and orthogonal approach to classifying sequencing data. This was applied to cells in the G&T-seq data, micro-dissected spots generated by LCM and spatial transcriptomics spots. Therefore, it can likely also be applied to other tumours with significant LOH and phased haplotypes available. Whilst this LOH was beneficial for identification of the tumour, it did lead to difficulties in bulk subclonal reconstruction due to the associated high purity. This was overcome by *in-silico* spike in of normal reads to reduce purity and demonstrates the sample specific modifications required for current tools.

4.6.3 Limitations of this study

Previous studies have mostly focused on profiling multiple tumours through one modality in the hopes of classifying tumours into subtypes rather than in-depth characterisation of individual tumours. This work demonstrates the level of detail to which current technologies are capable of resolving phylogenies. However, it is difficult to draw meaningful conclusions about mechanisms of tumour biology from just one tumour. For example, the origin and underlying biology of the cluster of cells expressing high levels of ribosomal proteins with high MAZ activity remains unclear. As MAZ is often overexpressed in tumours and thought to promote tumour progression, this cluster may represent a subset of cells with an important phenotype (Zheng *et al.*, 2022). Furthermore, a recent unpublished study in neuroblastoma has identified extrachromosomal DNA carrying *MYCN* mediating ribosome biogenesis in a subset of cells potentially describing a similar cell state (Stöber *et al.*, 2023). Therefore, this subset of cells warrants further investigation in other MPNST samples.

Whilst the extreme copy number heterogeneity was exceptionally advantageous for the purposes of lineage tracing, this did result in difficulties sampling the same subclones. Even adjacent samples taken from the same region may contain a highly localised subclone which is present in only one sample. For example, the R1_4 and R1_5 subclones seen in the scDNA-seq cells did not appear to be present in the bulk R1 sample or the scRNA-seq. Furthermore, most G&T-seq cells from R4 were sampled from an additional subclone with a chr10 loss, which significantly affected integration performance as the corresponding R4 subclone in the scDNA-seq was not present. Therefore, ideally single nuclei from the same dissociation should be used for these different modalities and ensure the same cell populations are present. This would have several advantages in confirming SNVs and CNAs across modalities and significantly improve the G&T-seq anchor integration method I applied. However, logistically this is very challenging given the multiple modalities used in this study, and profiling of cells rather than nuclei from frozen tissue may be required.

Newly developed spatial methods have been applied to tumour samples, revealing how tumours grow and interact with the TME. However, their resolution often remains limited, and the mixture of cell types confounds analysis. This was particularly evident in my spatial transcriptomic analysis, as despite displaying the same CNAs, spots were placed in different clusters based on gene expression, possibly explained by different cell composition of the spots. Although we attempted to build upon this by performing spatial transcriptomics at the single-cell level, this was ultimately unsuccessful. Indeed, this demonstrates the challenges associated with utilising a nascent technology which is not yet robust enough for widespread adoption. In addition, whilst CNA inference was achieved in this tumour, this is likely only possible due to the very high tumour purity present in these samples and may not be transferable to most tumour types which are typically less pure.

Although I was able to detect subclones in a particular cell state, their underlying controlling regulators cannot be directly identified and would require experimental validation. If these states are the result of epigenetic modifications, additional profiling of the tumour's epigenome may be required. For example, detection of H3K27me3 status at the single-cell level may allow identification of the genes released from transcriptional repression due to its loss. Furthermore, the extent to which these cell states are due to copy number events is also unclear and would be intriguing to investigate experimentally.

Finally, whilst CNAs are indeed detectable across data modalities, diagonal integration was challenging and unreliable. As the number of scDNA-seq datasets increase, additional integration methods between genome and transcriptome will likely be developed to address these shortcomings. Alternatively, future multi-omic technologies may circumvent these integration difficulties and allow genotype and both transcriptomic and epigenetic phenotype to be directly recorded for the same cell.

In summary, this work moves us towards a spatio-temporal representation of tumour development which begins to dissect the relationship between genotype and phenotype. When applied to further datasets, integration of the methods used in this study will be pivotal in uncovering the ancestry and phenotype of important subclones. This may reveal mechanisms resulting in treatment resistance exploited by exceptional subclones and may lead to methodical subclone-informed therapeutic strategies.

Chapter 5. Detection of DTCs in metastatic disease through research autopsy

5.1 Introduction

The vast majority of cancer deaths can be attributed to tumour metastases (Chaffer & Weinberg, 2011). However, cancer in the advanced metastatic setting has been understudied. This is partly due to difficulties in acquiring samples from patients with widespread disease and challenges in developing effective treatments for metastatic patients. Recently, research autopsy studies have enabled the study of metastases, however, the dissemination process and resultant CTCs and DTCs has almost been completely ignored in this context (Iacobuzio-Donahue *et al.*, 2019). For example, it is unclear if the metastases seen at autopsy are derived from a particular subset of DTCs or if there was a general failure in suppression of all DTCs. Whether all metastases can shed and seed DTCs is also unknown. Therefore, there is a need to study the evolutionary history of metastases in relation to DTCs and to profile the role of DTCs in the metastatic setting.

The PEACE study is a prospective research autopsy study with the purpose of facilitating tissue donation from multiple metastases in the post-mortem setting. Early efforts from this study have already contributed key findings in several studies profiling metastasis from various tumour types (Abbosh *et al.*, 2017; Turajlic *et al.*, 2018a; Litchfield *et al.*, 2020; Spain *et al.*, 2023). In addition to tumour tissue, normal tissues can be collected during autopsy from organs which cannot be easily sampled in living patients. Therefore, I collected macroscopically normal tissues in an attempt to detect and profile DTCs in metastatic patients. The main tumour types investigated in this study were clear cell renal cell cancer and melanoma patients – two cancer types with high metastatic potential.

5.1.1 Clinical studies of DTCs in renal cell carcinoma and melanoma

In contrast to breast cancer, very little is known about the prevalence or prognostic impact of DTCs in solid tumours. In renal cell carcinoma, only two studies have investigated the presence and prognostic value of DTCs. In non-metastatic patients, DTCs were detected with an antibody against cytokeratin 18 and, depending on the stage of disease, identified in 23-31% of patients (Buchner *et al.*, 2003). This study reviewed 2 million mononuclear cells from bone marrow aspirates with a median of 1 cell per million detected in patients with cytokeratin-positive cells. However, there was no significant difference in survival between patients with cytokeratin-positive cells and those without. The same group also performed a similar study in metastatic patients, which detected cytokeratin-positive cells in 42% of patients (Buchner *et al.*, 2006). In contrast to limited stage disease, the presence of three or more positive cells was significantly associated with poor prognosis in the metastatic setting. However, despite reviewing double the number of cells from the bone marrow, this study reported a similar distribution of cytokeratin-positive cells with only one or two detected in most patients. To date, no studies have performed detailed characterisation of isolated DTCs in ccRCC and little is known about their relationship to the primary tumour or other metastases.

In contrast, the high relapse rates of melanoma following resection have resulted in multiple studies of DTCs. The lymph nodes have been a particularly active area of research in melanoma due to the nature of lymphatic spread in melanoma. The sentinel node is critical as it is the lymph node that directly receives lymph from the melanoma. Presence of melanoma cells in the sentinel node biopsy confers poor prognosis and is part of the American Joint Committee on Cancer (AJCC) 2017 staging system (Morton *et al.*, 2006; Amin *et al.*, 2017). However, the role of complete lymph node dissections has been contentious, with two randomised trials failing to demonstrate a benefit in overall survival (Bello & Faries, 2020). Thus, how CTCs interact and progress through lymph nodes and disseminate to target tissues to become DTCs is an important area of study.

Studies investigating the presence of DTCs in lymph nodes have detected and consistently associated increasing numbers of lymph node DTCs with worse prognosis (Ulmer *et al.*, 2014, 2018). One study utilised a size-based microfluidics device to isolate melanoma cells and performed RT-PCR based transcript detection and array CGH for copy number profiling, revealing copy number changes characteristic of melanoma (Weidele *et al.*, 2019). Finally, another study also utilised array CGH which revealed limited shared copy number aberrations between DTCs and the primary tumour (Werner-Klein *et al.*, 2018). The lack of a direct ancestral relationship suggested that either these cells are present in the primary tumour but at low prevalence, or that dissemination of DTCs to the lymph nodes occurs early.

For melanoma cells in the bone marrow, a small study found four out of 20 patients with DTCs although this included three patients with stage IV disease (Thybusch-Bernhardt *et al.*, 1999). More recently, DTCs detected by flow cytometry were reported in 28-65% of melanoma patients, depending on stage (Chernysheva *et al.*, 2019). However, as in most DTC studies, they did not investigate prognosis or further characterise the cells which were identified as DTCs. In addition, another set of studies reports the presence of DTCs in uveal melanoma – a disease genomically distinct from cutaneous melanoma (Eide *et al.*, 2009, 2013, 2019). However, there was with conflicting data on whether the presence of DTCs confers good or poor prognosis.

5.1.2 Selection of markers for DTC detection

The choice of a tumour marker requires careful consideration. As the two tumour types profiled in this study are not of epithelial origin or do not reliably express epithelial markers, EpCAM (the most widely used marker in DTC studies due to the lack of epithelial expression in haematological tissues) cannot be used for the detection of DTCs.

For ccRCC, EpCAM based approaches for isolating CTCs have proven unreliable as EpCAM expression is variable in RCCs (Went *et al.*, 2005; Zimpfer *et al.*, 2014; Maertens *et al.*, 2017). CAIX is expressed in 90% of clear cell RCCs and was chosen for this study, as it is the most widely used IHC marker clinically. In addition, CAIX has been successfully used to isolate CTCs and is expected to fill the gap of a lack of appropriate surface markers in CTC research (Takacova *et al.*, 2012; Liu *et al.*, 2016). CAIX is a marker of hypoxia as its promoter contains a hypoxia responsive element where the Hypoxia-Inducible Factor 1 (HIF-1) binds. In ccRCC, VHL loss leads to dysregulated hypoxia pathways and the resulting accumulation of HIF-1 increases CAIX transcription.

For detection of DTCs in melanoma, EpCAM also cannot be used as melanocytes are of neural crest origin. For the isolation of CTCs, CD146 or melanoma cell adhesion molecule (MCAM) and MCSP has been used as a surface marker as part of the CELLSEARCH CTC detection system in a two-step process where CD146 cells are isolated first and MCSP used to select melanoma cells from this population (Rao *et al.*, 2011; Khoja *et al.*, 2013). MCSP was chosen for this study as CD146 is expressed on many endothelial cells which are present in the normal lung and liver tissues collected from autopsy. Other antibodies such as HMB45 or those against gp100 are used clinically and have also been used for CTC or DTC detection, but they require fixation of the already fragile cells derived from research autopsy so were not used during this study (Werner-Klein *et al.*, 2018; Chernysheva *et al.*, 2019).

5.2 Aims

The aim of this project was to:

- Optimise dissociation of tissues from autopsy and generate viable single cell suspensions.
- Detect and isolate DTCs or micrometastases from histologically normal tissues and map their relationship to the primary tumour and metastases.
- Confirm the existence of aberrant cells in histologically normal tissues and characterise their genomes.

5.3 Summary of autopsy samples collected

In total, tissues deemed normal upon inspection by the pathologist at time of autopsy was collect from 31 patients. In addition, two melanoma metastases were also collected from the TRACERx Melanoma trial to corroborate flow cytometry staining of melanocytes. These samples were collected from patients MX172 and MX417 and were resected solitary thigh metastases that had developed following surgical removal of the primary. A summary of patients whose tissues were used for analysis and their clinical histories is shown in Table 5.1.

Patient	Sex	Age	Primary	Metastases	Treatment				Samples Collected			
					SR	RT	TT	IO	Bone	Lung	Liver	Tumour
PEA213	M	93	ccRCC	Lung, liver, bone, LN		✓	✓		✓			-
PEA227	F	52	ccRCC	Brain, bone, adrenal, LN					✓			-
PEA279	M	66	ccRCC	Pancreas, bone	✓	✓	✓	✓	✓	✓	✓	Pancreas
PEA183	M	70	ccRCC	Brain, lung, liver, bone, LN	✓	✓	✓	✓	✓	✓	✓	Liver
PEA142	F	71	ccRCC	Liver, bone		✓	✓	✓	✓	✓	✓	Kidney
PEA182	M	59	ccRCC	Lung, bone, LN		✓	✓	✓	✓	✓	✓	Kidney, lung, LN
PEA293	F	70	ccRCC	Bone, LN	✓		✓	✓	✓	✓	✓	LN
PEA329	F	49	ccRCC	Lung, bone, LN, subcutaneous		✓	✓	✓	✓	✓	✓	LN
PEA314	M	71	ccRCC	Lung, bone, adrenal, LN	✓	✓	✓	✓	✓	✓	✓	LN
PEA192	M	91	ccRCC	Lung, liver, LN	✓	✓	✓	✓	✓	✓	✓	Lung
PEA352	M	55	ccRCC	Lung, LN	✓	✓	✓	✓	✓	✓	✓	LN
PEA276	F	50	Collecting duct carcinoma	Lung, liver, bone, LN	✓	✓	✓		✓	✓	✓	Lung, liver
PEA294	M	57	Melanoma	Brain, lung, adrenal, LN	✓	✓	✓	✓	✓	✓	✓	Brain, LN
PEA312	M	29	Melanoma	Brain	✓		✓	✓	✓	✓	✓	Brain
PEA128	F	59	Melanoma	Brain, LN	✓		✓	✓	✓	✓	✓	Brain, LN
PEA126	F	31	Melanoma	Brian, bone, LN, subcutaneous, muscle	✓		✓	✓	✓			-
PEA203	F	78	Melanoma	Brain, Lung, adrenal, LN	✓			✓	✓			-
PEA212	M	80	Melanoma	Lung, bone, LN	✓			✓	✓			-

PEA230	F	56	Melanoma	Brain	✓		✓		✓			-
PEA249	M	82	Melanoma	Liver, LN	✓			✓	✓	✓	✓	-
PEA206	F	55	Melanoma	Lung, liver, adrenal, spleen, bone, LN, subcutaneous	✓	✓	✓	✓	✓	✓	✓	Liver, lung
PEA286	M	56	Melanoma	Lung, adrenal, bone, LN	✓			✓	✓	✓	✓	Lung, LN
PEA306	F	61	Melanoma	Brain, bone		✓	✓	✓	✓	✓	✓	-
PEA304	M	65	Melanoma	Brain, lung, LN	✓			✓	✓	✓	✓	Lung
PEA347	M	85	Melanoma	Brain, liver, LN	✓	✓		✓	✓	✓	✓	Liver
PEA151	M	53	Uveal Melanoma	Bone, liver, LN		✓		✓	✓			-
PEA224	M	57	Uveal Melanoma	Lung, liver, adrenal, LN		✓		✓	✓	✓	✓	✓
PEA122	M	79	Uveal Melanoma	Lung, liver, bone, subcutaneous		✓		✓	✓			-
PEA254	M	51	Uveal Melanoma	Brain, lung, liver, subcutaneous, LN		✓		✓	✓	✓	✓	-
PEA245	F	72	Uveal Melanoma	Lung, liver, cardiac, bone, subcutaneous		✓		✓	✓	✓	✓	Lung, liver
PEA344	F	76	Uveal Melanoma	Lung, liver, bone, LN		✓		✓	✓	✓		-

Table 5.1 Clinical background and samples collected through research autopsy. LN, lymph node; SR, surgery; RT, radiotherapy; TT, targeted treatment; IO, immunotherapy.

5.4 Digestion and generation of single-cell suspension from research autopsy tissue

Single-cell suspensions were generated by enzymatic digestion using collagenase. Bone and lung samples were noted to frequently have large clumps of cells, unlike suspensions from liver samples. Treatment with DNase significantly improved this clumping. Therefore, it is likely that DNA released from the breakdown of dead cells was causing this clumping. However, DNA extraction is likely reduced or partially inhibited following DNase treatment, therefore only samples without DNase treatment were used, as genomic analysis of single cells was a key aim of this project.

5.5 Detection and genomic profiling of CAIX$^+$ cells in normal tissues from patient with ccRCC

3 patients (PEA142, PEA182, and PEA293) were selected for preliminary detection and subsequent analysis of DTCs. A detailed graphical representation of their clinical history is shown in Figure 5.1.

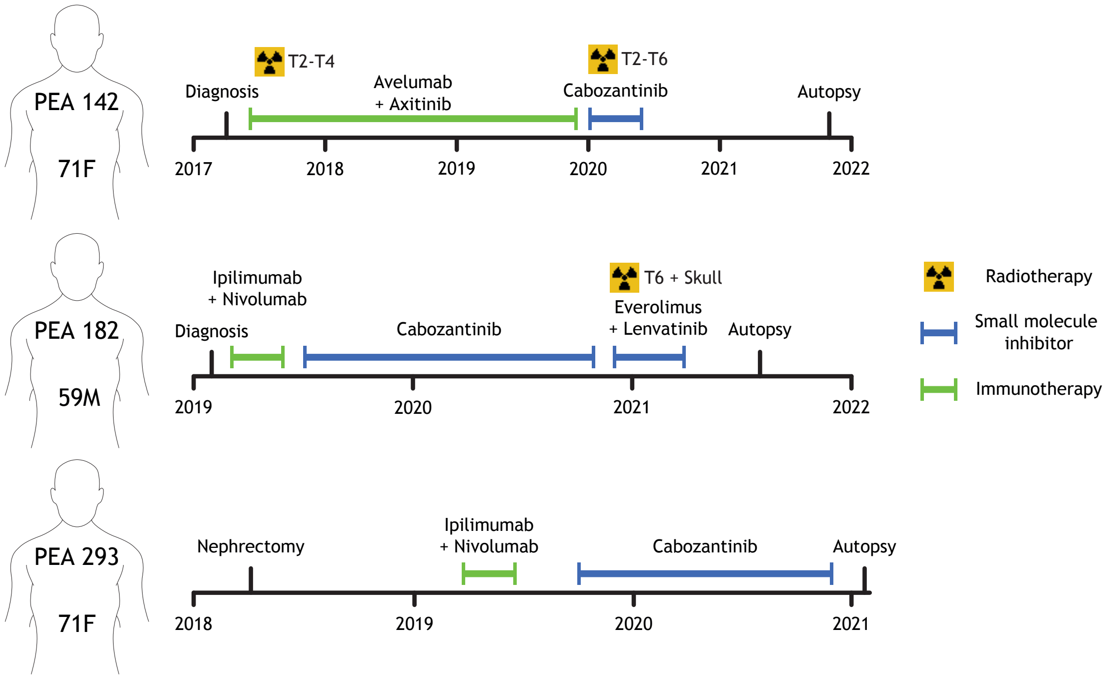


Figure 5.1 Clinical histories of three pilot autopsy patients.

5.5.1 Flow cytometry analysis of ccRCC

To confirm staining of CAIX in ccRCC cells, flow cytometry was performed on an *in situ* primary kidney tumour from PEA182. As expected, a large number of CAIX$^+$ cells were seen for the primary collected for PEA182 (Figure 5.2). Furthermore, a separate population of EpCAM$^+$ cells was observed, and these cells were mostly CAIX$^-$. They could be either EpCAM$^+$ RCC cells or cells from the tumour microenvironment, in keeping with previous studies showing that ccRCCs are variable in their EpCAM expression despite their epithelial origin. Further FACS experiments performed for isolating single cells for scRNA-seq (see section 5.7) also detected large proportions of CAIX$^+$ cells in other primary tumour or metastasis samples (Appendix 8.2.4).

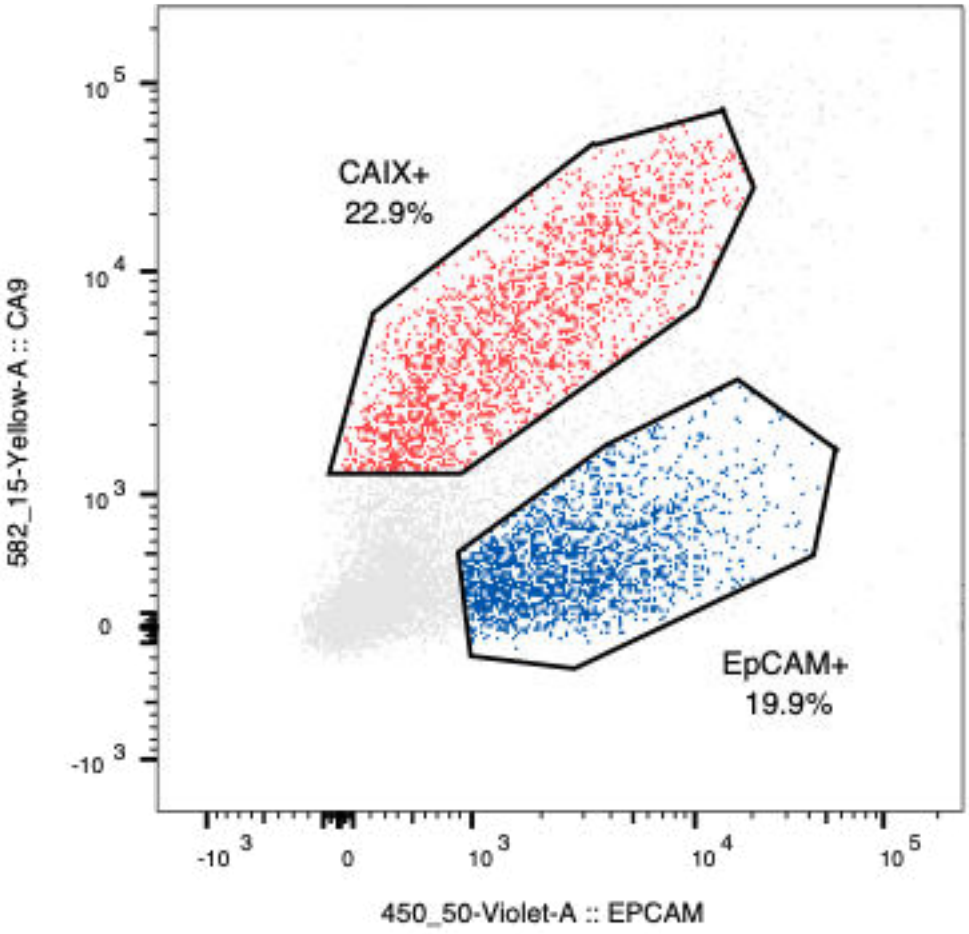
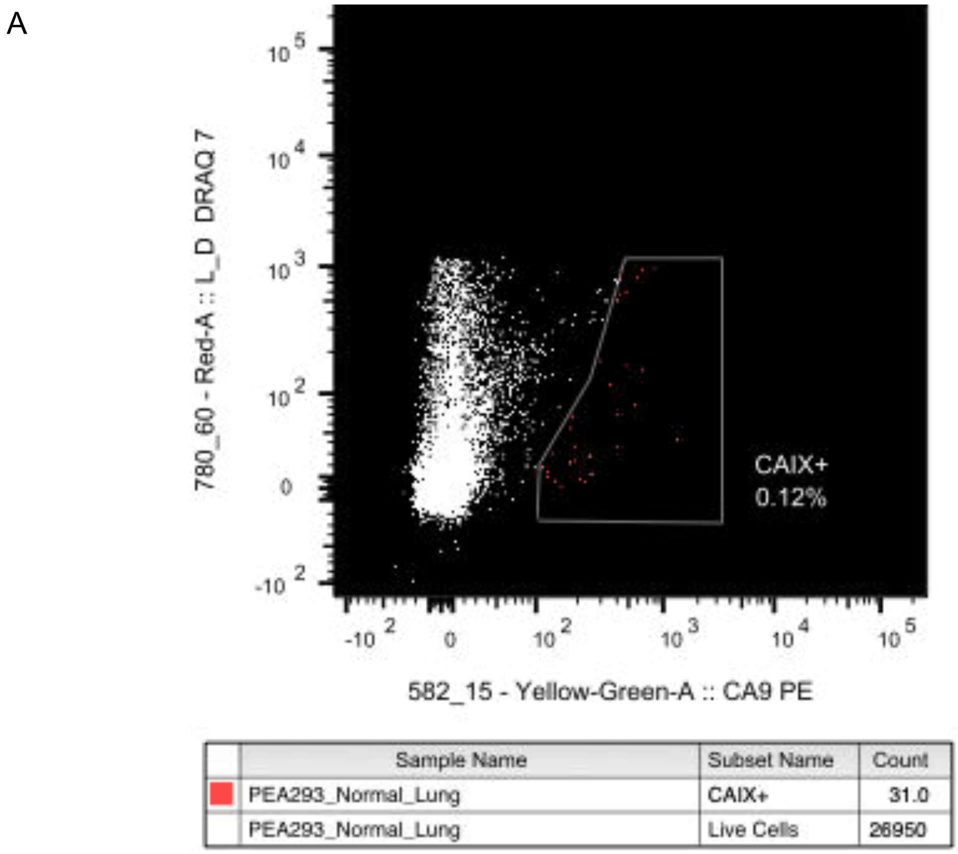


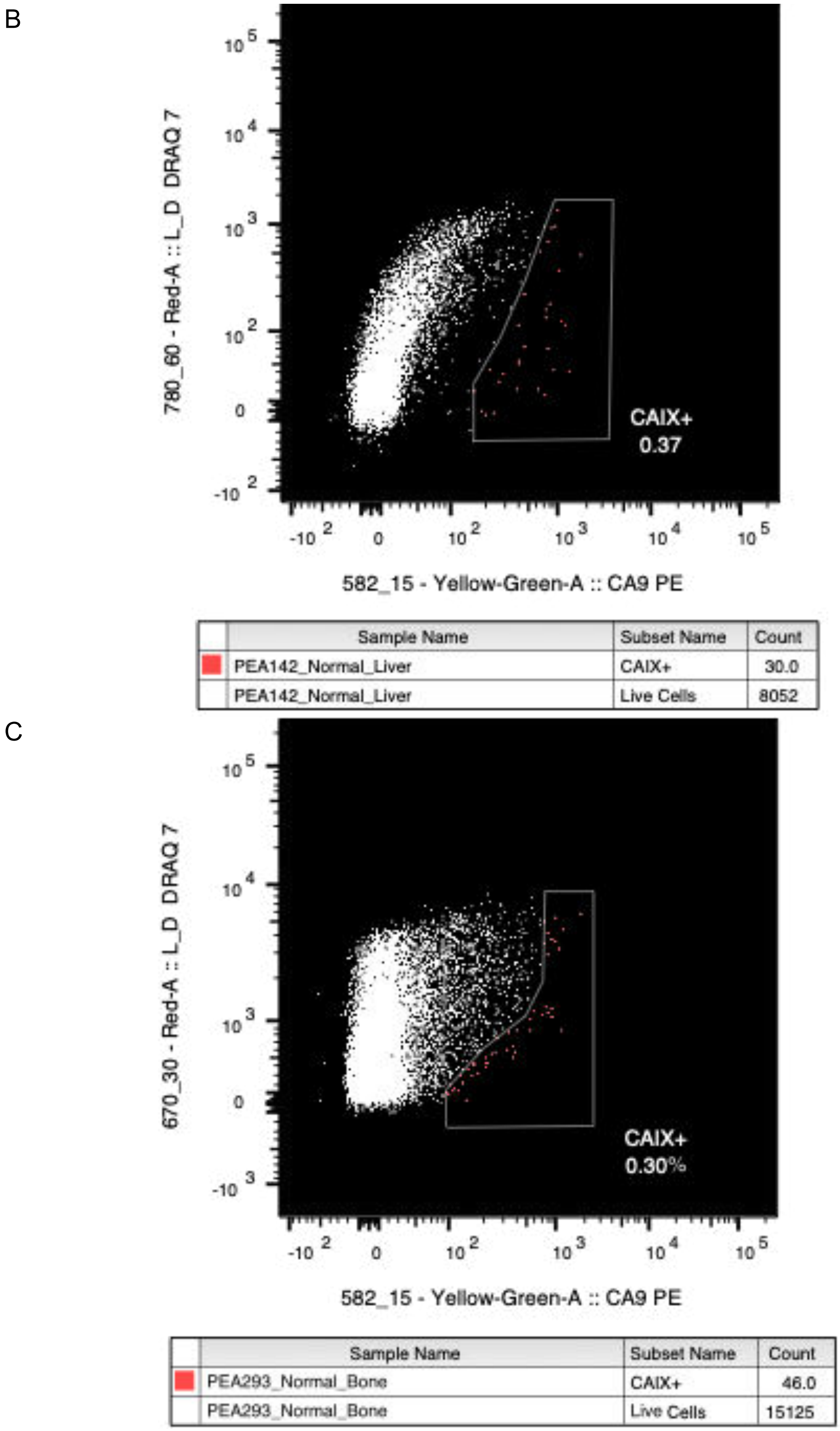
Figure 5.2 Flow cytometry of kidney primary tumour of PEA182.

5.5.2 Flow cytometry analysis of normal tissues

To detect putative DTCs in normal tissues, flow cytometry was performed on single cell suspensions derived from macroscopically normal tissues. Around 0.1-0.5% of cells were found to be CAIX$^+$ in normal tissues across patients (flow cytometry experiments of normal liver from PEA142 is shown in Figure 5.3B and normal lung and bone from PEA293 shown in Figure 5.3A & C). I noted that the prevalence of these cells was far higher than that of the 1 in a 100,000 to 1,000,000 DTC reported in ccRCC or other DTC studies. Thus, the population of CAIX$^+$ cells would almost certainly contain some normal diploid cells. As there is no simple method for determining the identity of these CAIX$^+$ cells whilst retaining the genome, sequencing is required for confirmation. Therefore, these positive single cells were then sorted into plates for downstream analysis.



	Sample Name	Subset Name	Count
🟥	PEA293_Normal_Lung	CAIX+	31.0
	PEA293_Normal_Lung	Live Cells	26950



	Sample Name	Subset Name	Count
🟥	PEA142_Normal_Liver	CAIX+	30.0
	PEA142_Normal_Liver	Live Cells	8052

	Sample Name	Subset Name	Count
🟥	PEA293_Normal_Bone	CAIX+	46.0
	PEA293_Normal_Bone	Live Cells	15125

Figure 5.3 Flow cytometry of normal lung, liver, and bone samples. Samples from PEA142 and PEA293 as labelled with gate and percentage of CAIX$^+$ cells shown.

5.5.3 Copy number profiling of CAIX$^+$ cells

DNA was amplification was performed following isolation of CAIX$^+$ single cells. However, we observed a high failure rate of amplification in these autopsy samples, resulting in successful generation of only 19 DNA libraries from over 100 cells. Indeed, only cells from the kidney primary and lung underwent successful amplification and sequencing.

Total copy number calling was performed on cells which were successfully amplified. One profile was removed due to excessive noise. As expected, the cell from the primary tumour PEA182 displayed a highly aneuploid profile in keeping with a tumour cell (Figure 5.4A). Although whole-genome amplification of both molecules of the original DNA template at a particular locus is very rare in single-cells, reads amplified from the two alleles of each SNP can be detected at a number of heterozygous SNP positions, given the vast size of the human genome. A reduction in the number of heterozygous SNP positions in this cell featuring reads reporting different alleles of chr3p suggested LOH, consistent with what's expected for RCC tumour cell (Figure 5.4B).

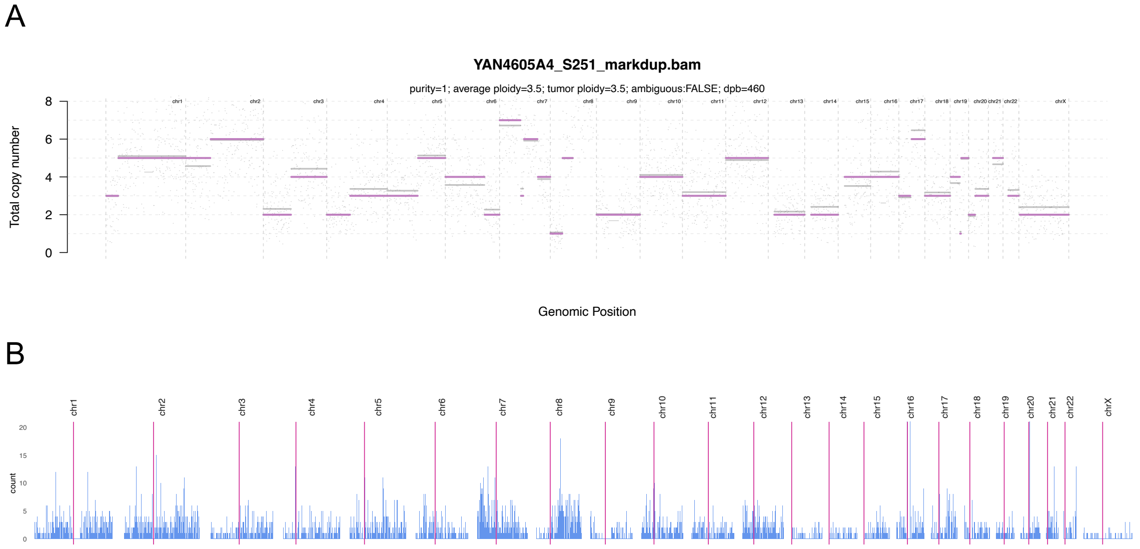


Figure 5.4 Genomic profiling of a CAIX$^+$ cell from the primary tumour of PEA182. Copy number profile shown in A with copy number for each bin shown in grey points, raw segment copy number shown as grey bars and rounded integer copy number as purple bars. The number of 1000 Genomes Project heterozygous SNPs with both alleles genotyped are shown in B with centromere position denoted by the red line. Depth per bin, dpb.

Of the CAIX$^+$ cells selected from normal tissues; one cell displayed a highly aneuploid copy number profile (Figure 5.5A). Although this cell did not exhibit 3p loss, the large number of copy number gains were suggestive of a malignant cell. In chromosomes 1, 3, 4, 6 and 9, there was an absence of reads reporting different alleles at 1000 Genomes Project heterozygous SNP positions, suggestive of LOH in this cell (Figure 5.5B). Therefore, this cell likely underwent chr3 loss followed by a gain of the remaining copy of chr3. As chr3p loss is a key early event in the evolution of ccRCC, the genome of this cell is in keeping with a ccRCC cell and could be confirmed by genotyping for SNVs called from the bulk tumour. Interestingly, this patient did not have any lung disease detected on imaging or at the time of autopsy suggesting this cell is a true DTC or is part of a micro-metastasis.

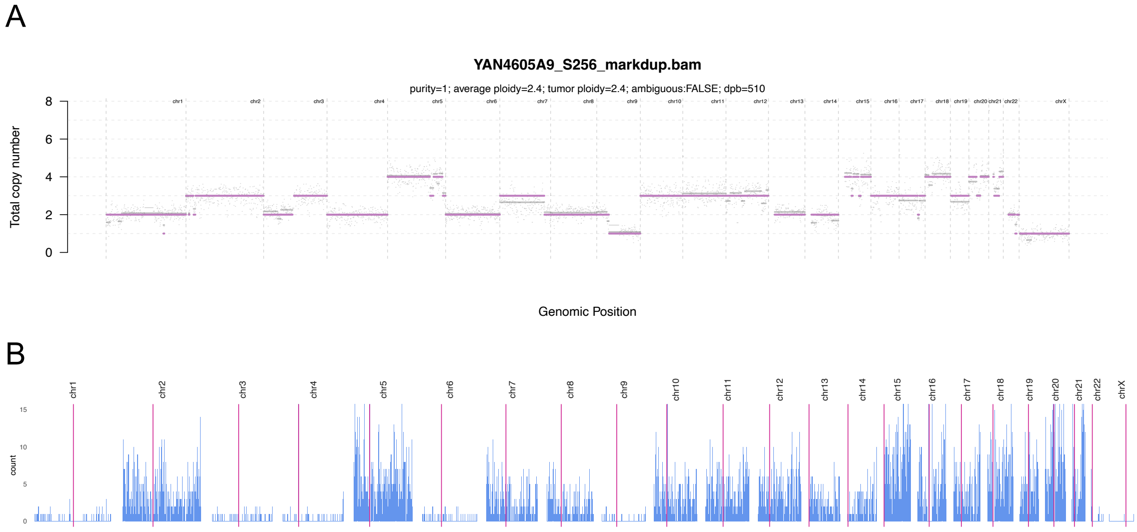


Figure 5.5 Genomic profile of a CAIX$^+$ cell from the normal lung of PEA142. Copy number profile shown in A and number of heterozygous SNPs shown in B as previous.

However, none of the 16 remaining CAIX$^+$ cells displayed aneuploid copy number profiles. This suggests that, whilst these cells were selected for being CAIX$^+$, they do not appear to be tumour cells (Appendix 8.2.1). Therefore, CAIX was not sufficiently specific for isolating ccRCC tumour cells through flow cytometry.

5.5.4 Aberrant cells of unknown origin

In addition to diploid cells, several cells were noted to display a gain or loss in part of a single chromosome arm. For example, one cell had a loss of part of chr15q (Figure 5.6A) and another lost part of chr6q (Figure 5.6C). I noted a reduction of reads reporting more than one allele at 1000 Genomes Project SNP positions for these segments, suggesting these losses resulted in LOH and were real (Figure 5.6B and D).

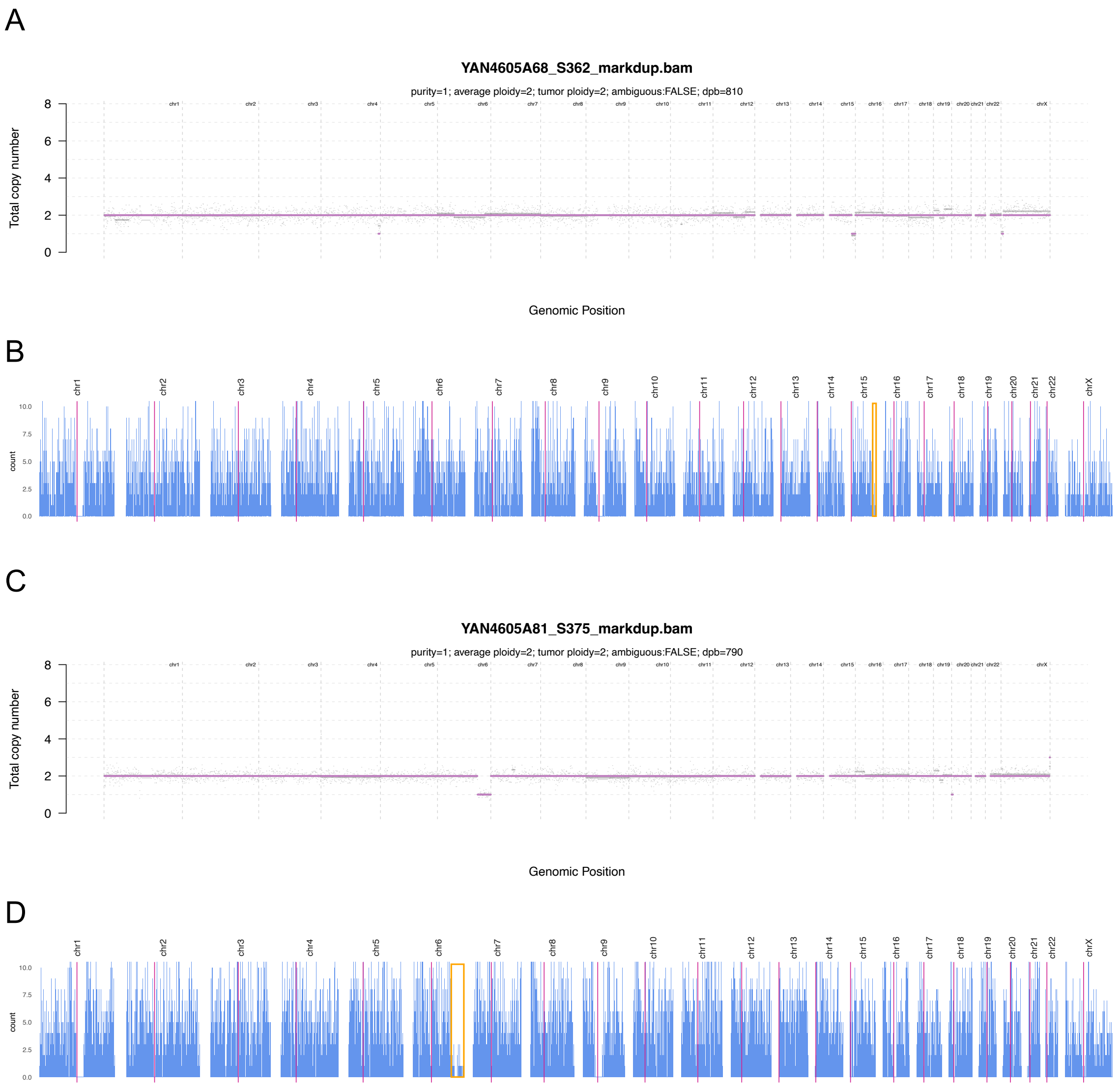


Figure 5.6 Copy number profile of aberrant unknown cells. Copy number profiles of cells with focal copy number gains or losses from the normal lung of PEA142 and PEA293 shown in A and C respectively. The number of heterozygous SNPs for these cells is shown in B and D, with segments featuring LOH highlighted in orange.

5.6 Immunohistochemical detection of micro-metastases

To confirm the presence of CAIX$^+$ cells histologically, IHC was performed on selected samples from this pilot. The primary kidney sample from PEA142 stained strongly positive for CAIX in the cell membrane and cytoplasm and for PAX8 (a renal lineage marker) in the nucleus, as expected (Figure 5.7).

Single staining of CAIX was performed on the same sample of normal lung from PEA142 which revealed focal areas with microscopic clusters CAIX$^+$ cells (Figure 5.8). These nests of enlarged atypical cells contained around 50-100 CAIX$^+$ cells which displayed morphology typical of ccRCC with prominent nucleoli in keeping with a micrometastasis. No other positively stained cells which appeared to be tumour were found in any of the other samples or patients. Whilst CAIX$^+$ cells were seen in liver samples, these cells were located around biliary ducts and are likely cholangiocytes, which are known to express CAIX (Figure 5.9). This demonstrates the lack of specificity of CAIX as a marker and explains the minute population of positive staining cells seen in liver samples during flow cytometry.

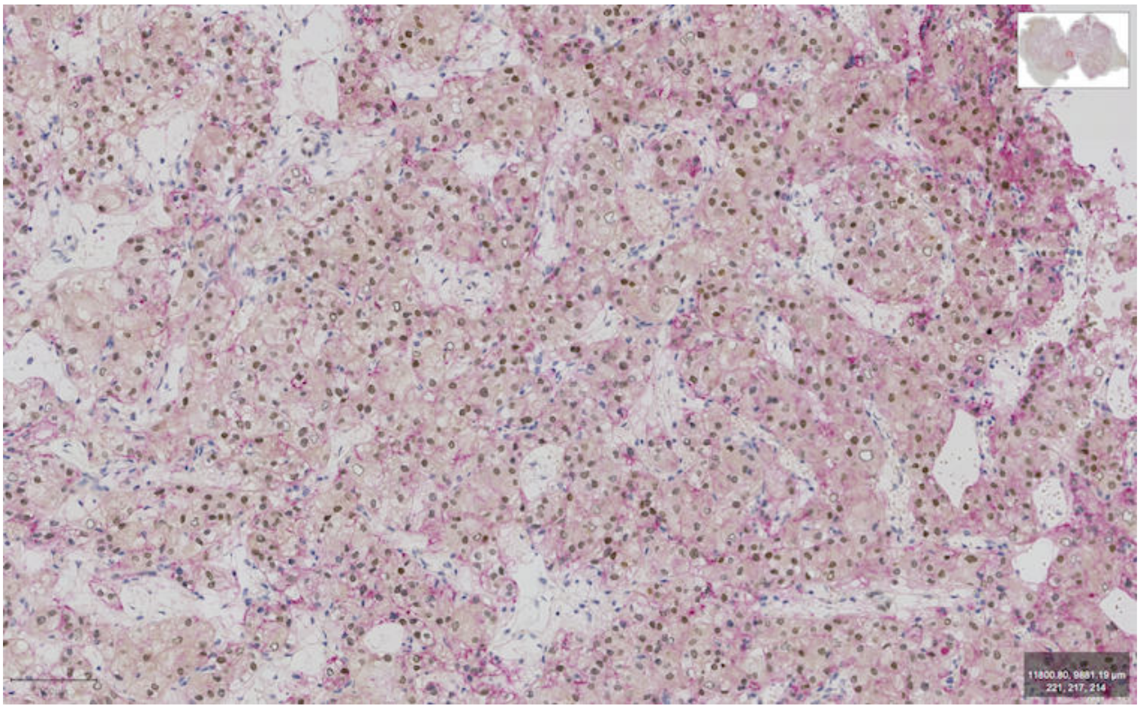


Figure 5.7 Dual CAIX/PAX8 staining in the primary kidney tumour of PEA142. CAIX and PAX8 stained in pink and brown respectively.

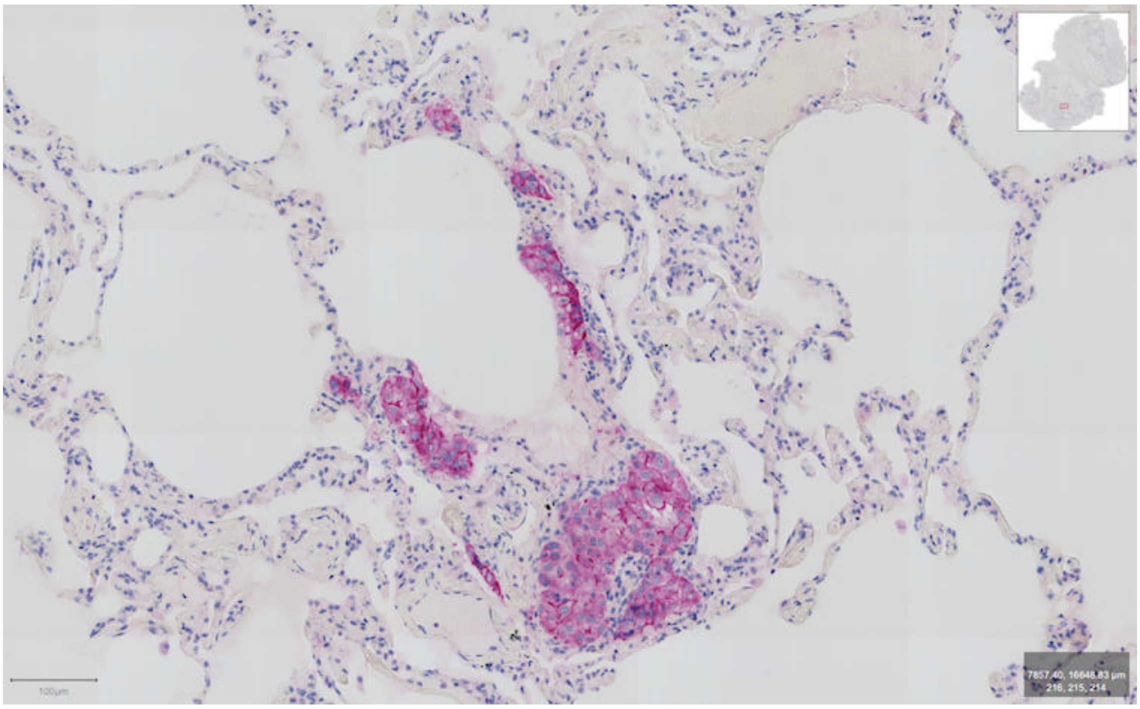


Figure 5.8 CAIX staining of normal lung from PEA142. Microscopic cluster of CAIX$^+$ cells stained in pink with prominent nucleoli.

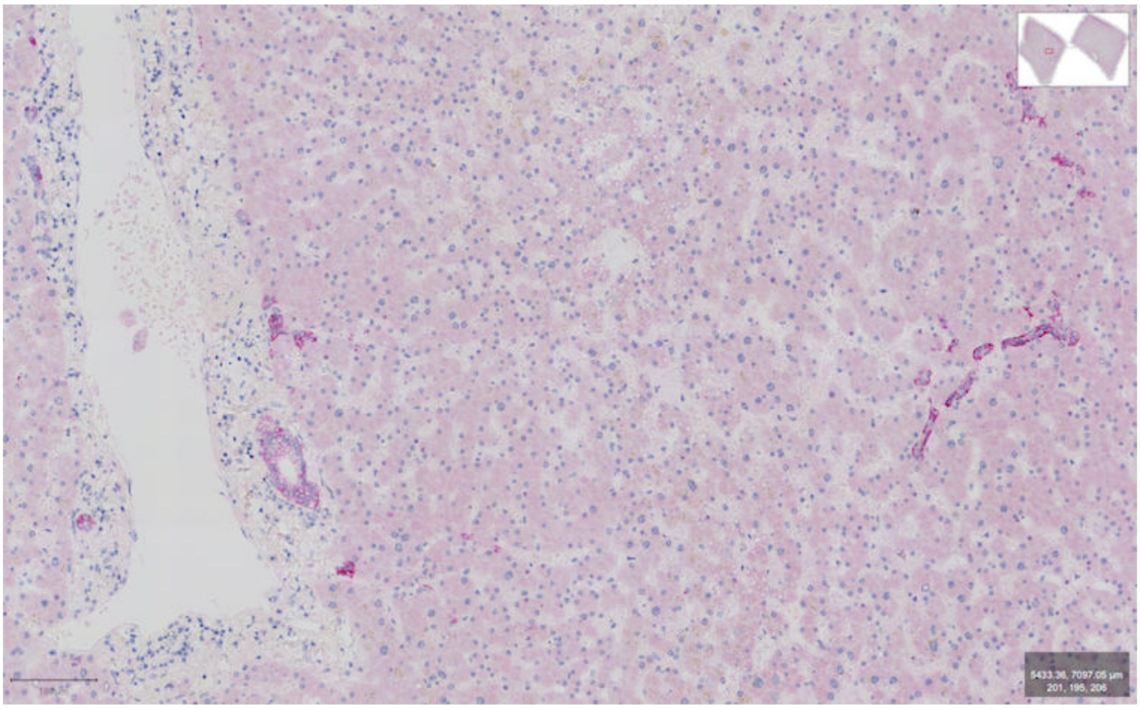


Figure 5.9 CAIX staining of normal liver from PEA142. Positively staining CAIX$^+$ cells are seen as part of portal triads and are likely cholangiocytes which make up a bile duct.

5.7 scRNA-seq of CAIX$^+$ cells

Given the low specificity of CAIX as a tumour marker, I attempted to use scRNA-seq which has a significantly higher throughput to profile positive cells. As the number of positive cells per sample was very low, multiple samples must be used to meet the minimum number of cells required for profiling. This made this experiment logistically challenging given the number of samples and the fact that a sorting step must be performed for each sample to select for CAIX$^+$ cells. Therefore, a split-pool barcoding scRNA-seq protocol was used to overcome this. This involves fixation of cells and enables freezing of sorted samples at different time points, before undergoing barcoding together in the same plate. Furthermore, samples from different patients were multiplexed together to meet the minimum number of cells required.

A summary of samples multiplexed together is shown in Table 5.2. I also included two bone marrow aspirate samples from TRACERx Renal patients in this experiment (discussed in Chapter 5).

Sample	Cell type	Cell type	Cell type	Cell type
1	PEA142_KidPrim_Pos	RK1007_BMA_Neg	RK1020_BMA_Neg	
2	PEA142_BMA_Pos			
3	PEA182_BMA_Pos	PEA293_BMA_Pos	PEA192_Bone_Neg	PEA352_Bone_Neg
4	PEA182_Lung_Pos	PEA293_Lung_Pos	PEA192_Lung_Neg	PEA352_Lung_Neg
5	PEA182_LungMet_Pos	PEA192_LungMet_Pos	PEA329_Lung_Neg	PEA314_Lung_Neg
6	PEA329_Bone_Pos	PEA314_Bone_Pos	PEA182_BMA_Neg	PEA293_BMA_Neg
7	PEA329_LNMet_Pos	PEA314_LNMet_Pos	PEA293_LNMet_Neg	PEA352_LNMet_Neg
8	PEA329_Lung_Pos	PEA314_Lung_Pos	PEA182_LungMet_Neg	PEA192_LungMet_Neg
9	PEA192_Bone_Pos	PEA352_Bone_Pos	PEA329_Bone_Neg	PEA314_Bone_Neg
10	PEA192_Lung_Pos	PEA352_Lung_Pos	PEA182_Lung_Neg	PEA293_Lung_Neg
11	PEA293_LNMet_Pos	PEA352_LNMet_Pos	PEA329_LNMet_Neg	PEA314_LNMet_Neg
12	RK1007_BMA_Pos	RK1020_BMA_Pos	PEA142_BMA_Neg	

Table 5.2 Multiplexed samples undergoing scRNA-seq. Samples are named by the patient, tissue, and CAIX positivity. BMA, bone marrow aspirate; KidPrim, kidney primary; LungMet, lung metastases; LNMet, lymph node metastasis; Pos, positive; Neg; negative.

Unfortunately, only ~12% of reads reported a valid barcode and subsequently only 468 cells passed quality control. Most of these cells originated from two multiplexed samples (samples 4 and 8) suggesting for most of my samples, the quality was extremely poor, resulting in failure to generate usable single-cell transcriptomes (Figure 5.10A). Nevertheless, Louvain clustering identified 5 clusters of cells which near all originate from samples 4 and 8 (Figure 5.10B).

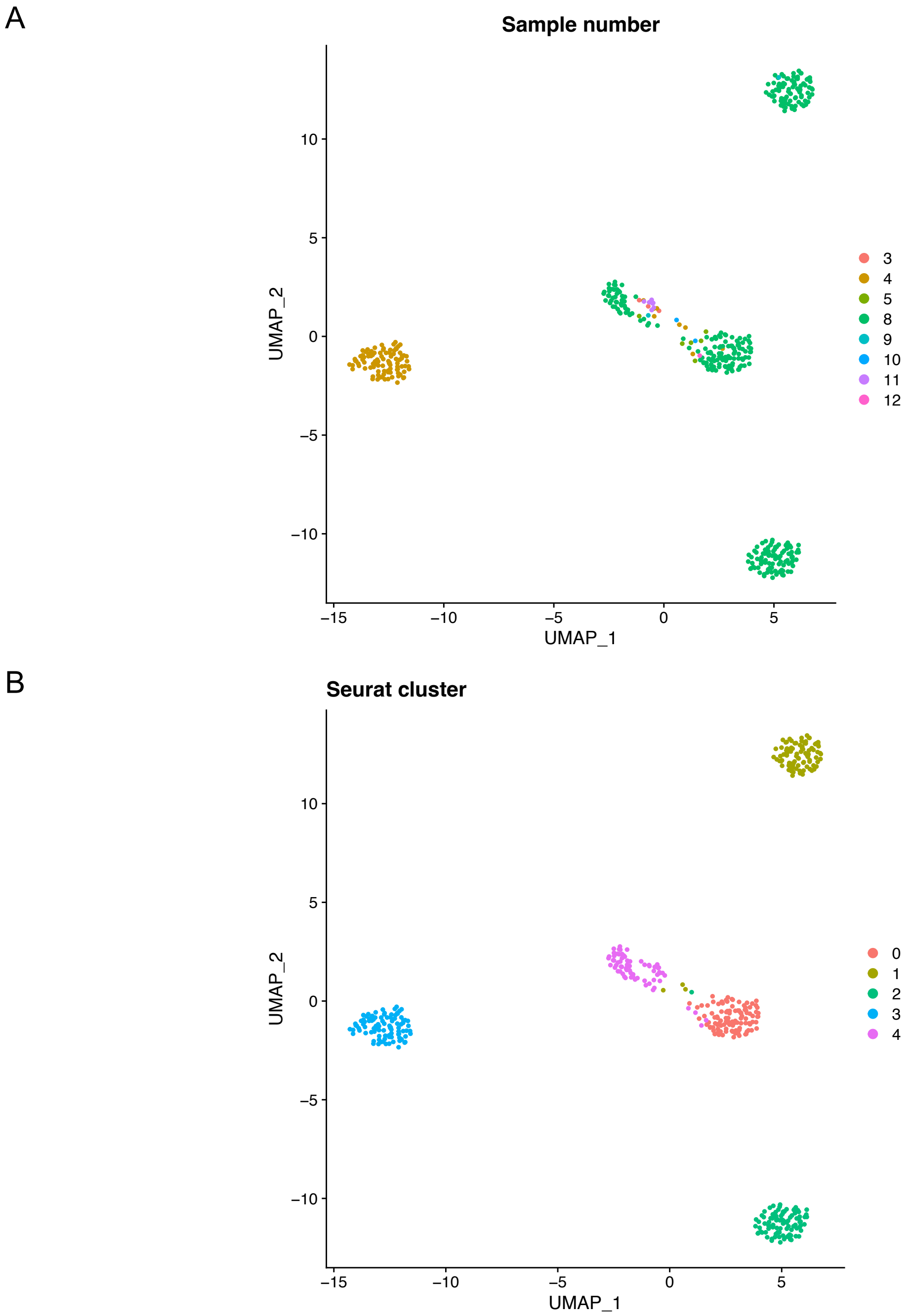


Figure 5.10 scRNA-seq UMAP of samples. Sample origin shown in A and Louvain cluster shown in B.

Despite my attempt at enriching for CAIX$^+$ through flow cytometry, none of the cells profiled expressed *CAIX*. Therefore, only CAIX$^-$ cells appear to be of sufficient quality for transcriptomic profiling. Next, I attempted to identify cell types from the various clusters of cells. Cells from cluster 0 expressed *PAX8* and *HIF1A* which are markers of ccRCC, lacked expression of *VHL*, but also expressed *EPCAM* (Figure 5.11). Given that most cells from cluster 0 originated from sample 8, this cluster of cells may represent EpCAM$^+$ CAIX$^-$ RCC tumour cells from the lung metastases of PEA182 or PEA192. In addition, clusters 1 and 3 expressed *DNAH11* and *CHST9* which are markers of respiratory ciliated cells. Finally, cluster 2 expressed *LAMA2* and *FBLN1* which are fibroblast markers.

Overall, scRNA-seq only generated useable transcriptomes from samples 4 and 8. Surprisingly, the two fresh samples from TRACERx bone marrow aspirates did not successfully generate single-cell transcriptomes. This could be due to the fixation and sorting steps causing samples quality degradation. For samples 4 and 8, I attempted to demultiplex these samples so that cells could be assigned to different donors based on alleles at SNP positions. However, the number of cells was too low to perform computational demultiplexing using population genotypes. Matching genotype information of SNPs in these subjects derived from other sequencing projects may allow de-multiplexing of these samples for further analysis.

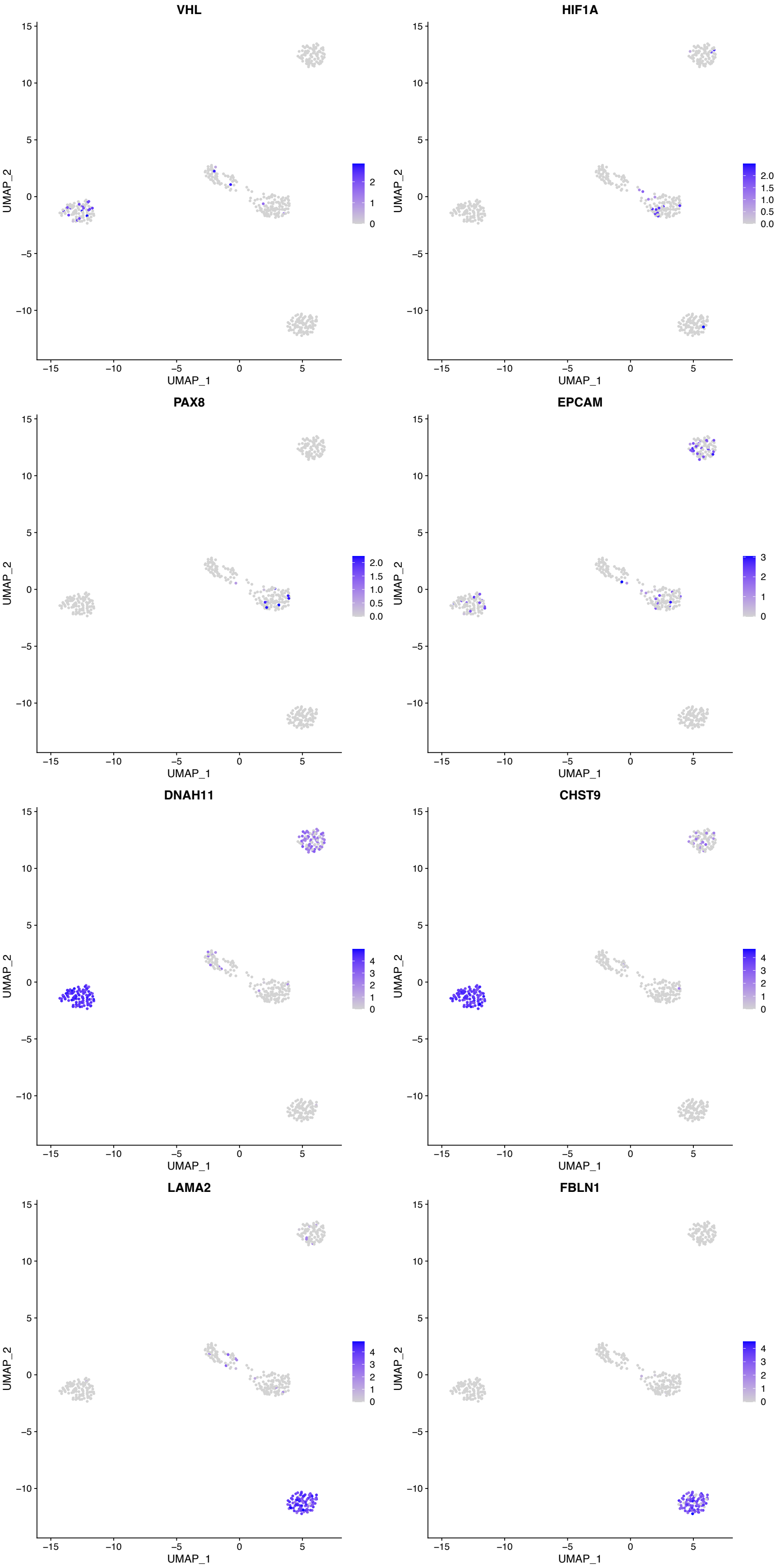
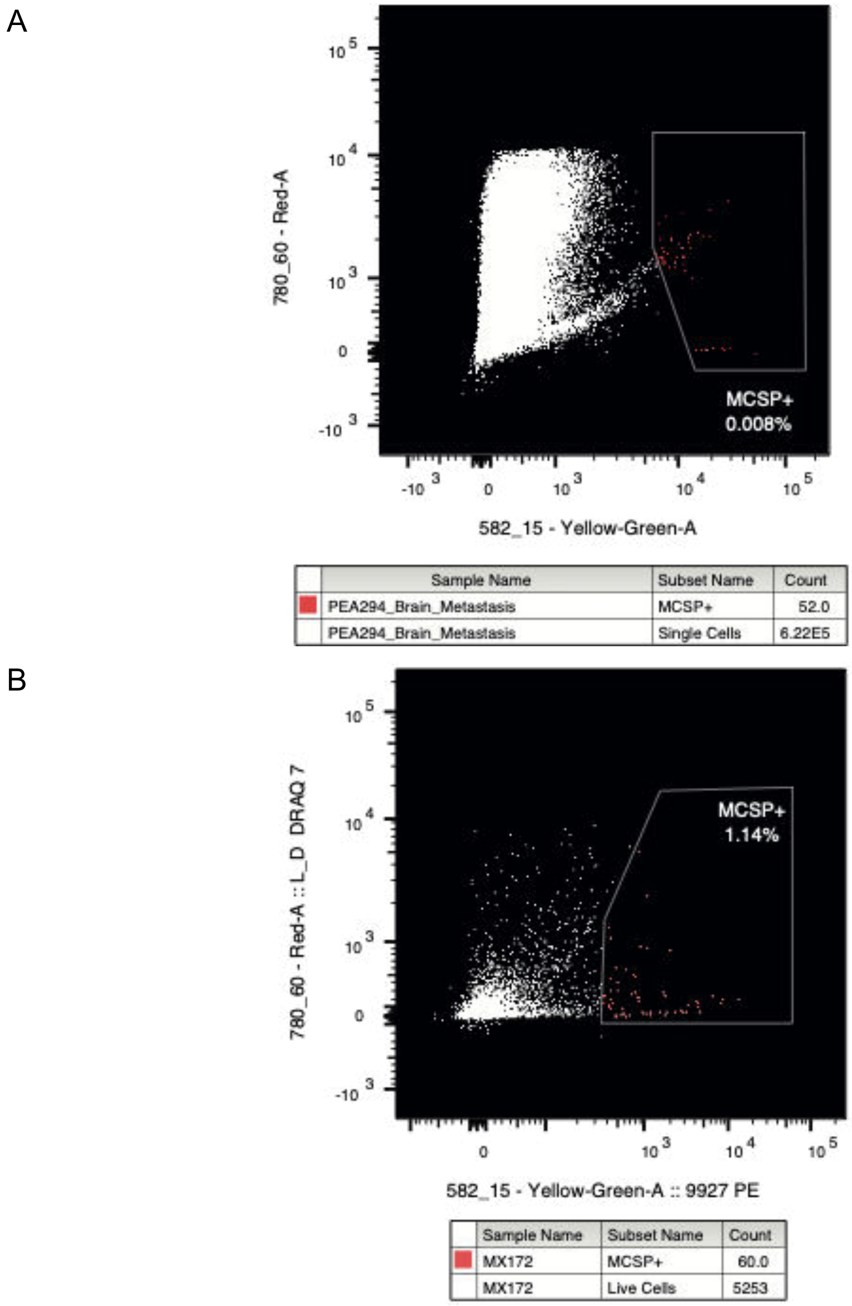


Figure 5.11 Expression of marker genes in scRNA-seq.

5.8 Detection of melanoma cells in normal tissues

5.8.1 Flow cytometry analysis of melanoma metastases

Although a large number of MCSP$^+$ cells were detected in the melanoma cell line SK-MEL28, very few MCSP$^+$ cells were seen in melanoma metastasis samples from PEACE, such as the brain metastases from PEA294 or the two freshly resected metastases from MX172 and MX417 (Figure 5.12 and Appendix 8.2.2).

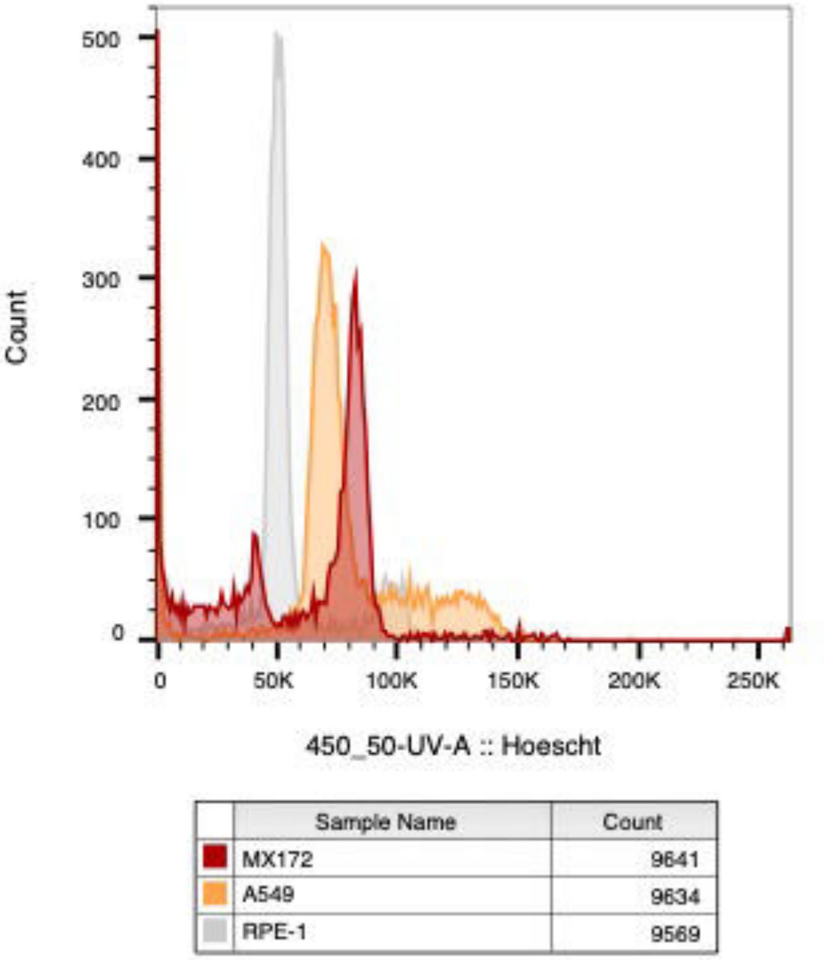


	Sample Name	Subset Name	Count
🟥	PEA294_Brain_Metastasis	MCSP+	52.0
	PEA294_Brain_Metastasis	Single Cells	6.22E5

	Sample Name	Subset Name	Count
🟥	MX172	MCSP+	60.0
	MX172	Live Cells	5253

Figure 5.12 Flow cytometry analysis of melanoma metastases. Brain metastasis from PEA294 shown in A and thigh metastasis from MX172 shown in B.

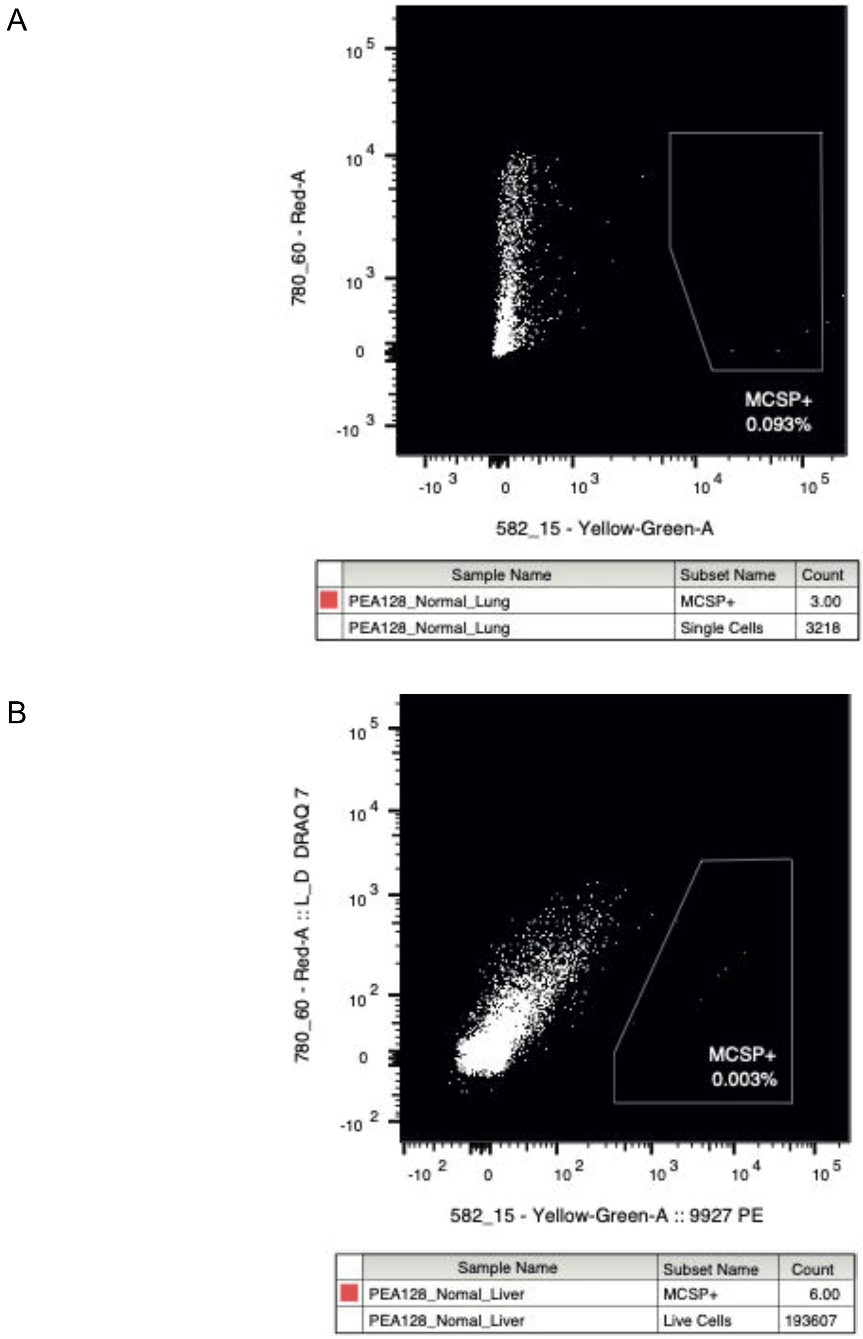
To confirm that the metastatic samples collected were tumour samples, ploidy analysis was also performed. This confirmed that the metastasis sample from MX172 consisted of mostly tumour cells, as it contained a small population of diploid cells and a much larger population of cells with far higher ploidy, likely representing cells that have undergone WGD (Figure 5.13). Therefore, MCSP does not appear to be a sensitive marker for the detection of melanoma cells in our samples. CD146 was also used as a marker for melanoma, but again did not detect a significant number of positive cells (Appendix 8.2.3).



	Sample Name	Count
🟥	MX172	9641
🟧	A549	9634
⬜	RPE-1	9569

Figure 5.13 Flow cytometry ploidy analysis of resected metastases. The diploid RPE-1 cell line was used to establish a baseline ploidy of 2 and the lung cancer cell line A549 is hypotriploid.

Although MCSP may not be a sensitive marker of melanoma cells, those which are positive may still be worth profiling. Therefore, MCSP$^+$ cells were isolated from normal samples from PEA128, PEA294 and PEA312 (normal lung and liver from PEA128 shown in Figure 5.14).



	Sample Name	Subset Name	Count
🟥	PEA128_Normal_Lung	MCSP+	3.00
	PEA128_Normal_Lung	Single Cells	3218

	Sample Name	Subset Name	Count
🟥	PEA128_Nomal_Liver	MCSP+	6.00
	PEA128_Nomal_Liver	Live Cells	193607

Figure 5.14 Flow cytometry analysis of normal bone. Normal lung and liver from PEA128 shown in A and B respectively.

DNA was successfully amplified for 5 sorted MCSP$^+$ single cells with three cells exhibiting diploid copy number profiles. One MCSP$^+$ cell sorted from normal lung tissue of PE1A128 displayed a noisy copy number profile, which could be in keeping with an aneuploid cell, although genotyping of SNVs would be required to confirm this (Figure 5.15A). In addition, a cell with a small gain of chr19q from the same normal lung tissue was detected which represents another aberrant cell (Figure 5.15B). Overall, these results again confirm that the tissue quality from these autopsies, independent of tumour type, is not sufficient to perform single-cell profiling and allow interpretable downstream analysis.

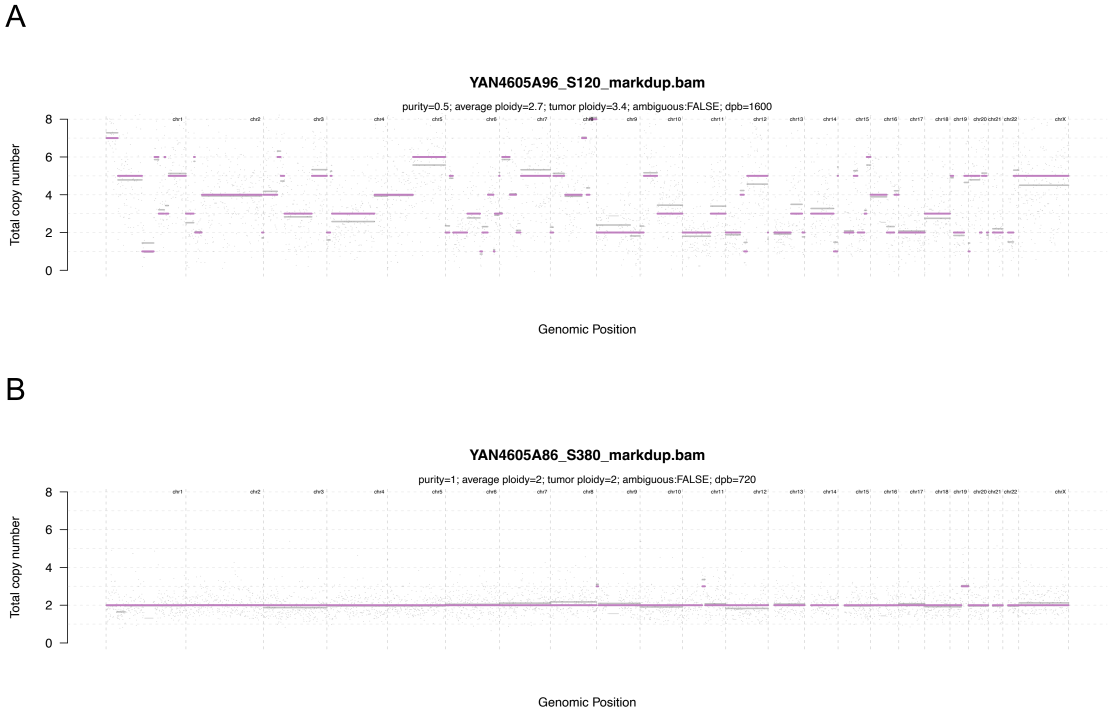


Figure 5.15 Copy number profile of two MCSP$^+$ cells from normal lung of PEA128.

5.9 Discussion

In this study, I attempted to detect and genomically profile tumour cells in normal tissues from patients with metastatic ccRCC and melanoma. One aneuploid cell was found in the normal lung of a ccRCC patient without detectable lung disease and a micrometastasis was found in the same sample using IHC staining. Interestingly, this micrometastasis did not display features of necrosis or increased mitotic activity suggesting it was not actively proliferating. In addition, several aberrant cells of unknown origin were also profiled from normal tissues. However, I was unable to increase the number of cells profiled. This was due to two main difficulties encountered in this study: poor sample quality and lack of specificity in tumour cell markers.

5.9.1 Sample quality is poor in research autopsy tissues

Recent studies have performed DNA sequencing on bulk samples from rapid autopsy studies held within hours of death. In this study, samples were collected from research autopsies often held several days after death resulting in mostly poor sample qualities. Nevertheless, I demonstrate that shallow coverage WGS is possible for single cells derived from these autopsy samples. However, it appears that the dissociation process generates a large number of poor quality cells resulting in unreliable extraction and amplification of DNA with a high failure rate. Only cells from tumour tissues or normal lung successfully generated DNA. This could be explained by the presence of a tissue specific inhibitor released by dead cells from bone or liver as amplification in cells from lung in the same patient (and therefore having the same time to autopsy and refrigeration process) was possible.

scRNA-seq was also attempted for autopsy samples and could only identify cell types in high quality samples. This was partly expected given that RNA is a less stable molecule than DNA. However, in the absence of a high throughput method for DNA profiling, this was the only feasible technique available to me.

5.9.2 Marker-based tumour cell identification is unreliable

Although one aneuploid cell, likely originating from a micrometastasis, was profiled, CAIX-based detection was not sufficiently specific to identify tumour cells with high confidence. This may be due to normal cells expressing CAIX due to hypoxia after patient death or during sample processing. Alternative markers could be explored in combination with CAIX to improve specificity, although this would require fixation as there are few cell surface markers for ccRCC.

Concerningly, neither MCSP nor CD146 was detected in autopsy-derived metastases or freshly collected metastatic melanoma resections. Ploidy analysis confirmed these sample as having a significant number of tumour cells, confirming these markers are not reliable. MCSP and CD146 have been reported to be expressed by melanomas and is used by CELLSEARCH in combination for detecting CTCs. The lack of positive cells could be due to loss of expression of MCSP by tumour cells, which may be more common in metastatic samples which have undergone further evolution. Alternatively, this could also be explained by lack of binding by the specific antibody used in this study.

In addition, I also identified several CAIX$^+$ cells which exhibited the loss of a part of a chromosome. The nature and origin of these aberrant cells of unknown origin remains unclear, although they seem to be enriched in the selected population. These cells may be damaged cells with have suffered loss of (part of) a chromosome, and CAIX may select for these cells as it is also a hypoxic marker. However, these cells were also detected in fresh samples collected from surgery, therefore is not a phenomenon only seen in cells derived from autopsy samples. This will be discussed further in Chapter 5.

Additional SNV and CNA data from bulk samples of the primary tumour collected from surgery, and metastases collected during research autopsy, would confirm the identity of aneuploid or aberrant cells profiled in this study. It would also provide further information about their relationship to the primary and metastases collected at autopsy. However, as so few single cells were collected and profiled, this was not performed.

In summary, despite utilising different markers, the identification of tumour cells from tissues not expressing epithelial markers was extremely challenging. This is in contrast to nearly all previous DTC studies, which have relied upon epithelial markers such as EpCAM or cytokeratins. Although sample quality from research autopsies is poor, single-cell techniques were applied and enabled characterisation of DTCs in pilot cases. Methods with higher throughput have the potential to overcome these challenges and enable in-depth profiling of DTCs.

Chapter 6. Characterising bone marrow DTCs in early-stage ccRCC

6.1 Introduction

Kidney cancer is the 7th most common cancer in the UK with over 13,000 new cases diagnosed every year (Cancer Research UK, 2020). Over 90% of kidney tumours are renal cell carcinomas with three quarters of these being clear cell (Hsieh *et al.*, 2017). Unfortunately, despite the absence of metastases, up to a third of patients with limited disease go on to develop recurrence. Therefore, to prevent relapse and increase survival, there is a need to understand the metastatic process in ccRCC in greater detail.

6.1.1 Clinical management of ccRCC

For patients with localised disease, surgery to remove the primary tumour is the mainstay of curative treatment (Hsieh *et al.*, 2017). Partial nephrectomy is performed to preserve healthy renal parenchyma with radical nephrectomy considered if there are multiple renal tumours or extension into the vasculature. In patients with advanced disease who have a significant disease burden at the primary site but limited metastatic disease, cytoreductive nephrectomy can be performed (Bex *et al.*, 2016). This is thought to reduce the source of metastasis-seeding cells, although whether this results in a benefit in overall survival since the introduction of targeted therapies and immunotherapies is under debate (Chakiryan *et al.*, 2022; Bakouny *et al.*, 2023).

As cytotoxic drugs are generally ineffective for RCC, medical management of metastatic disease consists of targeted therapies, including tyrosine kinase inhibitors such as sunitinib, pazopanib, axitinib, and cabozantinib (Escudier *et al.*, 2019). Intracellular mammalian target of rapamycin (mTOR) inhibitors such as everolimus and temsirolimus and the anti-VEGF (vascular endothelial growth factor) monoclonal antibody bevacizumab are also used. ccRCC has long been known to be responsive to the earliest forms of immunotherapy with interferon-α and high dose interleukin-2 used to treat metastatic disease since the 1990s (Fyfe *et al.*, 1995; McDermott *et al.*, 2005). More recently, immunotherapies such as the anti-PD1 (programmed cell death protein 1) and anti-CTLA-4 (cytotoxic T-lymphocyte-associated protein 4) monoclonal antibodies nivolumab and ipilimumab, administered in combination, have become first-line treatment for intermediate or high risk ccRCCs. Furthermore, pembrolizumab (another anti-PD1 antibody) has shown a benefit in disease-free survival in the adjuvant setting (Choueiri *et al.*, 2021).

6.1.2 Genomics of ccRCC

Recently, the initiating and secondary events in the development of ccRCC have been revealed by studies profiling genetic and chromosomal abnormalities (Jonasch *et al.*, 2021). Two key events are present in nearly every ccRCC: the loss of chromosome 3p and inactivation of *VHL*.

The loss of chromosome 3p is a critical early event in ccRCC (Gerlinger *et al.*, 2014; Mitchell *et al.*, 2018). Loss of 3p has several consequences due to the genes which reside on this segment of the genome. Critically, *VHL* is located on chr3p25 and following 3p loss, mutation or methylation of the remaining copy provide these cells with a proliferative advantage through accumulation of HIF1α and upregulation of VEGF and platelet-derived growth factor (PDGF). In addition, several chromatin remodelling genes which are frequently mutated in ccRCC such as *PBRM1*, *SETD2* and *BAP1* are also located on chromosome 3p (The Cancer Genome Atlas Research Network, 2013). Mutation of the remaining copy or haploinsufficiency in the case of *SETD2* renders these genes non-functional and contributes to tumour progression (Chiang *et al.*, 2018).

The TRACERx renal study has further characterised the genomic and evolutionary features of ccRCC by describing seven evolutionary subtypes and linking these to prognosis (Turajlic *et al.*, 2018b). Conserved patterns of co-occurrence of copy number events such as losses of 4q and 9p, and mutual exclusivity of *BAP1* and *SETD2/PBRM1* mutations in individual clones were noted. Furthermore, chromothripsis was shown to be the most common mechanism by which chromosome 3p is lost and occurs in association with 5q gain (Mitchell *et al.*, 2018). Interestingly, in patients with von-Hippel Lindau (VHL) disease, loss of chromosome 3p occurred at different breakpoint locations in independent synchronous tumours (Fei *et al.*, 2016). This highlights the importance of 3p loss in ccRCC tumourigenesis and is evidence of parallel evolution. Timing analysis has shown that the simultaneous 3p loss and 5q gain can occur decades before diagnosis, thus providing a window of opportunity for intervention (Mitchell *et al.*, 2018).

Studies of metastases have also contributed to our understanding of ccRCC (Turajlic *et al.*, 2018a). The pattern of metastatic dissemination was shown to be associated with the evolutionary subtype which may allow stratification of patients for surgical intervention. Furthermore, loss of chromosome 9p and 14q was present in significantly more metastatic subclones that in the primary tumour and is associated with worse prognosis. This suggests that these events may be key in providing these subclones with the ability to metastasise.

6.1.3 Pancreatic metastases in ccRCC

The most frequent organs that RCC metastasises to are the lung, lymph nodes, bone, liver and adrenals (Dudani *et al.*, 2021). In addition, the pancreas is a clinically intriguing distant site of metastasis in ccRCC. Despite removal of the primary tumour, oligometastatic disease may recur after many years (Yuasa *et al.*, 2015). A small number of these metastases have been analysed and shown to be seeded early in tumour evolutionary time (Turajlic *et al.*, 2018a). Therefore, it has been hypothesised that the pancreas may represent a privileged site of ccRCC dormancy, where DTCs are protected from the immune system and eventually lead to metastases. This phenomenon of late recurrence may be due to DTCs entering quiescence (possibly involving signalling from the host tissue) until reactivated by some stimuli. Alternatively, these cells may be dividing but increase overall mass at such a low rate that they can only be detected by imaging several years later, raising questions about their clinical significance and need for removal.

Whilst profiling metastases can inform us about the successful subclones which survived the dissemination process and subsequently proliferated, they cannot inform us about the role of other cancer cells of the primary tumour. To date, DTCs have not been profiled in ccRCC and only one study has detected putative cytokeratin positive DTCs in patients with RCC (Buchner *et al.*, 2003). Therefore, there is a need to isolate DTCs and study their relationship to the primary tumour.

6.1.4 Detection of DTCs in ccRCC

Given the lack of specificity demonstrated by CAIX as a marker of ccRCC tumour cells in Chapter 4, an alternative marker was selected. Recently, PAX8, a kidney lineage transcription factor, was reported to be essential for oncogenic signalling in ccRCC (Patel *et al.*, 2022). Furthermore, PAX8 was also shown to be co-opted to allow expression of amplifications of *MYC*. PAX8 is expressed in the kidney, thyroid, and organs derived from the mesonephric and Müllerian ducts and has been used as a clinical marker for ccRCC and other kidney cancer subtypes (Hu *et al.*, 2012). Crucially, it has not been detected in haematopoeitic tissues. As freshly collected bone marrow samples from TRACERx are of higher quality compared to PEACE, samples could be fixed to allow staining of PAX8 which is primarily located in the nucleus. Therefore, PAX8 was chosen as a marker for the detection of ccRCC DTCs.

6.1.5 Collection of bone marrow aspirates in surgery through TRACERx Renal

The TRACERx Renal trial is a prospective longitudinal cohort trial which has made significant contributions to our knowledge on the genomics of ccRCCs as highlighted above. Due to the close-collaborations and infrastructure fostered by this trial, tissue from surgical resections can be directly collected from theatre, undergo review by a pathologist, and remaining tissue can be immediately processed for research. This has allowed for techniques which require fresh samples with minimal degradation to be applied to the collected tissues, such as single-cell methods or organoid generation. Therefore, this trial represents an excellent setting to collect samples from patients to identify DTCs.

After discussion with the surgical and anaesthetic teams, I set up a sub-study within TRACERx Renal to allow the collection of BMAs from patients with ccRCC undergoing surgery. This involved amending the trial protocol, creating the patient information sheet and consent form, and gaining approval for the sub-study from the region ethics committee. For each sample collection, I recruited the patients to the sub-study, notified the theatre team and performed the BMA in theatre.

6.2 Aims

The aim of this project was to:

- Collect bone marrow aspirates from patients with ccRCC undergoing surgery.
- Identify, isolate and genomically profile DTCs from BMAs.
- Determine the relationship between DTCs and the primary ccRCC to trace the evolutionary origins of these cells.

6.3 Summary of clinical histories

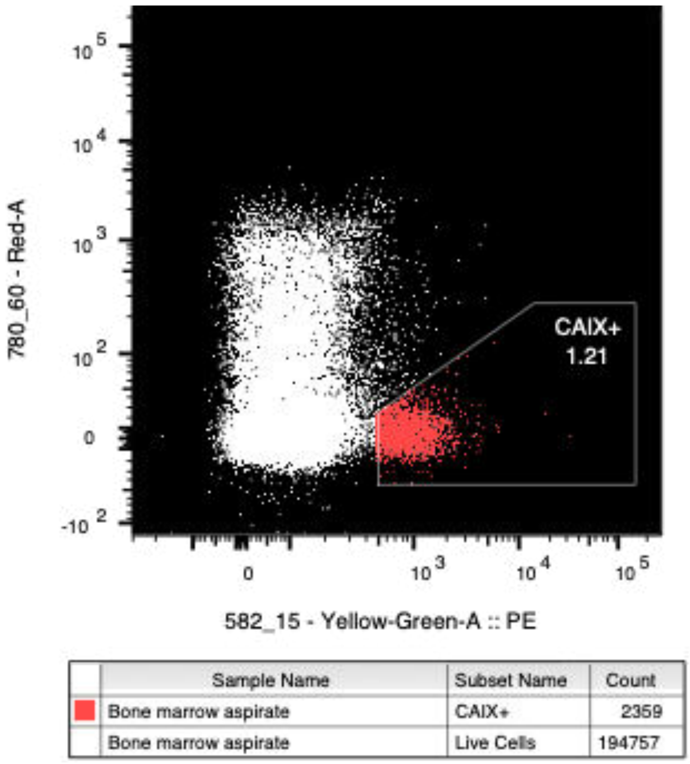
In total, BMA samples were collected from 7 patients undergoing surgery after the induction of anaesthesia. The procedure typically took less than 5 minutes and did not cause undue delay to the operation or the surgical list. This cohort featured patients with ccRCCs with a mixture of different stages of disease. It also included one sample taken from a patient with an oncoblastoma – a benign tumour which can require surgical resection due to complications from mass effects. In addition, dissociated single cells were also generated from the matched normal kidney of patient RK1040 collected during an earlier surgery, the primary tumour of RK1013, and the resected pancreatic metastases of RK1043. A summary of patients and their clinical histories and histology is shown in Table 6.1.

Patient	Sex	Age	Diagnosis	Operation	Pathological Stage	Fuhrman grade	Metastases & treatment	IHC		BMA site
								CAIX	PAX8	
RK1007	M	62	ccRCC	Caval thrombectomy	N/A	4	Caval thrombus	+	+	Right hip
RK1013	F	57	ccRCC	Left radical nephrectomy	pT3a pN0 pMx	3	Nil	n.d.	n.d.	Right hip
RK1020	F	71	Oncocytoma	Left radical nephrectomy	N/A	N/A	N/A	n.d.	n.d.	Right hip
RK1030	F	69	ccRCC	Right radical nephrectomy	pT3a pNx pMx	4	Nil	+	+	Right hip
RK1033	M	55	ccRCC (sarcomatoid differentiation)	Left radical nephrectomy	pT3a ypNx ypMx	4	Multiple in lung (neoadjuvant IO)	+	n.d.	Right hip
RK1036	M	73	ccRCC	Right radical nephrectomy	pT3a ypNx ypMx	3	Lung (neoadjuvant IO)	n.d.	n.d.	Right hip
RK1040	M	55	ccRCC	Resection of femur metastasis	N/A	3	Left femur	n.d.	n.d.	Left hip
RK1043	M	59	ccRCC	Pancrea-tectomy and splenectomy	N/A	3	Pancreas	n.d.	n.d.	N/A

Table 6.1 Clinical information for patients and samples collected at surgery. IO, immunotherapy; n.d., not done.

6.4 Detection and genomic profiling of CAIX$^+$ cells from bone marrow aspirates

Initially, CAIX was used as a marker for tumour cell detection as discussed in Chapter 4. However, a higher-than-expected proportion of CAIX+ cells of approximately 1% was detected (Figure 6.1). Of the 6 cells which were amplified and profiled, all displayed diploid copy number profiles, confirming they were normal cells (Appendix 8.3.1). Therefore, CAIX was not used for further attempts at isolating DTCs.



	Sample Name	Subset Name	Count
🟥	Bone marrow aspirate	CAIX+	2359
	Bone marrow aspirate	Live Cells	194757

Figure 6.1 Flow cytometry of CAIX$^+$ cells from the BMA of patient RK1007.

6.5 Immunocytochemistry detection and micromanipulation of PAX8$^+$ cells

In order to detect ccRCC DTCs, antibodies against PAX8 were used to stain cells. However, despite using three different fixation methods and two antibodies, detection of PAX8 through flow cytometry was not successful in cell lines or fresh tumour samples. Therefore, I attempted to perform immunocytochemistry staining of PAX8 instead. Although staining was successful for methanol fixed cells, the deposition of cells onto Cytospin slides meant that cells could not be isolated for downstream analysis. Therefore, I adapted a method for immunofluorescence staining used in flow cytometry (see section 2.5.4) so that cells could be stained in solution for PAX8. This staining was positive for the control cell line RPTEC which is derived from renal proximal tubule epithelial cells and negative for the breast cancer cell line MCF-7 (Appendix 8.3.2). In addition, positively stained cells were seen for cells from the normal kidney sample of RK1040 and also that of the pancreatic metastasis from patient RK1043 (Figure 6.2).

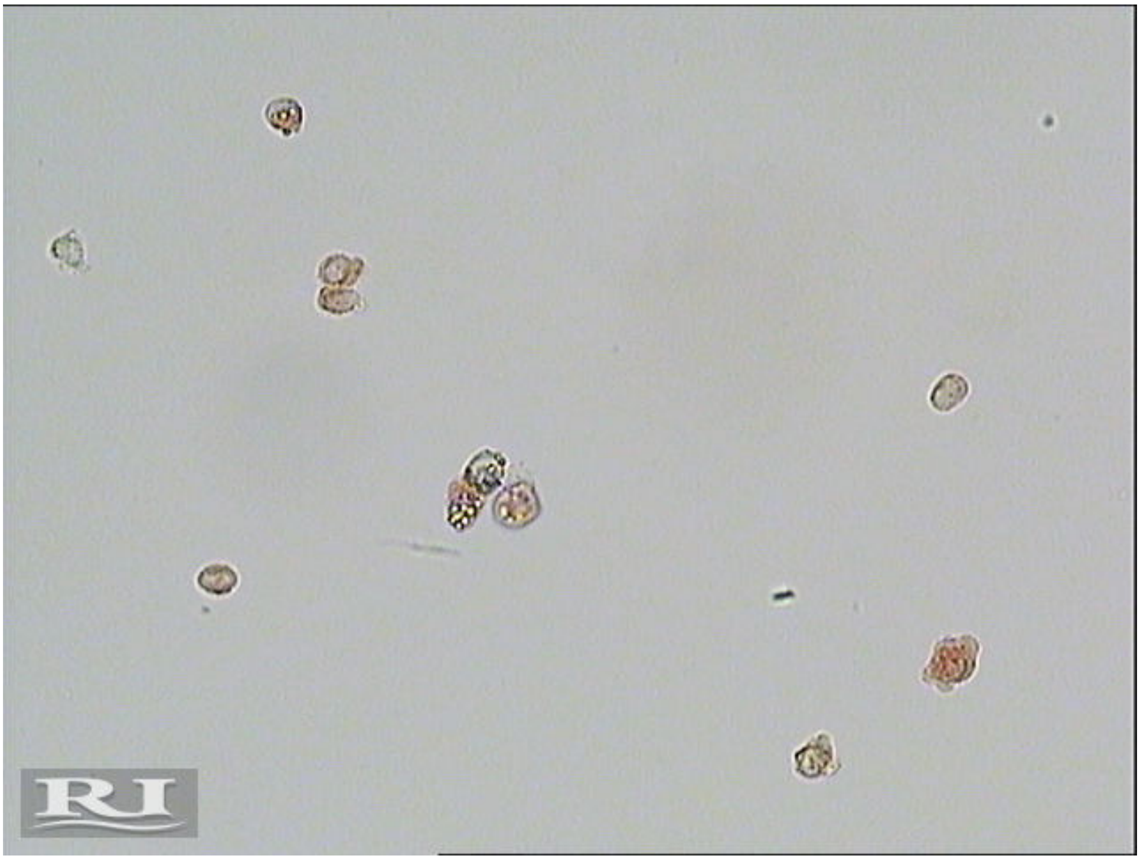


Figure 6.2 PAX8 ICC staining of pancreatic metastases from patient RK1043. Positively staining cells seen in pink.

A micromanipulator was then used to pick individual positively staining cells from the stained single cell suspension and deposited onto a small volume of PBS (Figure 6.3). After careful review that no additional cells were transferred, this PBS containing the single cell was pipetted into an Eppendorf tube. PAX8+ were seen in all patients at a rate of 1 to 5 in 50,000, except for RK1020 which did not have any positive cells.

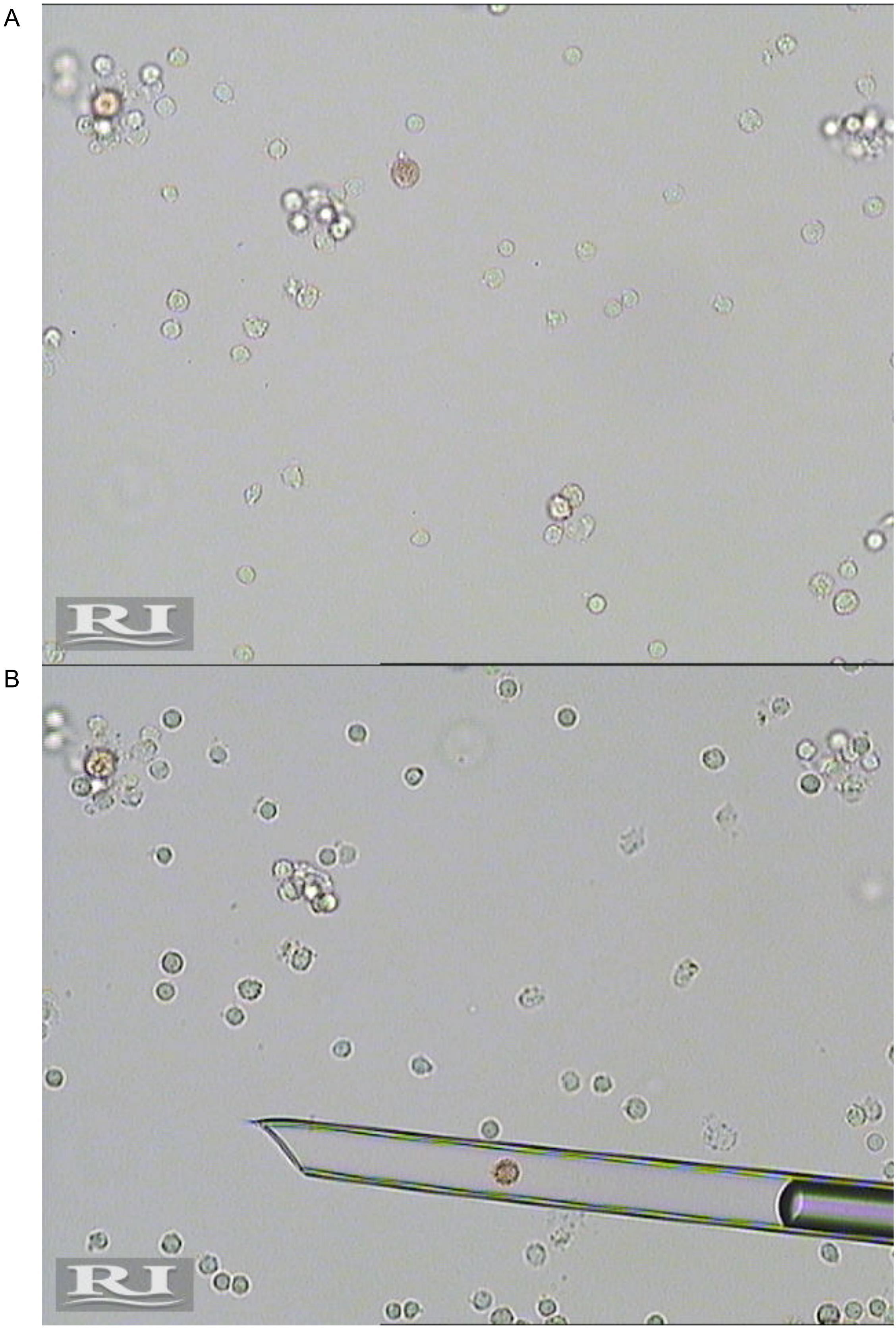


Figure 6.3 Micromanipulation of a PAX$^+$ cell. Cell stained in pink from the BMA of RK1013 before and after being aspirated into the glass capillary shown in A and B.

6.6 Genomic profiling of PAX8$^+$ cells

A summary of PAX8$^+$ cells isolated through micromanipulation which underwent successful amplification is show in Table 6.2. Overall, the success rate for amplification of isolated cells was 58%.

Patient	Sample	Cells amplified	Cell ID
RK1040	Normal Kidney	1/3	YAN5664A3
RK1013	Primary tumour	2/3	YAN5664A1
			YAN5664A2
RK1043	Pancreatic Met	5/6	YAN5664A11
			YAN5664A12
			YAN5664A13
			YAN5664A14
			YAN5664A15
RK1007	BMA	1/3	YAN5664A4
RK1013	BMA	1/5	YAN5664A5
RK1030	BMA	4/8	YAN5664A16
			YAN5664A17
			YAN5664A18
			YAN5664A19
RK1033	BMA	8/8	YAN5664A24
			YAN5664A25
			YAN5664A26
			YAN5664A27
			YAN5664A28
			YAN5664A29
			YAN5664A30
RK1036	BMA	4/8	YAN5664A20
			YAN5664A21
			YAN5664A22
			YAN5664A23
RK1040	BMA	5/9	YAN5664A6
			YAN5664A7
			YAN5664A8
			YAN5664A9
			YAN5664A10

Table 6.2 Cells isolated and amplified from TRACERx Renal samples.

Copy number calling was performed on all amplified cells with YAN5664A17, YAN5664A20 and YAN5664A22 removed due to high levels of noise resulting in their profiles being uninterpretable.

6.6.1 Copy number profiles of PAX8$^+$ cells from primary tumour and pancreatic metastasis

As expected, the copy number profile from a PAX8$^+$ cell isolated from the primary tumour of RK1013 was aneuploid with a classical chr3p loss (Figure 6.4). This cell displayed a loss in chr14 which is another common copy number event in ccRCC. In addition, a near complete loss of one copy of chr8 was also seen.

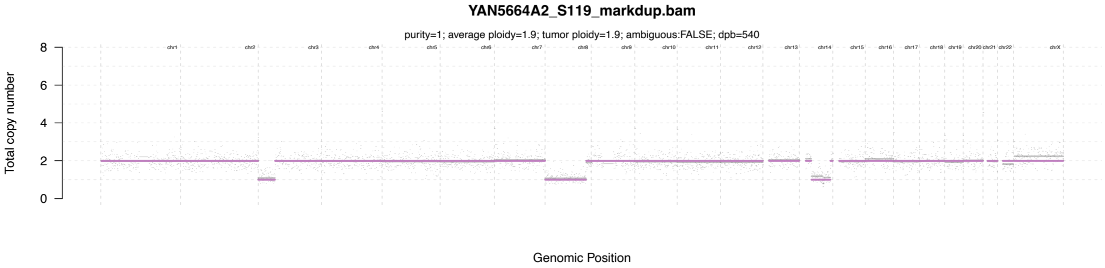


Figure 6.4 Copy number profile of a PAX8$^+$ cell from the primary tumour of patient RK1013.

Similarly, cells isolated from the pancreatic metastasis of RK1043 revealed copy number events including classic changes such as chr3p loss and chr5 gain (Figure 6.5). In addition, part of chr9q was lost in all three cells. There were some subtle copy number differences across these three cells, such as a loss of part of chr3q and chr12q in YAN5664A11. Surprisingly, two PAX8$^+$ cells from this pancreatic metastasis displayed diploid copy number profiles suggesting that not all positively staining cells were tumour cells (Appendix 8.3.3).

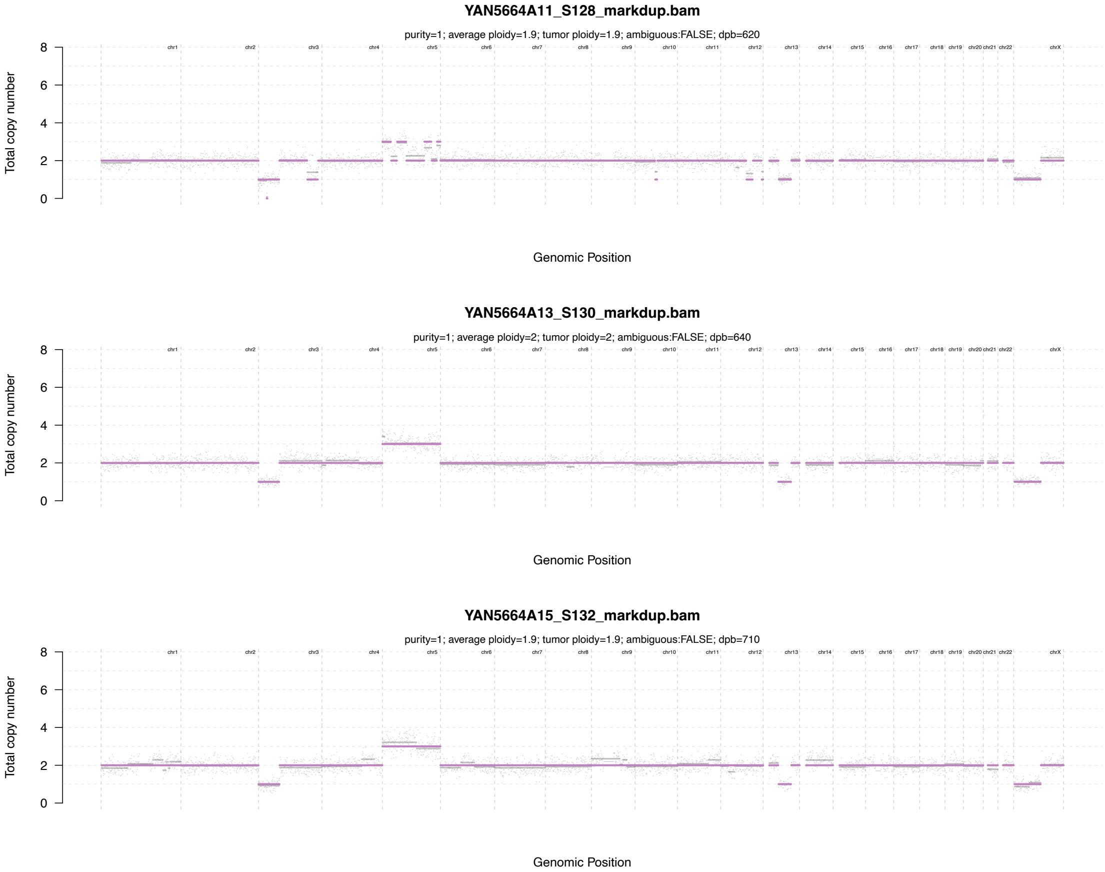


Figure 6.5 Single cells isolated from pancreatic metastases.

6.6.2 Copy number profiles of PAX8$^+$ cells in bone marrow

One cell did feature gains of two copies of chromosome 5 and one of chromosome 10 (Figure 6.6A). In addition, one cell displayed a subtle reduction in the relative number of reads for chr3q although the fitted profile would only be explained by being a doublet (Figure 6.6B). As chr3p loss and chr5q are common in ccRCCs, these cells may represent early DTCs. The remaining cells displayed diploid profiles with the exception of those described in section 6.6.3 below (Appendix 8.3.4).

6.6.3 Aberrant cells of unknown origin

As in the previous PEACE study, several cells with small chromosomal gains or losses were profiled in keeping with being aberrant cells of unknown origin. One PAX8$^+$ cell from the primary tumour of patient RK1013 displayed a focal gain of chr22 (Figure 6.6C). Two other cells from the bone marrow displayed losses of chr22 and part of chr2q (Figure 6.6D and E).

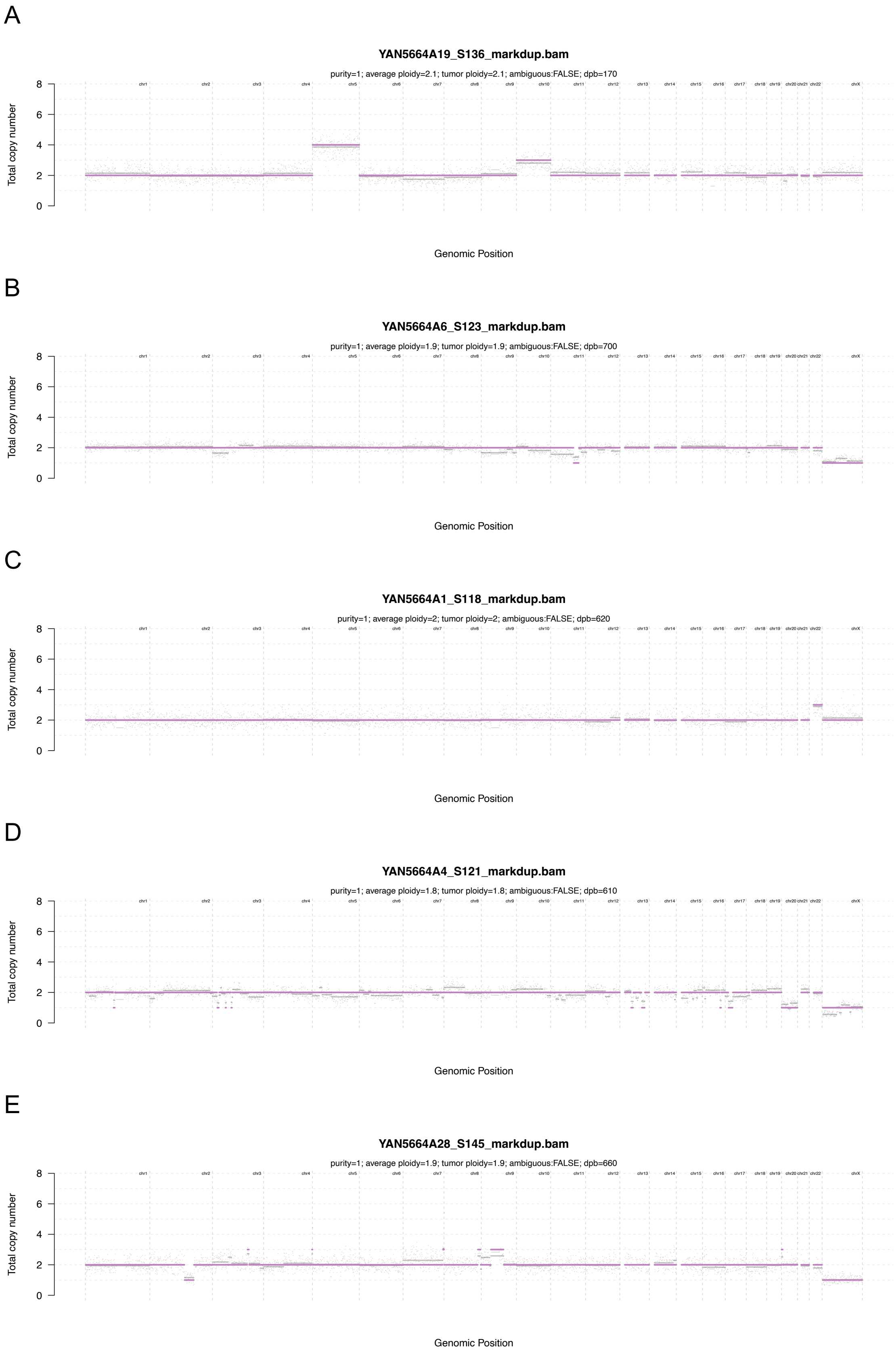


Figure 6.6 Aneuploid profile of PAX8^{+} cells from the bone marrow.

6.7 Discussion

In this study, I collected bone marrow aspirates from six patients with ccRCC undergoing surgery. I demonstrate that ICC staining of PAX8 can be performed in solution and that positively staining cells can be isolated using micromanipulation. Whole-genome amplification, sequencing, and copy number calling can be performed on the isolated single cells.

Copy number profiles from isolated cells revealed conserved chr3p loss in tumour cells isolated from the primary tumour and pancreatic metastasis. One aneuploid PAX8$^+$ cell with gains in chr5 and chr10 was detected in the bone marrow which could be in keeping with being a DTC. However, the lack of chr3p loss was unexpected as it is thought to be the earliest event in ccRCC tumourigenesis and often occurs simultaneously with chr5 gain. Genotyping SNVs found in the matched primary tumour in this cell would confirm whether this cell is truly a DTC.

Genomic profiling of pancreatic metastases in ccRCC has been limited and has not been done at the single-cell level. Although the three cells profiled here share most copy number aberrations, subtle differences in copy number between them demonstrates intra-tumour heterogeneity in this metastasis and suggests ongoing evolution.

6.7.1 Challenges in DTC detection and isolation

Several challenges were encountered during this study relating to correctly identifying tumour cells and isolating cells. Micromanipulation was utilised to pick and transfer cells from a single-cell suspension. However, transfer of the isolated cell into a PCR tube was technically challenging. This was due to a lack of suitable containers which were sufficiently shallow to allow deposition of the cell from the glass capillary. Furthermore, adhesion of the cell to the inside of the capillary was also encountered but could be managed by coating the capillary with 1% BSA. These difficulties likely increased the failure rate with the cell likely not transferred into the PCR tube for the ~40% of cases where whole-genome amplification failed.

Recently, automated robotic systems for micromanipulation such as the CellCelector have been developed and used successfully for CTC isolation through immunofluorescence of multiple markers (Acheampong *et al.*, 2023). This also has the added advantage of increasing the number of cells able to be profiled as manual micromanipulation is hugely time-consuming.

Similar to CAIX in the PEACE study, PAX8 also does not appear to be specific enough for detection of ccRCC tumour cells when profiling low numbers of cells. Whilst the presence of PAX^{+} diploid cells in the primary tumour could be explained by residual normal renal parenchymal cells, the two diploid cells isolated from the pancreatic metastasis suggest this stain is also not particularly specific. Therefore, alternative markers or combinations of markers may be needed to increase specificity of the cells detected as DTCs. However, as staining with two or more markers with ICC is difficult, immunofluorescence-based detection may be required. Unfortunately, a microscope with both fluorescence channel and micromanipulator was not accessible for this study.

Alternatively, a size-based microfluidic system such as Parsortix could also be considered for isolation of cells. This method is thought to be less traumatic and direct processing of samples without PBMC isolation may allow the more fragile DTCs to be collected and profiled. In addition, cell surface antigens may also be better preserved with this method.

6.7.2 Aberrant cells of unknown origin

Surprisingly, multiple cells with small chromosomal losses or gains were detected when selecting for cells with PAX8 or CAIX expression, as described in Chapter 4. These cells were also identified as making up nearly half of EpCAM^{+} cells in the bone marrow of breast cancer patients (Demeulemeester *et al.*, 2016). They were termed “aberrant cells of unknown origin” and thought to be epithelial cells which have travelled to the bone marrow or were of haematopoietic origin. Therefore, the aberrant cells profiled here may have similar origins.

Additional explanations for these cells could include that small genomic aberrations in normal tissues may be more common than previously thought. The focal chromosomal gains or losses observed here may mirror the recent findings of somatic driver mutations across multiple normal tissues. Alternatively, as these cells appear to express PAX8, they may truly be of renal origin which have migrated to the bone marrow. Recently, paradigm changing work has reported dissemination of non-malignant epithelial cells through the bloodstream mediated by the sodium leak channel nonselective protein (NALCN) (Rahrmann *et al.*, 2022). These cells go on to colonize and form morphologically normal structures in organs such as the lung and kidney. Metastases have always been thought to be exclusive feature of tumour cells, but these findings challenge this. However, perhaps this should not be considered surprising given that epithelial cells have migratory capacity which contributes to wound healing. In addition, there are many examples of cell migration in the body during development such as neural crest migration. The aberrant cells seen in this study may be another example of non-malignant epithelial cells, which have undergone dissemination to different organs.

Overall, I was able to isolate and profile the genomes of single cells from bone marrow samples. Several cells with copy number aberrations which warrant further study were detected although they were not confidently identified as DTCs.

Chapter 7. Discussion

This thesis has applied single-cell and spatial sequencing techniques and integrated data modalities to explore tumour evolution. I have also applied single-cell sequencing in the unique and challenging contexts of research autopsy and minimal disease to profile rare cells of interest. Together, these demonstrate the potential use cases of the latest sequencing technologies and our ability to profile tumours at a single-cell resolution.

7.1 Summary

In contrast to many large cohort studies that are limited to one modality, Chapters 3 and 4 investigated one tumour at high resolution with multiple omics modalities. Single-cell data was used to profile the composition of this tumour both in terms of tumour subclones and the tumour microenvironment. Bulk subclonal reconstruction techniques were validated, and a phylogenetic tree was reconstructed to the single-cell level. The impact of copy number changes on subclones through changes in gene expression and cell states was also explored.

In addition, spatial information was incorporated to provide contextual information for the different cells profiled. Recently advances in technologies have permitted subclones to be spatially mapped based on SNVs. However, the resolution permitted by these techniques is limited by the number of SNV-based subclones we can define. Instead, copy number heterogeneity in this tumour was leveraged to extend this to dozens of spatial subclones and map out their relationships to recapitulate growth in a human tumour.

Multiple different omics modalities were integrated, primarily through copy number profiling. Whilst this was successful between copy number profiles derived from DNA-based datasets such as bulk WGS, LCM, and scDNA-seq, this was more challenging for copy number changes inferred across omics layers. Nevertheless, diagonal integration has allowed spatial phylogenies to be reconstructed from transcriptomic data in minute detail for the first time.

In Chapter 5, single-cell sequencing technologies were applied in a highly challenging setting. I demonstrate that genomic profiling of single cells isolated from research autopsy derived tissues is feasible and identify a cell from a histologically confirmed ccRCC micrometastasis. However, due to lack of specificity of tumour markers and a high rate of attrition for profiling cells, more extensive profiling of additional samples was not performed. scRNA-seq of flow-sorted autopsy samples was attempted, but only two samples produced analysable single-cell transcriptomes.

Finally, I set up a sub-study in a translational clinical trial which allowed me to collect bone marrow aspirates from a small cohort of renal cell carcinoma patients with limited disease which is described in Chapter 6. I then optimised a protocol for staining of cells which allows for the isolation of cells of interest through manual micromanipulation. Although no obvious DTCs were ultimately profiled, several cells featuring copy number gains and/or losses were detected.

7.2 Cross-cutting themes

7.2.1 Genomic profiling of single cells

The ability to sequence the genome of individual tumour cells holds the potential to revolutionise our understanding of cancer evolution. Two key genetic alterations read out from single cells have proven informative in this thesis: copy number aberrations and SNVs.

Across my work, integer copy number values have been reliably called in single cells from both high and low throughput shallow-coverage WGS. This was also extended to allele-specific calling if haplotype phasing information was available. Due to their consistency with which these copy number changes can be inferred, they can be used as used to identify aneuploidy in cells and to perform lineage tracing (if there is a sufficient number and diversity of events).

In contrast to copy number profiling, I found that *de novo* SNV calls from single cells appear to be far less trustworthy, likely to false positives generated by DNA polymerase errors during PCR amplification. However, genotyping positions of variants (SNPs or SNVs) was highly informative. SNP positions from the 1000

Genomes Project were used to determine allelic imbalance and provided orthogonal validation on tumour cells' identity in all three projects. Similarly, genotyping known SNVs called from bulk data in single cells was used to identify specific subclones. Therefore, to maximise information derived from single-cell genomes which have undergone whole-genome amplification, mutation calls from paired bulk sequencing is currently still necessary.

As new technologies overcome the technical limitations of current single-cell sequencing methods, the quantity and quality of information we can extract from single cells will improve. Tagmentation technologies such as DLP and ACT allow for library preparation before minimal PCR amplification, resulting in unique sequencing inserts derived from the original template (Zahn *et al.*, 2017; Minussi *et al.*, 2021). This allows for duplicates introduced by subsequent PCR amplification to be computationally removed. When combined with technologies that can dispense nanolitre volumes of reagents, thereby greatly reducing library preparation costs, large numbers of cells can be profiled. SNV calling can then be performed on pseudo-bulks by merging reads from single cells together and could be potentially implemented for the MPNST project in future. Alternatively, a novel method which specifically amplifies the primary template DNA is able to provide high coverage breadth with significantly increased SNV calling sensitivity (Gonzalez-Pena *et al.*, 2021). When applied to neurons from patients with Alzheimer's disease, SNVs with a specific mutational signature associated with oxidative damage to guanine nucleotides were detected from single-cell data alone (Miller *et al.*, 2022a).

Single-cell sequencing has greatly expanded the possibilities of lineage tracing in tumours. These phylogenetic methods currently mostly rely upon copy number changes. However, in copy number quiet or tumours without copy number heterogeneity, alternative methods are required. Lineage tracing in human tumours has been achieved with mitochondrial mutations found in either the genome or transcriptomes of single cells (Ludwig *et al.*, 2019; Miller *et al.*, 2022b). In model organisms, lineage tracing systems can be introduced experimentally and recent work tracking tumour evolution at high resolution has revealed tumours growing under different selection pressures, and tumours with rare subclones as the source of expansion (Yang *et al.*, 2022). In addition, if epigenetic alterations or epimutations are inherited by descendent tumour cells, these could serve as alternative markers for lineage tracing. Indeed, heritable epimutations have been successfully used to perform lineage tracing in chronic lymphocytic leukaemia and these phylogenies were validated with genetic mutations genotyped in the same cells (Gaiti *et al.*, 2019).

7.2.2 Early detection of aberrant cells

Early detection and prevention of cancer is a major goal of cancer research. Recent large-scale sequencing efforts have determined the timing of events in tumourigenesis for different cancer types (Gerstung *et al.*, 2020). This knowledge could be translated clinically by eliminating cells featuring these early events. This preventative framework could be applied for both MPNST and for ccRCC tumours.

Across all three projects, I encountered cells with subtle chromosomal abnormalities which occur during the development of cancer. For ccRCC, loss of chr3p, followed by inactivation of the remaining copy *VHL*, has been shown to be one of the first events in tumourigenesis (Gerlinger *et al.*, 2014; Mitchell *et al.*, 2018). Therefore, detection and elimination of cells with chr3p loss could be vital in preventing these cells from clonally expanding and acquiring further mutations required for development of outright ccRCC. This strategy may potentially be effective for patients with von Hippel-Lindau syndrome, who already possess a mutated copy of the *VHL* gene inherited through their germline. Any cells which then also develop chr3p loss can expand and may be present in significant numbers in the normal tissues of these patients. Elimination of these cells would not only prevent ccRCC formation but also various other tumours which commonly develop in von Hippel-Lindau patients.

In the case of MPNSTs with H3K27me3 loss, multiple different chromosomes appear to be lost after loss of *NF1*, eventually resulting in cells with a near-haploid genome. It is possible that *NF1* plays a critical role in detecting chromosomal losses and preventing their accumulation, and that cells with chromosomal losses may be present in other non-malignant peripheral nerves or nerve sheaths. WGD could then rescue them from this low-fitness state and set them on the path to malignant transformation. Therefore, eliminating these near-haploid cells or prevented them from undergoing WGD could be a viable strategy for preventing malignancy.

In order to provide further support for this hypothesis, these near-haploid cells must be isolated and further studied. This could be achieved in a similar fashion as the DTC project from tumour tissue but using flow cytometry ploidy analysis to identify haploid cells. Alternatively, knowledge of a tumour's evolutionary history could be leveraged to identify these rare cells. Current models of tumour evolution detail clonal sweeps where ancestral clones before the MRCA are (near-)completely out-competed, rendering them unavailable for further analysis. However, there may still be residual cells from these ancestral populations present in the tumour, and therefore should also be present in single-cell sequencing samples. This could be done by identifying cells that only harbour SNVs truncal to the entire tumour (i.e., those that before the MRCA) and reviewing their copy number profiles. Thus, single-cell sequencing could potentially be used to find cells which are ancestral to the MRCA previously considered to be extinct. If found, these ancestral cells can provide direct evidence of a particular evolutionary path or resolve ambiguities in reconstruction of the phylogenetic tree.

Finally, aberrant cells of unknown origin, characterised by losses of different chromosomes in normal lung or bone marrow were also observed, similar to previous studies (Knouse *et al.*, 2014; Demeulemeester *et al.*, 2016). The importance of these cells remains unknown although they were found more frequently than expected in our study. One could hypothesise that these sporadic losses are reasonably common, but are detected and eliminated (e.g., by cell-intrinsic mechanisms, or by the immune system) by the body before tumorigenesis. Our previous study has correlated their presence with increasing age and may represent failure of the aging immune system to remove them. The true malignant potential of these cells is then only seen in cases where there are existing oncogenic events that accelerate tumourigenesis, such as in ccRCC or MPNST, or in older patients where these cells have accumulated in sufficient numbers over time. Further profiling of these cells could uncover their importance as well as the mechanisms behind these losses. Recently, field-cancerization effects have also been noted, where non-tumour cells (either parenchymal or stromal) neighbouring tumour cells also develop chromosomal abnormalities (Jakubek *et al.*, 2020; Zhou *et al.*, 2020). These findings suggest the need for more comprehensive profiling of cells surrounding tumours to understand the mechanisms behind this phenomenon and identify early markers of cancer development.

7.2.3 Challenges in adopting new sequencing technologies

Despite the successes from applying cutting-edge techniques to tumour samples, several challenges were encountered. These primarily relate to implementation of newly developed technologies or a need for novel computational methods to be developed for these techniques.

Although a new technique may work reliably when first developed, successful adoption in new settings is not guaranteed. For example, despite Slide-seq being implemented by an experienced team with technical expertise, I was not able to perform any meaningful analysis on the data. This was likely due to subtle differences in puck generation with barcodes of beads incorrectly identified, possibly due to a different microscope being used. For DTC isolation from TRACERx Renal bone marrow samples, although several groups have used micromanipulation to isolate cells from Cytospin slides, this was technically challenging for me to perform, possibly due to differences in slides, microcapillaries, and other equipment. More advanced options such as robotic micromanipulation machines have been developed and successfully used to isolated CTCs but were unfortunately not available in the UK. Therefore, difficulties in implementation and availability of equipment can delay or prevent techniques from fulfilling their potential.

Computational tools are often rapidly developed for new techniques which are expected to be widely used. However, there can be an imbalance resulting in certain techniques not expected to be as popular being neglected. Currently, there is an unmet need for more computational tools for single-cell DNA sequencing analysis. In contrast, there is a large selection of tools for analysing scRNA-seq, making it difficult to determine which tools should be used and which perform the best. In the MPNST project, single-cell analyses were limited by a lack of available packages such as those for SNV calling and matching scRNA-seq cells to subclones. In addition, a phylogeny building method combining both SNVs and CNAs has to our knowledge not yet been developed. The development of these methods is likely hampered by scarcity of scDNA-seq tumour datasets and will likely progress as more datasets become available. Nevertheless, this presented an opportunity to develop my own methodologies for SNV filtering and integration of single-cell datasets.

7.3 Future work

7.3.1 MPNST acts as a reference for further multi-omic profiling

For the MPNST study, the application of such a wide range of techniques and the discoveries derived from each form of sequencing forms a valuable basis to make decisions on the most informative sequencing technologies to apply to larger cohort studies. For example, advances in LCM techniques and sequencing now allow for PCR-free library preparation and could therefore allow sequencing reads of LCM spots from the same region to be merged and treated as a pseudo-bulk for downstream analysis (Ellis *et al.*, 2021). This would allow much of the same information to be derived as bulk WGS without the associated significant financial costs of bulk sequencing.

Not only does this project provide a reference for information to be gained through each technique, the design of individual experiments can also be improved or optimised. For example, in future MPNST single-cell experiments, nuclei can be profiled from aliquots prepared from the same piece of tissue to prevent sampling differences resulting from the surprisingly extreme degree of heterogeneity seen in these tumours. Furthermore, the discovery of populations of cells with different ploidy complicates copy number fitting, as profiles can always be doubled leading to difficulty in identifying WGD cells (Tarabichi *et al.*, 2021). However, if a ploidy sort is performed before single-cell sequencing, this ambiguity in WGD status can be resolved, allowing further insights into WGD cells to be made. Alternatively, single-cell sequencing with DLP+ also captures microscopic images of each nucleus undergoing sequencing, which could be used to inform ploidy. Finally, in a tumour with high purity such as this MPNST, sections with some adjacent normal tissue would be beneficial for inferring copy number changes from spatial transcriptomics, as a normal reference is required by most methods.

7.3.2 Applications of technological advances

Advances in both the technology and reduction in sequencing costs will enable more scientific questions to be addressed. More extensive profiling of normal tissues could partially overcome the issue of low tumour marker specificity encountered in the TRACERx Renal DTC project and may even make profiling of autopsy-derived tissues feasible. Furthermore, development of systems to experimentally induce specific copy number gains or losses could be leveraged to profile cellular responses to these events and their underlying mechanisms (Sheltzer *et al.*, 2017). These experimental systems may also shed light on how cells sense aneuploidy and prevent their genomes becoming more aberrant. Together, larger scale profiling and experimental studies may reveal insights into the nature of the aberrant cells of unknown origin found in the DTC projects.

More broadly, the development of new sequencing technologies has been unrelenting. Recently, new technologies have moved from profiling different single omics layers to multi-omic profiling and incorporating spatial information. For example, the epigenome and transcriptome can now be jointly profiled using spatial ATAC-RNA-seq and spatial CUT&Tag–RNA-seq (Zhang *et al.*, 2023). Other exciting new techniques such as Live-seq involves cytoplasmic biopsies from more than one time point for transcriptomic profiling whilst keeping the cell alive. This would allow for functional profiling before and after the application of interventions such as therapeutics or radiation in the same cell (Chen *et al.*, 2022). Meanwhile, Light-seq enables *in situ* photocrosslinking of barcodes to cDNA for transcriptomic sequencing of areas of interest (Kishi *et al.*, 2022). If successfully implemented, these technologies will allow creative experiments to be designed with applications only limited by our imagination.

7.3.3 Determining the importance of individual subclones

The importance of heterogeneity has been shown with the advent of personalised medicine. Here, differences between patients, or *inter*-tumour heterogeneity, has enabled the discovery of subgroups of patients with tumours which may benefit from specific treatments. Despite this, many treatments are still ineffective due to the degree of complexity in tumours, which arises from individual subclones which make up the tumour or *intra*-tumour heterogeneity. We are now entering into an age where tumours will be profiled at increasing detail and number of modalities, as exemplified by the MPNST project. Multi-region or representative sampling, single-cell or spatial sequencing techniques, and advances in computational methods will all contribute to characterising the complexity of cancer. Together, these approaches will allow the cellular makeup of subclones to be determined along with their detailed genotypic, phenotypic, and spatial features.

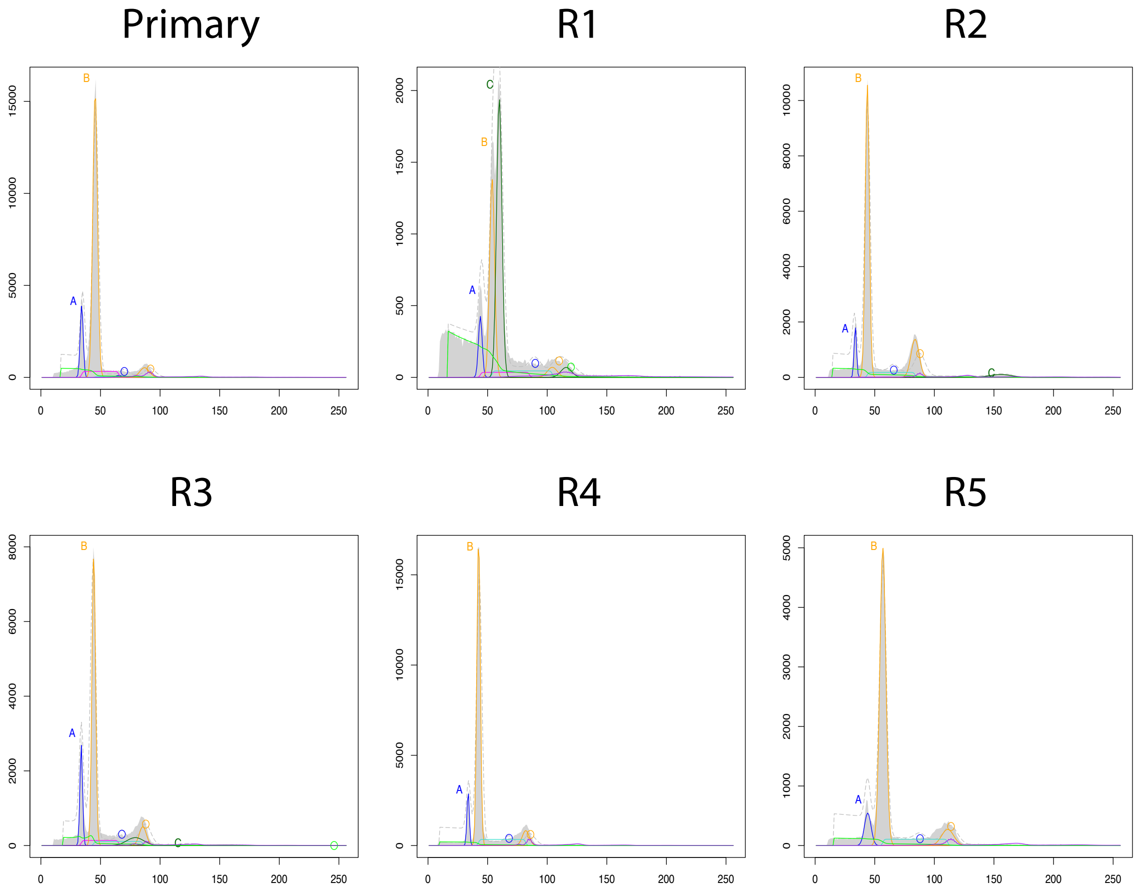
The clinical implications of our ability to generate phenomenal amounts of data from each tumour are unclear and will be an exciting area of development. From a personalised medicine perspective, identification of subclones lacking genomic alterations targeted by precision oncology would explain treatment failure. Furthermore, multi-modal information from exceptional subclones could inform us about the mechanism behind treatment resistance or immune evasion and can be exploited to design new therapies or increase the efficacy of current ones. Conversely, although there is hope that identification and elimination of subclones will improve outcomes for patients, the clinical importance of subclones of interest remains unclear. For example, even if a particularly aggressive subclone was eliminated, would this have an discernible impact on survival or would the remaining subclones still result in the patient's death? As the complexity within individual tumours has been understood, new therapies under development have incorporated this knowledge. These include therapies which specifically target clonal alterations, which are shared by all tumour cells, and not subclonal alterations which would only be effective against those subclones. Finally, whether this type of detailed subclonal profiling for each patient is required or whether novel findings from larger discovery cohorts can be generalised and applied to new patients still needs to be determined.

In conclusion, the work presented in this thesis is a showcase of current single-cell and spatial techniques and demonstrates the level of detail with which tumour evolution can be recapitulated.

Chapter 8. Appendix

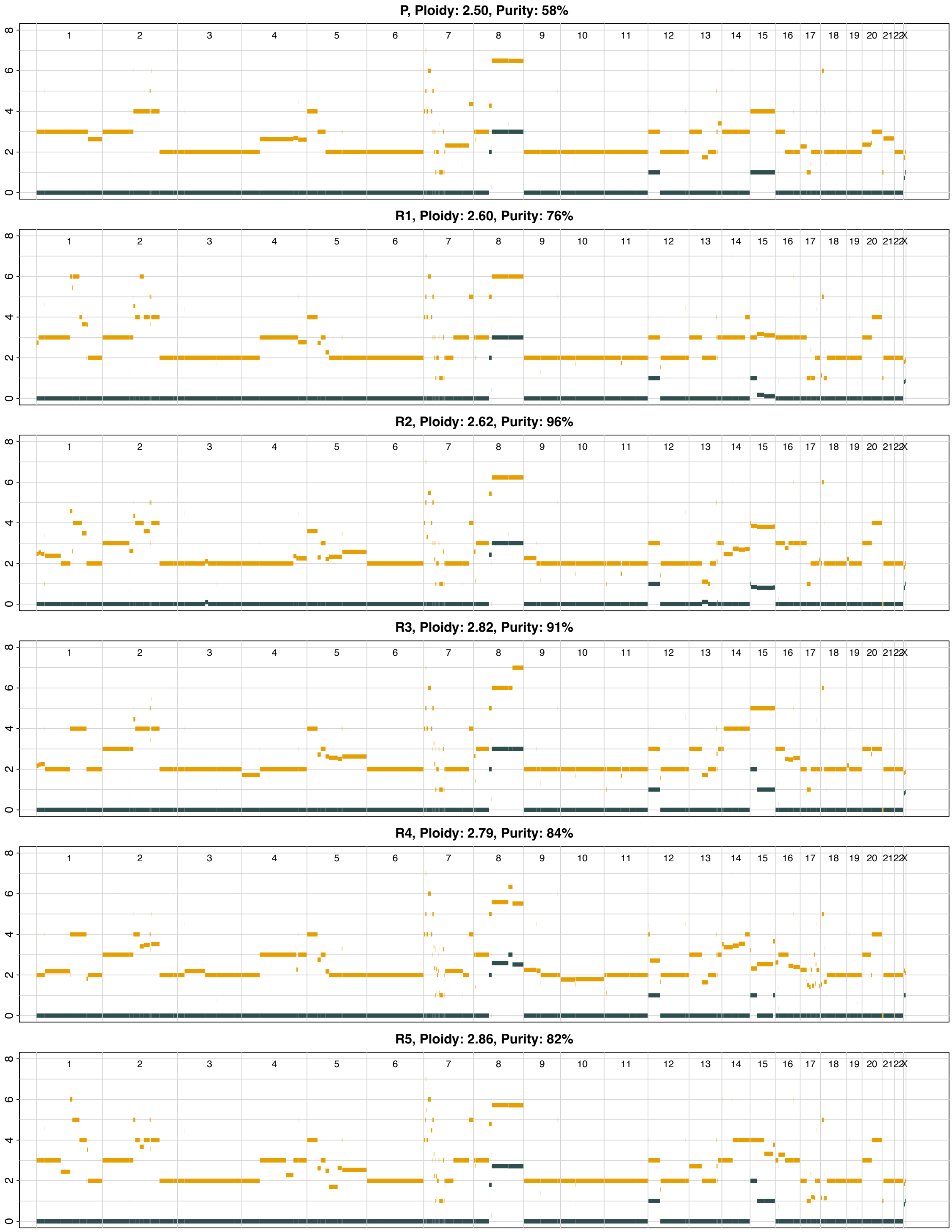
8.1 MPNST Appendix

8.1.1 Experimental ploidy analysis



Diploid cells shown by blue peak and tumour cells by yellow peak. Note the additional higher yellow peak which suggests the presence of an extra population of cells having undergone further WGD. These cells appear to have a ploidy of approximately 5.

8.1.2 Bulk Battenberg copy number profiles


P, Ploidy: 2.50, Purity: 58%

R1, Ploidy: 2.60, Purity: 76%

R2, Ploidy: 2.62, Purity: 96%

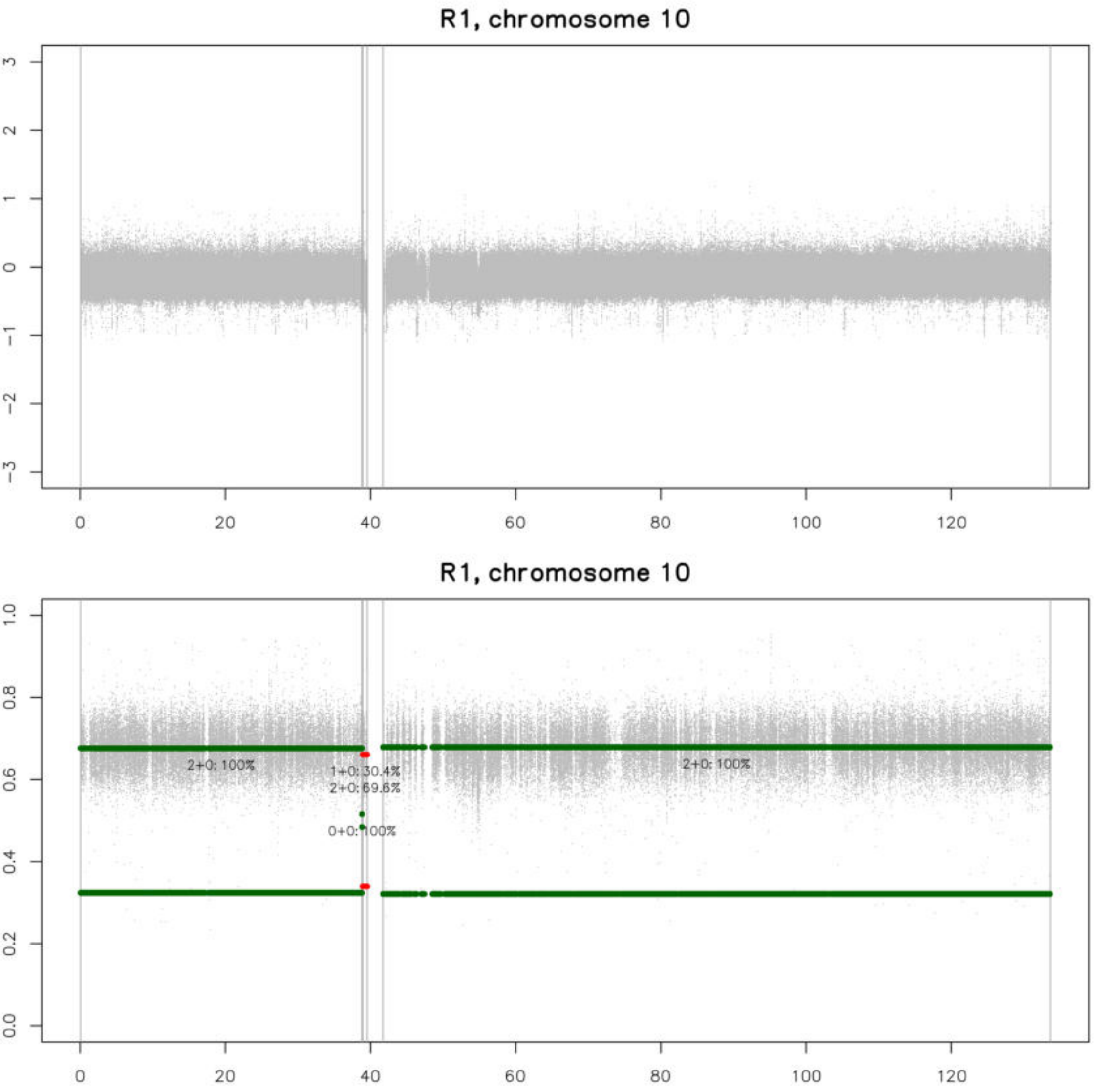
R3, Ploidy: 2.82, Purity: 91%

R4, Ploidy: 2.79, Purity: 84%

R5, Ploidy: 2.86, Purity: 82%

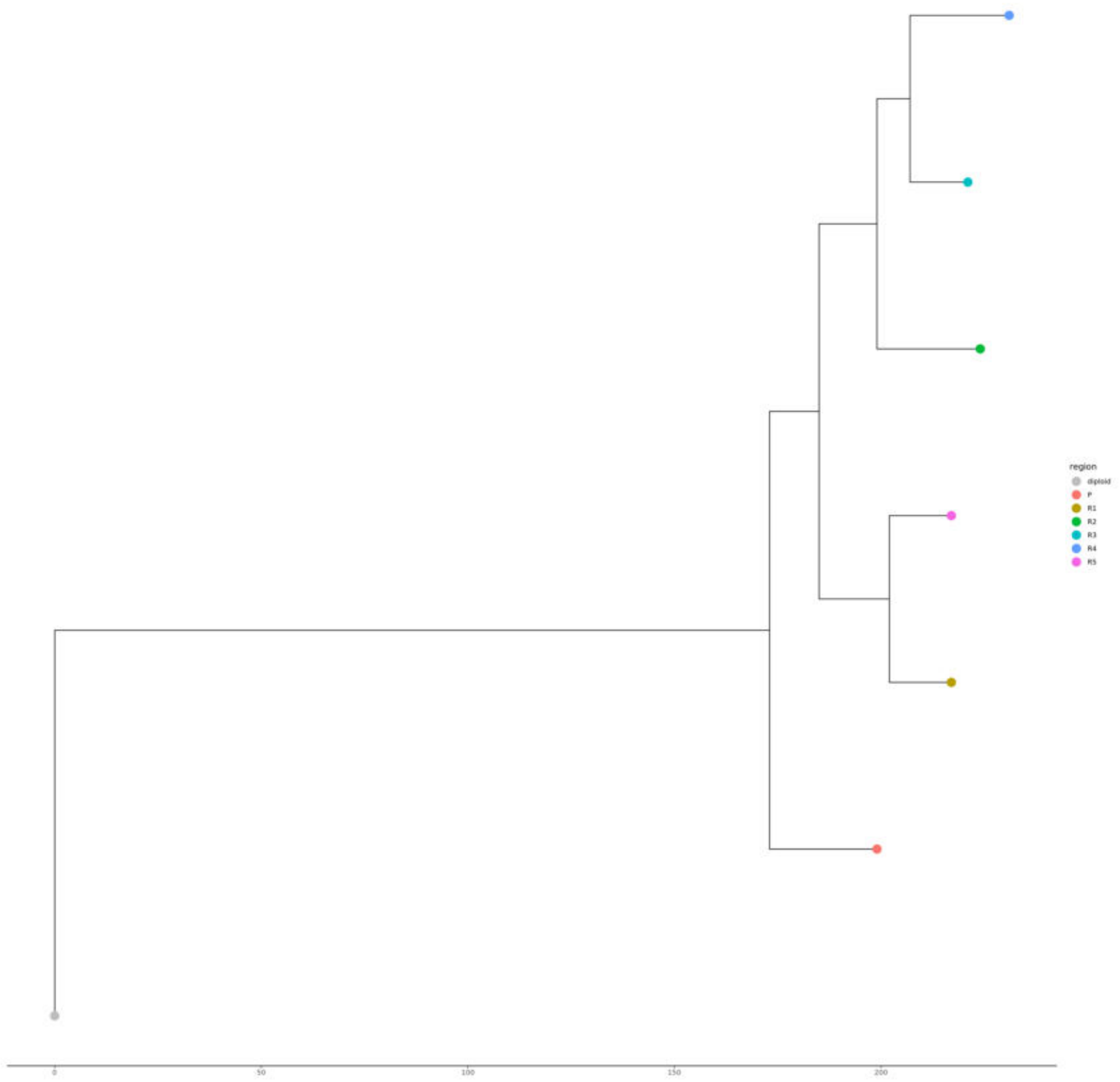
Copy-number values for all chromosomes with major alleles coloured by region and the minor alleles in teal. Non-integer copy-number values indicate the presence of a subclone with the size of each subclone proportional to the distance between integer states.

8.1.3 LogR and BAF tracks of chromosome 10 of the R1 recurrence



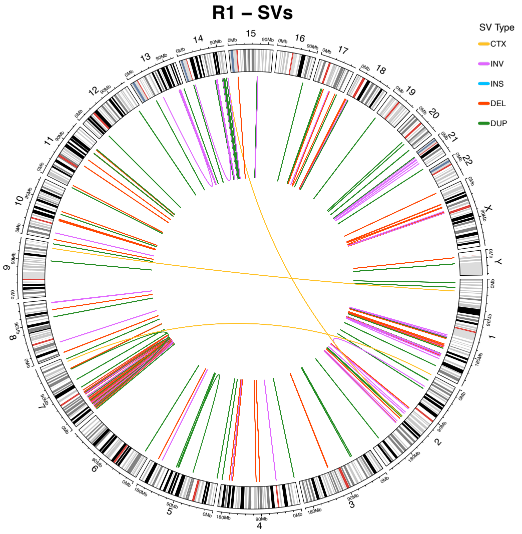
No deviation was observed in the logR (top) or BAF tracks (bottom) from Battenberg that would signify the presence of a subclone with chr10p loss that could be detected in bulk sequencing.

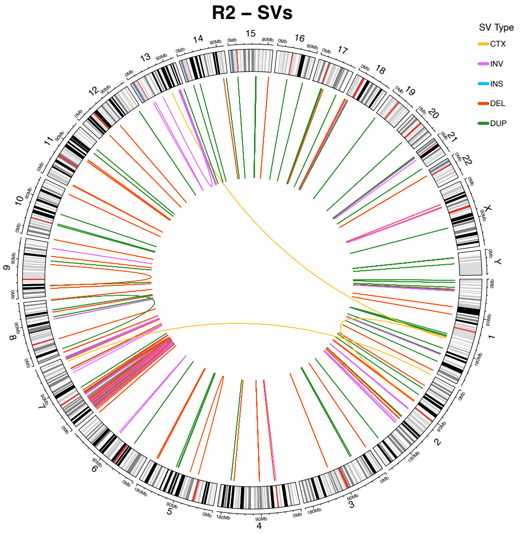
8.1.4 Copy number-based maximum parsimony sample tree

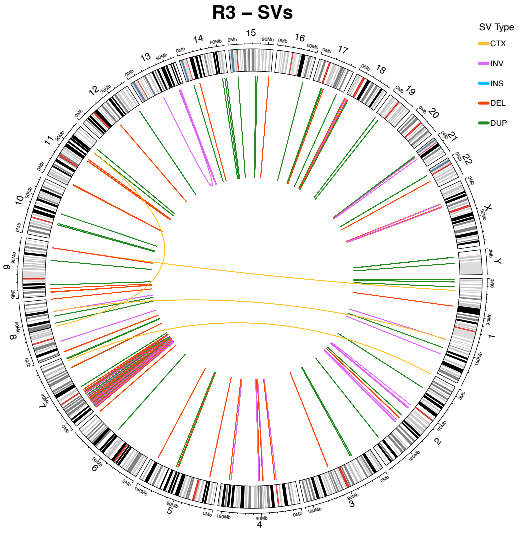


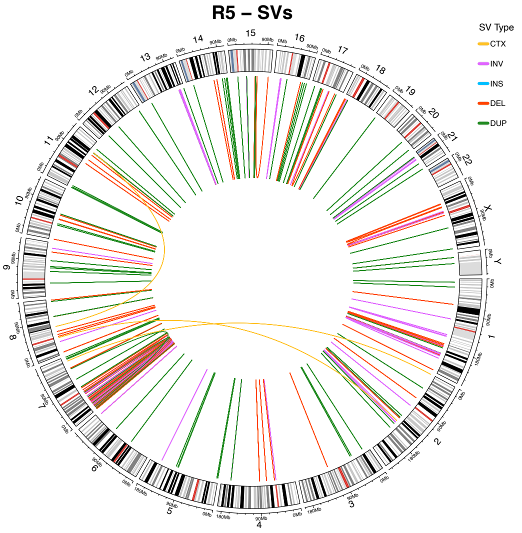
This sample tree is almost identical to the SNV-based tree with the exception of R3 and R4 being sister clones, compared to the SNV-based tree, where R2 and R4 are closest together.

8.1.5 Structural variants of additional regions

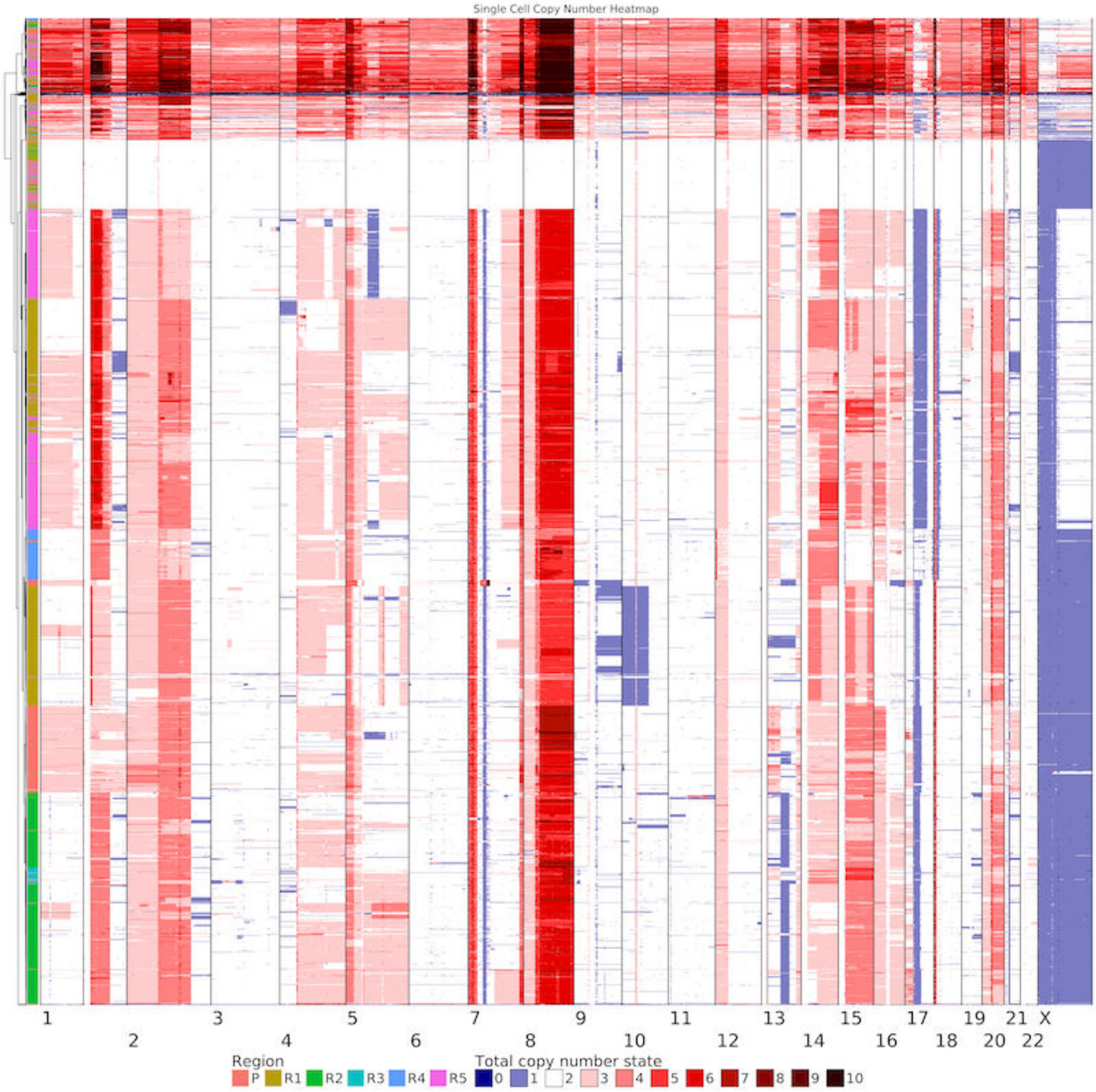

R1 – SVs


R2 – SVs


R3 – SVs

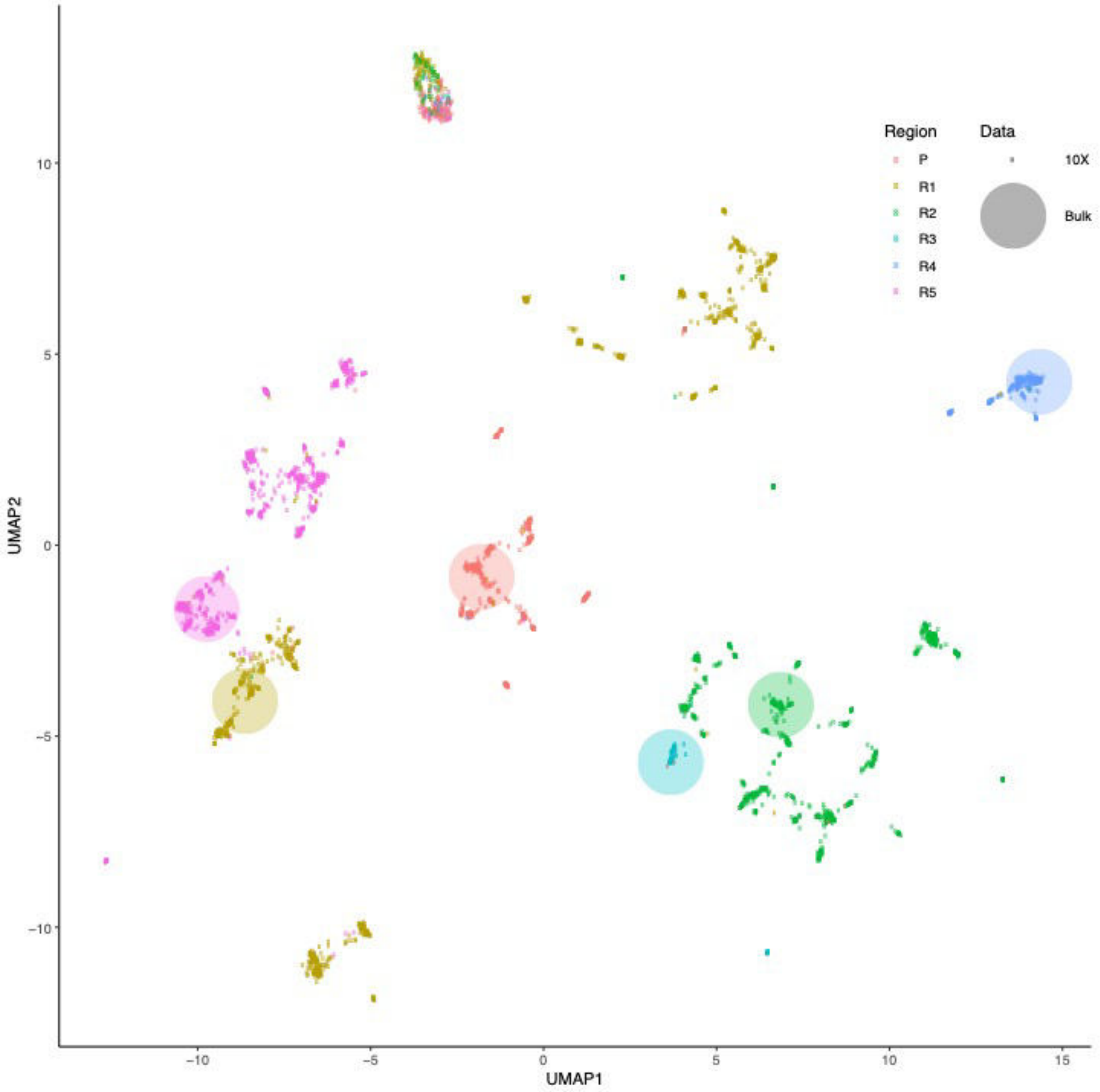

R5 – SVs

8.1.6 Raw single-cell copy number profiles

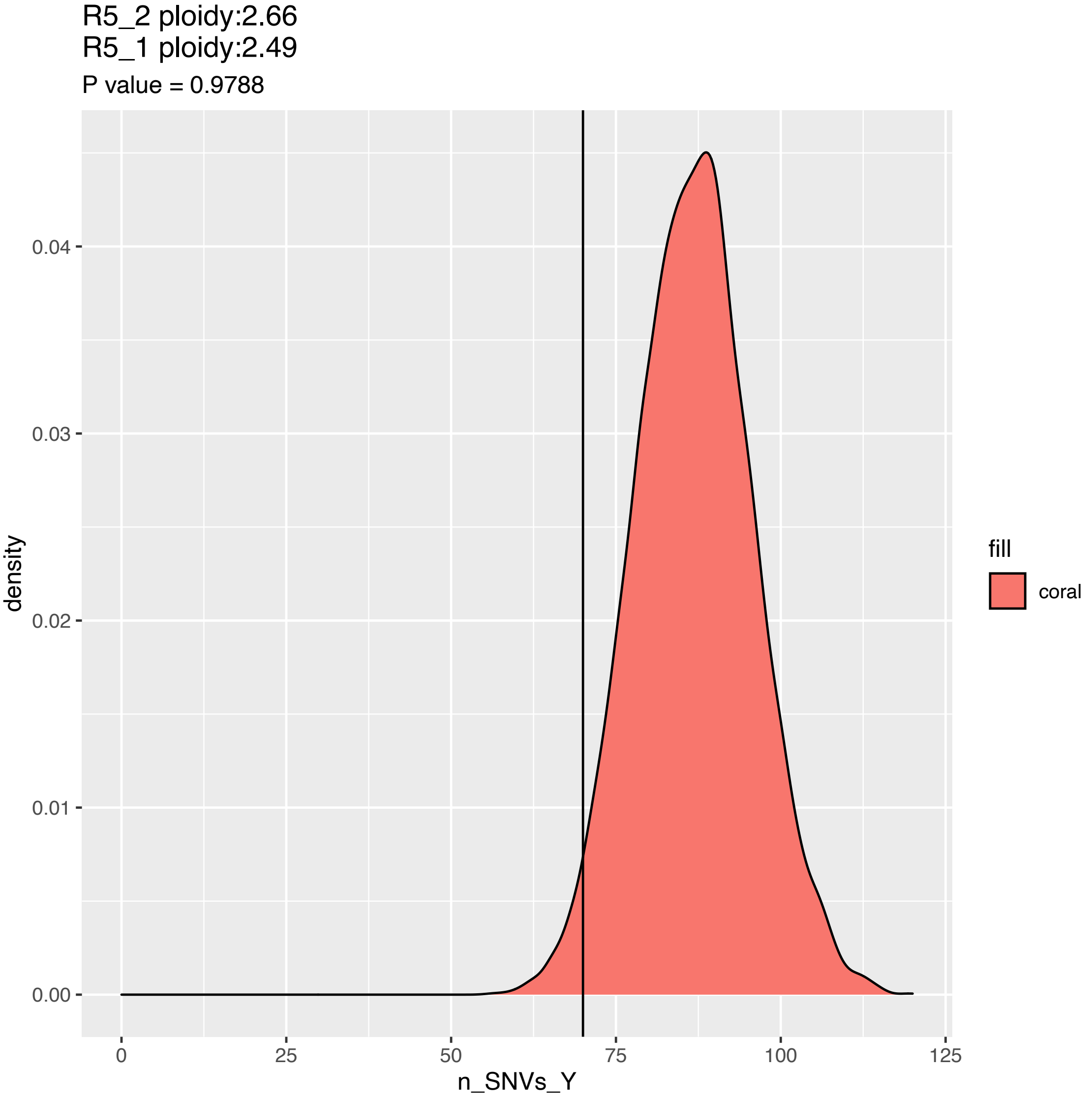


All cells including diploid and cells with a whole-genome doubled profile are shown. Cells are ordered through hierarchical clustering and annotated by their region of origin.

8.1.7 UMAP of bulk and single-cell total copy number profiles

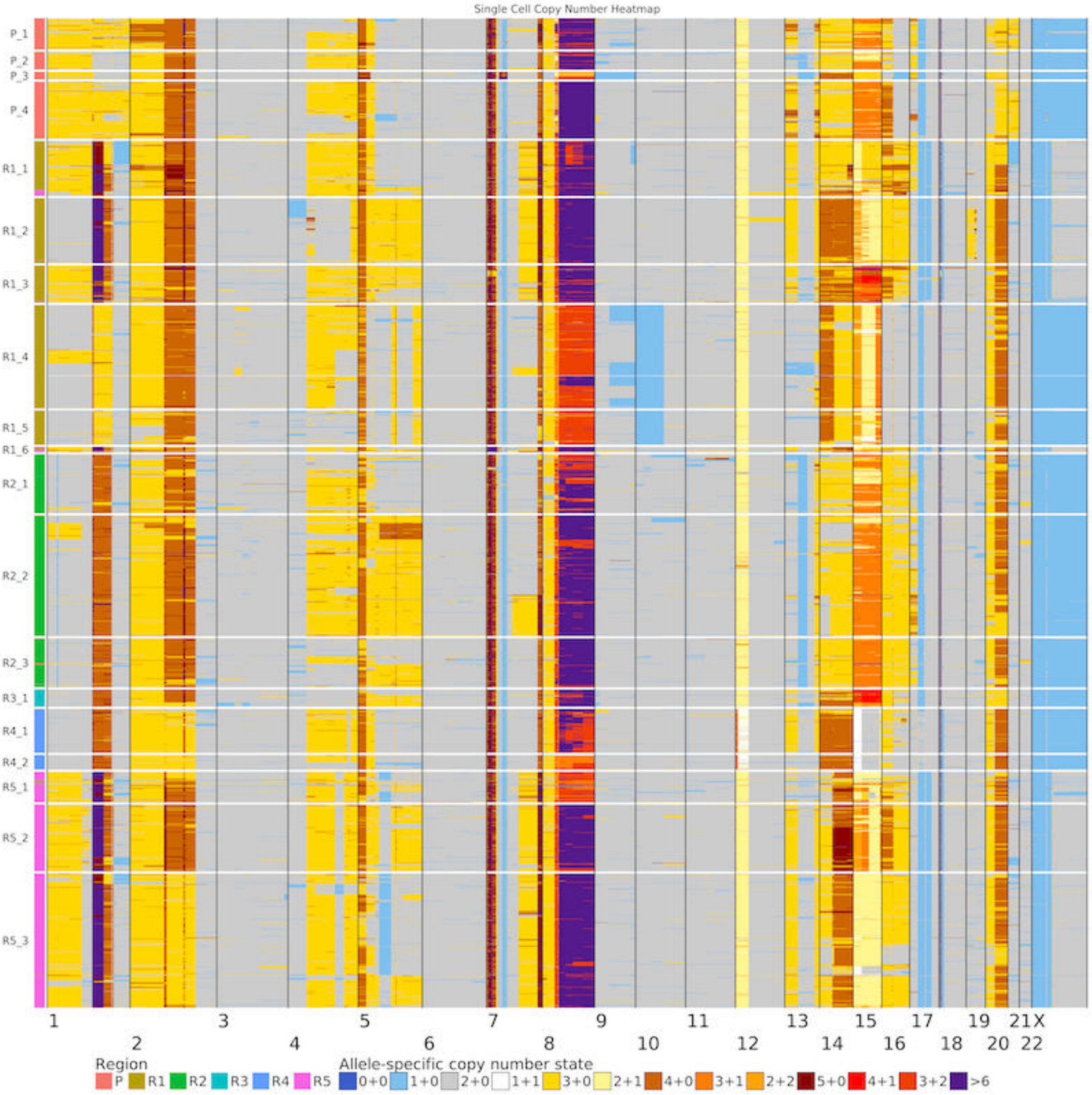


8.1.8 Simulating number of subclone-specific SNVs in cluster



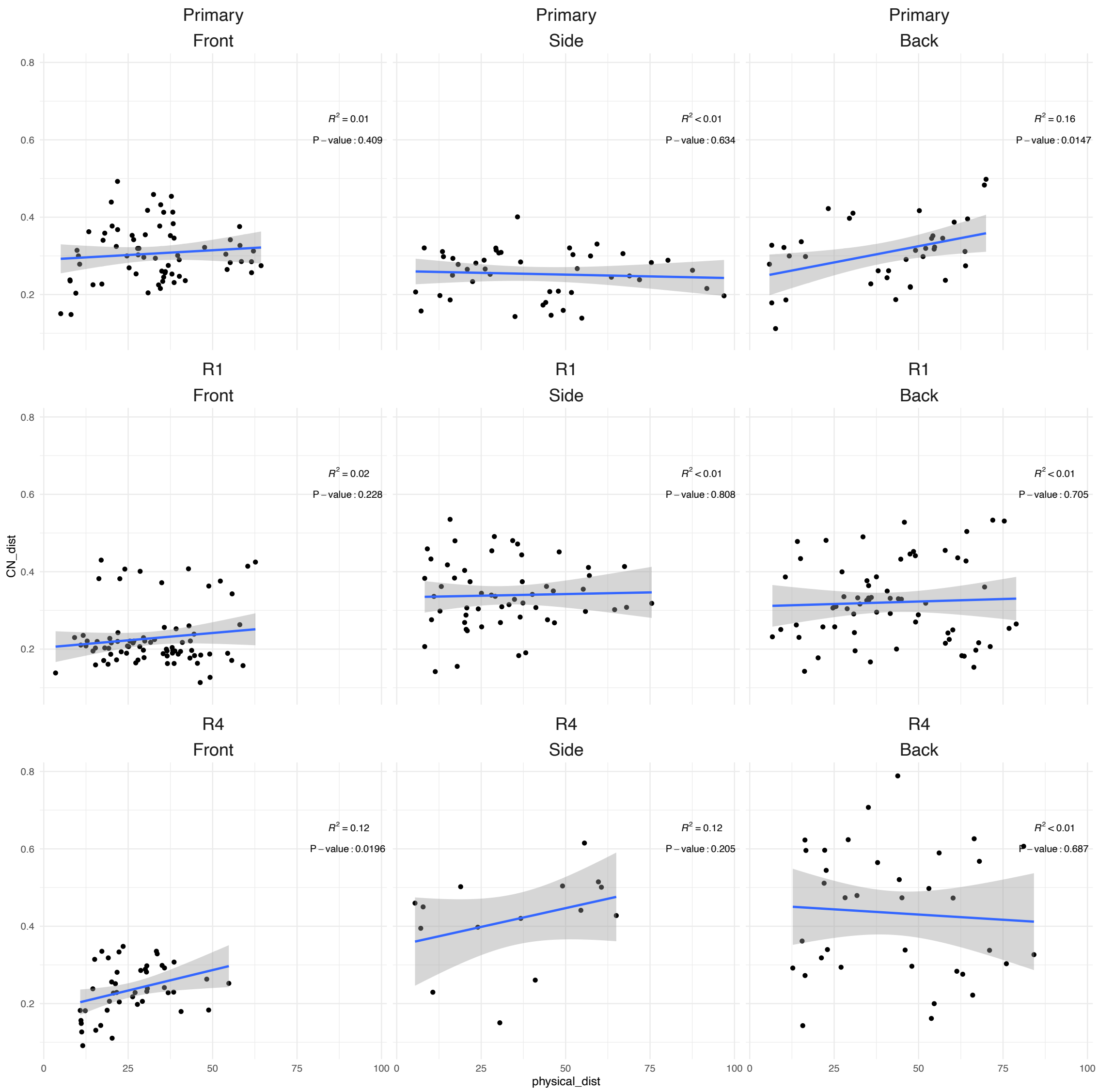
Distribution of number SNVs expected to be seen only in R5_1 by chance over 10,000 random samplings due to low coverage if R5_1 did not have any exclusive SNVs and all SNVs were, in fact, also present in R5_2. Vertical line showing the number of SNVs for R5_1 observed.

8.1.9 Allele-specific copy number profiles of scDNA-seq cells

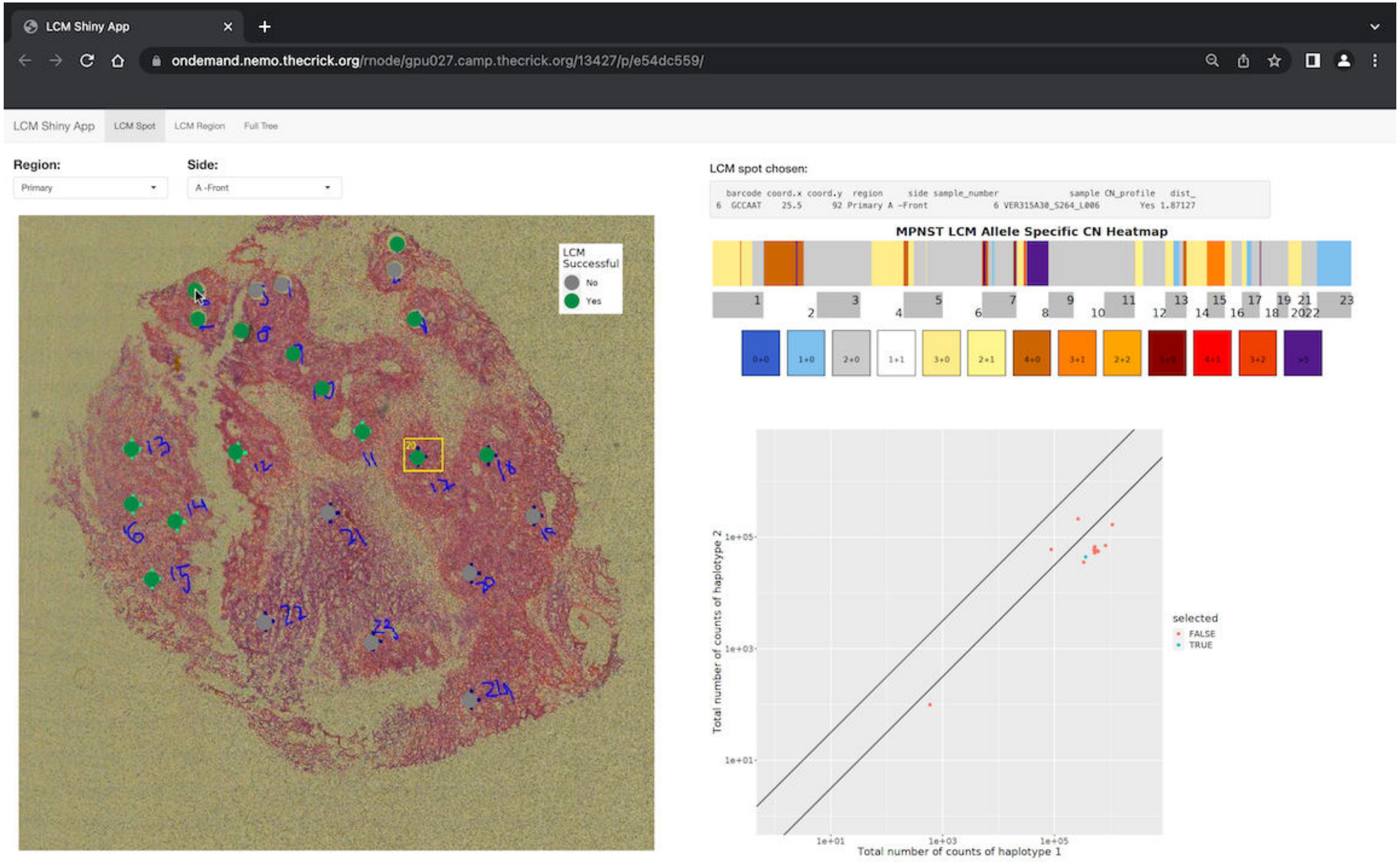


Cells are annotated by their region of origin and copy number values of segments are shown as copies of the major + minor allele.

8.1.10 Physical and genetic distances correlation in LCM sections



8.1.11 Interactive R Shiny app for LCM analysis

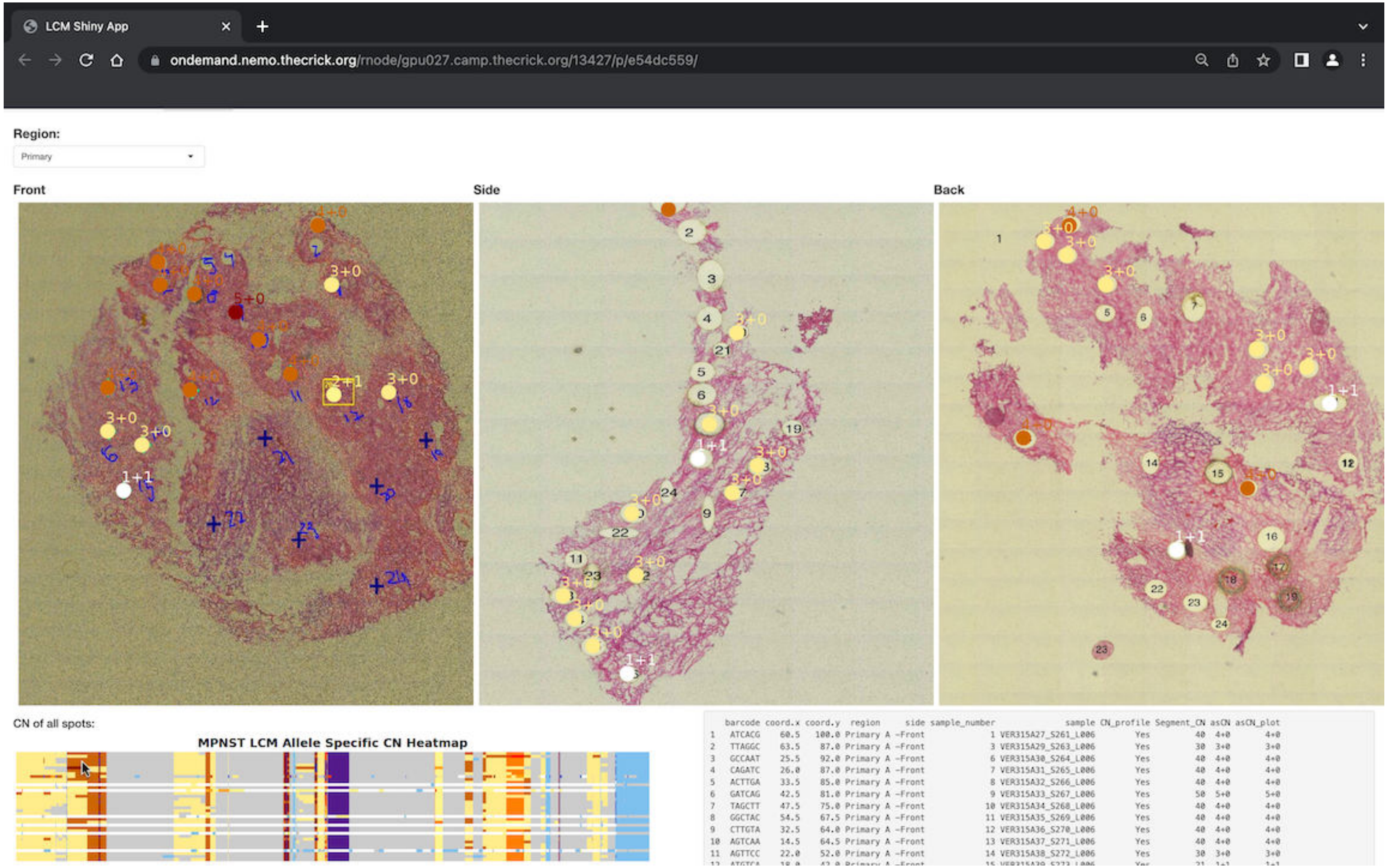


Region: Primary

Side: A -Front

LCM spot chosen:

```
  barcode coord.x coord.y  region     side sample_number              sample CN_profile   dist_
6  GCCAAT    25.5      92 Primary A -Front             6 VER315A30_S264_L006        Yes 1.87127
```

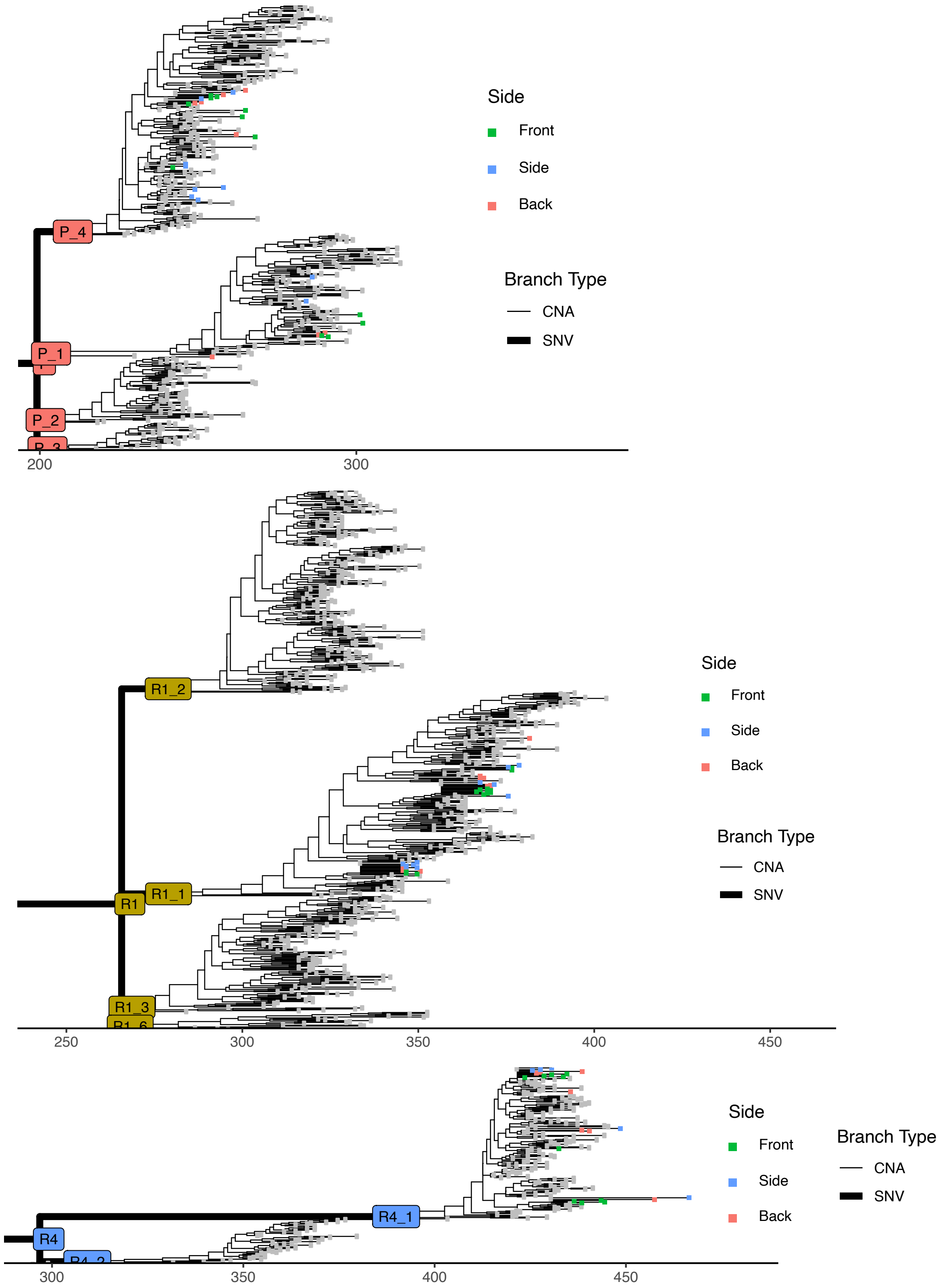


LCM Shiny App
ondemand.nemo.thecrick.org/rnode/gpu027.camp.thecrick.org/13427/p/e54dc559/
Front
Side
Back
CN of all spots:
MPNST LCM Allele Specific CN Heatmap
4+0
3+0
5+0
2+1
1+1

Region: Primary

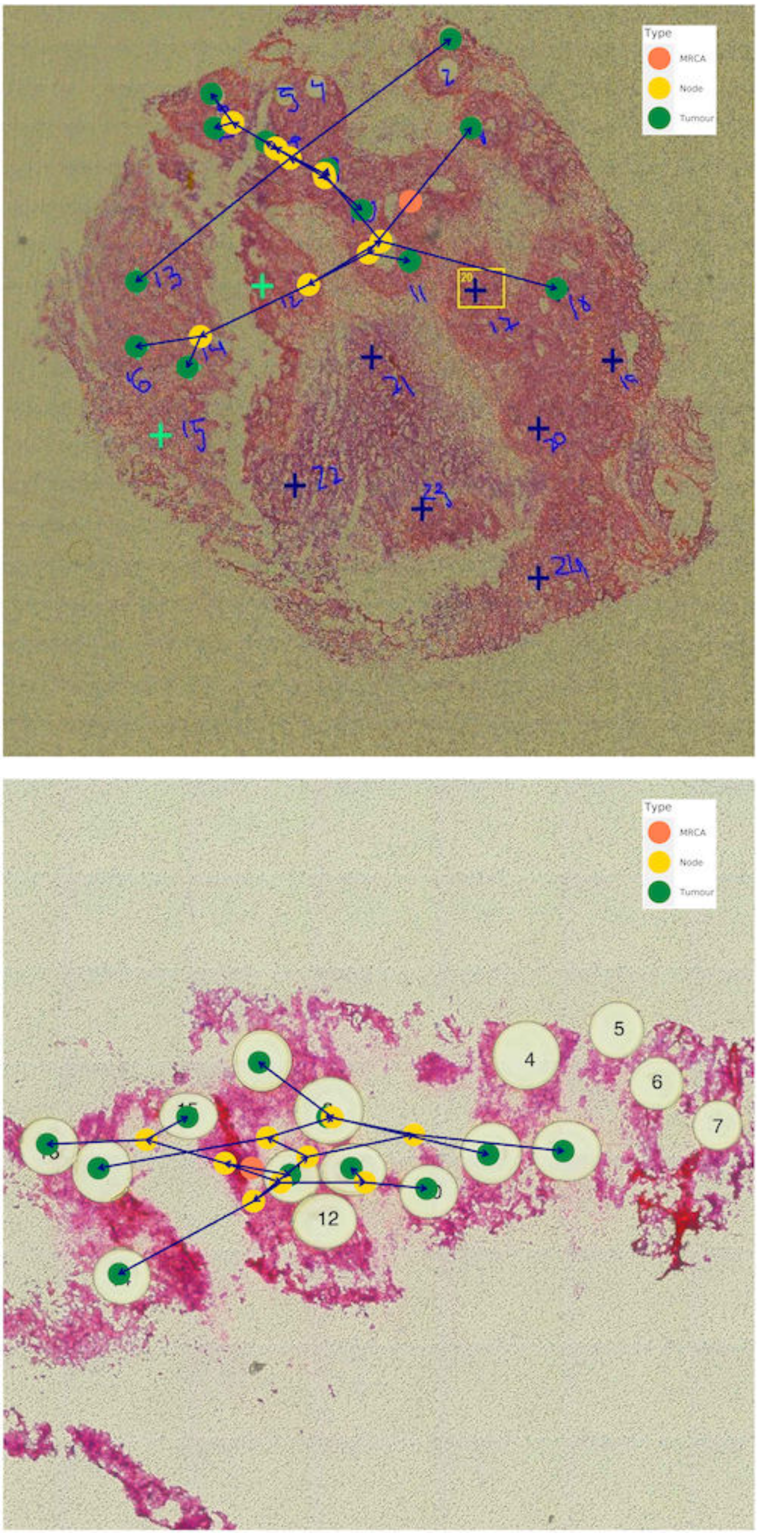
```
   barcode coord.x coord.y  region     side sample_number              sample CN_profile Segment_CN asCN asCN_plot
1   ATCACG    60.5   100.0 Primary A -Front             1 VER315A27_S261_L006        Yes         40  4+0       4+0
2   TTAGGC    63.5    87.0 Primary A -Front             3 VER315A29_S263_L006        Yes         30  3+0       3+0
3   GCCAAT    25.5    92.0 Primary A -Front             6 VER315A30_S264_L006        Yes         40  4+0       4+0
4   CAGATC    26.0    87.0 Primary A -Front             7 VER315A31_S265_L006        Yes         40  4+0       4+0
5   ACTTGA    33.5    85.0 Primary A -Front             8 VER315A32_S266_L006        Yes         40  4+0       4+0
6   GATCAG    42.5    81.0 Primary A -Front             9 VER315A33_S267_L006        Yes         50  5+0       5+0
7   TAGCTT    47.5    75.0 Primary A -Front            10 VER315A34_S268_L006        Yes         40  4+0       4+0
8   GGCTAC    54.5    67.5 Primary A -Front            11 VER315A35_S269_L006        Yes         40  4+0       4+0
9   CTTGTA    32.5    64.0 Primary A -Front            12 VER315A36_S270_L006        Yes         40  4+0       4+0
10  AGTCAA    14.5    64.5 Primary A -Front            13 VER315A37_S271_L006        Yes         40  4+0       4+0
11  AGTTCC    22.0    52.0 Primary A -Front            14 VER315A38_S272_L006        Yes         30  3+0       3+0
12  ATGTCA    18.0    42.0 Primary A -Front            15 VER315A39_S273_L006        Yes         21  1+1       1+1
```

8.1.12 LCM sections on phylogenetic tree.



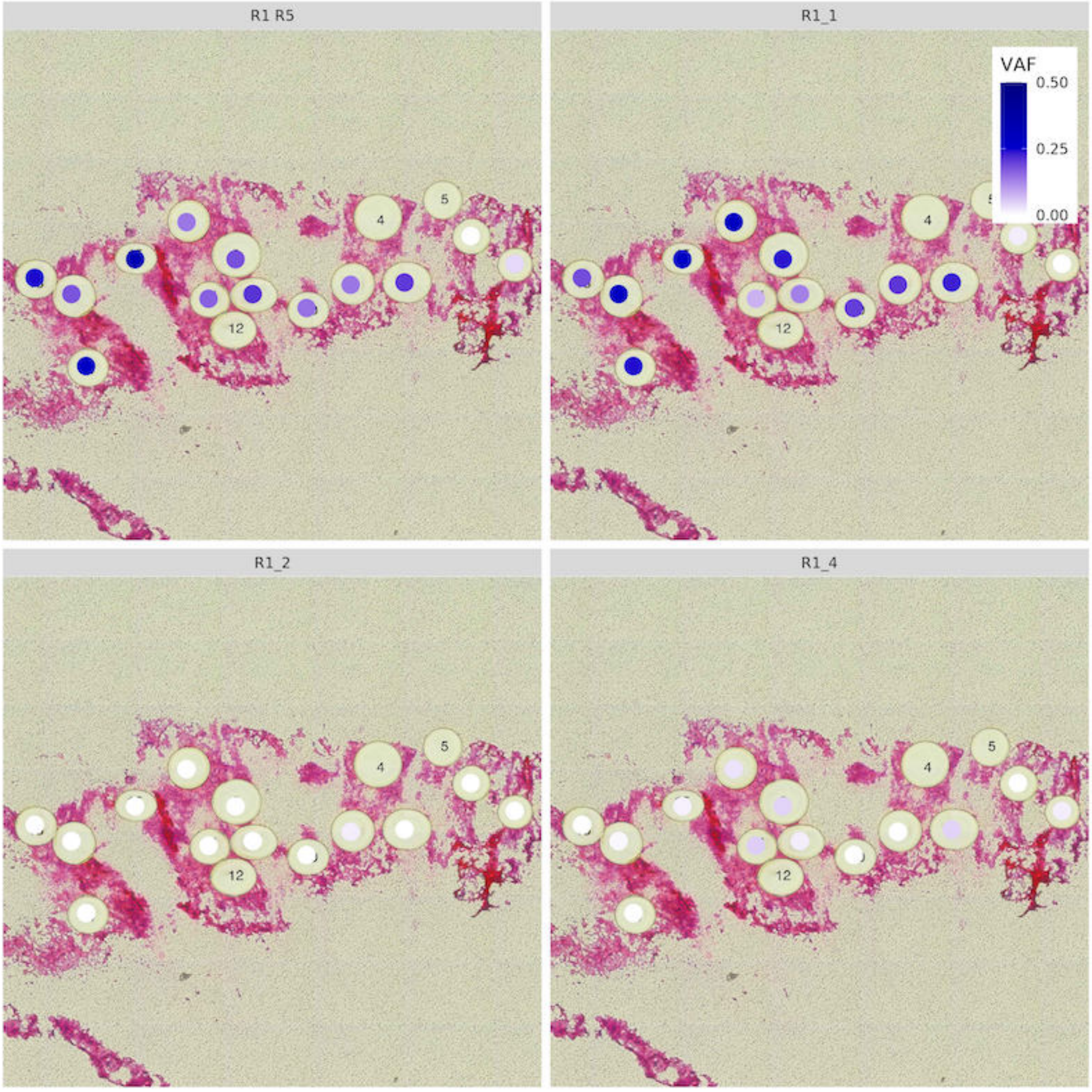
Note that spots from the same side or section do not always cluster together and are scattered across the subclone they are most similar to.

8.1.13 Additional LCM spatial trees



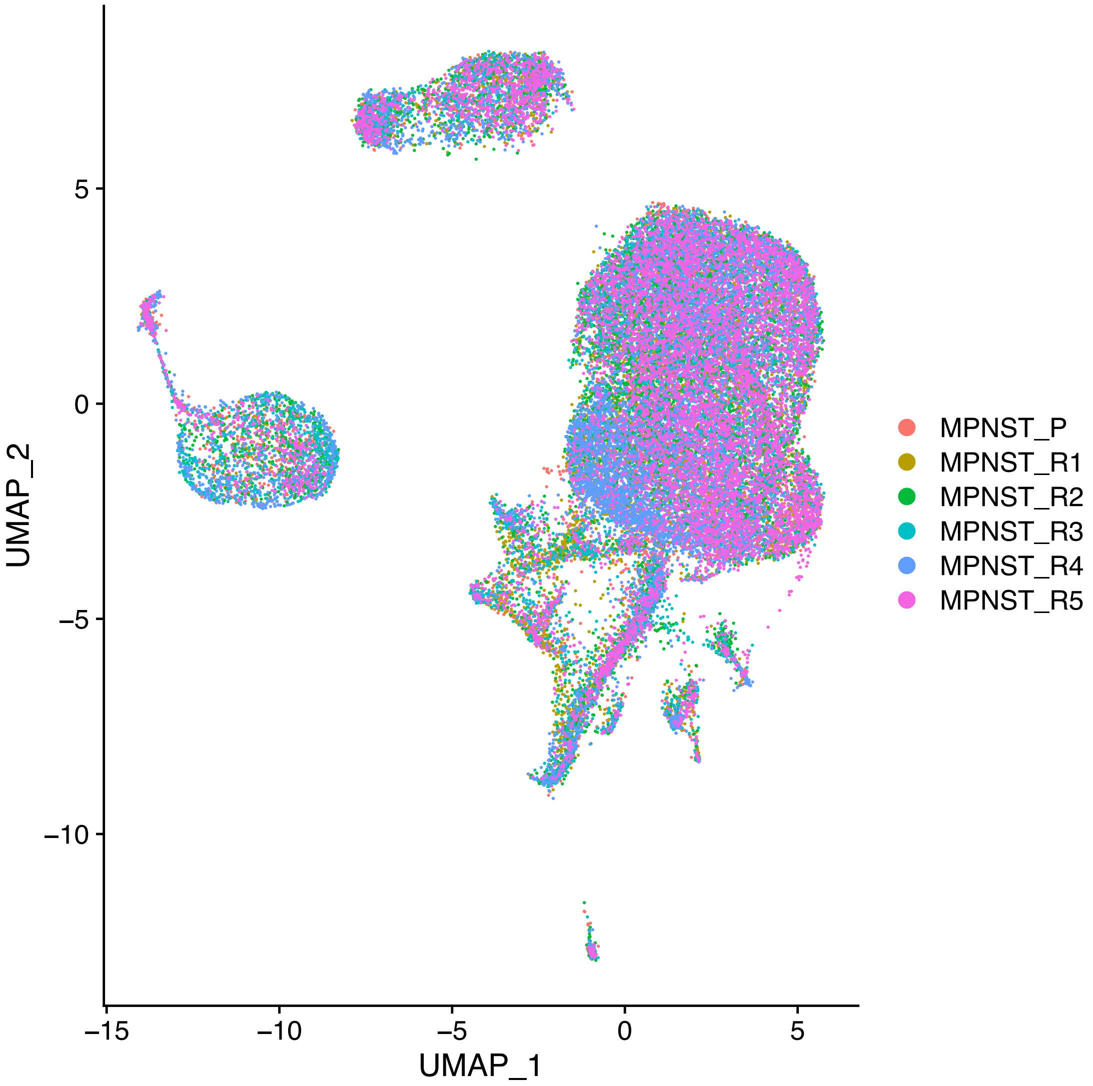
Phylogenetic trees overlaid on LCM images of the Primary “Front” (upper) and R1 “Side” (lower) slides.

8.1.14 Genotyping SNVs in LCM spots



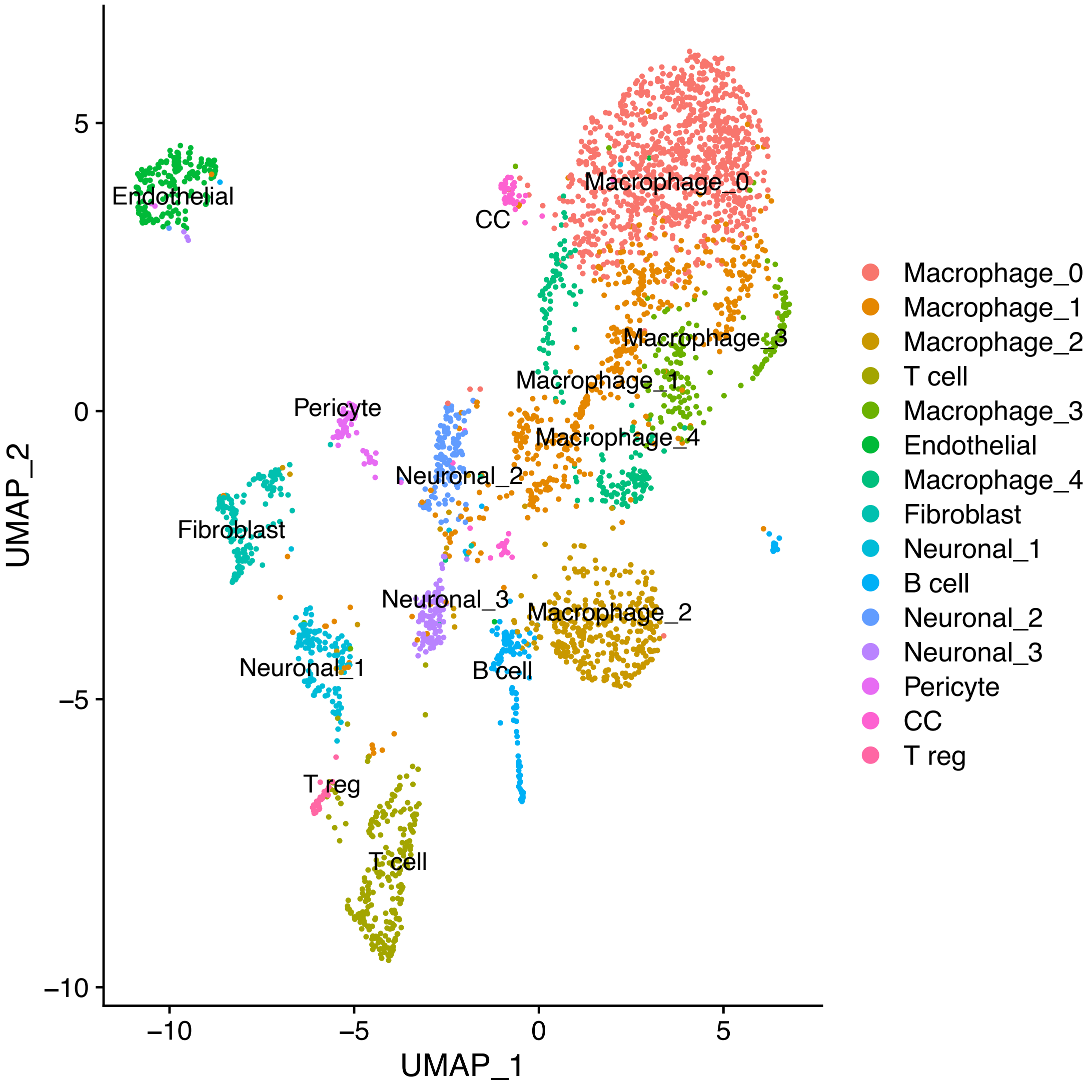
Variant allele frequency (VAF) shown for subclonal clusters of SNVs from R1 for each spot from the R1_Side section. Note two spots (which displayed diploid copy number profiles) on the far right of the section do not bear any SNVs.

8.1.15 Batch correction removes transcriptomic effects of CNAs



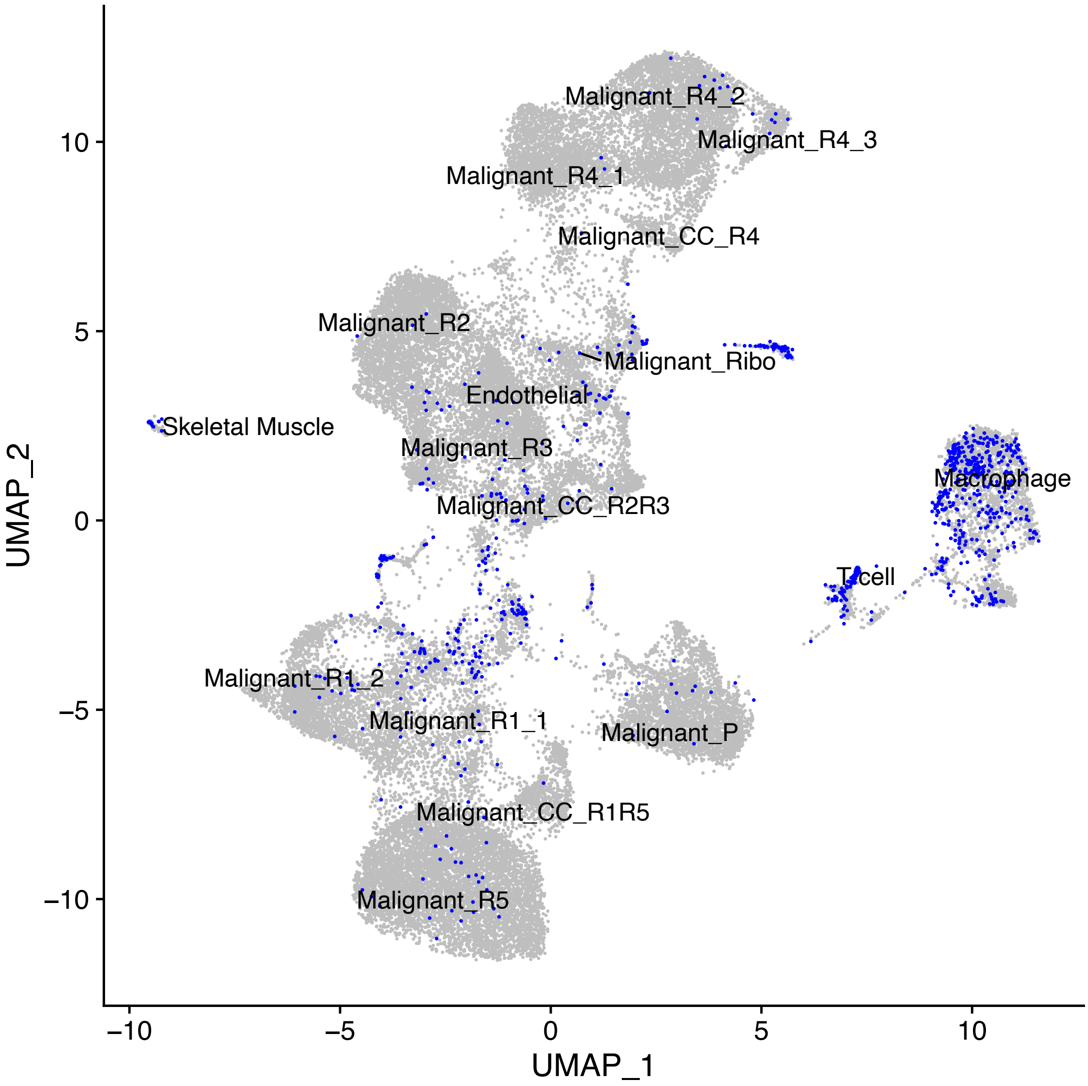
Cells from all tumour regions overlap and the biological effects of CNAs on gene expression are overcorrected as technical noise.

8.1.16 Identifying immune cell types through sub-clustering.



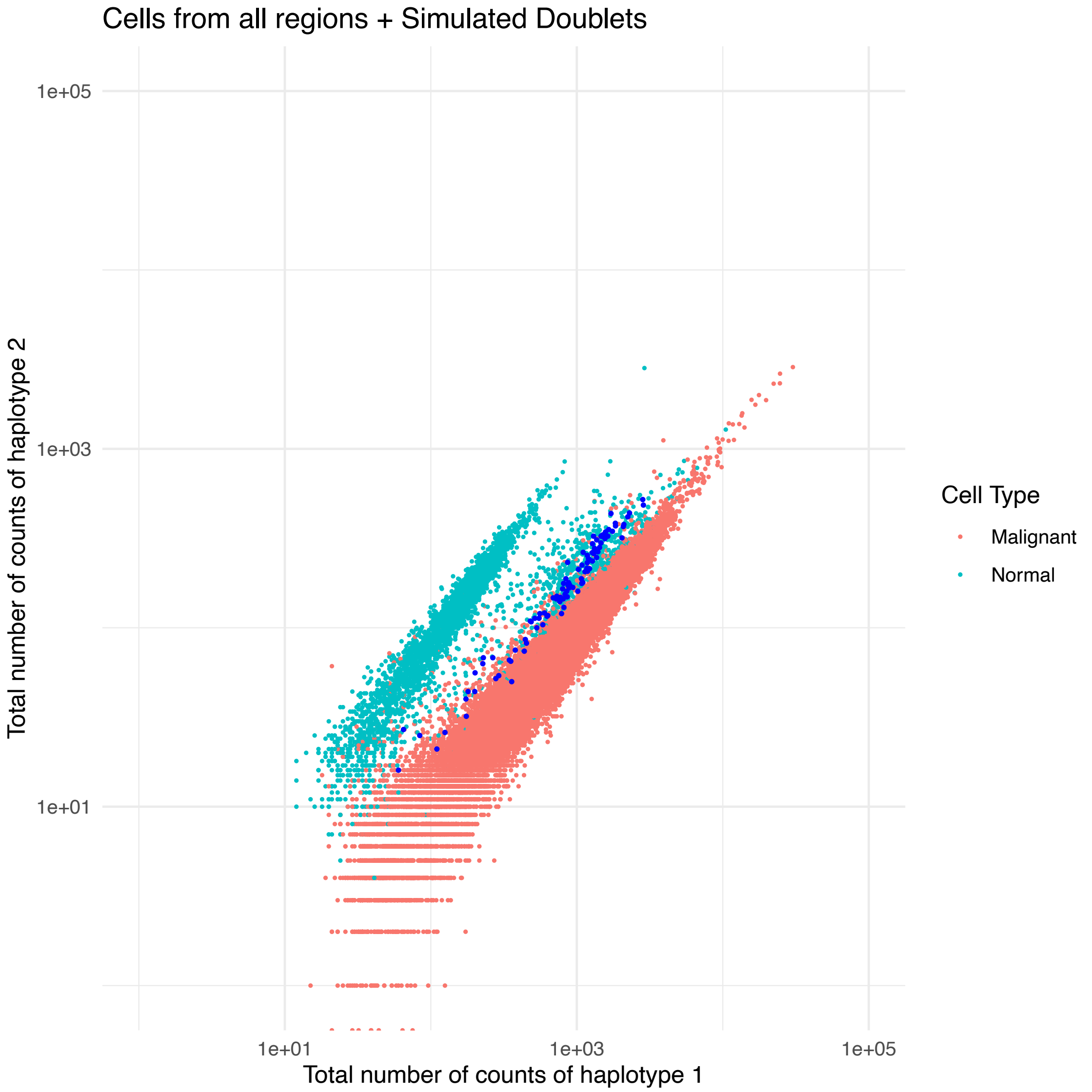
Cycling cells are annotated as CC. Several clusters originally annotated as macrophages expressing markers of neuronal cells were also detected.

8.1.17 Gene expression clustering of putative doublets



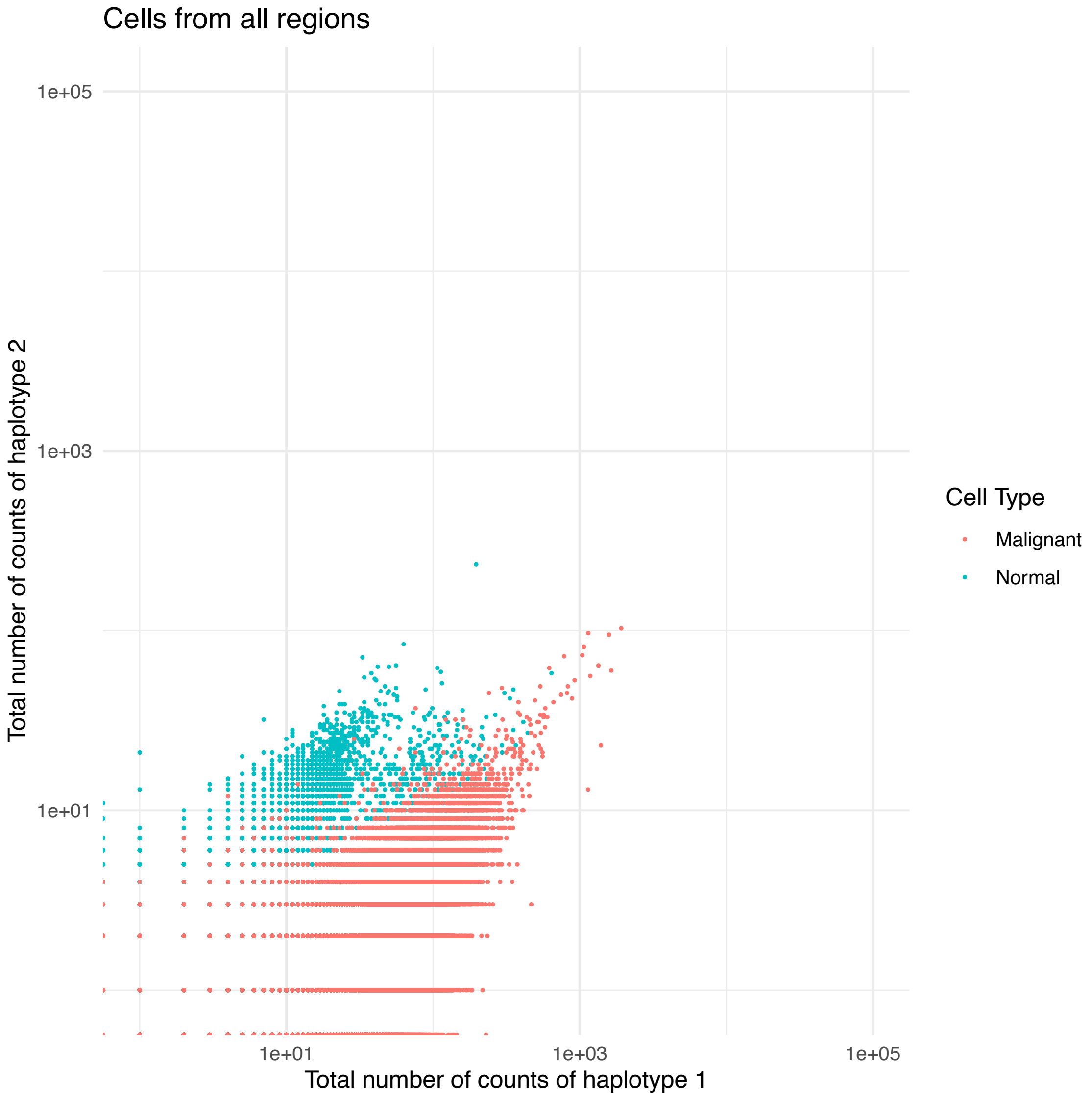
UMAP projection of scRNA-seq cells with suspected doublets are shown in navy.

8.1.18 Haplotype counts of simulated doublets



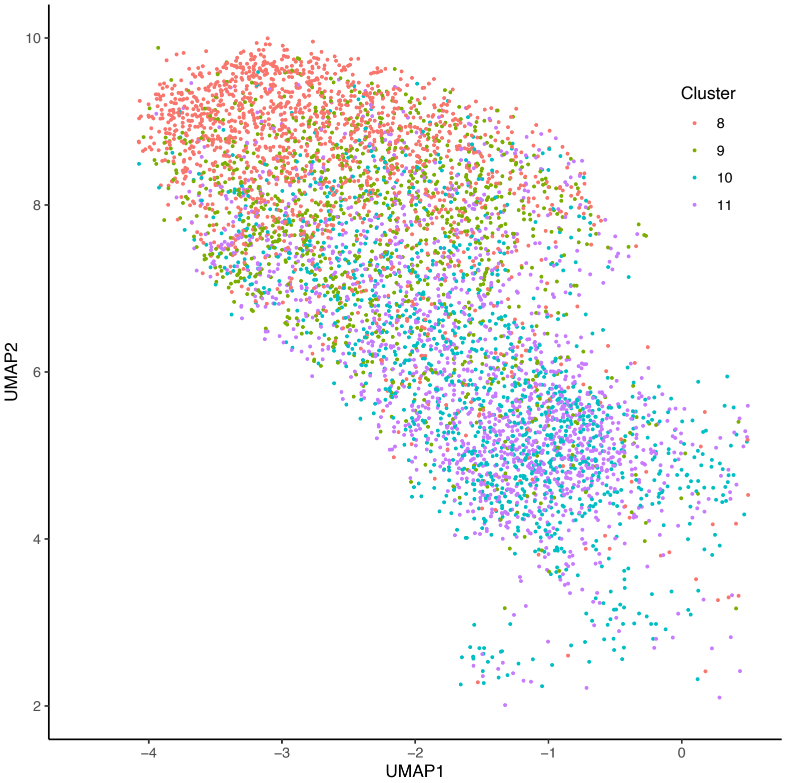
Simulated doublets, shown in dark blue, have a similar haplotype count ratio to those suspected to be doublets from the scRNA-seq data.

8.1.19 Classification of cells using SNPs from chromosome 3



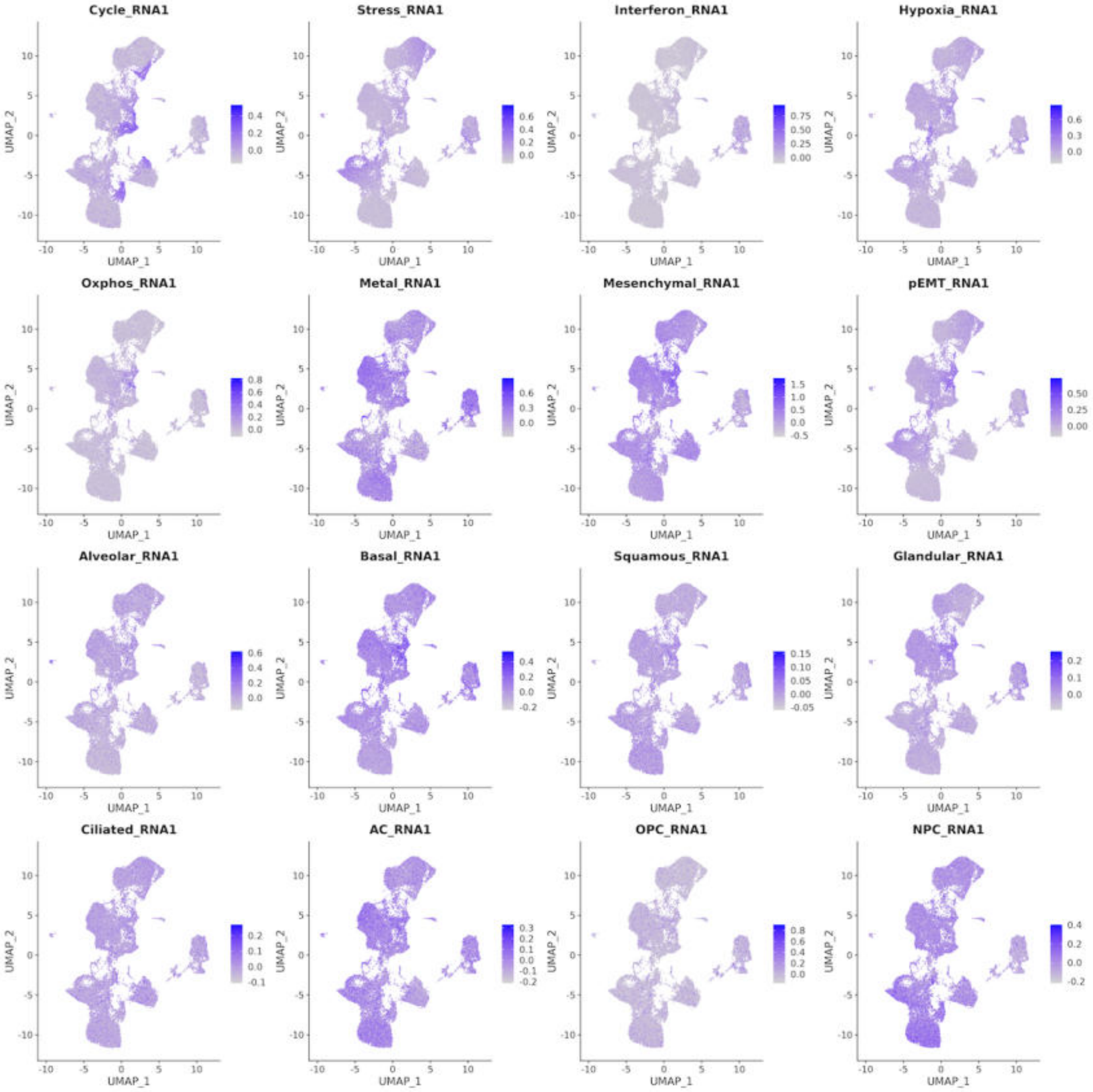
Cell types can be accurately annotated based on SNPs from only one chromosome.

8.1.20 GRN-based UMAP projection of 4 inferCNV clusters from R4



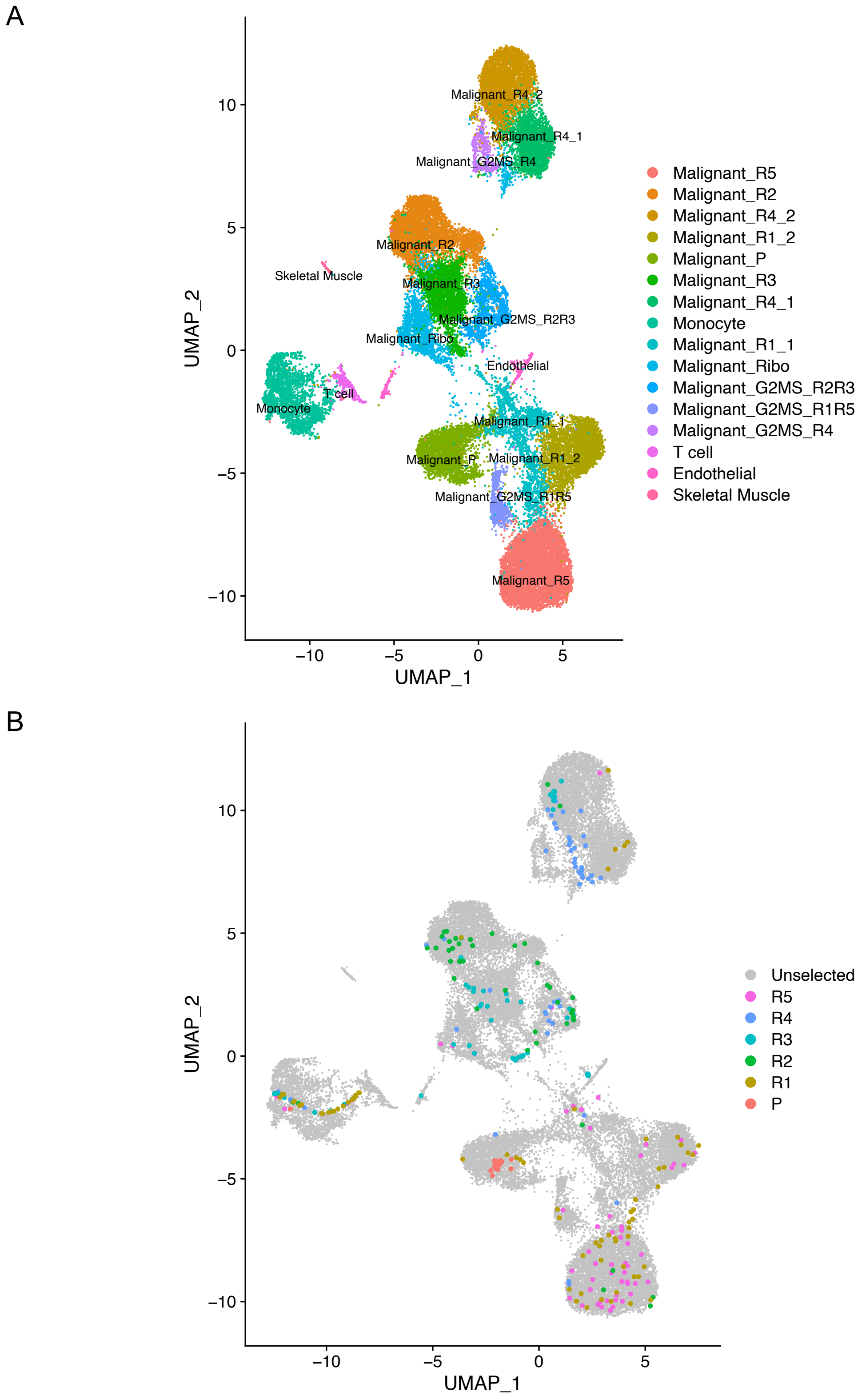
As in the gene expression-based UMAP, cells cluster together by inferCNV clusters.

8.1.21 Cancer cell states in scRNA-seq



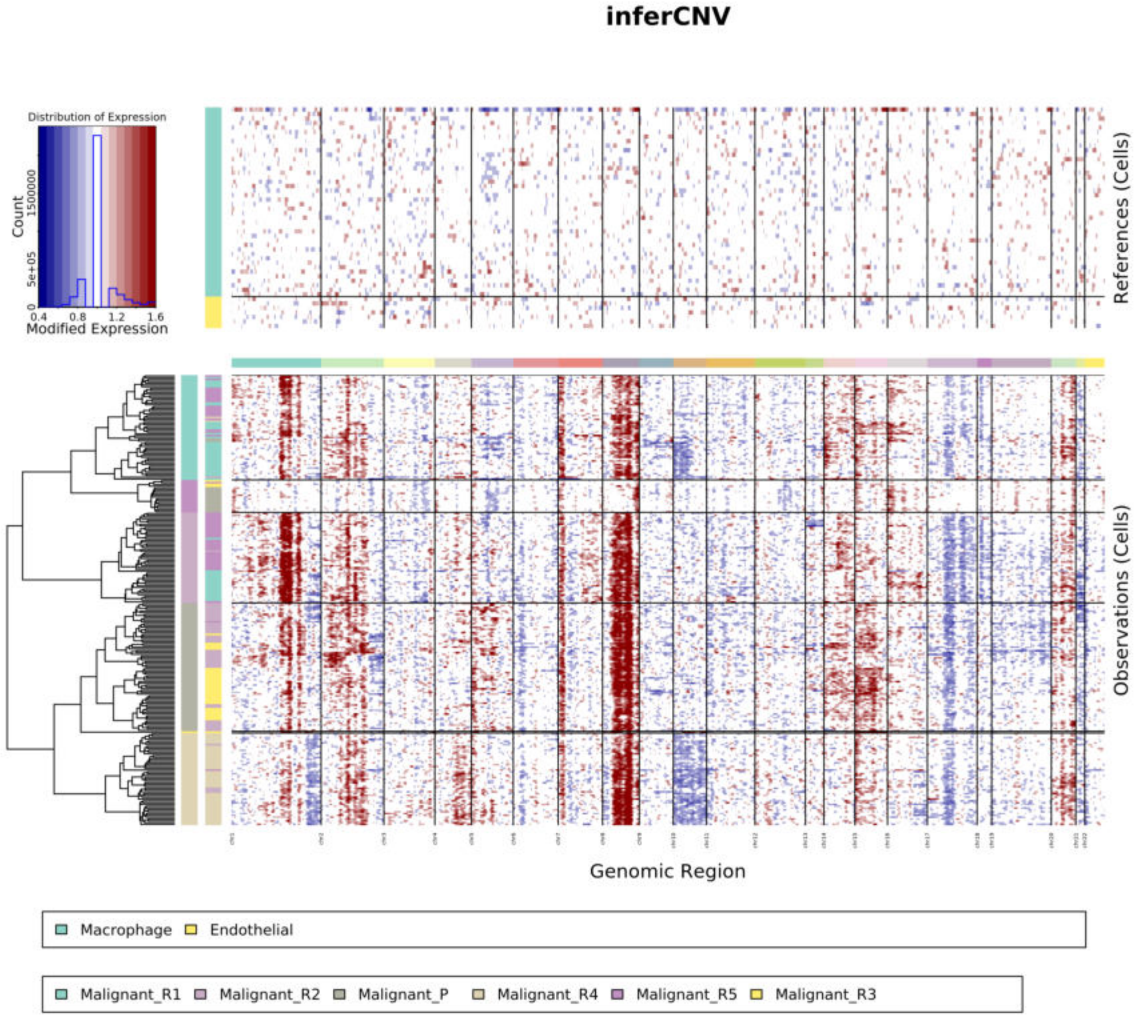
Activity of all 16 cell states in cells profiled with scRNA-seq.

8.1.22 UMAP projection of integrated G&T-seq and scRNA-seq cells

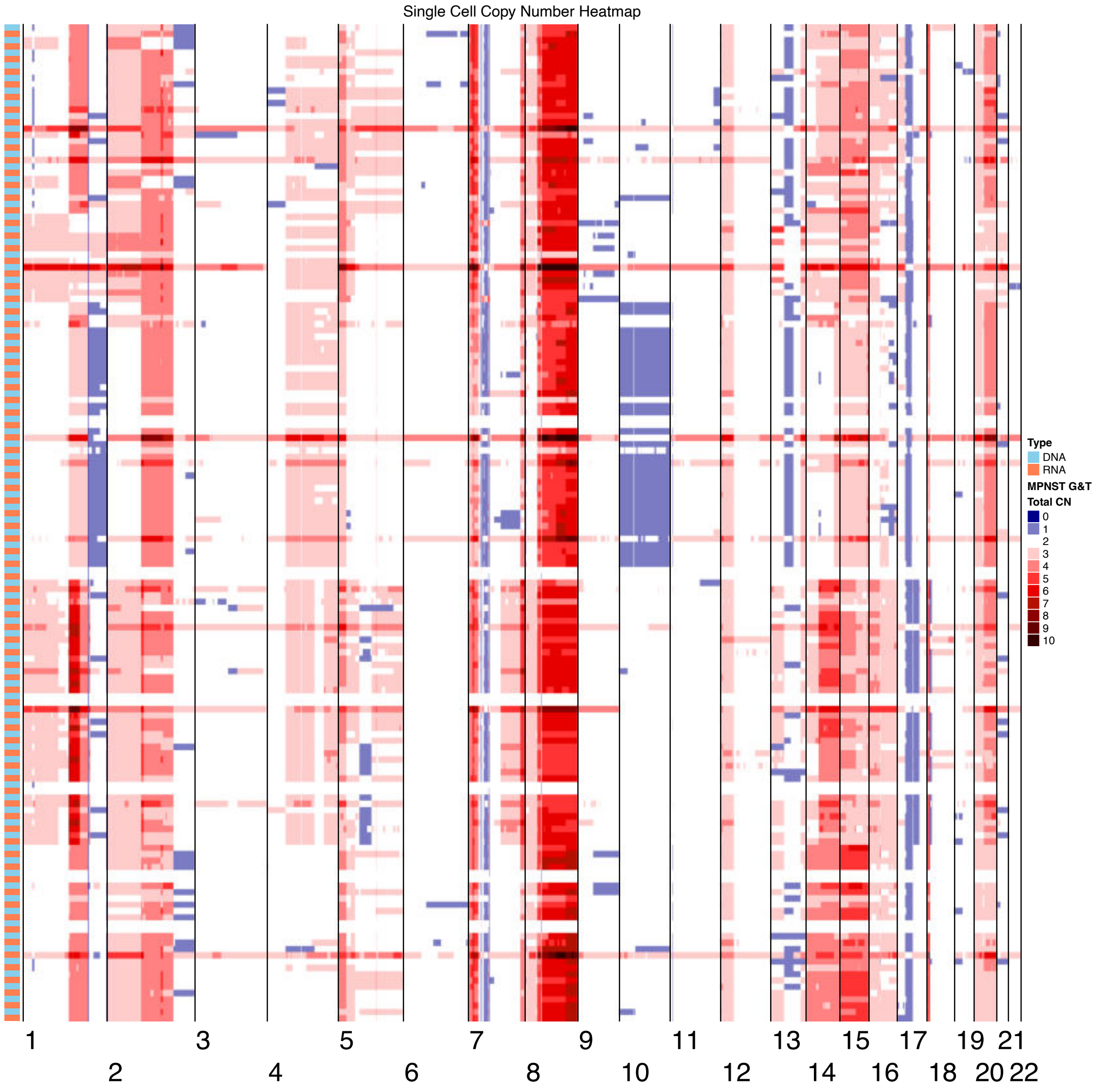


UMAP projection of clusters of cells identified in scRNA-seq shown in panel A and G&T-seq cells highlighted by their region of origin shown in panel B.

8.1.23 InferCNV profiles of G&T-seq cells

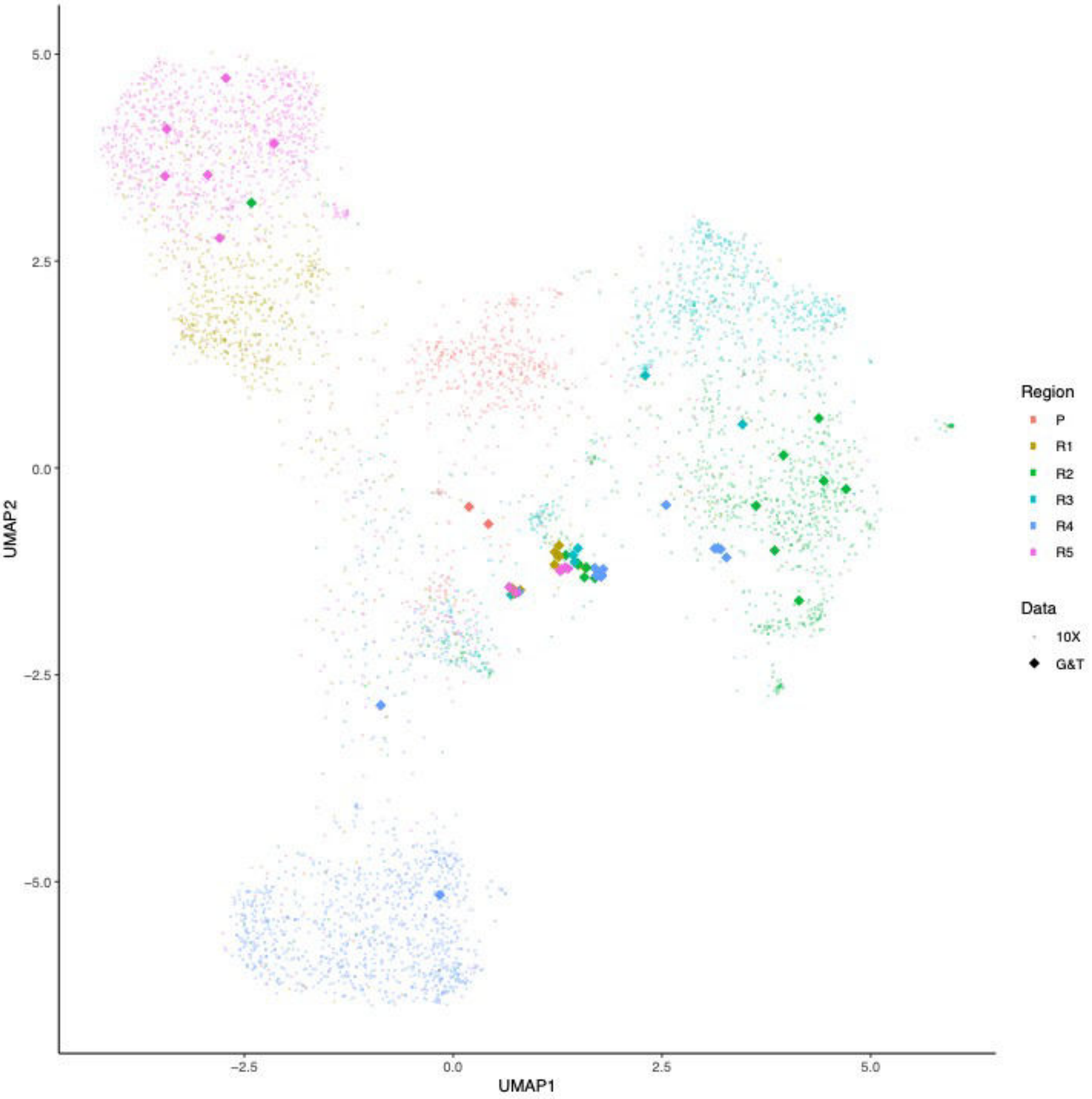


8.1.24 DNA and RNA derived total copy number profiles



Total copy number profiles called from DNA and batch corrected from RNA from the same cell shown on alternate rows. This allows for visual inspection of differences between batch-corrected copy number states and ground-truth copy number states for genomic segments.

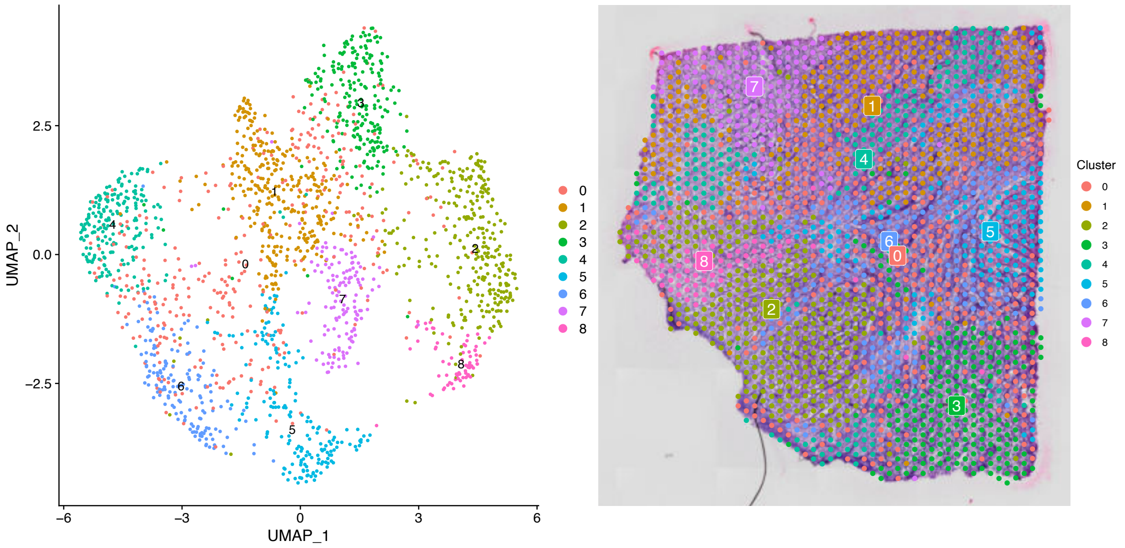
8.1.25 UMAP projection of G&T-seq anchors and scRNA-seq cells



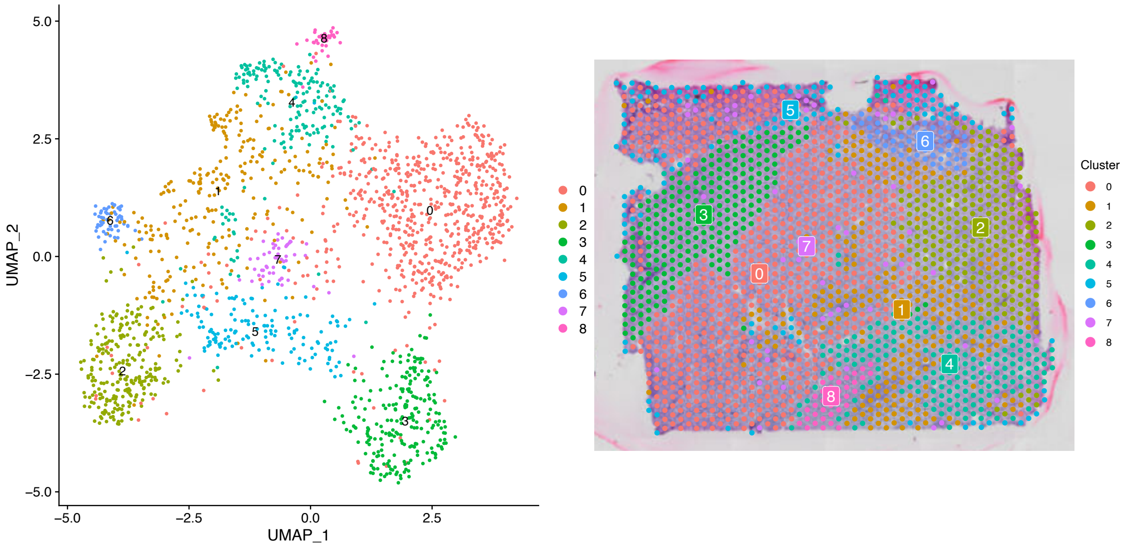
Note that many G&T-seq anchors from R4 are placed near scRNA-seq cells from R2 due to being sampled from a different subclone. Furthermore, there is an absence of G&T-seq anchors for many areas of the UMAP occupied by scRNA-seq cells.

8.1.26 Additional spatial transcriptomic slides

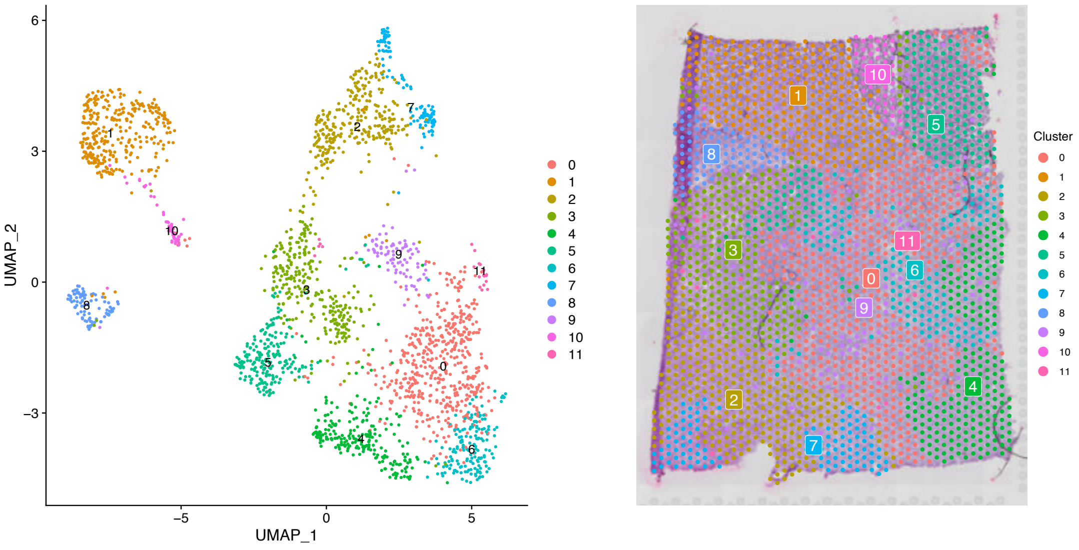
Primary_B



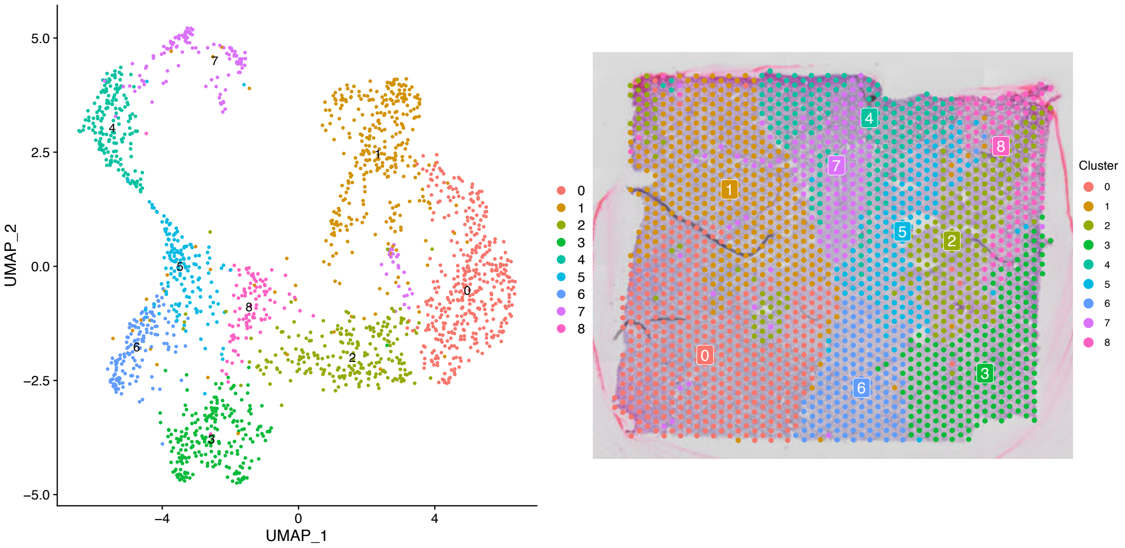
R2



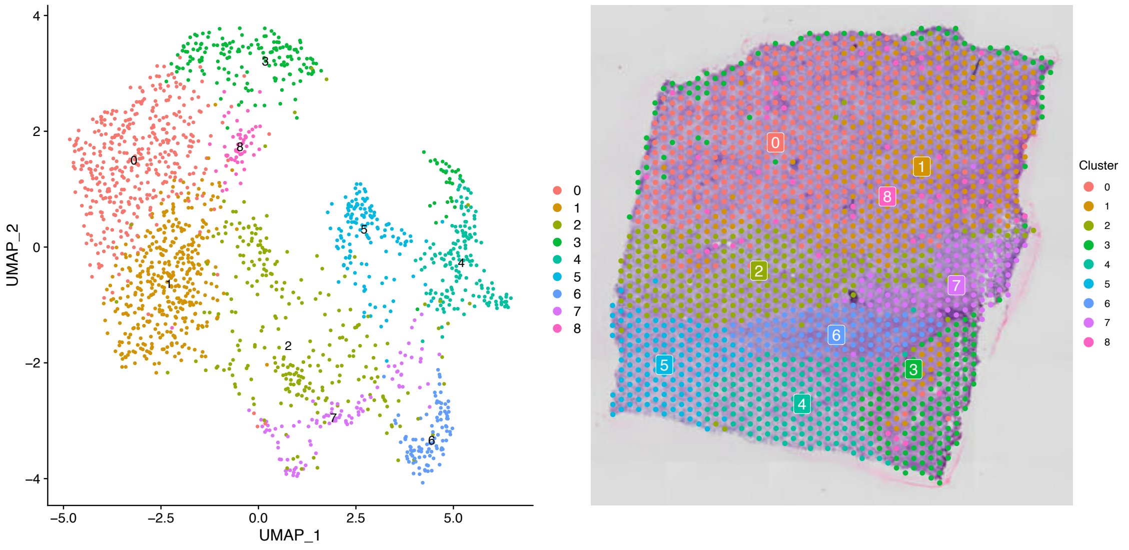
R3


6
3
0
−3
UMAP_2
−5
0
5
UMAP_1
0
1
2
3
4
5
6
7
8
9
10
11
Cluster

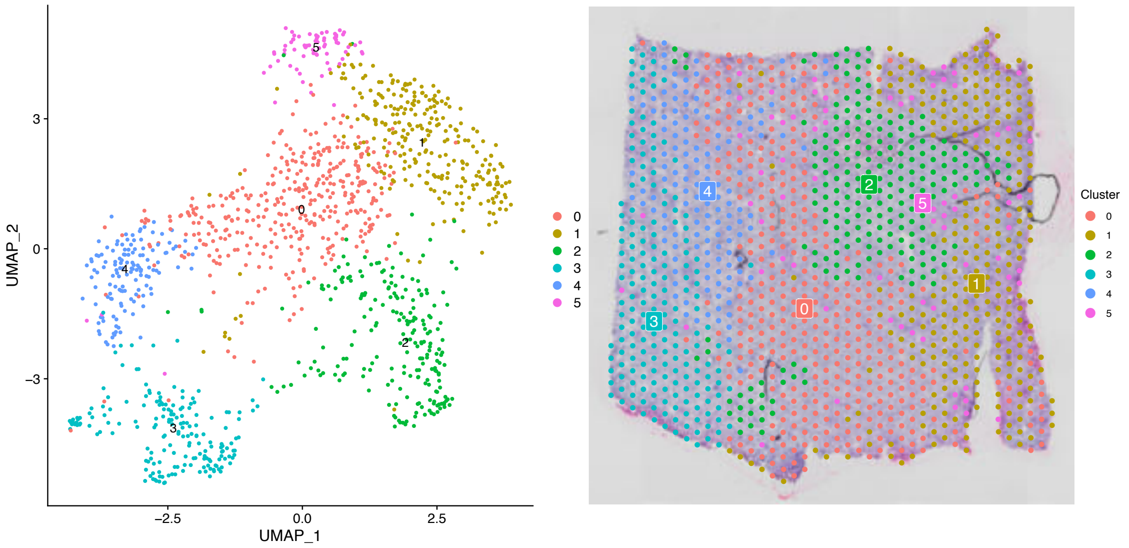
R4


5.0
2.5
0.0
−2.5
−5.0
UMAP_2
−4
0
4
UMAP_1
0
1
2
3
4
5
6
7
8
Cluster

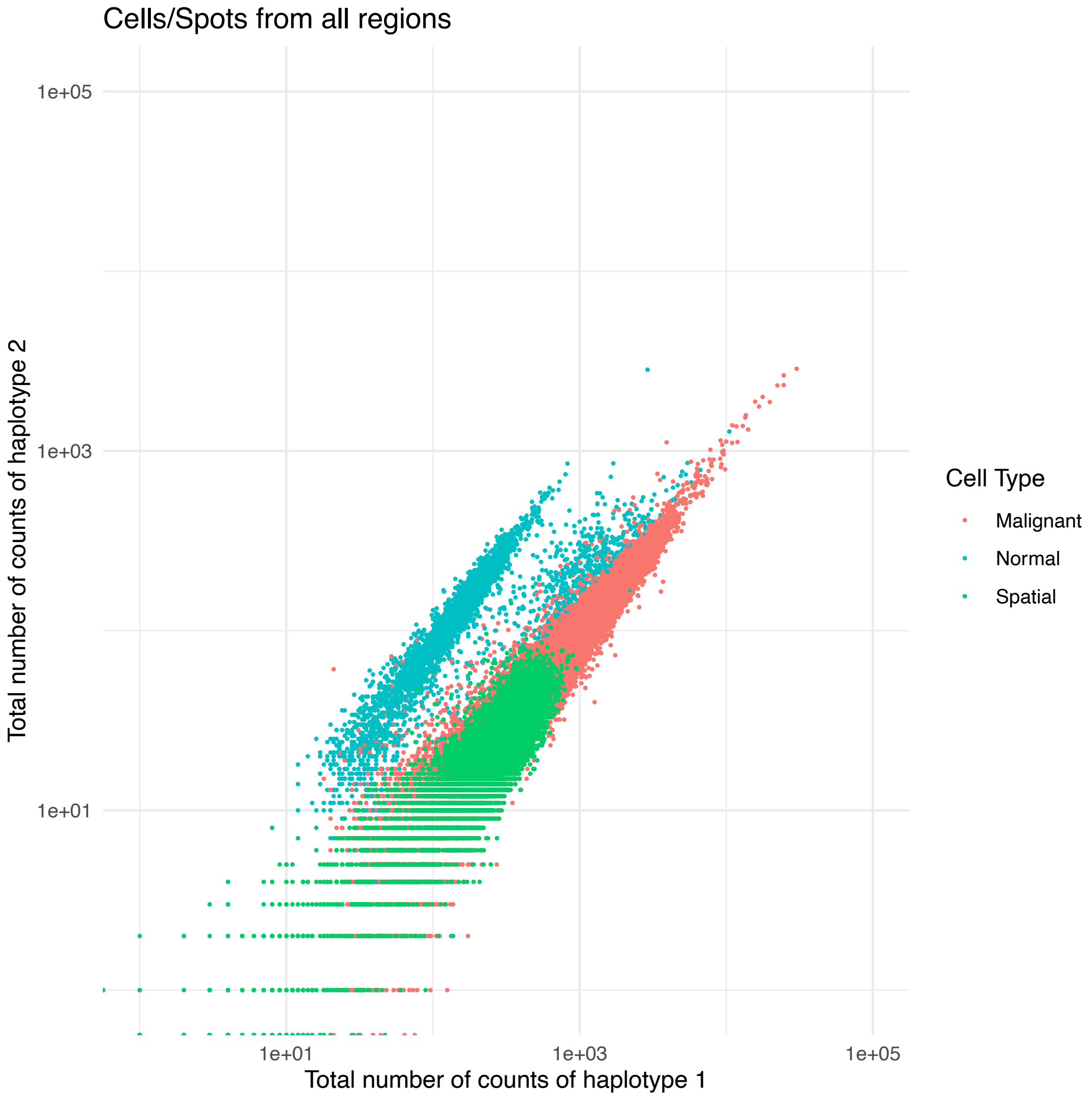
R5_A


UMAP_2
4
2
0
−2
−4
−5.0
−2.5
0.0
2.5
5.0
UMAP_1
0
1
2
3
4
5
6
7
8
Cluster
0
1
2
3
4
5
6
7
8

R5_B

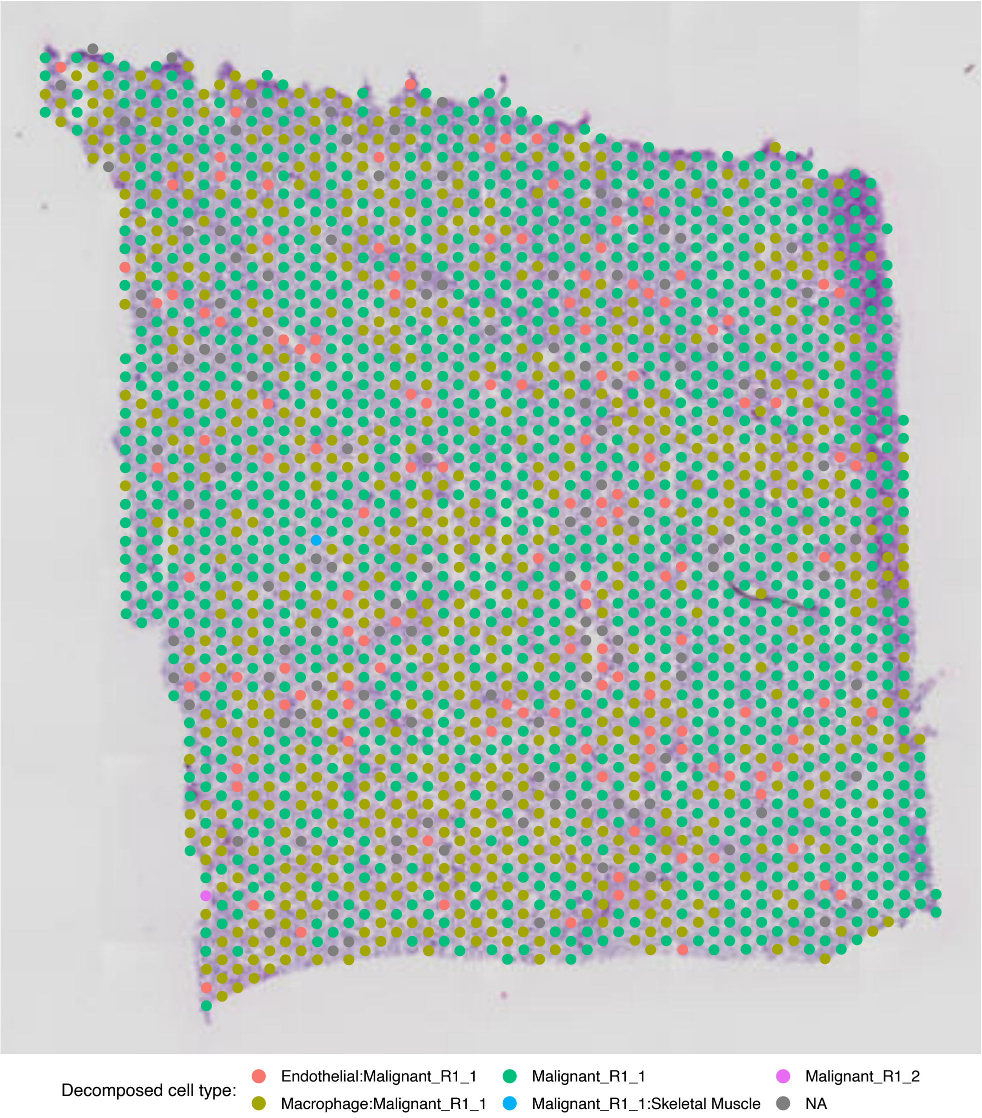

UMAP_2
3
0
−3
−2.5
0.0
2.5
UMAP_1
0
1
2
3
4
5
Cluster
0
1
2
3
4
5

8.1.27 Haplotype counts for spatial transcriptomic slides

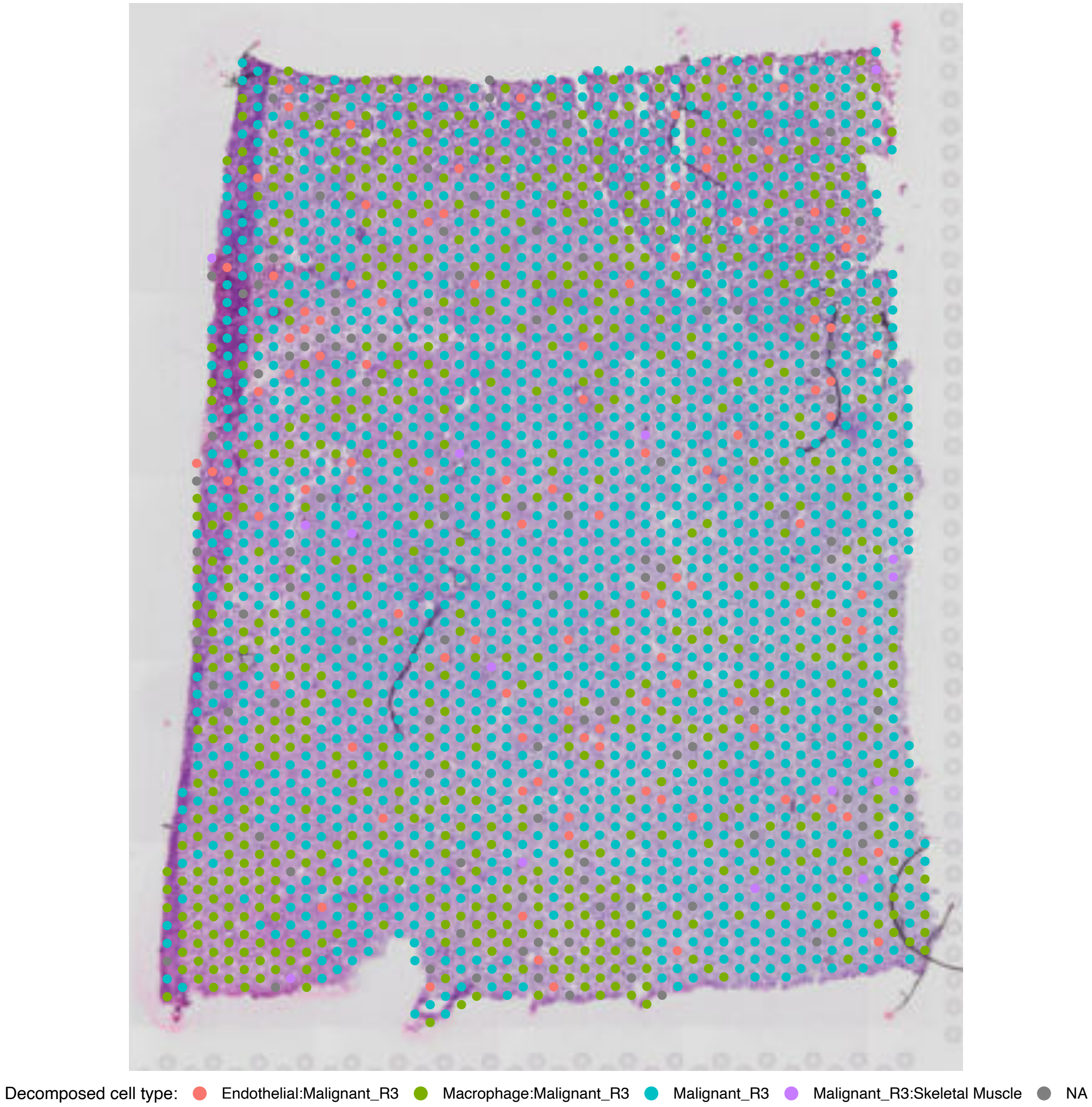


Counts of the two haplotypes of spots from all eight Visium slides and scRNA-seq cells from all regions.

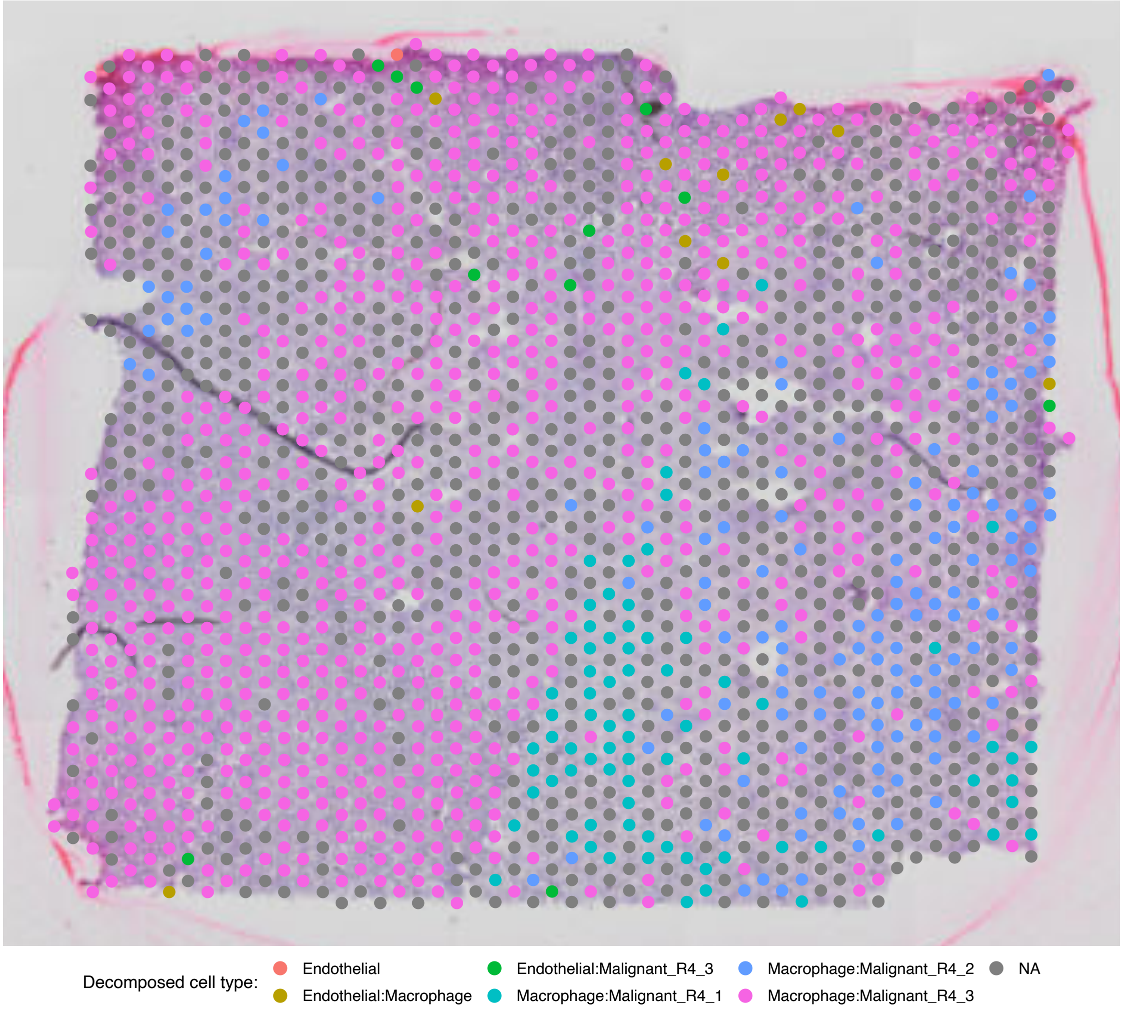
8.1.28 Selected additional deconvolved spatial transcriptomic slides


Decomposed cell type:
● Endothelial:Malignant_R1_1
● Macrophage:Malignant_R1_1
● Malignant_R1_1
● Malignant_R1_1:Skeletal Muscle
● Malignant_R1_2
● NA

Deconvolved cell types of R1.

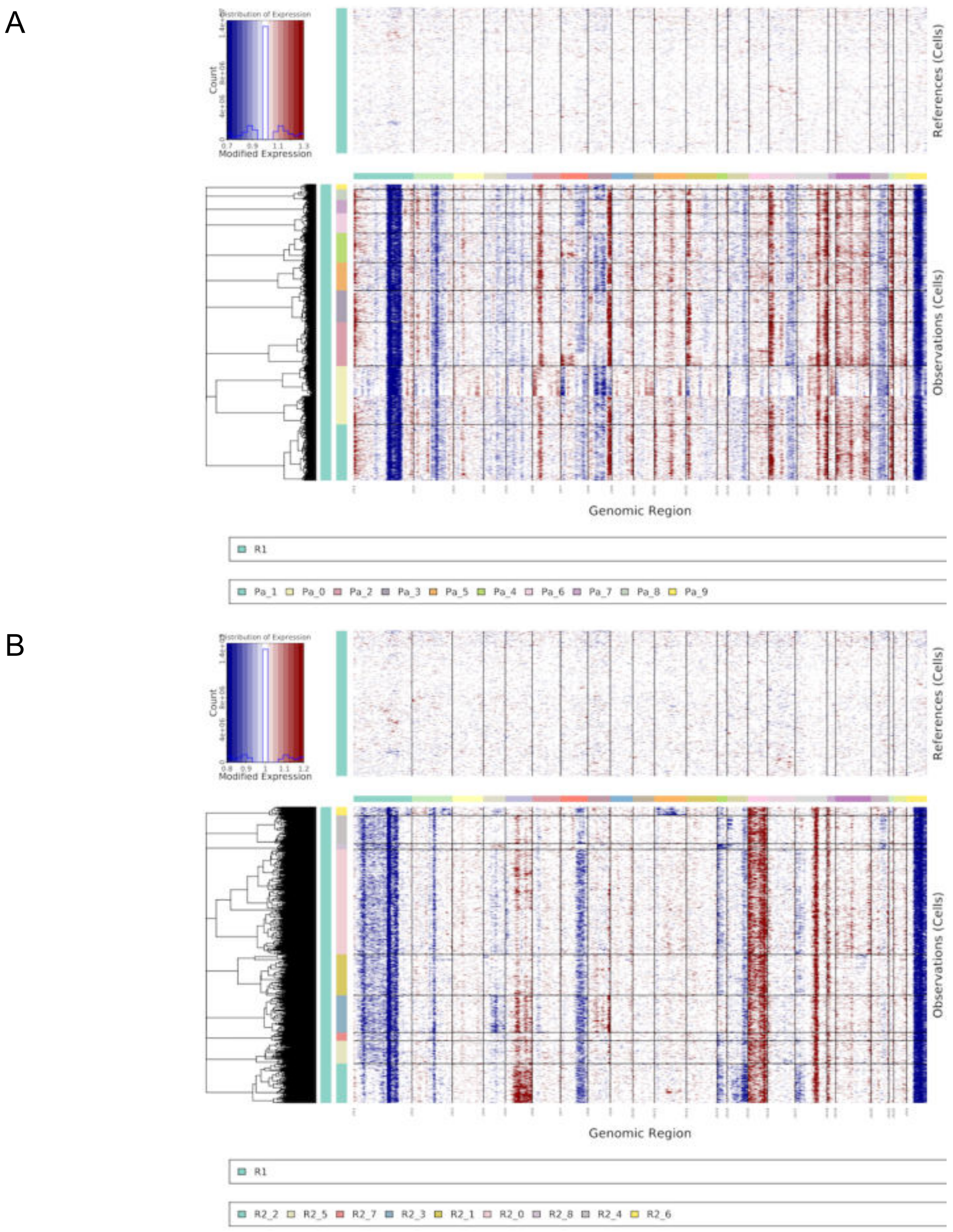

Decomposed cell type:
● Endothelial:Malignant_R3
● Macrophage:Malignant_R3
● Malignant_R3
● Malignant_R3:Skeletal Muscle
● NA

Deconvolved cell types of R3.


Decomposed cell type:
Endothelial
Endothelial:Macrophage
Endothelial:Malignant_R4_3
Macrophage:Malignant_R4_1
Macrophage:Malignant_R4_2
Macrophage:Malignant_R4_3
NA

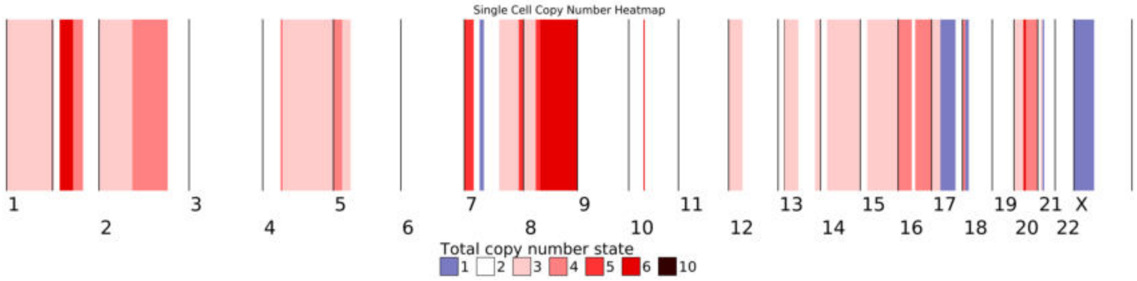
Deconvolved cell types of R4.

8.1.29 Additional relative copy number differences in Visium slides

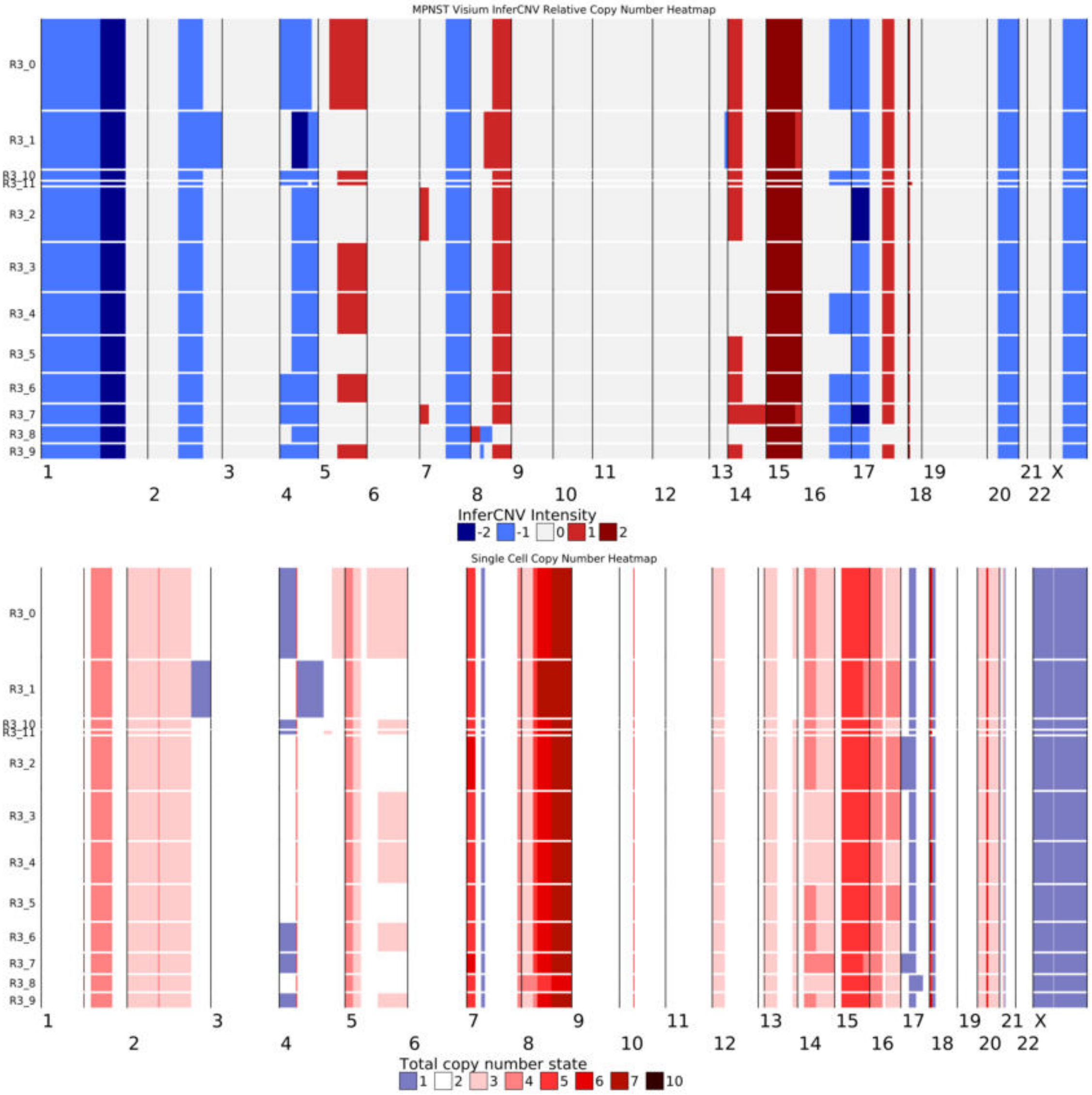


Relative copy number changes detected by inferCNV for the Primary_A and R2 slide compared to R1 shown in panels A and B respectively. These relative changes (such as the presence of a loss of chr1p and gain in chr5 in R2 but not Primary_A) were used to identify which subclone was present on the R1 slide.

8.1.30 Consensus copy number profile of scDNA-seq R1 subclone

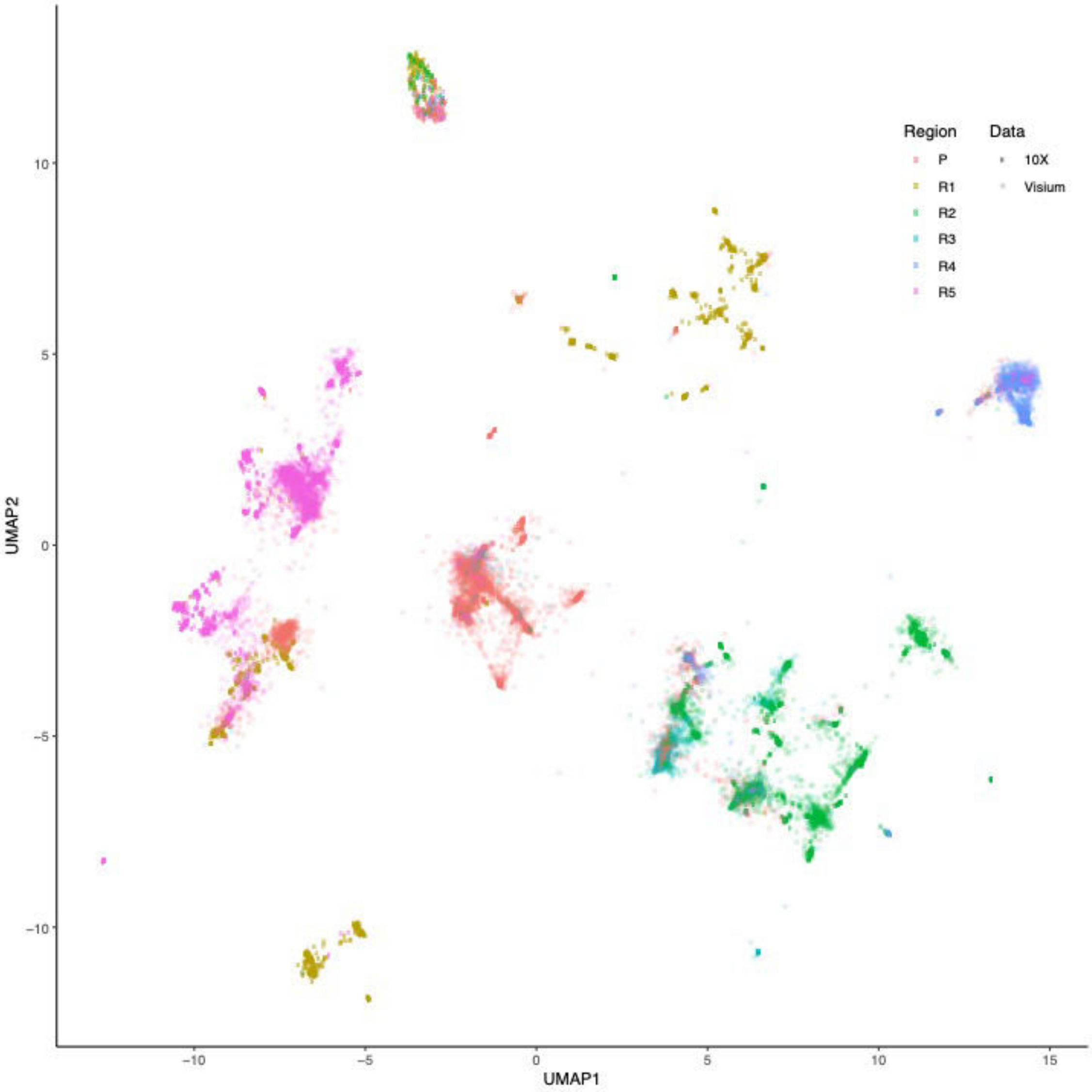


8.1.31 Deriving total copy number state from relative changes



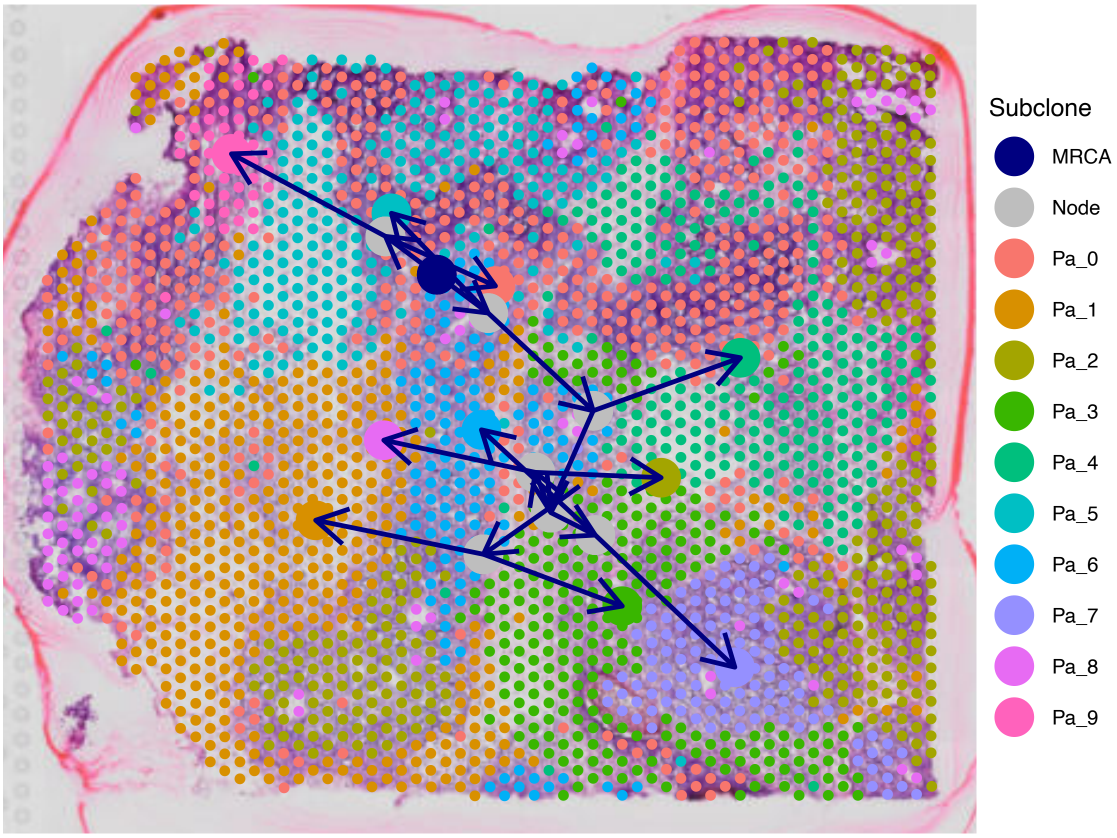
Example relative copy number gains/losses derived from inferCNV values (upper), and total copy number profile (lower) shown for R3.

8.1.32 UMAP projection of individual spatial transcriptomic spots

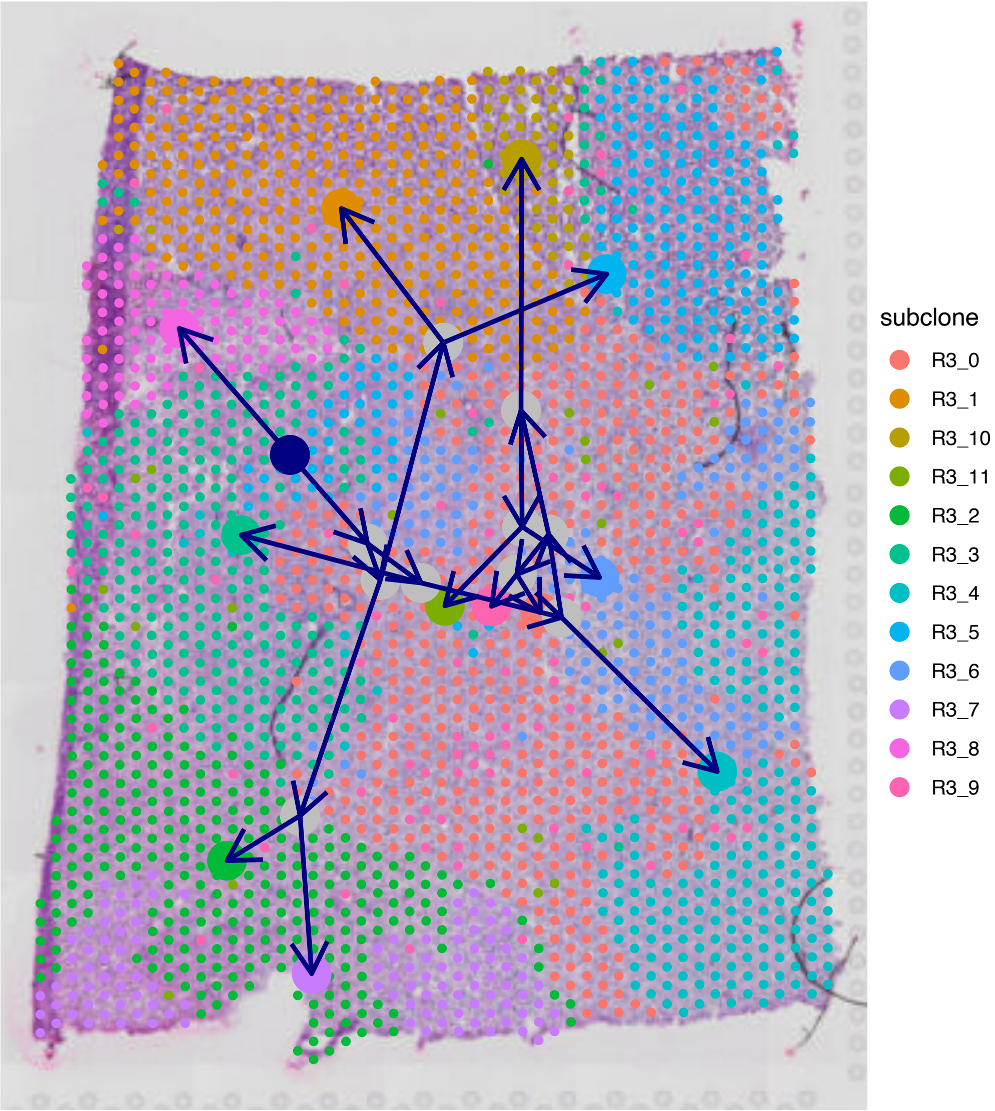


Some rare spots from the primary can be seen clustering with cells from the recurrence regions.

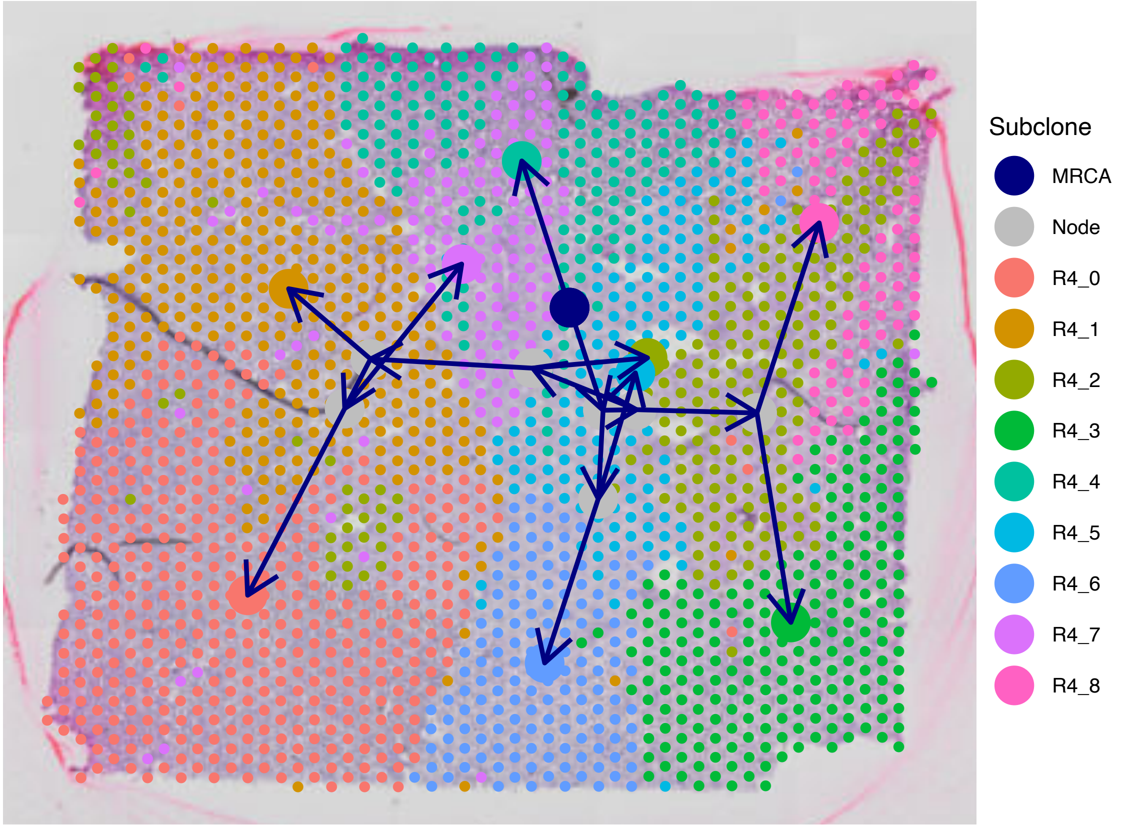
8.1.33 Selected additional spatial transcriptomic phylogenetic trees



Phylogenetic trees of Primary_A spatial transcriptomic subclones.

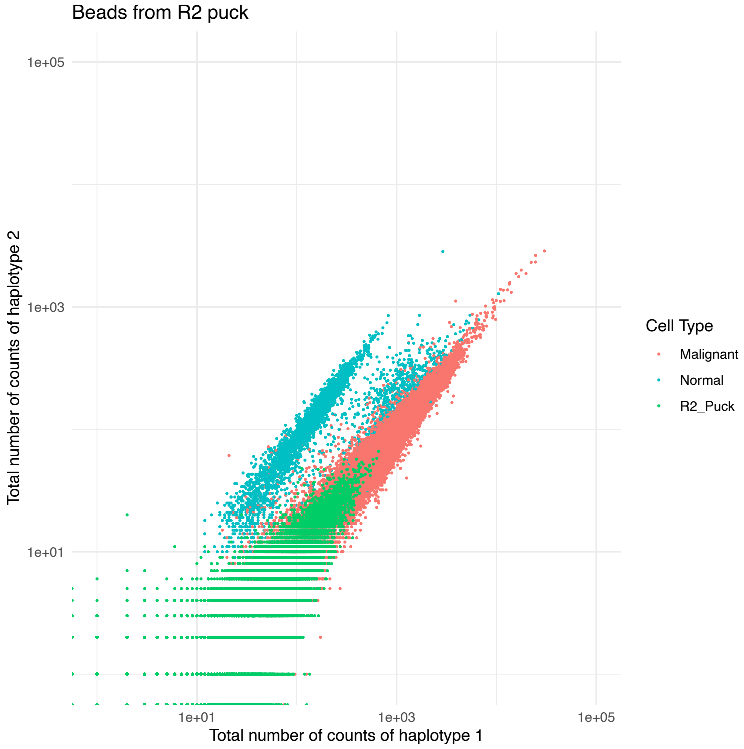


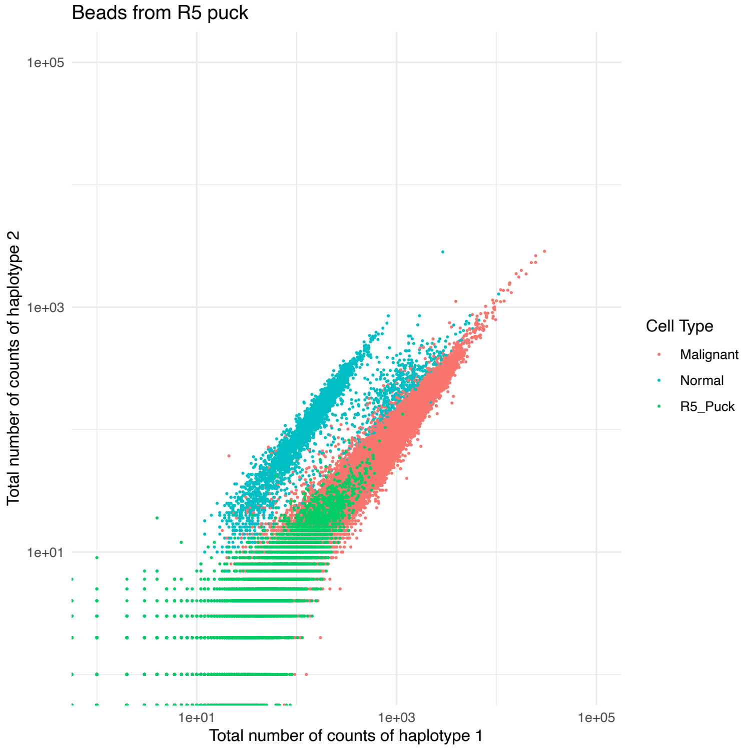
Phylogenetic trees of R3 spatial transcriptomic subclones.



Phylogenetic trees of R4 spatial transcriptomic subclones.

8.1.34 Haplotype counts for Slide-seq pucks

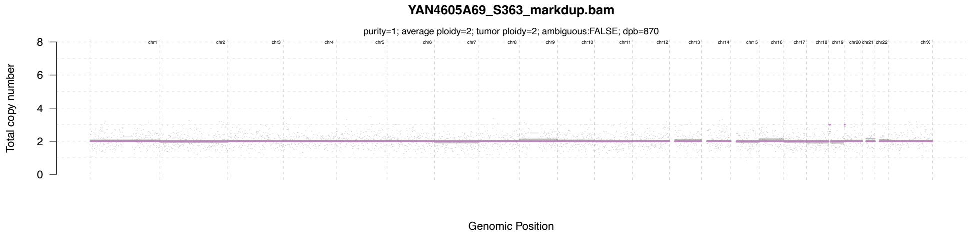


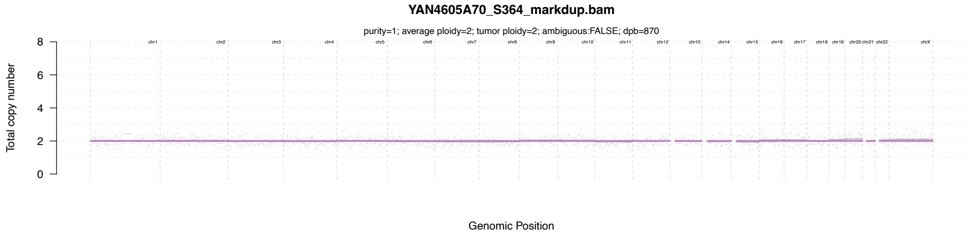


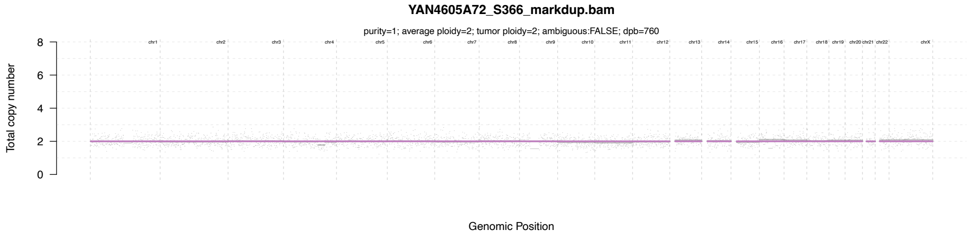
Counts of the two haplotypes of beads from Slide-seq pucks of R2 and R5 and scRNA-seq cells from all regions.

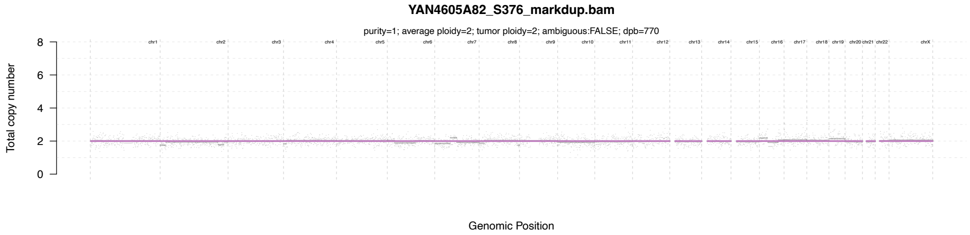
8.2 PEACE Appendix

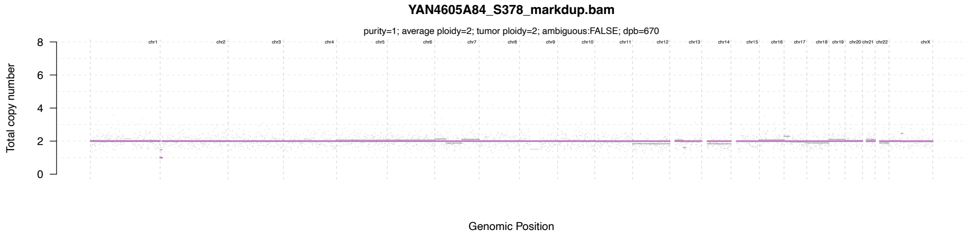
8.2.1 Example copy number profiles of CAIX$^+$ cells from normal lung

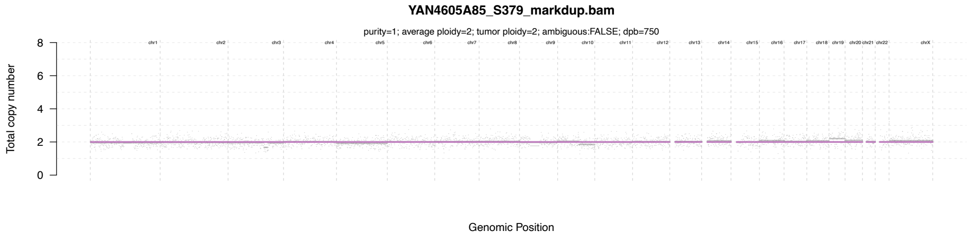






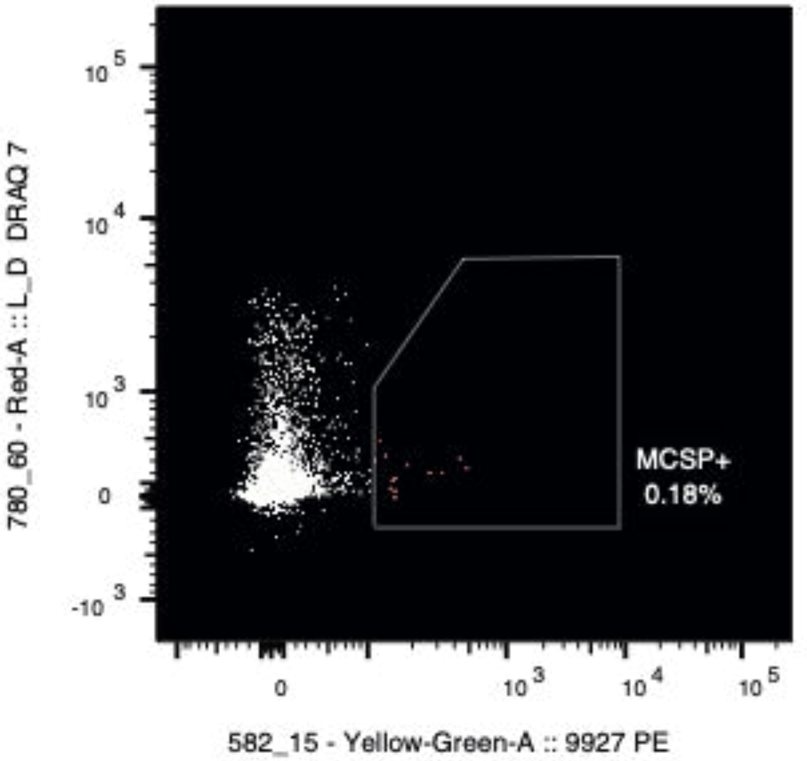


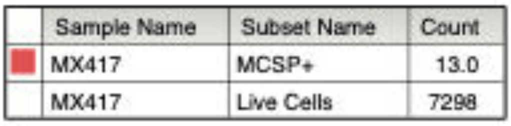




Cells YAN4605A69, 70, and 72 are from PEA142 and 82, 84, and 85 from PEA293.

8.2.2 Flow cytometry of MCSP$^+$ cells from a melanoma metastasis

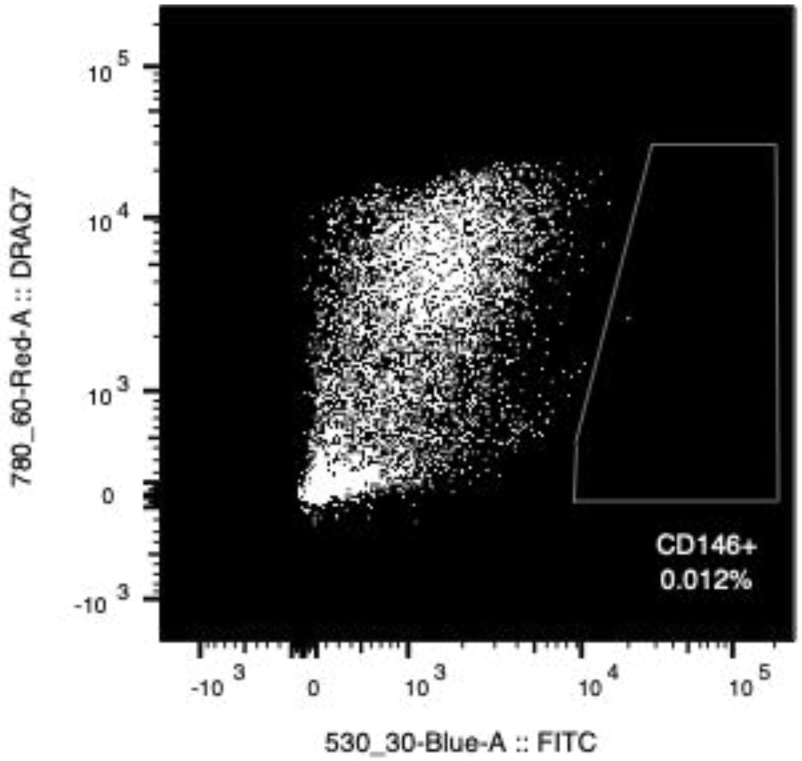


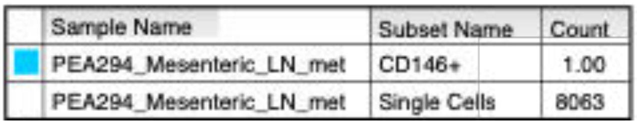


	Sample Name	Subset Name	Count
🟥	MX417	MCSP+	13.0
	MX417	Live Cells	7298

Flow cytometry of cells from a thigh metastasis from patient MX417.

8.2.3 Flow cytometry of CD146$^+$ cells from a melanoma metastasis

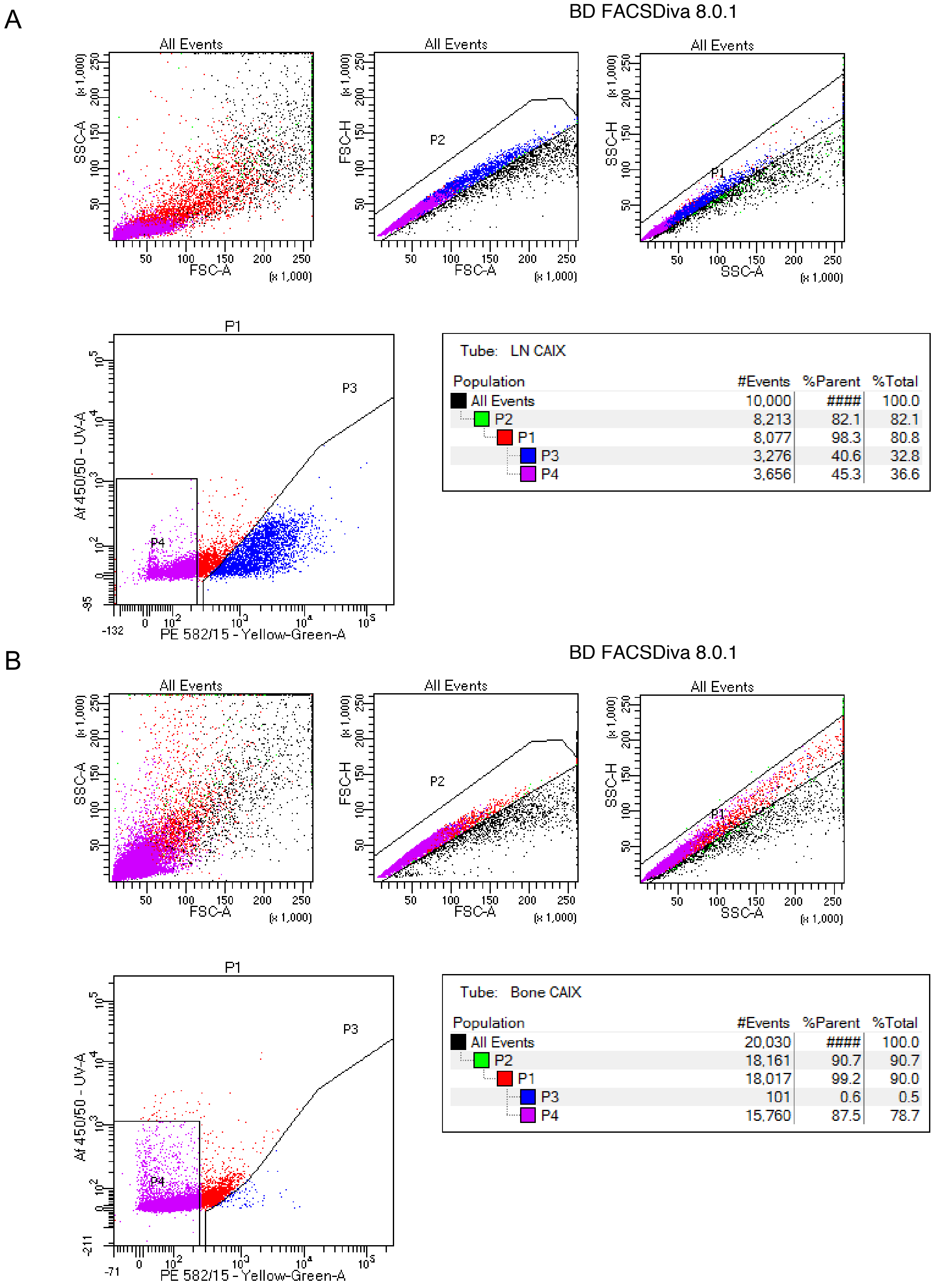




	Sample Name	Subset Name	Count
🟦	PEA294_Mesenteric_LN_met	CD146+	1.00
	PEA294_Mesenteric_LN_met	Single Cells	8063

Flow cytometry of cells from a lymph node metastasis from PEA294.

8.2.4 Sorting of CAIX$^+$ cells for scRNA-seq



Tube: LN CAIX

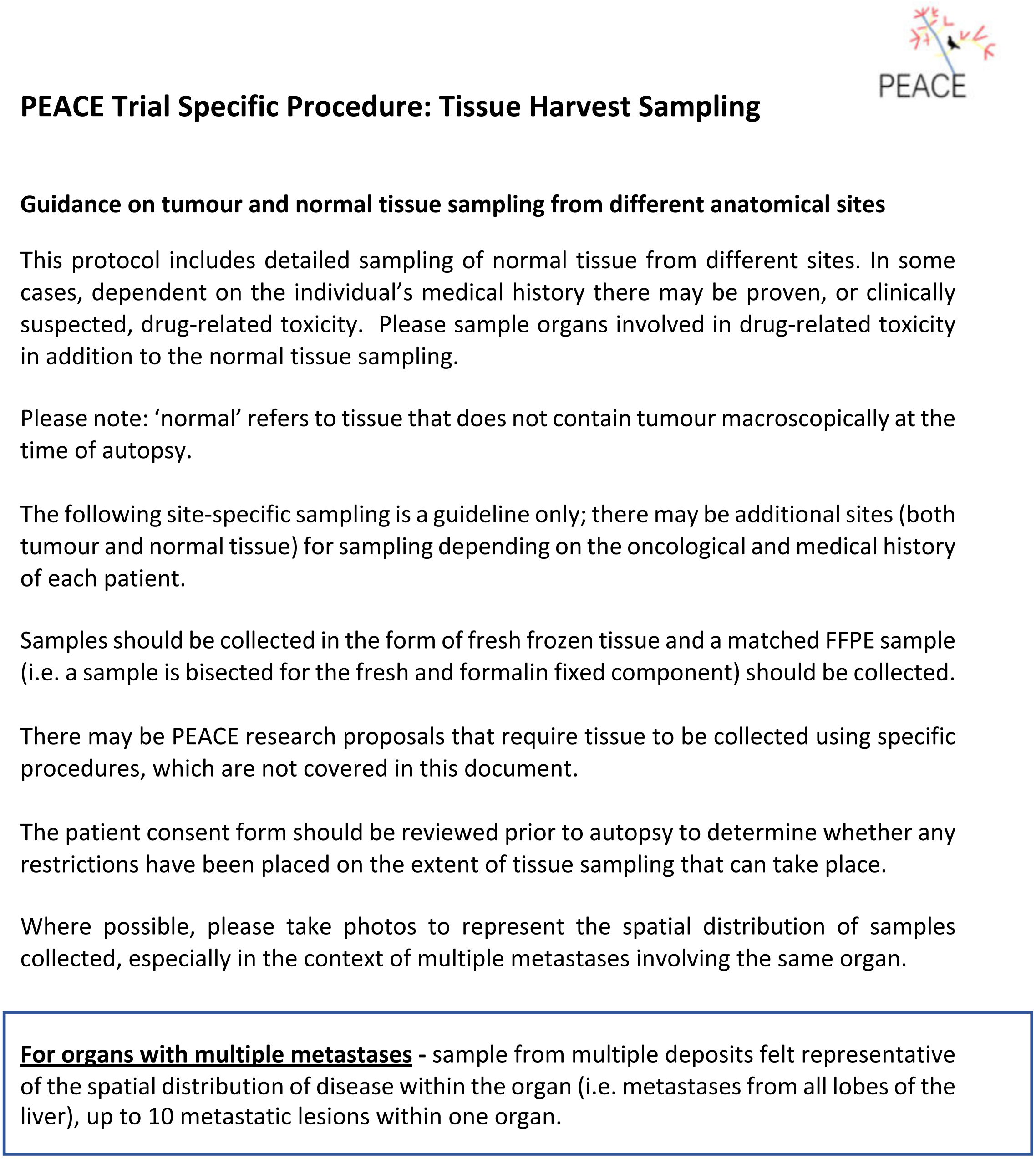
Population	#Events	%Parent	%Total
■ All Events	10,000	####	100.0
🟩 P2	8,213	82.1	82.1
🟥 P1	8,077	98.3	80.8
🟦 P3	3,276	40.6	32.8
🟪 P4	3,656	45.3	36.6

Tube: Bone CAIX

Population	#Events	%Parent	%Total
■ All Events	20,030	####	100.0
🟩 P2	18,161	90.7	90.7
🟥 P1	18,017	99.2	90.0
🟦 P3	101	0.6	0.5
🟪 P4	15,760	87.5	78.7

Flow sorting of cells from the lymph node metastasis and normal bone of PEA314 and PEA329 multiplexed together shown in panel A and B respectively. Singlets are gated in P1, CAIX$^+$ cells in P3 and CAIX$^-$ cells in P4.

8.2.5 PEACE Tissue Collection Guidance



PEACE Trial Specific Procedure: Tissue Harvest Sampling

PEACE

Guidance on tumour and normal tissue sampling from different anatomical sites

This protocol includes detailed sampling of normal tissue from different sites. In some cases, dependent on the individual's medical history there may be proven, or clinically suspected, drug-related toxicity. Please sample organs involved in drug-related toxicity in addition to the normal tissue sampling.

Please note: 'normal' refers to tissue that does not contain tumour macroscopically at the time of autopsy.

The following site-specific sampling is a guideline only; there may be additional sites (both tumour and normal tissue) for sampling depending on the oncological and medical history of each patient.

Samples should be collected in the form of fresh frozen tissue and a matched FFPE sample (i.e. a sample is bisected for the fresh and formalin fixed component) should be collected.

There may be PEACE research proposals that require tissue to be collected using specific procedures, which are not covered in this document.

The patient consent form should be reviewed prior to autopsy to determine whether any restrictions have been placed on the extent of tissue sampling that can take place.

Where possible, please take photos to represent the spatial distribution of samples collected, especially in the context of multiple metastases involving the same organ.

For organs with multiple metastases - sample from multiple deposits felt representative of the spatial distribution of disease within the organ (i.e. metastases from all lobes of the liver), up to 10 metastatic lesions within one organ.

Material and equipment

Equipment suggested for collection of tissue:

- Liquid nitrogen in appropriate canister (if this method is being used for snap freezing, otherwise a bucket with dry ice)
- 15ml falcon tubes filled with formalin to fix tissue
- Cryovials for storage of snap frozen foil wrapped tissue specimens, labelled according to the PEACE labelling SOP
- An appropriate device for photography with annotation at the time of processing
- Disposable forceps
- Wet ice
- Sample summary sheet/ Notebook to record samples taken (anonymised)
- Ruler for measurements made (if needed)
- Pre-labelled cassettes for pathology samples (if needed)
- Disposable scalpels
- Punch biopsies (if needed)
- Large weigh boats
- 20ml syringe for collection of fluids
- 50ml falcon tubes for collection of fluid
- 15ml falcon tubes for collection of fluid
- Lab pens
- EDTA tubes for collection of germline blood- if not collected prior to death

See page 3 for sampling guide for normal tissue, tumour tissue, body fluids, blood and additional notes.

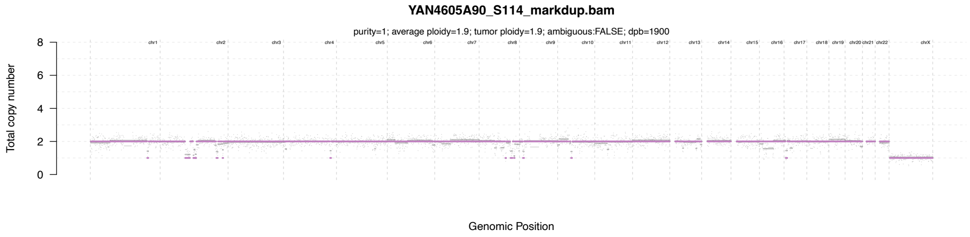
Sampling Guide for Organ Sites & Bodily Fluids

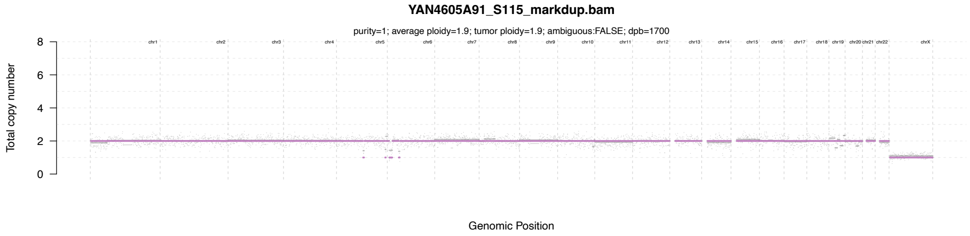
Brain	Left	Frontal lobe	1x sample of normal tissue * (see additional note)	Consider collecting one sample per cm^2 of tumour	*If tumour present, collect additional normal samples 5mm and 10mm away from tumour (i.e. - tumour samples of parietal lobe metastasis according to size, and 2x normal samples relating to the tumour deposit)
		Parietal			
		Temporal			
		Occipital			
	Right	Frontal lobe			
		Parietal			
		Temporal			
		Occipital			
	N/A	Cerebellum	1x sample of normal tissue	Consider collecting one sample per cm^2 of tumour	
		Pituitary gland	Take adequate tissue to assess anterior & posterior gland, and stalk	Take adequate tissue to assess tumour, anterior & posterior gland, and stalk	
Thyroid	Left		1x sample of normal tissue* (see additional note)	Consider collecting one sample per cm^2 of tumour* (see additional note)	*For renal cell carcinoma cases extensive sampling (large sample representative of organ following normal and tumour sampling) should be considered
	Right				
Breast	Left/Right	Breast outer quadrant	1x sample of normal tissue* (see additional note)	Consider collecting one sample per cm^2 of tumour	*Only sample if evidence of metastasis
Lungs	Left	Upper lobe	3x samples of normal tissue * (see additional note)	Consider collecting one sample per cm^2 of tumour	*For lung cancer cases extensive sampling should be considered, up to 10 samples per lobe
		Lower lobe			
		Main bronchus	1x sample of normal tissue	N/A	
		Secondary bronchi			
		Terminal bronchi			
		Visceral / parietal pleura	1 x sample of normal tissue* (see additional note)		*Skim the thinnest amount possible to enrich for mesothelial cells
	Right	Upper lobe	3x samples of normal tissue* (see additional note)	Consider collecting one sample per cm^2 of tumour	*For Lung cases extensive sampling should be considered, up to 10 samples per lobe
		Middle lobe			
		Lower lobe			
		Main bronchus	1x sample of normal tissue	N/A	
		Secondary bronchi			
		Terminal bronchi			
		Visceral / parietal pleura	1 x sample of normla tissue* (see additional note)	Take whole tumour nodules	*Skim the thinnest amount possible to enrich for mesothelial cells
	Mediastinum	Superior Trachae	1x sample of normal tissue	Consider collecting one sample per cm^2 of tumour	
		lymph node			
Heart	Left	Ventricle (myocardium)	1x sample of normal tissue	Consider collecting one sample per cm^2 of tumour	
Stomach	N/A	N/A	1x sample of normal tissue	Consider collecting one sample per cm2 of tumour	
Pancreas	N/A	Head	1x sample of normal tissue* (see additional note)	Consider collecting one sample per cm^2 of tumour (see additional note)	*For renal cell carcinoma cases extensive sampling (large sample representative of organ following normal and tumour sampling) should be considered
		Body			
		Tail			

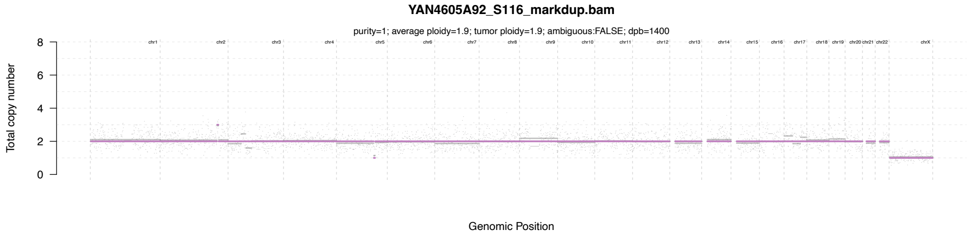
Organ	Laterality	Organ site	Normal tissue	Tumour	Additional note
Large bowel	N/A	Ascending colon	1x sample 1x sample of normal tissue (full thickness - lumen to serosa)	Consider collecting one sample per cm^2 of tumour	
		Transverse colon			
		Descending colon			
Small bowel	N/A	Terminal Ileum	1x sample 1x sample of normal tissue (full thickness - lumen to serosa)	Consider collecting one sample per cm^2 of tumour	
		Mid small bowel			
		Jejunum			
Spleen	N/A	N/A	1x sample of normal tissue	Consider collecting one sample per cm^2 of tumour	
Liver	Left lobe	Lobe	1x sample of normal tissue	Consider collecting one sample per cm^2 of tumour	
	Right lobe		1x sample of normal tissue		
Kidney	Left	Upper pole	1x sample of normal tissue* (see additional note)	Consider collecting one sample per cm^2 of tumour* (see additional note)	*For renal cell carcinoma cases extensive sampling (large sample representative of organ following normal and tumour sampling) should be considered
		Mid-pole			
		Lower pole			
	Right	Upper pole			
		Mid-pole			
		Lower pole			
Adrenal	Left	N/A	1x sample of normal tissue* (see additional note)	Consider collecting one sample per cm^2 of tumour* (see additional note)	*For renal cell carcinoma cases extensive sampling (large sample representative of organ following normal and tumour sampling) should be considered
	Right				
Bladder	N/A	N/A	1x sample of normal tissue (full thickness- lumen to serosa)	Consider collecting one sample per cm^2 of tumour	
Prostate	N/A	Prostate	Nil routine	Consider collecting one sample per cm^2 of tumour	
Fallopian Tube	Left/ Right	N/A	1x sample of normal tissue	Consider collecting one sample per cm^2 of tumour	
Bone	Left / Right	Involved sites only	1xsample of normla tissue paired with metastatic deposit * (see additional note)	Consider collecting one sample per cm^2 of tumour	* 5 cm away from involved area
Muscle	Right (approx. at level of L3)	Psoas	1x sample of normal tissue	N/A	
	N/A	Involved sites only	1xsample of normal tissue paired with metastatic deposit) * (see additional note)	Consider collecting one sample per cm^2 of tumour	* 5 cm away from involved area
Fat	N/A	Abdominal fat	1x sample of normal tissue	N/A	
Subcutaneous deposits	N/A	Involved sites only	N/A	Consider collecting one sample per cm^2 of tumour	
Lymph nodes		Involved sites only	N/A	Consider collecting one sample per cm^2 of tumour	
		Uninvolved lymph nodes	At the discretion of harvest team (aim 2-5 macroscopically uninvolved lymph nodes	N/A	
Lymphoid Tissue	N/A	Tonsillar tissue	1x sample of normal tissue	Consider collecting one sample per cm^2 of tumour	
		Spleen			
Body fluids	Left / Right	Pleural fluid	If fluid present collect each of the left in 15ml/50ml falcon tube and process after Tissue Harvest		
	N/A	Pericardial fluid			
	N/A	CSF			
	N/A	Ascites			
Bloods			If due to unforeseen circumstances blood samples were not collected prior to death, up to 73ml of blood (refer to baseline sampling for breakdown) may be taken at the time of the tissue sampling after death, if possible.		

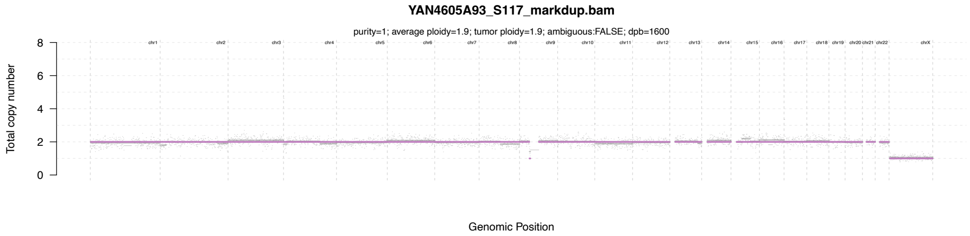
8.3 TRACERx Renal Appendix

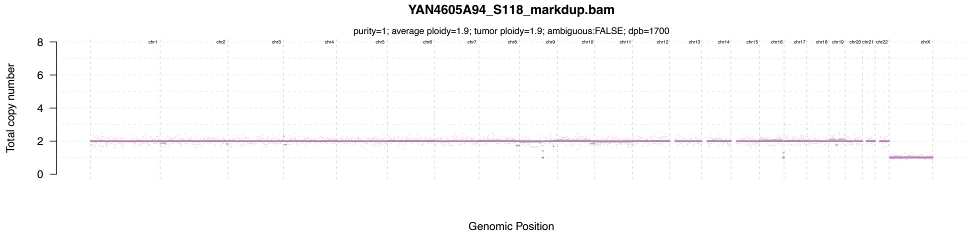
8.3.1 Copy number profiles of bone marrow aspirate CAIX$^+$ cells from RK1033

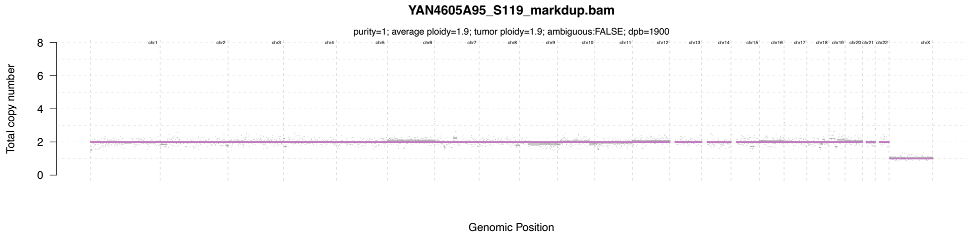




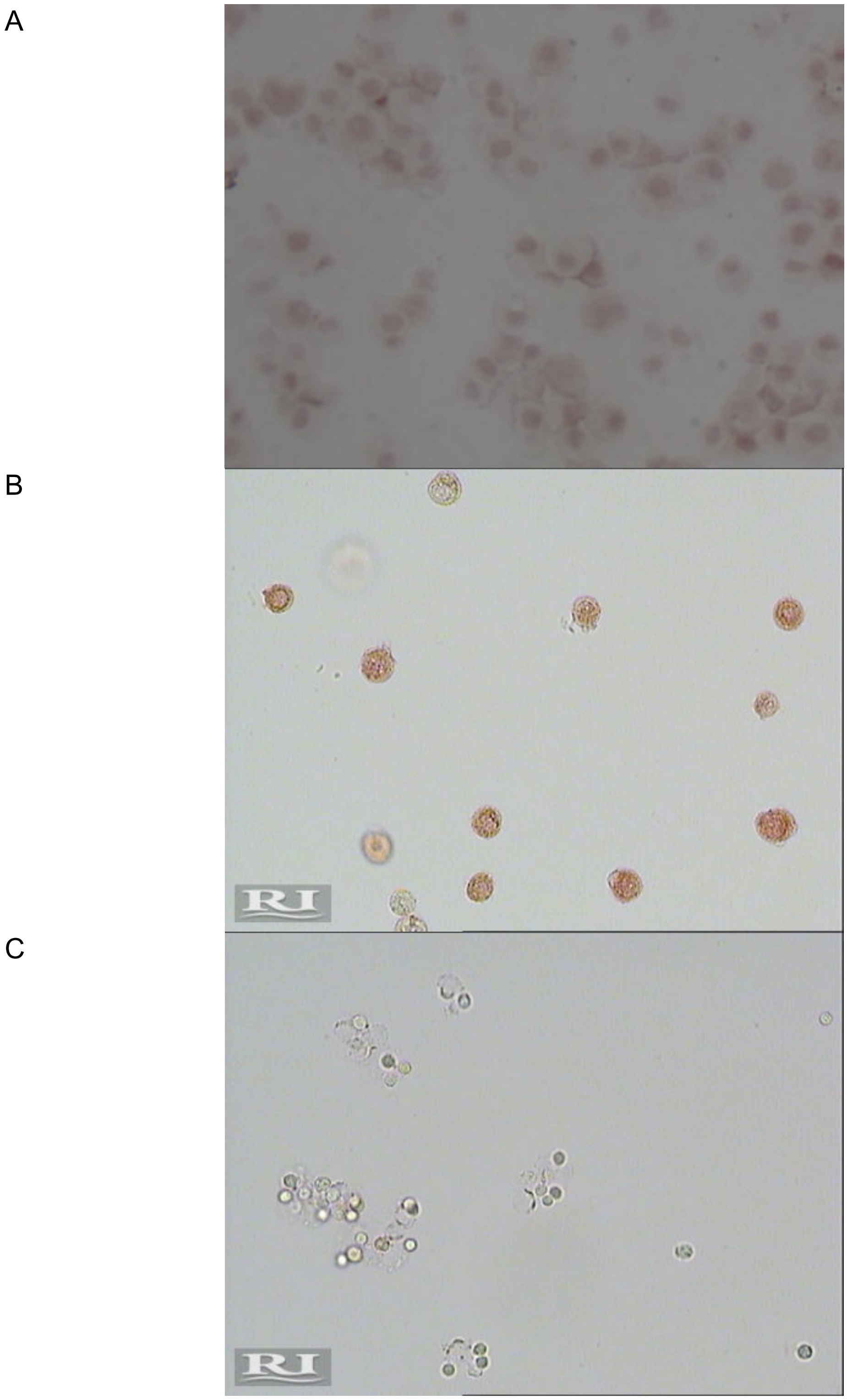






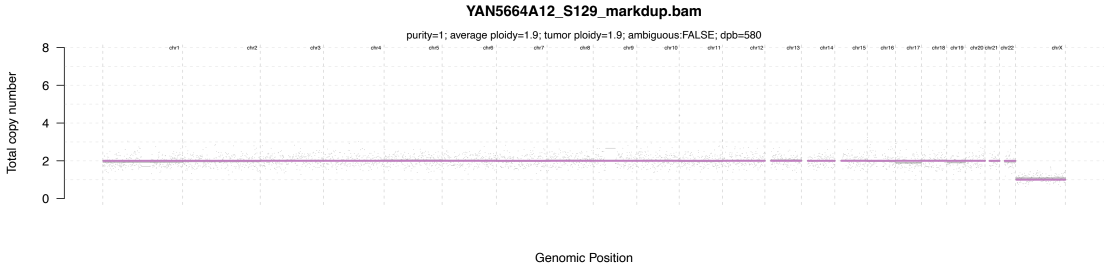


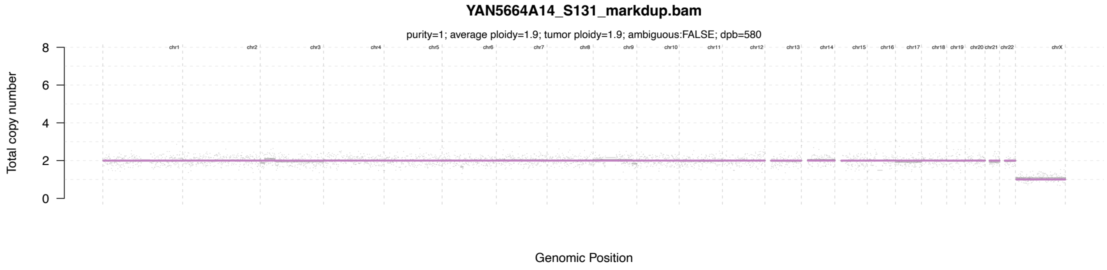
8.3.2 Immunocytochemistry detection of PAX8+ cells in cell lines



Cytospin of RPTEC cells stained for PAX8 shown in panel A. RPTEC cells and MCF-7 cells stained in solution for PAX8 shown in panels B and C respectively.

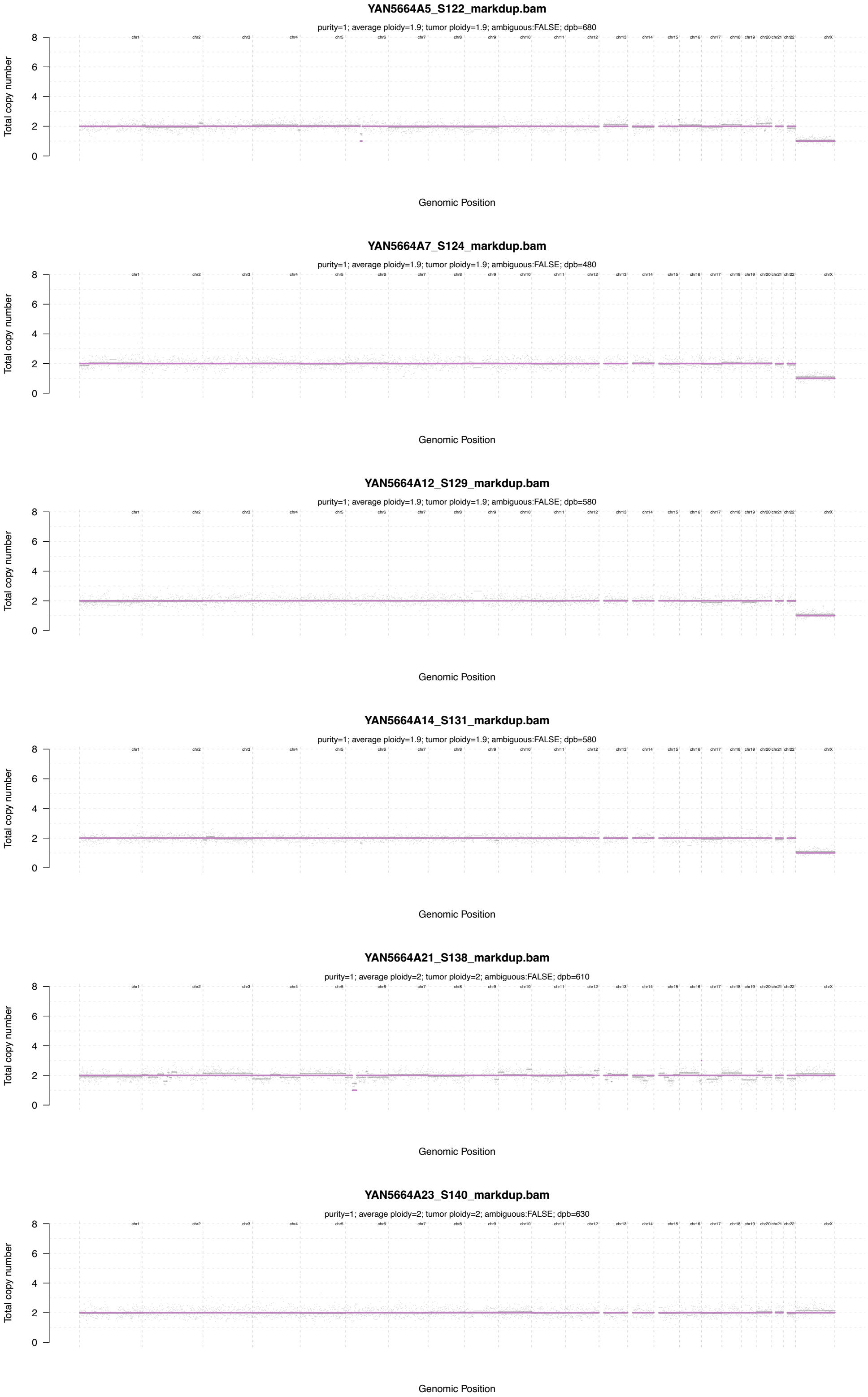
8.3.3 Additional copy number profiles of PAX8$^+$ cells from a pancreatic metastasis





Diploid copy number profiles of two cells collected from the pancreatic metastasis of RK1043.

8.3.4 Additional copy number profiles of PAX8^{+} cells from bone marrow aspirates



Example diploid copy number profiles of PAX8^{+} cells from bone marrow aspirates.

8.3.5 TRACERx Renal bone marrow sub-study patient information sheet

Patient Information Sheet: Bone Marrow Aspiration

TRACERx Renal (**TRA**cking Renal Cell Carcinoma **E**volution Through Therapy (**Rx**)

RMH Protocol No.3723

Introduction

To study how cancers spread and to prevent this, researchers need to be able to study cancer cells which have left the main cancer. In other types of cancer, cancer cells can be found in the bone marrow at an early stage, however, we do not currently know if this happens in kidney cancer.

In addition to the main TRACERx Renal study, we are asking your permission (consent) to collect a sample from your bone marrow to look for and analyse any abnormal cells found. This will be performed at the time of your cancer surgery: after you are asleep and before your surgery begins.

We are asking permission (consent) for your participation in this research project. Consent is a freely given agreement, based on a full understanding of what is to happen. Researchers can only use the samples after their research has been approved by independent researchers and by an independent Research and Ethics Committee. This is to make sure that the research is in the interest of patients and is carried out safely.

What is the purpose of the study?

Unfortunately, kidney cancer in some patients can come back after surgery. This is thought to be due to small numbers of cancer cells which have already spread to other parts of the body. We are seeking (a) to understand when this spread occurs and (b) to understand these cancer cells in more detail. We hope that insights from our work may

What will happen to me if I take part?

If you decide to participate in the study you will be asked to sign the attached consent form. We will collect a sample from your bone marrow during your cancer surgery after you are asleep. This procedure is known as a bone marrow aspiration.

What does a bone marrow aspiration involve?

Bone marrow is the sponge-like material found in the middle of your bone where your blood cells are made. A bone marrow aspiration is a procedure in which a liquid sample of bone marrow is sucked out through a small hollow needle into a syringe. Bone marrow tissue is usually taken from the back of the hip bone where the hip is closest to the surface. This procedure takes about 10 to 15 minutes.

Is it painful?

You will not feel pain during the procedure as you will be asleep. We will also give local anaesthetic to lessen any remaining discomfort when you wake up.

What are the possible risks associated with having a bone marrow aspiration?

Bone marrow biopsies are frequently performed and are a low risk procedure. The main risks include pain, bleeding and bruising, but these are usually temporary. Developing an infection at the site of the biopsy is extremely rare.

What happens afterwards?

The dressing should be kept on for 24 hours and kept dry to minimise risk of getting an infection; after this time you can remove the dressing.

Where will my tissue sample be kept and who will have access to it?

Any samples you donate may be stored, processed and/or analysed at The Royal Marsden NHS Foundation Trust or the following laboratories, including but not limited to; The Francis Crick Institute, University College London (UCL) and Institute of Cancer Research. Samples may also be sent to other collaborators such as commercial companies, after approval from The Royal Marsden NHS Foundation Trust (RM). These collaborators may not be located in the United Kingdom, and please note data protection guidelines may differ outside of the UK. In each laboratory where tissue is or a medical consultant. The Chief Executive of the Trust has overall responsibility for the tissue.

We would also like to collect samples to be used for future research by researchers interested in cancer research. The projects would need to be approved by The Royal Marsden NHS Foundation Trust including the Chief Investigator of this study, and a Research Ethics Committee. The samples would not include any personal information therefore researchers will not be able to identify you from your samples. These projects may be carried out by researchers at institutions other than The Royal Marsden NHS Foundation Trust and The Francis Crick Institute or laboratories listed above, including researchers working for commercial companies outside of the UK.

What will happen to the samples taken for this study?

The bone marrow samples will be used to look for abnormal cells that may have spread from the main cancer. This research will include studies of the genetic material (DNA) in these cells, and also of normal genetic material (DNA) found in the white cells in the blood, to identify where these cells have come from. We will then perform further studies to try understand these cells in more detail by looking at what they are doing and if they cause cancer to return.

How will this study benefit me?

This research is done to help us understand more about how cancers spread. The purpose of the research is not to benefit you, but to increase our understanding of cancer and help us to treat it better in the future.

What about confidentiality?

Samples will be given a unique code so that only The Royal Marsden NHS Foundation Trust know your name or any other personal details. Samples can only be traced back to the patient, by the patient’s own clinical team. Tissue may be transferred to other external organisations. This will be done under written agreement, which guarantees the use and safe keeping of samples, which will be kept anonymous. This means that no information

Will you be able to tell me the results of any research on my tissue sample?

It can often take a long time before results are known and it will not be possible to discuss the results of individual tissue samples. Although results will not affect your care now, they may help you and people like you in the future. They may help us learn more about what causes kidney cancer and other diseases, how to prevent them and how to treat them.

If I agree, what do I have to do?

We will ask you to sign a consent form agreeing to take part in this project. You can still change your mind at any time, even after you have signed a consent form. Your GP will be informed of your taking part, but otherwise all information collected about you during the course of the study will be kept strictly confidential.

What if I do not wish to take part and change my mind?

Before you decide whether to take part, it is important for you to understand why the research is being done and what it will involve. Please take time to read this information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. You may wish to discuss the study with your GP or others before deciding. You can change your mind at any time, even if you are no longer under the care of The Royal Marsden NHS Foundation Trust. This will not affect your medical care.

Who is organising and funding the research?

This study is being organised by The Royal Marsden NHS Foundation Trust Renal Unit. You will not receive any reimbursement for participating in this research.

How will my data be processed?

The Royal Marsden NHS Foundation Trust will be using information from you and/or your medical records in order to undertake this study and will act as the data controller for this study. This means that we are responsible for looking after your information and using it properly. The Royal Marsden NHS Foundation Trust will keep identifiable information about you for at least 5 years after the study has finished, in line with local policies and legal requirements. Your rights to access, change or move your information we have already obtained. To safeguard your rights, we will use the minimum personally-identifiable information possible.

The Royal Marsden NHS Foundation Trust will use your name, NHS number and contact details to contact you about the research study, and make sure that relevant information about the study is recorded for your care, and to oversee the quality of the study. Individuals employed by the Royal Marsden, and regulatory organisations, may look at your medical and research records to check the accuracy of the research study. Your NHS site will pass these details to these individuals along with the information collected from you and/or your medical records.

When you agree to take part in a research study, the information about your health and care may be provided to researchers running other research studies in this organisation and in other organisations. These organisations may be universities, NHS organisations or companies involved in health and care research in this country or abroad. Your information will only be used by organisations and researchers to conduct research in accordance with the UK Policy Framework for Health and Social Care Research**.**

This information will not identify you and will not be combined with other information in a way that could identify you. The information will only be used for the purpose of healthcare research, and cannot be used to contact you or to affect your care. It will not be used to make decisions about future services available to you, such as insurance.

You can find out more about how The Royal Marsden uses your information by contacting the Data Protection Officer at The Royal Marsden. Email: dpo@rmh.nhs.uk.

What will happen to the results of the research study?

The results of this clinical research study will be published in a scientific journal. Results may also be presented at scientific conferences. Results from an analysis of your tumour may be made available, in an anonymous format, for use by other researchers. No details

This study has been reviewed and approved by a research ethics committee.

Who do I contact if I have any questions?

If you have any questions or you no longer want us to use your tissue samples, please consult your Consultant or other members of your clinical team.

Thank you for your time spent reading this information sheet and considering this clinical study.

8.3.6 TRACERx Renal bone marrow sub-study consent form

Consent form: Bone Marrow Aspiration

TRACERx Renal (**TRA**cking Renal Cell Carcinoma **E**volution Through Therapy (**Rx**)

RMH Protocol No. 3723

Please initial boxes t

1. I have read and understood the Bone Marrow Aspiration patient information sheet ver 0.1 dated 25-FEB-2020 for the above study and have had the opportunity to ask questions and discuss it with my doctor. □

2. I agree to a sample of bone marrow aspirate to be taken during my surgery for use in the above research project. □

3. I agree for analysis of DNA and other relevant material as part of the research project □

4. I agree that the samples and information collected about me will be stored on behalf of The Royal Marsden NHS Foundation Trust for use in future projects. □

5. I understand that my samples may be used for projects carried out by researchers at institutions, including researchers working for commercial (including pharmaceutical) companies and results of the analysis of my tumour may be made available to other researchers in an anonymous format. □

6. I understand that I am free to withdraw my approval for use of the sample at any time without giving a reason and without my medical care or legal rights being affected. □

7. I understand that research using the samples I give may include research aimed at understanding cancer, but the results of these investigations are unlikely to have any implications for me personally. □

8. I understand that I shall not benefit financially if future research leads to the development of new treatments or medical tests. □

9. I agree that that the samples and information collected can be used for future research by researchers at other institutions or laboratories including researchers working for commercial companies. □

______________________	_______________	____________________
Name of Patient	Date	Signature
______________________	_______________	____________________
Name of Person taking consent (Principal Investigator or authorised delegate)	Date	Signature

(3 Copies: 1 for patient, 1 for researcher and 1 to be kept with hospital notes)

8.3.7 TRACERx Renal bone marrow collection guidance document

RENAL AND MELANOMA UNIT GUIDANCE FOR CCR 3723 TRACERx RENAL BONE MARROW SUB-STUDY

PURPOSE: To document the sample collection and shipment process

SCOPE: This guidance document is applicable to the CCR 3723 TRACERx Renal Bone Marrow sub-study

PROCEDURE:

Screening process

- Potential patient identified on surgical lists up to 2 weeks prior by Clinical Fellow. Research team notified of potential biopsy.
- Patient approached and consented by research team if seen in clinic prior to surgery.
- If not seen before surgery, consent taken on the day prior to surgery by Clinical Fellow.
- Signed ICF given to Research team. RK number allocated if not already on TRACERx Renal.
- Research team notified of confirmed biopsy collection as soon as possible.

Prior to surgery

- Surgeon and anaesthetic team notified 1-3 days ahead of procedure using *appendix 1 Email Template.*
- Theatre team/Anaesthetist to confirm all required consumables are available.
- BSC to ensure all consumables to be provided by the Research team are available.
- Once the RK number has been assigned, collection vials to be clearly labelled with the following information: CCR 3723, RK number, Patient initials, Date of Collection, Time of Collection.

Consumables:

Consumables	Provided by:
Bone marrow aspirate needle (large)	R&M
Bone marrow aspirate needle (small)	R&M
Lithium Heparin Blood Tube (8 x 6ml)	R&M
Trolley	Theatres
Sterile Field Pack	Theatres
Sterile Gloves	Theatres
Chloraprep sticks	Theatres
2% Lidocaine (5ml)	Theatres
25G (orange) needle	Theatres
21G (green) needle	Theatres
10/20ml syringe	Theatres
Gauze and dressing (e.g. Transparent Tegaderm)	Theatres

Day of procedure:

- Clinical Fellow and 1 x BSC will attend if required. Clinical Fellow to attend the theatre briefing in the morning to explain study and procedure.
- A copy of the consent form and SOP to be present for reference.

Sample Collection

- Procedure to take place in Anaesthetic room prior to surgery.
- Patient will be positioned in lateral decubitus position with assistance of theatre team. The Clinical Fellow will perform the bone marrow aspiration from the posterior superior iliac spine and collect up to 40ml of sample into lithium heparin tubes. A suitable dressing will be applied after the procedure.
- This is anticipated to take approximately 5-10 minutes. If there is bleeding after the procedure, then a period of time for compression on the wound will be required.
- The patient will be repositioned after bone marrow aspiration. Surgery will then commence.
- Complete *Appendix 2 Requisition Form.*

Shipment

- Samples will be transferred on wet ice to the Crick via courier in the presence of the Clinical Fellow. Complete the top section of a *Appendix 3 Sample Transport Form* and include in shipment.
- Samples will be processed on same day at the Crick and undergo quality check by the Crick team, including confirmation of viability and suitability, prior to inclusion.
-
- Germline blood collected as part of CCR3723 TRACERx Renal protocol shall be included in this shipment also.

Tracking and confirmation of receipt

- Once samples have arrived at their destination, the bottom section of the Sample Transport Form must be completed and emailed back to RMG.Tissue.Collector@RMH.NHS.UK as confirmation of receipt to comply with HTA requirements.
- Sample details and movement of tissue to be recorded on sample management system Freezerpro.
- Requisition form to be filed in relevant folder.

Appendices

- Appendix 1 Email Template
- Appendix 2 Requisition Form
- Appendix 3 Sample Transport Form

Reference List

Abbosh, C., Birkbak, N.J., Wilson, G.A., Jamal-Hanjani, M., Constantin, T., Salari, R., *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**: 446–451. 2017.

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J.T., *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**: 194. 2019.

Acha-Sagredo, A., Ganguli, P. & Ciccarelli, F.D. Somatic variation in normal tissues: friend or foe of cancer early detection? *Ann. Oncol.* **33**: 1239–1249. 2022.

Acheampong, E., Morici, M., Abed, A., Bowyer, S., Asante, D.-B., Lin, W., *et al.* Powering single-cell genomics to unravel circulating tumour cell subpopulations in non-small cell lung cancer patients. *J. Cancer Res. Clin. Oncol.* **149**: 1941–1950. 2023.

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**: 1083–1086. 2017.

Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**: 94–101. 2020.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A. V., *et al.* Signatures of mutational processes in human cancer. *Nature* **500**: 415–421. 2013.

Ali, H.R., Jackson, H.W., Zanotelli, V.R.T., Danenberg, E., Fischer, J.R., Bardwell, H., *et al.* Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat. Cancer* **1**: 163–175. 2020.

Amin, M.B., Edge, S.B., Greene, F.L., Schilsky, R.L., Brookland, R.K., Washington, M.K., *et al. American Joint Committee on Cancer (AJCC). AJCC Cancer Staging Manual*. 2017.

Andre, F., Filleron, T., Kamal, M., Mosele, F., Arnedos, M., Dalenc, F., *et al.* Genomics to select treatment for patients with metastatic breast cancer. *Nature* **610**: 343–348. 2022.

Argelaguet, R., Cuomo, A.S.E., Stegle, O. & Marioni, J.C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**: 1202–1215. 2021.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**: 371-385.e18. 2018.

Baker, T.G., Alden, J., Dubuc, A.M., Welsh, C.T., Znoyko, I., Cooley, L.D., *et al.* Near haploidization is a genomic hallmark which defines a molecular subgroup of giant cell glioblastoma. *Neuro-Oncology Adv.* **2**: 1–10. 2020.

Bakouny, Z., El Zarif, T., Dudani, S., Connor Wells, J., Gan, C.L., Donskov, F., *et al.* Upfront Cytoreductive Nephrectomy for Metastatic Renal Cell Carcinoma Treated with Immune Checkpoint Inhibitors or Targeted Therapy: An Observational Study from the International Metastatic Renal Cell Carcinoma Database Consortium. *Eur. Urol.* **83**: 145–151. 2023.

Banys, M., Krawczyk, N. & Fehm, T. The role and clinical relevance of disseminated tumor cells in breast cancer. *Cancers (Basel).* **6**: 143–152. 2014.

Barkley, D., Moncada, R., Pour, M., Liberman, D.A., Dryg, I., Werba, G., *et al.* Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.* **54**: 1192–1201. 2022.

Baron, M., Tagore, M., Hunter, M. V., Kim, I.S., Moncada, R., Yan, Y., *et al.* The Stress-Like Cancer Cell State Is a Consistent Component of Tumorigenesis. *Cell Syst.* **11**: 536-546.e7. 2020.

Bello, D.M. & Faries, M.B. The Landmark Series: MSLT-1, MSLT-2 and DeCOG (Management of Lymph Nodes). *Ann. Surg. Oncol.* **27**: 15–21. 2020.

Bex, A., Ljungberg, B., van Poppel, H. & Powles, T. The Role of Cytoreductive Nephrectomy: European Association of Urology Recommendations in 2016. *Eur. Urol.* **70**: 901–905. 2016.

Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**: 1352–1362. 2021.

Birkbak, N.J., Eklund, A.C., Li, Q., McClelland, S.E., Endesfelder, D., Tan, P., *et al.* Paradoxical relationship between chromosomal instability and survival outcome in cancer. *Cancer Res.* **71**: 3447–3452. 2011.

Biswas, D., Birkbak, N.J., Rosenthal, R., Hiley, C.T., Lim, E.L., Papp, K., *et al.* A clonal expression biomarker associates with lung cancer mortality. *Nat. Med.* **25**: 1540–1548. 2019.

Bolli, N., Avet-Loiseau, H., Wedge, D.C., Van Loo, P., Alexandrov, L.B., Martincorena, I., *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**: 2997. 2014.

Borcherding, N., Vishwakarma, A., Voigt, A.P., Bellizzi, A., Kaplan, J., Nepple, K., *et al.* Mapping the immune environment in clear cell renal carcinoma by single-cell genomics. *Commun. Biol.* **4**: 122. 2021.

Boveri, T. Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J. Cell Sci.* **121**: 1–84. 2008.

Boveri, T. Zur frage der entstehung maligner tumoren. *Fischer*. 1914.

Bowes, A.L., Tarabichi, M., Pillay, N. & Van Loo, P. Leveraging single-cell sequencing to unravel intratumour heterogeneity and tumour evolution in human cancers. *J. Pathol.* **257**: 466–478. 2022.

Braun, S., Pantel, K., Müller, P., Janni, W., Hepp, F., Kentenich, C.R.M., *et al.* Cytokeratin-Positive Cells in the Bone Marrow and Survival of Patients with Stage I, II, or III Breast Cancer. *N. Engl. J. Med.* **342**: 525–533. 2000.

Braun, S., Vogl, F.D., Naume, B., Janni, W., Osborne, M.P., Coombes, R.C., *et al.* A Pooled Analysis of Bone Marrow Micrometastasis in Breast Cancer. *N. Engl. J. Med.* **353**: 793–802. 2005.

Broad Institute. Picard toolkit. *https://broadinstitute.github.io/picard/*. 2019.

Brown, G.R., Simon, M., Wentling, C., Spencer, D.M., Parker, A.N. & Rogers, C.A. A review of inherited cancer susceptibility syndromes. *JAAPA* **33**: 10–16. 2020.

Browning, B.L., Tian, X., Zhou, Y. & Browning, S.R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**: 1880–1890. 2021.

Buchner, A., Riesenberg, R., Kotter, I., Crispin, A., Hofstetter, A. & Oberneder, R. Detection and prognostic value of cytokeratin positive tumor cells in bone marrow of patients with renal cell carcinoma. *J. Urol.* **170**: 1747–1751. 2003.

Buchner, A., Riesenberg, R., Kotter, I., Hofstetter, A., Stief, C. & Oberneder, R. Frequency and prognostic relevance of disseminated tumor cells in bone marrow of patients with metastatic renal cell carcinoma. *Cancer* **106**: 1514–1520. 2006.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**: 486–490. 2015.

Cable, D.M., Murray, E., Zou, L.S., Goeva, A., Macosko, E.Z., Chen, F., *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **40**: 517–526. 2022.

Cancer Research UK. Cancer Incidence for Common Cancers. *http://www.cancerresearchuk.org/cancer-info/cancerstats/incidence/commoncancers/uk-cancer-incidence-statistics-for-common-cancers*. 2020.

Carter, S.L., Eklund, A.C., Kohane, I.S., Harris, L.N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38**: 1043–1048. 2006.

Casasent, A.K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., *et al.* Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* **172**: 205-217.e12. 2018.

CASM/Cancer IT Wellcome Sanger Institute. alleleCount. *http://cancerit.github.io/alleleCount/*. 2020.

Caswell, D.R. & Swanton, C. The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. *BMC Med.* **15**: 133. 2017.

Chaffer, C.L. & Weinberg, R.A. A perspective on cancer cell metastasis. *Science* **331**: 1559–1564. 2011.

Chakiryan, N.H., Gore, L.R., Reich, R.R., Dunn, R.L., Jiang, D.D., Gillis, K.A., *et al.* Survival Outcomes Associated with Cytoreductive Nephrectomy in Patients with Metastatic Clear Cell Renal Cell Carcinoma. *JAMA Netw. Open* **5**: 1–12. 2022.

Chen, T. & Guestrin, C. XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. 2016.

Chen, W., Guillaume-Gentil, O., Rainer, P.Y., Gäbelein, C.G., Saelens, W., Gardeux, V., *et al.* Live-seq enables temporal transcriptomic recording of single cells. *Nature* **608**: 733–740. 2022.

Chen, W., Hoffmann, A.D., Liu, H. & Liu, X. Organotropism: new insights into molecular mechanisms of breast cancer metastasis. *npj Precis. Oncol.* **2**. 2018.

Chernysheva, Markina, Demidov, Kupryshina, Chulkova, Palladina, *et al.* Bone Marrow Involvement in Melanoma. Potentials for Detection of Disseminated Tumor Cells and Characterization of Their Subsets by Flow Cytometry. *Cells* **8**: 627. 2019.

Chiang, Y.-C., Park, I.-Y., Terzo, E.A., Tripathi, D.N., Mason, F.M., Fahey, C.C., *et al.* SETD2 Haploinsufficiency for Microtubule Methylation Is an Early Driver of Genomic Instability in Renal Cell Carcinoma. *Cancer Res.* **78**: 3135–3146. 2018.

Chkhaidze, K., Heide, T., Werner, B., Williams, M.J., Huang, W., Caravagna, G., *et al.* Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *PLOS Comput. Biol.* **15**: e1007243. 2019.

Chopra, N., Tovey, H., Pearson, A., Cutts, R., Toms, C., Proszek, P., *et al.* Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative breast cancer. *Nat. Commun.* **11**: 1–12. 2020.

Choueiri, T.K., Tomczak, P., Park, S.H., Venugopal, B., Ferguson, T., Chang, Y.-H., *et al.* Adjuvant Pembrolizumab after Nephrectomy in Renal-Cell Carcinoma. *N. Engl. J. Med.* **385**: 683–694. 2021.

Chu, T., Wang, Z., Pe'er, D. & Danko, C.G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer* **3**: 505–517. 2022.

Clapp, G. & Levy, D. A review of mathematical models for leukemia and lymphoma. *Drug Discov. Today Dis. Model.* **16**: 1–6. 2015.

Clark, S.J., Argelaguet, R., Kapourani, C.-A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**: 781. 2018.

Cortes-Ciriano, I., Steele, C.D., Piculell, K., Al-Ibraheemi, A., Eulo, V., Bui, M.M., *et al.* Genomic Patterns of Malignant Peripheral Nerve Sheath Tumor (MPNST) Evolution Correlate with Clinical Outcome and Are Detectable in Cell-Free DNA. *Cancer Discov.* **13**: 654–671. 2023.

Czyż, Z.T., Hoffmann, M., Schlimok, G., Polzer, B. & Klein, C.A. Reliable Single Cell Array CGH for Clinical Samples. *PLoS One* **9**: e85907. 2014.

Das, P. & Taube, J.H. Regulating Methylation at H3K27: A Trick or Treat for Cancer Cell Plasticity. *Cancers (Basel).* **12**: 2792. 2020.

Davies, H., Bignell, G.R., Cox, C., Stephens, P., Edkins, S., Clegg, S., *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**: 949–954. 2002.

Davis, A., Gao, R. & Navin, N. Tumor evolution: Linear, branching, neutral or punctuated? *Biochim. Biophys. Acta - Rev. Cancer* **1867**: 151–161. 2017.

Degasperi, A., Zou, X., Dias Amarante, T., Martinez-Martinez, A., Koh, G.C.C., Dias, J.M.L., *et al.* Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. *Science* **376**. 2022.

Demeulemeester, J., Kumar, P., Møller, E.K., Nord, S., Wedge, D.C., Peterson, A., *et al.* Tracing the origin of disseminated tumor cells in breast cancer using single-cell sequencing. *Genome Biol.* **17**: 1–15. 2016.

Deng, Y., Bartosovic, M., Ma, S., Zhang, D., Kukanja, P., Xiao, Y., *et al.* Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609**: 375–383. 2022.

Dentro, S.C., Leshchiner, I., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A.G., *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**: 2239-2254.e39. 2021.

Dentro, S.C., Wedge, D.C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harb. Perspect. Med.* **7**. 2017.

Dhanasekaran, R., Deutzmann, A., Mahauad-Fernandez, W.D., Hansen, A.S., Gouw, A.M. & Felsher, D.W. The MYC oncogene — the grand orchestrator of cancer growth and immune evasion. *Nat. Rev. Clin. Oncol.* **19**: 23–36. 2022.

Dölle, C., Bindoff, L.A. & Tzoulis, C. 3,3′-Diaminobenzidine staining interferes with PCR-based DNA analysis. *Sci. Rep.* **8**: 1272. 2018.

Dong, X., Zhang, L., Milholland, B., Lee, M., Maslov, A.Y., Wang, T., *et al.* Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* **14**: 491–493. 2017.

Dudani, S., De Velasco, G., Wells, J.C., Gan, C.L., Donskov, F., Porta, C., *et al.* Evaluation of Clear Cell, Papillary, and Chromophobe Renal Cell Carcinoma Metastasis Sites and Association With Survival. *JAMA Netw. Open* **4**: e2021869–e2021869. 2021.

Dunn, G.P., Spiliopoulos, K., Plotkin, S.R., Hornicek, F.J., Harmon, D.C., Delaney, T.F., *et al.* Role of resection of malignant peripheral nerve sheath tumors in patients with neurofibromatosis Type 1: Clinical article. *J. Neurosurg.* **118**: 142–148. 2013.

Duò, A., Robinson, M.D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7**: 1141. 2020.

Eide, N., Faye, R.S., Høifødt, H.K., Øvergaard, R., Jebsen, P., Kvalheim, G., *et al.* Immunomagnetic detection of micrometastatic cells in bone marrow in uveal melanoma patients. *Acta Ophthalmol.* **87**: 830–836. 2009.

Eide, N., Faye, R.S., Høifødt, H.K., Sandvik, L., Qvale, G.A., Faber, R., *et al.* The Results of Stricter Inclusion Criteria in an Immunomagnetic Detection Study of Micrometastatic Cells in Bone Marrow of Uveal Melanoma Patients - Relevance for Dormancy. *Pathol. Oncol. Res.* **25**: 255–262. 2019.

Eide, N., Hoifødt, H.K., Nesland, J.M., Faye, R.S., Qvale, G.A., Faber, R.T., *et al.* Disseminated tumour cells in bone marrow of patients with uveal melanoma. *Acta Ophthalmol.* **91**: 343–348. 2013.

Ellis, P., Moore, L., Sanders, M.A., Butler, T.M., Brunner, S.F., Lee-Six, H., *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**: 841–871. 2021.

Erickson, A., He, M., Berglund, E., Marklund, M., Mirzazadeh, R., Schultz, N., *et al.* Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature* **608**: 360–367. 2022.

Escudier, B., Porta, C., Schmidinger, M., Rioux-Leclercq, N., Bex, A., Khoo, V., *et al.* Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **30**: 706–720. 2019.

Espina, V., Wulfkuhle, J.D., Calvert, V.S., VanMeter, A., Zhou, W., Coukos, G., *et al.* Laser-capture microdissection. *Nat. Protoc. 2006 12* **1**: 586–603. 2006.

Evans, D.G.R., Huson, S.M. & Birch, J.M. Malignant peripheral nerve sheath tumours in inherited disease. *Clin. Sarcoma Res.* **2**: 1–5. 2012.

Farid, M., Demicco, E.G., Garcia, R., Ahn, L., Merola, P.R., Cioffi, A., *et al.* Malignant Peripheral Nerve Sheath Tumors. *Oncologist* **19**: 193–201. 2014.

Fehrmann, R.S.N., Karjalainen, J.M., Krajewska, M., Westra, H.J., Maloney, D., Simeonov, A., *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47**: 115–125. 2015.

Fei, S.S., Mitchell, A.D., Heskett, M.B., Vocke, C.D., Ricketts, C.J., Peto, M., *et al.* Patient-specific factors influence somatic variation patterns in von Hippel–Lindau disease renal tumours. *Nat. Commun.* **7**: 11588. 2016.

Fittall, M.W. & Van Loo, P. Translating insights into tumor evolution to clinical practice: Promises and challenges. *Genome Med.* **11**: 1–14. 2019.

Fyfe, G., Fisher, R.I., Rosenberg, S.A., Sznol, M., Parkinson, D.R. & Louie, A.C. Results of treatment of 255 patients with metastatic renal cell carcinoma who received high-dose recombinant interleukin-2 therapy. *J. Clin. Oncol.* **13**: 688–696. 1995.

Gaiti, F., Chaligne, R., Gu, H., Brand, R.M., Kothen-Hill, S., Schulman, R.C., *et al.* Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569**: 576–580. 2019.

Gangnus, R., Langer, S., Breit, E., Pantel, K. & Speicher, M.R. Genomic profiling of viable and proliferative micrometastatic cells from early-stage breast cancer patients. *Clin. Cancer Res.* **10**: 3457–3464. 2004.

Gao, R., Bai, S., Henderson, Y.C., Lin, Y., Schalck, A., Yan, Y., *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* **39**: 599–608. 2021.

Gao, R., Davis, A., McDonald, T.O., Sei, E., Shi, X., Wang, Y., *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**: 1119–1130. 2016.

Gao, T., Soldatov, R., Sarkar, H., Kurkiewicz, A., Biederstedt, E., Loh, P.-R., *et al.* Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. *Nat. Biotechnol.* **41**: 417–426. 2023.

Gawad, C., Koh, W. & Quake, S.R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**: 175–88. 2016.

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A.J., Salm, M.P., Varela, I., *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**: 225–233. 2014.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.* **366**: 883–892. 2012.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**: 122–128. 2020.

Giesen, C., Wang, H.A.O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**: 417–422. 2014.

Gilje, B., Nordgård, O., Tjensvoll, K., Borgen, E., Synnestvedt, M., Smaaland, R., *et al.* Comparison of molecular and immunocytochemical methods for detection of disseminated tumor cells in bone marrow from early breast cancer patients. *BMC Cancer* **14**: 1–8. 2014.

Gonzalez-Pena, V., Natarajan, S., Xia, Y., Klein, D., Carter, R., Pang, Y., *et al.* Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc. Natl. Acad. Sci.* **118**: e2024176118. 2021.

Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv*, doi: 10.1101/372896. 2020.

Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**: 2847–2849. 2016.

Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**: 2811–2812. 2014.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L. & Tang, F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**: 2126–2135. 2013.

Gutmann, D.H., Ferner, R.E., Listernick, R.H., Korf, B.R., Wolters, P.L. & Johnson, K.J. Neurofibromatosis type 1. *Nat. Rev. Dis. Prim.* **3**: 17004. 2017.

Gužvić, M., Braun, B., Ganzer, R., Burger, M., Nerlich, M., Winkler, S., *et al.* Combined genome and transcriptome analysis of single disseminated cancer cells from bone marrow of prostate cancer patients reveals unexpected transcriptomes. *Cancer Res.* **74**: 7383–7394. 2014.

Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., *et al.* Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.* **22**: 1995–2007. 2012.

Haghverdi, L., Lun, A.T.L., Morgan, M.D. & Marioni, J.C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**: 421–427. 2018.

Hanahan, D. & Weinberg, R.A. Hallmarks of Cancer: The Next Generation. *Cell* **144**: 646–674. 2011.

Hansemann, D. Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung. *Arch. für Pathol. Anat. und Physiol. und für Klin. Med.* **119**: 299–326. 1890.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**: 3573-3587.e29. 2021.

Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep.* **2**: 666–673. 2012.

Hicks, J., Krasnitz, A., Lakshmi, B., Navin, N.E., Riggs, M., Leibu, E., *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16**: 1465–1479. 2006.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., *et al.* Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26**: 304–319. 2016.

Howard, G., Eiges, R., Gaudet, F., Jaenisch, R. & Eden, A. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. *Oncogene* **27**: 404–408. 2008.

Hsieh, J.J., Purdue, M.P., Signoretti, S., Swanton, C., Albiges, L., Schmidinger, M., *et al.* Renal cell carcinoma. *Nat. Rev. Dis. Prim.* **3**: 1–19. 2017.

Hu, Y., Hartmann, A., Stoehr, C., Zhang, S., Wang, M., Tacha, D., *et al.* PAX8 is expressed in the majority of renal epithelial neoplasms: an immunohistochemical study of 223 cases using a mouse monoclonal antibody. *J. Clin. Pathol.* **65**: 254–256. 2012.

Hunter, M. V., Moncada, R., Weiss, J.M., Yanai, I. & White, R.M. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nat. Commun. 2021 121* **12**: 1–16. 2021.

Hvichia, G.E., Parveen, Z., Wagner, C., Janning, M., Quidde, J., Stein, A., *et al.* A novel microfluidic platform for size and deformability based separation and the subsequent molecular characterization of viable circulating tumor cells. *Int. J. Cancer* **138**: 2894–2904. 2016.

Iacobuzio-Donahue, C.A., Michael, C., Baez, P., Kappagantula, R., Hooper, J.E. & Hollman, T.J. Cancer biology as revealed by the research autopsy. *Nat. Rev. Cancer* **19**: 686–697. 2019.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**: 163–166. 2014.

Jaiswal, S., Natarajan, P., Silver, A.J., Gibson, C.J., Bick, A.G., Shvartz, E., *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377**: 111–121. 2017.

Jakubek, Y.A., Chang, K., Sivakumar, S., Yu, Y., Giordano, M.R., Fowler, J., *et al.* Large-scale analysis of acquired chromosomal alterations in non-tumor samples from patients with cancer. *Nat. Biotechnol.* **38**: 90–96. 2020.

Jamal-Hanjani, M., Wilson, G.A., McGranahan, N., Birkbak, N.J., Watkins, T.B.K., Veeriah, S., *et al.* Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**: 2109–2121. 2017.

Janni, W., Vogl, F.D., Wiedswang, G., Synnestvedt, M., Fehm, T., Jückstock, J., *et al.* Persistence of disseminated tumor cells in the bone marrow of breast cancer patients predicts increased risk for relapse - A European pooled analysis. *Clin. Cancer Res.* **17**: 2967–2976. 2011.

Jolly, C. & Van Loo, P. Timing somatic events in the evolution of cancer. *Genome Biol.* **19**: 1–9. 2018.

Jonasch, E., Walker, C.L. & Rathmell, W.K. Clear cell renal cell carcinoma ontogeny and mechanisms of lethality. *Nat. Rev. Nephrol.* **17**: 245–261. 2021.

Jones, P.A. & Baylin, S.B. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **3**: 415–428. 2002.

Jones, R.C., Karkanias, J., Krasnow, M.A., Pisco, A.O., Quake, S.R., Salzman, J., *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**. 2022.

Juric, D., Castel, P., Griffith, M., Griffith, O.L., Won, H.H., Ellis, H., *et al.* Convergent loss of PTEN leads to clinical resistance to a PI(3)Kα inhibitor. *Nature* **518**: 240–244. 2015.

Kahn, J., Gillespie, A., Tsokos, M., Ondos, J., Dombi, E., Camphausen, K., *et al.* Radiation therapy in management of sporadic and neurofibromatosis type 1-associated malignant peripheral nerve sheath tumors. *Front. Oncol.* **4**: 324. 2014.

Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**: 89–94. 2018.

Kaufmann, T.L., Petkovic, M., Watkins, T.B.K., Colliver, E.C., Laskina, S., Thapa, N., *et al.* MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. *Genome Biol.* **23**: 241. 2022.

Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., *et al.* CUT&amp;Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**: 1930. 2019.

Khoja, L., Lorigan, P., Zhou, C., Lancashire, M., Booth, J., Cummings, J., *et al.* Biomarker utility of circulating tumor cells in metastatic cutaneous melanoma. *J. Invest. Dermatol.* **133**: 1582–1590. 2013.

Kim, S.M., Bhonsle, L., Besgen, P., Nickel, J., Backes, A., Held, K., *et al.* Analysis of the Paired TCR α- and β-chains of Single Human T Cells. *PLoS One* **7**: e37338. 2012.

Kishi, J.Y., Liu, N., West, E.R., Sheng, K., Jordanides, J.J., Serrata, M., *et al.* Light-Seq: light-directed in situ barcoding of biomolecules in fixed cells and tissues for spatially indexed sequencing. *Nat. Methods* **19**: 1393–1402. 2022.

Klein, C.A. Parallel progression of primary tumours and metastases. *Nat. Rev. Cancer* **9**: 302–312. 2009.

Klein, C.A., Blankenstein, T.J.F., Schmidt-Kittler, O., Petronio, M., Polzer, B., Stoecklein, N.H., *et al.* Genetic heterogeneity of single disseminated tumour cells in minimal residual cancer. *Lancet* **360**: 683–689. 2002.

Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H.W., Li, T., *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**: 661–671. 2022.

Knouse, K.A., Wu, J., Whittaker, C.A. & Amon, A. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc. Natl. Acad. Sci.* **111**: 13409–13414. 2014.

Knudson, A.G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proc. Natl. Acad. Sci.* **68**: 820–823. 1971.

Korbel, J.O. & Campbell, P.J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**: 1226–1236. 2013.

Kucab, J.E., Zou, X., Morganella, S., Joel, M., Nanda, A.S., Nagy, E., *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**: 821-836.e16. 2019.

Laks, E., Mcpherson, A., Zahn, H., Hansen, C., Shah, S.P. & Aparicio, S. Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell* **179**. 2019.

Lee, W., Teckie, S., Wiesner, T., Ran, L., Prieto Granada, C.N., Lin, M., *et al.* PRC2 is recurrently inactivated through EED or SUZ12 loss in malignant peripheral nerve sheath tumors. *Nat. Genet.* **46**: 1227–1232. 2014.

Letouzé, E., Allory, Y., Bollet, M.A., Radvanyi, F. & Guyon, F. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biol.* **11**: P25. 2010.

Lewinsohn, M.A., Bedford, T., Müller, N.F. & Feder, A.F. State-dependent evolutionary models reveal modes of solid tumour growth. *Nat. Ecol. Evol.* **7**: 581–596. 2023.

Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.

Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**: 112–121. 2020.

Lin, E., Rivera-Báez, L., Fouladdel, S., Yoon, H.J., Guthrie, S., Wieger, J., *et al.* High-Throughput Microfluidic Labyrinth for the Label-free Isolation of Circulating Tumor Cells. *Cell Syst.* **5**: 295-304.e4. 2017.

Litchfield, K., Stanislaw, S., Spain, L., Gallegos, L.L., Rowan, A., Schnidrig, D., *et al.* Representative Sequencing: Unbiased Sampling of Solid Tumor Tissue. *Cell Rep.* **31**. 2020.

Liu, S., Tian, Z., Zhang, L., Hou, S., Hu, S., Wu, J., *et al.* Combined cell surface carbonic anhydrase 9 and CD147 antigens enable high-efficiency capture of circulating tumor cells in clear cell renal cell carcinoma patients. *Oncotarget* **7**: 59877–59891. 2016.

Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C.C., *et al.* High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell* **183**: 1665-1681.e18. 2020.

Loeb, L.A. & Harris, C.C. Advances in Chemical Carcinogenesis: A Historical Review and Prospective. *Cancer Res.* **68**: 6863–6872. 2008.

Lomakin, A., Svedlund, J., Strell, C., Gataric, M., Shmatko, A., Rukhovich, G., *et al.* Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature* **611**: 594–602. 2022.

Ludwig, L.S., Lareau, C.A., Ulirsch, J.C., Christian, E., Muus, C., Li, L.H., *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**: 1325-1339.e22. 2019.

Luecken, M.D. & Theis, F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**. 2019.

Lyskjær, I., Lindsay, D., Tirabosco, R., Steele, C.D., Lombard, P., Strobl, A.C., *et al.* H3K27me3 expression and methylation status in histological variants of malignant peripheral nerve sheath tumours. *J. Pathol.* **252**: 151–164. 2020.

Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., *et al.* G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**: 519–522. 2015.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**: 1202–1214. 2015.

Maertens, Y., Humberg, V., Erlmeier, F., Steffens, S., Steinestel, J., Bögemann, M., *et al.* Comparison of isolation platforms for detection of circulating renal cell carcinoma cells. *Oncotarget* **8**: 87710–87717. 2017.

Makohon-Moore, A.P., Zhang, M., Reiter, J.G., Bozic, I., Allen, B., Kundu, D., *et al.* Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. *Nat. Genet.* **49**: 358–366. 2017.

Mallory, X.F., Edrisi, M., Navin, N. & Nakhleh, L. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biol.* **21**: 1–22. 2020.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**: 880–886. 2015.

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**: 555–572. 2020.

Mathiesen, R.R., Fjelldal, R., Liestøl, K., Due, E.U., Geigl, J.B., Riethdorf, S., *et al.* High-resolution analyses of copy number changes in disseminated tumor cells of patients with breast cancer. *Int. J. Cancer* **131**: 405–416. 2012.

McBride, K.A., Ballinger, M.L., Killick, E., Kirk, J., Tattersall, M.H.N., Eeles, R.A., *et al.* Li-Fraumeni syndrome: cancer risk assessment and clinical management. *Nat. Rev. Clin. Oncol.* **11**: 260–271. 2014.

McDermott, D.F., Regan, M.M., Clark, J.I., Flaherty, L.E., Weiss, G.R., Logan, T.F., *et al.* Randomized Phase III Trial of High-Dose Interleukin-2 Versus Subcutaneous Interleukin-2 and Interferon in Patients With Metastatic Renal Cell Carcinoma. *J. Clin. Oncol.* **23**: 133–141. 2005.

McGinnis, C.S., Murrow, L.M. & Gartner, Z.J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**: 329-337.e4. 2019.

McGranahan, N., Rosenthal, R., Hiley, C.T., Rowan, A.J., Watkins, T.B.K., Wilson, G.A., *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171**: 1259-1271.e11. 2017.

McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**: 613–628. 2017.

Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D.S., Kloiber, K., *et al.* Deep learning-based cell composition analysis from tissue expression profiles. *Sci. Adv.* **6**. 2020.

Merugu, S., Chen, L., Gavens, E., Gabra, H., Brougham, M., Makin, G., *et al.* Detection of circulating and disseminated neuroblastoma cells using the imageStream flow cytometer for use as predictive and pharmacodynamic biomarkers. *Clin. Cancer Res.* **26**: 122–134. 2020.

Miller, M.B., Huang, A.Y., Kim, J., Zhou, Z., Kirkham, S.L., Maury, E.A., *et al.* Somatic genomic changes in single Alzheimer's disease neurons. *Nature* **604**: 714–722. 2022a.

Miller, T.E., Lareau, C.A., Verga, J.A., DePasquale, E.A.K., Liu, V., Ssozi, D., *et al.* Mitochondrial variant enrichment from high-throughput single-cell RNA sequencing resolves clonal populations. *Nat. Biotechnol.* **40**: 1030–1034. 2022b.

Minussi, D.C., Nicholson, M.D., Ye, H., Davis, A., Wang, K., Baker, T., *et al.* Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature* **592**: 302–308. 2021.

Mitchell, T.J., Turajlic, S., Rowan, A., Nicol, D., Farmery, J.H.R., O'Brien, T., *et al.* Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. *Cell* **173**: 611-623.e17. 2018.

Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D., *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**. 2018.

Moffitt, J.R., Lundberg, E. & Heyn, H. The emerging landscape of spatial profiling technologies. *Nat. Rev. Genet.* **23**: 741–759. 2022.

Morgan, T.M., Lange, P.H., Porter, M.P., Lin, D.W., Ellis, W.J., Gallaher, I.S., *et al.* Disseminated tumor cells in prostate cancer patients after radical prostatectomy and without evidence of disease predicts biochemical recurrence. *Clin. Cancer Res.* **15**: 677–683. 2009.

Morton, D.L., Thompson, J.F., Cochran, A.J., Mozzillo, N., Elashoff, R., Essner, R., *et al.* Sentinel-Node Biopsy or Nodal Observation in Melanoma. *N. Engl. J. Med.* **355**: 1307–1317. 2006.

Nam, A.S., Chaligne, R. & Landau, D.A. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat. Rev. Genet.* **22**: 3–18. 2021.

Nam, A.S., Kim, K.-T., Chaligne, R., Izzo, F., Ang, C., Taylor, J., *et al.* Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571**: 355–360. 2019.

Naume, B., Synnestvedt, M., Falk, R.S., Wiedswang, G., Weyde, K., Risberg, T., *et al.* Clinical outcome with correlation to disseminated tumor cell (DTC) status after DTC-guided secondary adjuvant treatment with docetaxel in early breast cancer. *J. Clin. Oncol.* **32**: 3848–3857. 2014.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**: 90–94. 2011.

Naxerova, K. & Jain, R.K. Using tumour phylogenetics to identify the roots of metastasis in humans. *Nat. Rev. Clin. Oncol.* **12**: 258–272. 2015.

Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**: 835-849.e21. 2019.

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**: 773–782. 2019.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**: 979–993. 2012a.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., *et al.* The Life History of 21 Breast Cancers. *Cell* **149**: 994–1007. 2012b.

Nilsen, G., Liestøl, K., Loo, P. Van, Moen Vollan, H.K., Eide, M.B., Rueda, O.M., *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**. 2012.

Ning, L., Li, Z., Wang, G., Hu, W., Hou, Q., Tong, Y., *et al.* Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons. *Sci. Rep.* **5**: 11415. 2015.

Noble, R., Burri, D., Le Sueur, C., Lemant, J., Viossat, Y., Kather, J.N., *et al.* Spatial structure governs the mode of tumour evolution. *Nat. Ecol. Evol.* **6**: 207–217. 2021.

Nowell, P.C. The Clonal Evolution of Tumor Cell Populations. *Science* **194**: 23–28. 1976.

Pantel, K., Alix-Panabières, C. & Riethdorf, S. Cancer micrometastases. *Nat. Rev. Clin. Oncol.* **6**: 339–351. 2009.

Patel, S.A., Hirosue, S., Rodrigues, P., Vojtasova, E., Richardson, E.K., Ge, J., *et al.* The renal lineage factor PAX8 controls oncogenic signalling in kidney cancer. *Nature* **606**: 999–1006. 2022.

Pedersen, M., Larsen, A., Stoltenberg, M. & Penkowa, M. The role of metallothionein in oncogenesis and cancer prognosis. *Prog. Histochem. Cytochem.* **44**: 29–64. 2009.

Pemov, A., Li, H., Presley, W., Wallace, M.R. & Miller, D.T. Genetics of human malignant peripheral nerve sheath tumors. *Neuro-Oncology Adv.* **2**: i50–i61. 2020.

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G. & Sandberg, R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**: 1096–1098. 2013.

Piranlioglu, R., Lee, E., Ouzounova, M., Bollag, R.J., Vinyard, A.H., Arbab, A.S., *et al.* Primary tumor-induced immunity eradicates disseminated tumor cells in syngeneic mouse model. *Nat. Commun.* **10**: 1430. 2019.

Pliner, H.A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**: 983–986. 2019.

Pommier, A., Anaparthy, N., Memos, N., Kelley, Z.L., Gouronnec, A., Yan, R., *et al.* Unresolved endoplasmic reticulum stress engenders immune-resistant, latent pancreatic cancer metastases. *Science* **360**: eaao4908. 2018.

Potter, S.S. Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.* **14**: 479–492. 2018.

Rahrmann, E.P., Shorthouse, D., Jassim, A., Hu, L.P., Ortiz, M., Mahler-Araujo, B., *et al.* The NALCN channel regulates metastasis and nonmalignant cell dissemination. *Nat. Genet.* **54**: 1827–1838. 2022.

Rao, C., Bui, T., Connelly, M., Doyle, G., Karydis, I., Middleton, M.R., *et al.* Circulating melanoma cells and survival in metastatic melanoma. *Int. J. Oncol.* **38**: 755–760. 2011.

Reddy, E.P., Reynolds, R.K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300**: 149–152. 1982.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., *et al.* The human cell atlas. *Elife* **6**. 2017.

Risson, E., Nobre, A.R., Maguer-Satta, V. & Aguirre-Ghiso, J.A. The current paradigm and challenges ahead for the dormancy of disseminated tumor cells. *Nat. Cancer* **1**: 672–680. 2020.

Rodriguez-Brenes, I.A., Komarova, N.L. & Wodarz, D. Tumor growth dynamics: insights into evolutionary processes. *Trends Ecol. Evol.* **28**: 597–604. 2013.

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**: 1463–1467. 2019.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**: 176–182. 2018.

Rotem, A., Ram, O., Shoresh, N., Sperling, R.A., Goren, A., Weitz, D.A., *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**: 1165–1172. 2015.

Rowley, J.D. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature* **243**: 290–293. 1973.

Rubin, M.A., Putzi, M., Mucci, N., Smith, D.C., Wojno, K., Korenchuk, S., *et al.* Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. *Clin. Cancer Res.* **6**: 1038–45. 2000.

Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**: 547–554. 2019.

Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., Szabo, S., *et al.* High Frequency of Mutations of the PIK3CA Gene in Human Cancers. *Science* **304**: 554–554. 2004.

Sanchez-Luque, F.J., Richardson, S.R. & Faulkner, G.J. Analysis of somatic LINE-1 insertions in neurons. *Neuromethods* **131**: 219–251. 2017.

Schmidt-Kittler, O., Ragg, T., Daskalakis, A., Granzow, M., Ahr, A., Blankenstein, T.J.F., *et al.* From latent disseminated cells to overt metastasis: Genetic analysis of systemic breast cancer progression. *Proc. Natl. Acad. Sci.* **100**: 7737–7742. 2003.

Schumacher, S., Bartenhagen, C., Hoffmann, M., Will, D., Fischer, J.C., Baldus, S.E., *et al.* Disseminated tumour cells with highly aberrant genomes are linked to poor prognosis in operable oesophageal adenocarcinoma. *Br. J. Cancer* **117**: 725–733. 2017.

Schürch, C.M., Bhate, S.S., Barlow, G.L., Phillips, D.J., Noti, L., Zlobec, I., *et al.* Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell* **182**: 1341-1359.e19. 2020.

Schwartz, R. & Schäffer, A.A. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* **18**: 213–229. 2017.

Sharma, S., Kelly, T.K. & Jones, P.A. Epigenetics in cancer. *Carcinogenesis* **31**: 27–36. 2010.

Sheltzer, J.M., Ko, J.H., Replogle, J.M., Habibe Burgos, N.C., Chung, E.S., Meehl, C.M., *et al.* Single-chromosome Gains Commonly Function as Tumor Suppressors. *Cancer Cell* **31**: 240–255. 2017.

Shi, H., Williams, M.J., Satas, G., Weiner, A.C. & Mcpherson, A. Exploiting allele-specific transcriptional effects of subclonal copy number alterations for genotype-phenotype mapping in cancer cell populations. *bioRxiv*. 2023.

Shi, K., Wang, G., Pei, J., Zhang, J., Wang, J., Ouyang, L., *et al.* Emerging strategies to overcome resistance to third-generation EGFR inhibitors. *J. Hematol. Oncol.* **15**: 94. 2022.

Singer, J., Kuipers, J., Jahn, K. & Beerenwinkel, N. Single-cell mutation identification via phylogenetic inference. *Nat. Commun.* **9**: 5144. 2018.

Slyper, M., Porter, C.B.M., Ashenberg, O., Waldman, J., Drokhlyansky, E., Wakiro, I., *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**: 792–802. 2020.

Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I. & Forbes, S.A. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**: 696–705. 2018.

Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**: 209–216. 2015.

Spain, L., Coulton, A., Lobon, I., Rowan, A., Schnidrig, D., Shepherd, S.T.C., *et al.* Late-stage metastatic melanoma emerges through a diversity of evolutionary pathways. *Cancer Discov.*, doi: 10.1158/2159-8290.CD-22-1427. 2023.

Srivastava, S., Ghosh, S., Kagan, J. & Mazurchuk, R. The Making of a PreCancer Atlas: Promises, Challenges, and Opportunities. *Trends in Cancer* **4**: 523–536. 2018.

Steele, C.D., Tarabichi, M., Oukrif, D., Webster, A.P., Ye, H., Fittall, M., *et al.* Undifferentiated Sarcomas Develop through Distinct Evolutionary Pathways. *Cancer Cell* **35**: 441-456.e8. 2019.

Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**: 400–404. 2012.

Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., Di Bella, D.J., *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**: 313–319. 2021.

Stöber, M.C., González, R.C., Brückner, L., Conrad, T., Wittstruck, N., Szymansky, A., *et al.* Intercellular extrachromosomal DNA copy number heterogeneity drives cancer cell state diversity. *bioRxiv* 2023.01.21.525014. 2023.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**: 865–868. 2017.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazlsy, C., Thorne, N., *et al.* Relative impact of nucleotide and copy number variation on gene phenotypes. *Science* **315**: 848–853. 2007.

Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**: 719–724. 2009.

Stuart, T., Butler, A., Hoffman, P., Stoeckius, M., Smibert, P., Satija, R., *et al.* Comprehensive Integration of Single-Cell Data Resource Comprehensive Integration of Single-Cell Data. *Cell* **177**: 1888-1902.e21. 2019.

Sun, D., Xie, X.P., Chipman, M.E., Shern, J.F., Parada, L.F., Zhang, X., *et al.* Stem-like cells drive NF1-associated MPNST functional heterogeneity and tumor progression. *Stem Cell* **28**: 1397-1410.e4. 2021.

Sun, R., Hu, Z., Sottoriva, A., Graham, T.A., Harpak, A., Ma, Z., *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.* **49**: 1015–1024. 2017.

Sztanka-Toth, T.R., Jens, M., Karaiskos, N. & Rajewsky, N. Spacemake: processing and analysis of large-scale spatial transcriptomics data. *Gigascience* **11**: 1–14. 2022.

Tabin, C.J., Bradley, S.M., Bargmann, C.I., Weinberg, R.A., Papageorge, A.G., Scolnick, E.M., *et al.* Mechanism of activation of a human oncogene. *Nature* **300**: 143–149. 1982.

Takacova, M., Bartosova, M., Skvarkova, L., Zatovicova, M., Vidlickova, I., Csaderova, L., *et al.* Carbonic anhydrase IX is a clinically significant tissue and serum biomarker associated with renal cell carcinoma. *Oncol. Lett.* **5**: 191–197. 2012.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**: 377–382. 2009.

Tarabichi, M. ASCAT.sc. *https://github.com/VanLoo-lab/ASCAT.sc*. 2020.

Tarabichi, M. ClusterID-based consensus clustering. *https://github.com/galder-max/CICC*. 2018.

Tarabichi, M., Martincorena, I., Gerstung, M., Leroi, A.M., Markowetz, F., Spellman, P.T., *et al.* Neutral tumor evolution? *Nat. Genet.* **50**: 1630–1633. 2018.

Tarabichi, M., Salcedo, A., Deshwar, A.G., Ni Leathlobhair, M., Wintersinger, J., Wedge, D.C., *et al.* A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat. Methods* **18**: 144–155. 2021.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., *et al.* COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**: D941–D947. 2019.

The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**: 330–337. 2012.

The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**: 43–49. 2013.

The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**: 609–615. 2011.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Cancer Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**: 82–93. 2020.

Thorban, S., Roder, J.D. & Siewert, J.R. Detection of micrometastasis in bone marrow of pancreatic cancer patients. *Ann. Oncol.* **10**. 1999.

Thybusch-Bernhardt, A., Klomp, H.J., Maas, T., Kremer, B. & Juhl, H. Immunocytological detection of isolated tumour cells in the bone marrow of malignant melanoma patients: A new method for the detection of minimal residual disease. *Eur. J. Surg. Oncol.* **25**: 498–502. 1999.

Tickle, T., Tirosh, I., Georgescu, C., Brown, M. & Haas, B. inferCNV of the Trinity CTAT Project. *https://github.com/broadinstitute/inferCNV.* 2019.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**: 381–386. 2014.

Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**: 404–416. 2019.

Turajlic, S., Xu, H., Litchfield, K., Rowan, A., Chambers, T., Lopez, J.I., *et al.* Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. *Cell* **173**: 581-594.e12. 2018a.

Turajlic, S., Xu, H., Litchfield, K., Rowan, A., Horswell, S., Chambers, T., *et al.* Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell* **173**: 595-610.e11. 2018b.

Ulmer, A., Dietz, K., Hodak, I., Polzer, B., Scheitler, S., Yildiz, M., *et al.* Quantitative Measurement of Melanoma Spread in Sentinel Lymph Nodes and Survival. *PLoS Med.* **11**. 2014.

Ulmer, A., Dietz, K., Werner-Klein, M., Häfner, H.-M., Schulz, C., Renner, P., *et al.* The sentinel lymph node spread determines quantitatively melanoma seeding to non-sentinel lymph nodes and survival. *Eur. J. Cancer* **91**: 1–10. 2018.

van de Haar, J., Hoes, L.R., Roepman, P., Lolkema, M.P., Verheul, H.M.W., Gelderblom, H., *et al.* Limited evolution of the actionable metastatic cancer genome under therapeutic pressure. *Nat. Med.* **27**: 1553–1563. 2021.

Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., *et al.* A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**: 2247–2276. 2020.

Van der Auwera, G., O'Connor, B. & Safari, an O.M.C. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. *Genomics in the Cloud* 300. 2020.

Van Loo, P., Nordgard, S.H., Lingjærde, O.C., Russnes, H.G., Rye, I.H., Sun, W., *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**: 16910–16915. 2010.

Vanharanta, S. & Massagué, J. Cancer Cell Review Origins of Metastatic Traits. *Cancer Cell* **24**: 410–421. 2013.

Vendramin, R., Litchfield, K. & Swanton, C. Cancer evolution: Darwin and beyond. *EMBO J.* **40**: e108389. 2021.

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods* **16**: 987–990. 2019.

Voet, T., Kumar, P., Van Loo, P., Cooke, S.L., Marshall, J., Lin, M.L., *et al.* Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic Acids Res.* **41**: 6119–6138. 2013.

Vogelstein, B., Fearon, E.R., Hamilton, S.R., Kern, S.E., Preisinger, A.C., Leppert, M., *et al.* Genetic Alterations during Colorectal-Tumor Development. *N. Engl. J. Med.* **319**: 525–532. 1988.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A. & Kinzler, K.W. Cancer Genome Landscapes. *Science* **339**: 1546–1558. 2013.

Wala, J.A., Bandopadhayay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C., *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**: 581–591. 2018.

Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**: 155–60. 2014.

Weckermann, D., Polzer, B., Ragg, T., Blana, A., Schlimok, G., Arnholdt, H., *et al.* Perioperative Activation of Disseminated Tumor Cells in Bone Marrow of Patients With Prostate Cancer. *J. Clin. Oncol.* **27**: 1549–1556. 2009.

Weidele, K., Stojanović, N., Feliciello, G., Markiewicz, A., Scheitler, S., Alberter, B., *et al.* Microfluidic enrichment, isolation and characterization of disseminated melanoma cells from lymph node samples. *Int. J. Cancer* **241**: 232–241. 2019.

Went, P., Dirnhofer, S., Salvisberg, T., Amin, M.B., Lim, S.D., Diener, P.-A., *et al.* Expression of Epithelial Cell Adhesion Molecule (EpCam) in Renal Epithelial Tumors. *Am. J. Surg. Pathol.* **29**: 83–88. 2005.

Werner-Klein, M., Scheitler, S., Hoffmann, M., Hodak, I., Dietz, K., Lehnert, P., *et al.* Genetic alterations driving metastatic colony formation are acquired outside of the primary tumour in melanoma. *Nat. Commun.* **9**: 595. 2018.

Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**: 238–244. 2016.

Wu, S.Z., Al-Eryani, G., Roden, D.L., Junankar, S., Harvey, K., Andersson, A., *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**: 1334–1347. 2021a.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov.* **2**: 100141. 2021b.

Yang, D., Jones, M.G., Naranjo, S., Rideout, W.M., Min, K.H. (Joseph), Ho, R., *et al.* Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell* **185**: 1905-1923.e25. 2022.

Yates, L.R. & Campbell, P.J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13**: 795–806. 2012.

Yates, L.R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**: 751–759. 2015.

Yates, L.R., Knappskog, S., Wedge, D., Farmery, J.H.R., Gonzalez, S., Martincorena, I., *et al.* Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell* **32**: 169-184.e7. 2017.

Yu, L., Wang, X., Mu, Q., Sing, S., Tam, T., Shek, D., *et al.* scONE-seq: A single-cell multi-omics method enables simultaneous dissection of phenotype and genotype heterogeneity from frozen tumors. *Sci. Adv.* **9**. 2023.

Yuasa, T., Inoshita, N., Saiura, A., Yamamoto, S., Urakami, S., Masuda, H., *et al.* Clinical outcome of patients with pancreatic metastases from renal cell cancer. *BMC Cancer* **15**: 1–6. 2015.

Zaccaria, S. & Raphael, B.J. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat. Biotechnol.* **39**: 207–214. 2021.

Zahn, H., Steif, A., Laks, E., Eirew, P., Vaninsberghe, M., Shah, S.P., *et al.* Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods* **14**: 167–173. 2017.

Zekavat, S.M., Viana-Huete, V., Matesanz, N., Jorshery, S.D., Zuriaga, M.A., Uddin, M.M., *et al.* TP53-mediated clonal hematopoiesis confers increased risk for incident atherosclerotic disease. *Nat. Cardiovasc. Res.* **2**: 144–158. 2023.

Zhang, A.W., O'Flanagan, C., Chavez, E.A., Lim, J.L.P., Ceglia, N., McPherson, A., *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat. Methods* **16**: 1007–1015. 2019.

Zhang, D., Deng, Y., Kukanja, P., Agirre, E., Bartosovic, M., Dong, M., *et al.* Spatial epigenome–transcriptome co-profiling of mammalian tissues. *Nature* **616**: 113–122. 2023.

Zhao, T., Chiang, Z.D., Morriss, J.W., LaFave, L.M., Murray, E.M., Del Priore, I., *et al.* Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* **601**: 85–91. 2022.

Zheng, C., Wu, H., Jin, S., Li, D., Tan, S. & Zhu, X. Roles of Myc-associated zinc finger protein in malignant tumors. *Asia. Pac. J. Clin. Oncol.* **18**: 506–514. 2022.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**: 14049. 2017.

Zhou, Y., Bian, S., Zhou, X., Cui, Y., Wang, W., Wen, L., *et al.* Single-Cell Multiomics Sequencing Reveals Prevalent Genomic Alterations in Tumor Stromal Cells of Human Colorectal Cancer. *Cancer Cell* **38**: 818-828.e5. 2020.

Zimpfer, A., Maruschke, M., Rehn, S., Kundt, G., Litzenberger, A., Dammert, F., *et al.* Prognostic and diagnostic implications of epithelial cell adhesion/activating molecule (EpCAM) expression in renal tumours: a retrospective clinicopathological study of 948 cases using tissue microarrays. *BJU Int.* **114**: 296–302. 2014.

Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**: 1622–1626. 2012.