Subclonal reconstruction of tumors using machine learning and population genetics

Giulio Caravagna[1], Timon Heide[1], Marc J. Williams[2], Luis Zapata[1], Daniel Nichol[1], Ketevan Chkhaidze[1], William Cross[2], George D. Cresswell[1], Benjamin Werner[1], Ahmet Acar[1], Louis Chesler[3], Chris P. Barnes[4], Guido Sanguinetti[5,6], Trevor A. Graham[2,§], Andrea Sottoriva[1,§].

[1]Evolutionary Genomics and Modelling Lab, Centre for Evolution and Cancer, The Institute of Cancer Research, London SM2 5NG, UK.
[2]Evolution and Cancer Lab, Barts Cancer Institute, School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, UK.
[3]Division of Clinical Studies, The Institute of Cancer Research, London SM2 5NG, UK.
[4]Department of Cell and Developmental Biology and UCL Genetics Institute, University College London, London WC1E 6BTCL, UK.
[5]School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB.
[6] International School for Advanced Studies - SISSA, Via Bonomea 265, Trieste 34136, IT.

[§]correspondence to Trevor A. Graham (t.graham@qmul.ac.uk) and Andrea Sottoriva (andrea.sottoriva@icr.ac.uk).

Abstract

The majority of cancer genomic data are generated from bulk samples composed of mixtures of cancer subpopulations, as well as normal cells. Subclonal reconstruction approaches based on machine learning aim to separate those subpopulations in a sample and reconstruct their evolutionary history. However, current approaches are entirely data-driven and agnostic to evolutionary theory. We demonstrate that systematic errors occur in the analysis if evolution is not accounted for, and this is exacerbated by multi-sampling of the same tumor. We present a novel approach for model-based tumor subclonal reconstruction (MOBSTER) that combines machine learning with theoretical population genetics. Using public whole-genome sequencing data from 2,606 samples from different cohorts, new data and synthetic validation, we show this method is more robust and accurate than current techniques in single sample, multi-region and longitudinal data. This approach minimizes the confounding factors of non-evolutionary methods, leading to more accurate recovery of the evolutionary history of human cancers.

Introduction

Cancers change over time through a process of clonal evolution[1], inevitably resulting in intra-tumor heterogeneity[2]. Genome sequencing of one or more bulk samples from tumors has become the most common way to study clonal evolution in human malignancies, and studies are dedicated to the identification of cancer (sub)clones[3]. A cancer "clone" remains a loosely defined entity, and its purest definition is "a group of cells within the tumor that share a common ancestor". In phylogenetic terms, this would represent a monophyletic clade. However, this implies that any ancestor in the entire phylogenetic tree of a tumor can be identified as the founder of a distinct "clone", even though it may show no biological difference from the rest of the cancer cells. This is why in the field we implicitly identify clones "of interest", such as those that have growth/survival advantage (an ancestor under positive selection), or those that generate metastases (an ancestor that arrived and grew at a given metastatic site). The limits in the definition of a clone are important to bear in mind when attempting to recover the tumor clonal architecture.

To identify clones in bulk cancer samples, the established approach is unsupervised clustering of variant read counts[4], with each of the resulting clusters defined as a clone. This procedure, called "subclonal reconstruction", leverages on variant read counts and associated variant allele frequency (VAF) of somatic mutations, adjusted for copy number status and tumor purity, to identify groups of variants with similar cellular proportions. Subclonal reconstruction allows tracing the "life history" of a tumor via determination of its phylogenetic tree (sometimes called a "clone tree")[3].

Current methodologies approach subclonal reconstruction with sophisticated mixture models[4], implemented via Dirichlet Processes[3,5,6] or Dirichlet finite mixtures[7]. These machine learning methods are entirely data-driven and are usually chosen because of their convenient statistical properties, rather than their adherence to the mechanisms of tumor evolution. They can be efficient and accurate, as long as the underlying assumptions are correct. All current subclonal reconstruction methods assume that variant read counts from bulk tumor samples present as a mixture of Binomial or Beta-Binomial mutational clusters, each one corresponding to a clone. However, these are not the only observable patterns in the data: the mutations that occur within each clone while it expands are also detectable. Given the size of the human genome, even with low mutations rates (e.g. 10^{-9}- nucleotide substitutions per base per division[8]), new mutations are expected at each cell division, and thus large numbers of passenger mutations inevitably accumulate within an expanding clone. The evolutionary dynamics of this passenger mutation accumulation are *neutral*, and give rise to a power-law distributed "tail" of ever more mutations at ever lower frequency. This has been mathematically demonstrated in theoretical population genetics[9-14] and is corroborated by genomic data at high resolution[15,16]. These within-clone neutral tails have not been directly addressed by previous methods, potentially confounding the measurement of clonal heterogeneity.

Here, we reconciled data-driven machine learning approaches to clustering VAFs and corresponding Cancer Cell Fractions (CCF), with the insight given by evolutionary theory. Specifically, we combined Dirichlet mixture models with the distributions predicted by theoretical population genetics models[9-12], producing a model-based method for subclonal reconstruction called MOBSTER (MOdel Based cluSTering in cancER). MOBSTER can process mutant allelic frequencies to identify and remove neutral tails from the input data, so that machine-learning subclonal reconstruction algorithms can be applied downstream to find subclones from read counts. We also expanded MOBSTER to analyze data from multiple samples of the same tumor, collected both in space and time.

Results

Mutation, drift and selection

Cancers grow from a single cell, and hence neutral mutations that occur in the first few cell divisions are present at high frequency in the final population, irrespective of the action of selection. In addition, stochastic fluctuations in population size of cell lineages can also increase the frequency of mutations in the absence of selection; this is called genetic drift[17]. The same is true within (sub)clones: a clone originates as a single cell, and neutral mutations that occur early within the clone are found in a large proportion of the clone's cells. Fundamental insight into the accumulation of mutations in the absence of positive selection came from the study of the Luria-Delbruck model in bacteria[18]. This has led to well-established population genetics theory describing the accumulation of mutations within neutrally growing populations[10,11]. The same theory applies to cancer clones[9,12] and can be extended to include positive selection[16]. Theory states that we should expect a tail of neutral passenger mutations within a clone (Figure 1a). Neutral tails only recently became evident in cancer data with the adoption of high-depth whole genome sequencing (WGS), as lower depth sequencing (e.g. <60x) is insufficient to detect tails reliably[16], and exome or panel sequencing often assay too few mutations to show a clear VAF spectrum.

Figure 1a shows the simplest example of a uniform 'neutral' tumor expansion. The corresponding clone tree has a single "truncal" node (Figure 1b). The VAF spectrum for this tumor consists of a "clonal peak" at high frequency, corresponding to the mutations that are present in all cells (i.e. in the most recent common ancestor, MRCA), and a neutral tail of mutations at lower VAF generated as the clone expands (Figure 1c). In the case where a subclone with selective advantage is present (Figure 1d,e), the data will present as two peaks at high frequency (one clonal and one subclonal) as well as a mixture of two overlapping neutral tails[16] (Figure 1f). Performing subclonal reconstruction on these data assuming a generative mixture of just Binomial or Beta-Binomial distributions will detect several clusters within the neutral tail that are erroneously identified as subclones, as illustrated in two simulated cases (neutral in Figure 1g, and with one selected subclone in Figure 1h). Importantly, mutations in neutral tails are not monophyletic, and hence grouping them together into clones is erroneous even under the strictest definition of a clone. Moreover, when these incorrect clones are used downstream for phylogenetic reconstruction, the resulting trees (Figure 1i) have a very different structure from the true trees (Figure 1b,e), thus propagating errors and uncertainty in the tree construction, with many equivalent (but wrong) trees potentially fitting the same data.

Moreover, low-depth sequencing and low purity data cause neutral tails to be under-sampled and likely to be mistaken for subclones, as they lose their characteristic power-law shape. Simulated WGS data (Figure 1j) show that with low coverage or purity, the signal of a neutral tail becomes statistically difficult to distinguish from that of a selected subclonal cluster (Figure 1k). This observation indicates that sequencing depth below 90x/100x and low purity prevents reliable subclonal reconstruction. We note that patterns of noisy subclonal

VAF distributions that may represent under-sampled tails (e.g. Figure 1k), are commonly observed in cancer sequencing data at depth <90x/100x.

Model-based clustering of variant allelic frequencies

The frequency f of newly acquired passenger mutations in an expanding population follows a Landau distribution[10], which at the frequency range detected by current sequencing standards can be approximated by a power law distribution $X{\sim}1/f^2$ (Figure 2a), as we previously reported[9]. Subclonal alleles under positive selection, together with their hitchhiking passengers, will instead form clusters in the VAF distribution as they rise in frequency due to positive selection[16,19].

We can model VAFs or fraction data via Beta distributions[7], and model read counts with Binomial or Beta-Binomial distributions[3,5-7]. In MOBSTER (Figure 2a), we model the evolutionary dynamics of a growing tumor containing subclones by combining Beta distributions (expected from subclones under selection) with a power law (expected from neutral tails). After fitting the VAF distribution, tail mutations can be removed and clustering of read counts from the remaining mutations can be performed via standard methods (Figure 2b). MOBSTER controls for tails while retaining the original variance of the data when clustering non-tail read counts downstream. Notably, MOBSTER always compares the fit of a mixture of clones with and without a neutral tail and uses a regularized model selection strategy to determine the best model fit to the data.

MOBSTER combines one Pareto Type-I random variable (a type of power-law) with k Beta random variables, resulting in a univariate finite mixture with $k + 1$ components. The likelihood for n datapoints x_i is

$$p(D|\boldsymbol{\theta},\boldsymbol{\pi}) = \prod_{i=1}^{n}\left[\pi_1 g(x_i|x_*,\alpha) + \sum_{w=2}^{k}\pi_w h(x_i|a_{w-1},b_{w-1})\right],$$

where g and h are density functions, $\boldsymbol{\theta} = (x_*,\alpha,a_1,..,a_k,b_1,\dots,b_k)$ is a vector of parameters and $\boldsymbol{\pi}$ are mixing proportions in a standard setting with $n \times (k+1)$ latent variables. The Pareto component follows $g(x|x_*,\alpha) \propto 1/x^{1+\alpha}$ for $x \geq x_*$, and the Beta follows $h(x|a,b) \propto x^{a-1}(1-x)^{b-1}$ in [0,1]. A derivation of MOBSTER, its relation to other approaches and technical comments are available in the Online Methods.

In the hypothetical example of a "functionally monoclonal" tumor with neutral subclonal dynamics (Figure 1a), MOBSTER fits $k = 1$ Beta clusters of truncal mutations (present in all cancer cells) plus a neutral tail (Figure 2c). Similarly, for a tumor with one selected subclone (Figure 1d), MOBSTER fits $k = 2$ Beta clusters and a tail (Figure 2d). When we identify and remove tail mutations from the data, subsequent clustering of read counts mutations identifies the true tumor clones and their correct clone trees (inner clone tree panels).

Synthetic validation of the method and confounding factors

We used synthetic data to validate MOBSTER and quantify the degree to which neutral tails confound subclonal deconvolution with standard methods (Supplementary Note, Supplementary Figures 1-9). We used a stochastic branching process[16] to simulate the growth of $n = 150$ tumors (Online Methods and Supplementary Data vignette "Example Subclonal Dynamics"). Out of these 150 cases, 30 tumors were neutral (as Figure 1a) and 120 contained one selected subclone (as Figure 1d). For each tumor we simulated bulk WGS at 120x median coverage and 100% purity. In every test, we always compared the fit of MOBSTER with and without a tail, retaining the best; we then recorded the predicted number of selected clones, k, and the fit precision (Supplementary Figure 3 and 4). We note that by applying further population genetics theory[16] to the output of MOBSTER, we can estimate the tumor evolutionary parameters, such as the mutation rate, the time of emergence of subclones, and their selection coefficients (Supplementary Figure 5). We also carried out several other tests for the detection of low-frequency subclones admixed with tails (Supplementary Figures 6 and 7).

By accounting for neutral tails, MOBSTER significantly outperformed standard approaches based on both Dirichlet variational mixtures and Dirichlet Processes (Extended Data Figure 1), two statistical frameworks at the core of subclonal reconstruction tools like sciClone[7], pyClone[5], DPclust[3] and many others. Results are consistent for various parameterizations, in particular of the concentration parameter $\alpha > 0$, which determines the propensity of adding clusters to the fit[3]. In Figure 2e we report the error rates for the inferred number of clones (k) with DPclust, pyClone (Binomial and Beta-Binomial) and sciClone. The detection of spurious extra clusters caused high uncertainty around the clone tree, with many solutions fitting the data equally well (Figure 2f). We tested the effects of sequencing coverage and purity on tail detection, and found that ~100x coverage and high purity were required to systematically identify tails. Higher coverage is required for samples with lower purity (Extended Data Figure 1). Additional synthetic tests with complex clonal architectures confirmed the robustness of the method (Supplementary Figures 8 and 9). These analyses indicate that the previously published moderate-depth WGS studies were underpowered to detect reliable subclonal architectures, since the signal used to distinguish a tail from a subclone deteriorates with lower sequencing depth (Figure 1j). With adequate data and controlling for neutral tails, we found the correct number of clones in the large majority of tests. Not considering neutral tails led to a systematic pattern of errors that, in the worst cases, could lead to a four-fold overestimation of the number of clones.

Not accounting for neutral tails also significantly impacts multi-region sequencing, as we discuss in the Supplementary Note. We found that multi-region bulk sequencing is affected by confounders that originate from the spatial effects of tumor growth and spatial sampling bias. In multi-sample analyses (Supplementary Note) we characterized a confounder termed the "hitchhikers mirage" (Extended Data Figure 2) caused by parts of neutral tails that spread in space, and that current methods mistake for selected subclones (Supplementary Figure 10). We also characterized two additional confounders due to the presence of locally sampled ancestors (Extended Data Figure 3) and admixing of multiple lineages (Extended Data Figure 4). These spatial confounders affect virtually all tumors (Supplementary Figures 11-13). Therefore, the joint use of MOBSTER and other heuristics is necessary to interpret subclonal deconvolution results from multi-region samples (Extended Data Figure 5, Supplementary Figure 14).

Analysis of genomic data from human samples

We applied MOBSTER to high coverage (>100x) WGS data available in the public domain (Supplementary Note). We first re-analyzed the breast cancer sample PD1420a sequenced at ~188x from Nik-Zainal *et al.*[3]. Compared to the original analysis, which found 3 subclones, MOBSTER fits two subclones ($k = 3$) and places a neutral tail for the lowest frequency cluster (Figure 3a). sciClone analysis of read counts for non-tail mutations confirmed $k = 3$ Binomial clusters (2 selected subclones). Both linear and branching phylogenies could be fit to the output, with the branching tree matching the original analysis[3]. The cluster that MOBSTER fits to a tail appears in multiple positions of the tumor tree in the original paper after phasing[3]. This is consistent with our analysis, as the tail is polyphyletic, and hence composed of a mixture of descendants of the different clones. We measured the evolutionary parameters of this tumor from the fits, finding concordant estimates with our previous work[16]. Mutation rate was $\mu = 3.5 * 10^{-7}$ mutations per base per tumor doubling, subclones emerged at $t = 5.5$ (smaller subclone) and $t = 10.4$ (larger subclone) doublings, and had selective coefficients of $s = 0.3$ and $s = 0.66$ respectively.

We reanalyzed the acute myeloid leukemia (AML) sample sequenced at 320x WGS by Griffith *et al.*[20]. MOBSTER identifies $k = 3$ clusters (2 subclones) and a neutral tail (Figure 3b). The two subclones were also detected by Griffith *et al.*[20], and were confirmed running sciClone after MOBSTER. However, MOBSTER simplified the clonal architecture by removing one spurious low-frequency "subclone". This observation likely improves the interpretation of these data, possibly explaining why the tail was the only cluster without a clear subclonal driver mutation. Measured mutation rate was $\mu = 9.9 * 10^{-10}$ per base per tumor doubling, subclones emerged at $t = 22$ and $t = 27$, and selection coefficients were $s = 1.3$ and $s = 3$, respectively.

We also generated new multi-region WGS data (median 100x) from spatially separated regions of two primary colorectal cancers previously analyzed at lower depth in Cross *et al.*[21]. In tumor Set06 we analyzed high-confidence single nucleotide variants (SNVs) in diploid segments consistent across samples, and ran a comparative analysis with and without MOBSTER (Supplementary Note). The analysis with MOBSTER did not find evidence of positive subclonal selection (Figure 3c, Supplementary Figure 15), corroborated by the lack of subclonal drivers and truncal *APC*, *KRAS*, *SMAD3* and *TP53* mutations, as originally reported[21]. The analysis without MOBSTER would have depicted a complex subclonal structure, with several Binomial clusters consistent with multiple clone trees (Supplementary Figure 16). The analysis of Set06 gave similar results (Figure 3d, Supplementary Figure 17). Consistent with Cross *et al.*[21], the clone tree depicted a tumor with only truncal driver events in *APC*, *KRAS*, *PIK3CA*, *ARID1A* and *TCF7L2*, and neutral subclonal dynamics. Again, a standard analysis would have identified a complex clonal architecture with multiple subclones (Supplementary Figure 18). Mutation rates were $\mu = 5.6 * 10^{-7}$ for Set07, and $\mu = 4.3 * 10^{-7}$ for Set06. Notably, orthogonal dN/dS analysis that uses the ratio of non-synonymous to synonymous mutations to detect selection[22,23], confirmed the lack of evidence for positive selection at the subclonal level in those tumors (Figure 3e, Supplementary Note).

We also applied MOBSTER to $n = 3$ non-small cell lung cancer samples sequenced at high depth (Figure 3f). These three tumors were those with the highest coverage and purity amongst a recently published cohort[24] (see also low-purity cases in Supplementary Figure 19).

Neutral evolution in 2,566 whole-genomes from PCAWG

We reanalyzed with MOBSTER one of the largest available cohorts of cancer WGS data to date, collated by the Pan-Cancer Analysis of Whole Genomes (PCAWG) international consortium and recently published in a series of studies[25], including the evolutionary history of more than 2,600 cancers[26]. The median depth of coverage in this dataset was 45x, with median purity of 65%. According to our power analysis, data at this resolution are not suitable for reliable subclonal reconstruction (Figure 1j, 1k and Extended Data Figure 1). Figure 4a shows a PCAWG case where a standard analysis called a selected subclone. The coverage was 55x and purity 66%, with a VAF distribution similar to the down-sampled synthetic neutral cases shown in Figure 1j. With these data, MOBSTER (Figure 4b, more cases in Supplementary Figure 20) cannot fit a neutral tail in the low-frequency portion of the VAF spectrum, and instead fits a subclone (Beta component). The ground-truth is not known, but given the resolution of the data we cannot exclude the likelihood that subclonal mutations in this sample are the result of a degenerate neutral tail (see Figure 1j,k). In cases where coverage and purity were higher, MOBSTER did identify neutral tails and resolved the remaining clonal structure (Figure 4c). As expected, standard approaches would have identified spurious clusters (Figure 4d), thus compromising the whole subclonal reconstruction.

We found widespread presence of neutral evolutionary patterns in PCAWG data using MOBSTER. We analyzed the VAF spectrum of 2,566 cancers (Supplementary Note). Theoretical population genetics predicts that, given enough power in the data, we should always expect to find a neutral tail, with or without selected subclones (Figure 2a). However, we consistently found neutral tails only in samples with higher coverage and purity (Figure 4e, red=cases with neutral tail, blue=cases without detectable tail), suggesting lack of power for subclonal inference in the majority of cases (Supplementary Figure 21).

To further validate the presence of neutral tail mutations in this cohort, we focused on $n = 902$ near-diploid cancers with >30x depth, >65% purity and where a tail was detected. From these cases we identified somatic mutations mapping to putative cancer driver genes[25,26] in neutral tails versus non-tail and performed dN/dS analysis[22] (Figure 4f). This orthogonal measurement confirmed that mutations in tails were likely neutral (dN/dS~1), aside from the caveats of interpreting dN/dS values in growing tumours[27], whereas non-tail mutations indicated selection (dN/dS>1).

We then focused on $n = 298$ diploid cases that were found to have at least 10% of the total mutation burden in the tail, indicating sufficient power to detect the clonal architecture with confidence. We measured the proportion of tumors with a selected subclone, defined by 2 or more Binomial clusters detected from non-tail mutations. We found evidence of ongoing subclonal selection only in $n = 9$ (3% of total, Supplementary Figure 22). In the remaining $n = 289$ cases, neutral evolutionary dynamics at the subclonal level were the adequate description of the data (Figure 4g). Lowering the threshold for proportion of tail mutations did not change the results (5% tail = 2.7% non-neutral cases; 2% tail = 3.7% non-neutral cases).

Our analysis suggests that for the majority of PCAWG cases, the data resolution was too low to conduct robust subclonal reconstruction. Moreover, neutral tails were detectable in higher coverage and purity samples, indicating that neutral dynamics are often an adequate description of the observed subclonal heterogeneity. Standard analyses of these data therefore risk systematically mistaking neutral tails for subclonal clusters, thus inflating the complexity of the inferred subclonal architectures and producing incorrect phylogenetic trees. Our analysis using MOBSTER hence demonstrates that neutral evolutionary patterns are prevalent in PCAWG data.

Analysis of longitudinal whole-genome datasets

We analyzed a cohort of $n = 35$ matched primary-relapse glioblastoma samples from 16 patients profiled using ~100x WGS in a recent study by Körber et al. 2019[28]. Our analysis identified 9 cases characterized only by neutral evolutionary dynamics at the subclonal level in both primary and relapse, while 7 patients had a detectable ongoing subclonal expansion (Supplementary Figure 23). We found cases where positively selected subclones were unique to the primary or the relapse (Figure 5a,b), but also cases where pre-existing subclones in the primary swept through the population in the relapse, likely due to positive selection from treatment (Figure 5c,d). In some cases, we found evidence of novel subclones at relapse (Figure 5e,f). MOBSTER also identified clusters of mutations that were due to whole-genome duplications, as in the case of a diploid primary tumor that became tetraploid at relapse (Figure 5g,h). We note that some of the confounding effects of neutral tails in multivariate analyses (Supplementary Note) were ubiquitous in these data and would have negatively impacted standard subclonal reconstruction (Supplementary Figure 23). Orthogonal analysis with dN/dS[22,23] methods suggested neutral values for tail mutations (dN/dS ~1) and positive selection for others (dN/dS >1) using a panel of glioma driver genes (Figure 5h). We note that the presence of subclones under positive selection in these data was also reported in the original study[28]. However, using MOBSTER we obtained simplified clonal architectures, pruning some of the clusters that were due to neutral tails. Indeed, a mixture of subclonal selection and neutral evolutionary dynamics through therapy has been recently reported in a large glioblastoma study[29].

Discussion

Subclonal reconstruction from cancer bulk sequencing data has paved the way to the study of cancer evolution[3,30]. Measurement of subclonal architectures have also clinical relevance: subclone multiplicity and other measures of intra-tumor heterogeneity have been reported as prognostic biomarkers[31-34]. Naturally therefore, there is the need to ensure that subclonal reconstruction is accurate.

Here we have presented a subclonal reconstruction method that combines data-driven machine learning with theoretical population genetics. This is in contrast to purely data-driven approaches that lack an underlying evolutionary model. Recently proposed standards for subclonal reconstruction[35] do not account for evolutionary dynamics, and hence this recommended best practice analysis is inherently flawed.

Moreover, we suggest that only high depth sequencing data of >90/100x is appropriate to infer subclonal architectures, and even higher depth is required for purity <75%. Subclonal reconstruction from lower depth data and lack of consideration for neutral tails risks a systematic over-calling of spurious subclones (Figure 1j,k), leading to incorrect inference of the life history of tumors. These problems affects multiple previously published studies (for example refs[3,34,36]) and prohibit the inference of subclonal structures in the large majority of PCAWG cases. Various issues arise also in multi-region sequencing data, resulting from biases that are intrinsic to spatial sampling (Supplementary Note) and thus affect several previous studies that had insufficient depth of sequencing to infer metastatic spread (for example refs[37-39]). These issues also lead to inflated estimates of positive subclonal selection from VAF distributions. Single-cell sequencing removes the problem of admixing of populations[40], however the underlying evolutionary dynamics described by theory remain valid for the frequency of mutations amongst the *N* cells sequenced[41].

The major impact of MOBSTER is that it controls for neutrally evolving cancer cell subpopulations, cleaning up the signal for downstream analyses that seek to focus on "functional" intra-tumor heterogeneity. Given the wide use of clustering methods for subclonal reconstruction, MOBSTER has the potential to impact intra-tumor heterogeneity studies that use bulk sequencing, and even that analyze the distribution of clade sizes in single-cell sequencing.

We also highlight the limitations of the definition of "clone" in cancer as a monophyletic clade with a most recent common ancestor, noting that in the clinic we are not interested in all the ancestors of a given group of cancer cells, but only in those few ancestors that drive progression, metastasis or treatment resistance. Importantly, even under this looser definition of a clone, clustering neutral tails with Binomial models is incorrect and leads to the identification of false clones, mistaking the polyphyletic branching process that gives rise to neutral tails for a monophyletic lineage.

This study highlights that there are intrinsic limitations to the information on tumor evolution encoded in current data, foremost because of the systematic confounding factors caused by sampling complex three-dimensional tumors. We propose that our analysis represents a step towards a more refined approach to subclonal reconstruction in bulk cancer data, a necessity for genomic-aided precision medicine.

Acknowledgements

A.S. is supported by the Wellcome Trust (202778/B/16/Z) and Cancer Research UK (A22909). T.G. is supported by the Wellcome Trust (202778/Z/16/Z) and Cancer Research UK (A19771). We acknowledge funding from the Medical Research Council (MR/P000789/1) to A.S. and from the National Institute of Health (NCI U54 CA217376) to A.S and T.A.G. C.P.B. acknowledges funding from the Wellcome Trust (209409/Z/17/Z). L.C. acknowledges funding by Cancer Research UK (A24566) and Children with Cancer UK

(17-235). This work was also supported a Wellcome Trust award to the Centre for Evolution and Cancer (105104/Z/14/Z). L.Z. is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Research Fellowship scheme (846614). We thank Nik Matthews and the Tumour Profiling Unit at the ICR for their support with Next- Generation Sequencing. We wish to thank Verena Körber and Thomas Höfer for sharing their data and for the fruitful discussion around the glioblastoma cohort.

Authors contribution

GC conceived, designed and implemented the method. TH and KC developed the spatial tumor growth simulations. TH and MW generated the data for synthetic tests, which were carried out and analyzed by GC, TH, MW and DN. GC, MW and LZ analyzed the data, with input and support from WC, GDC and AA. GS, CB, TAG and AS supervised method design. LC contributed to study supervision. AS and TAG conceived and supervised the study. All authors contributed to and approved the manuscript.

References (Main Text)

1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012).
2. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **27,** 1 (2019).
3. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
4. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harb Perspect Med* **7,** a026625 (2017).
5. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat Meth* **11,** 396–398 (2014).
6. Deshwar, A. G. *et al.* PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16,** 35 (2015).
7. Miller, C. A. *et al.* SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Comput. Biol.* **10,** e1003665 (2014).
8. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17,** 704–714 (2016).
9. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature Genetics* **48,** 238–244 (2016).
10. Kessler, D. A. & Levine, H. Large population solution of the stochastic Luria-Delbruck evolution model. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 11682–11687 (2013).
11. Kessler, D. A. & Levine, H. Scaling solution in the large population limit of the general asymmetric stochastic Luria-Delbrück evolution process. *J Stat Phys* **158,** 783–805 (2015).
12. Durrett, R. Population genetics of neutral mutations in exponentially growing cancer cell populations. *The Annals of Applied Probability* **23,** 230–250 (2013).
13. Nicholson, M. D. & Antal, T. Universal Asymptotic Clone Size Distribution for General Population Growth. *Bull Math Biol* **78,** 2243–2276 (2016).
14. Griffiths, R. C. & Tavaré, S. The age of a mutation in a general coalescent. *Communications in Statistics. Part C: Stochastic Models* **14,** 273–295 (1998).
15. Sun, R. *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nature Genetics* **49,** 1015–1024 (2017).
16. Williams, M. J. *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics* **50,** 895–903 (2018).
17. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics*. (Sinauer Associates, Inc., 2006).
18. Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28,** 491–511 (1943).

19. Graham, T. A. & Sottoriva, A. Measuring cancer evolution from the genome. *J. Pathol.* **241,** 183–191 (2017).

20. Griffith, M. *et al.* Optimizing Cancer Genome Sequencing and Analysis. *Cell Systems* **1,** 210–223 (2015).

21. Cross, W. *et al.* The evolutionary landscape of colorectal tumorigenesis. *Nat. ecol. evol.* **2,** 1661–1672 (2018).

22. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171,** 1–13 (2017).

23. Zapata, L. *et al.* Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol.* **19,** 924 (2018).

24. Lee, J. J.-K. *et al.* Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma. *Cell* **177,** 1842–1857.e21 (2019).

25. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578,** 82–93 (2020).

26. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578,** 122–128 (2020).

27. Williams, M. J. *et al.* Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dN/dS ratios. *eLife Sciences* **9,** 612 (2020).

28. Körber, V. *et al.* Evolutionary Trajectories of IDHWT Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. *Cancer Cell* **35,** 692–704.e12 (2019).

29. Barthel, F. P. *et al.* Longitudinal molecular trajectories of diffuse glioma in adults. *Nature* **135,** 1–9 (2019).

30. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–399 (2012).

31. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22,** 105–113 (2016).

32. Morris, L. G. T. *et al.* Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget* **7,** 10051–10063 (2016).

33. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *New England Journal of Medicine* **376,** 2109–2121 (2017).

34. Espiritu, S. M. G. *et al.* The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. *Cell* **173,** 1003–1013.e15 (2018).

35. Salcedo, A. *et al.* A community effort to create standards for evaluating tumor subclonal reconstruction. *Nature Biotechnology* **38,** 97–107 (2020).

36. Yang, L. *et al.* An enhanced genetic model of colorectal cancer progression history. *Genome Biol.* **20,** 1–17 (2019).

37. Yates, L. R. *et al.* Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell* **32,** 169–184.e7 (2017).

38. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520,** 353–357 (2015).

39. Noorani, A. *et al.* Genomic evidence supports a clonal diaspora model for metastases of esophageal adenocarcinoma. *Nature Genetics* **347,** 1–10 (2020).

40. Navin, N. E. The first five years of single-cell cancer genomics and beyond. *Genome Res.* **25,** 1499–1507 (2015).

41. Chkhaidze, K. *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).

Figure Legends

Figure 1. Theoretical predictions of cancer genomic data under different evolutionary dynamics. (a) A tumor formed by a single “functionally monoclonal” expansion follows neutral evolutionary dynamics driven only by mutation and drift. **(b)** The clone tree can be represented as a single “truncal” clone. **(c)** In diploid regions, the Variant Allele Frequency (VAF) distribution is characterized by one clonal cluster and a neutral $1/f^2$ tail of subclonal mutations. **(d)** In a tumor with one subclone under positive selection (functionally polyclonal) the evolutionary forces of mutation and drift are still at play within each clone. **(e)** The clone tree is represented as a truncal node giving rise to a selected subclone within it. **(f)** The VAF shows one extra cluster due to subclonal mutations in the subclone that have risen in frequency due to selection. **(g,h)** Standard subclonal deconvolution identifies clusters of neutral tail mutations that are not subclones, as they represent admixed polyphyletic lineages. **(i)** This causes inflated estimates of the number of clones that propagate errors and uncertainty downstream, with several incorrect phylogenetic trees fitting the data. **(j)** In these synthetic examples, the VAF distribution of a tumor with and without subclonal selection changes for different values of coverage and purity, affecting the ability to observe neutral tails. A neutral tail (grey) becomes difficult to detect at 40x depth. **(k)** The “degenerated tail” at 40x can be statistically indistinguishable from a positively selected subclonal cluster. Data at such resolution are not powered to distinguish true positive subclonal selection from neutral tail mutations.

Figure 2. Model-based tumor subclonal reconstruction. (a) MOBSTER combines a Pareto Type-I distribution with k Beta random variables into a univariate finite mixture with $k + 1$ components. The Pareto captures the frequency spectrum of neutral mutations predicted by theory (Landau distribution decaying as $1/f^2$), whereas Beta components detect alleles under positive selection. The histogram shows clustering assignments for a tumor with one selected subclone ($k = 2$). **(b)** MOBSTER filters out neutral tail mutations, and one can cluster the rest with any tool for subclonal reconstruction using read counts. CCF, cancer cell fraction. **(c, d)** MOBSTER applied to the examples in Figure 1a,b detects the clusters corresponding to the true selected clones, hence recovering the correct clonal architecture. WGS, whole genome sequencing **(e,f)** We used synthetic 120x WGS data from *n*=150 simulated tumors to compare current methods with MOBSTER (plots show mean and inter quartile range IQR, upper whisker is 3rd quartile +1.5*IQR and lower whisker is 1st quartile −1.5*IQR). We measured how many clusters (e) and clone trees we identify (f). Tests compare Binomial mixtures from DPclust, pyClone and sciClone, and Beta-Binomial mixtures from pyClone, parameterized by concentration $\alpha > 0$. DPclust and pyClone learn α from the data assuming a Gamma prior. sciClone is a variational method with hardcoded α. In (e) we report the logarithm of the ratio between the number of subclones found by MOBSTER (k_{fit}) and the true number of clones (k_{true}). Red dashed line represents $k_{\text{fit}} = k_{\text{true}}$. In (f) we plot the number of trees that can be fit by pigeonhole principle using the output of each tool.

Figure 3. Analysis of single sample and multi-region whole-genome data. (a) Breast carcinoma ~180x WGS sample from ref[3]. MOBSTER identified a neutral tail plus $k = 3$ Beta clusters (2 subclones, consistent with two clone trees). Analysis of non-tail mutations with sciClone confirmed 2 subclones. sciClone without MOBSTER would have fit one extra clone to the tail. Non-parametric bootstrap is used to estimate the 95% bootstrap confidence intervals for the parameters. **(b)** Leukemia ~320x WGS sample from ref[20]. MOBSTER found two subclones ($k = 3$), confirmed with sciClone, and 2 clone trees. **(c)** WGS data at 100x from 4 biopsies of colorectal cancer Set07. From VAF of diploid mutations we identified neutral tails and no subclonal selection; from non-tail mutations we found 5 clusters (multivariate clustering with $\alpha = 10^{-6}$, Supplementary Note). C1 is the truncal cluster; all other clusters are enriched mutations private to a biopsy, indicating ancestor effect (Supplementary Note). The clone tree depicts a neutrally expanding tumor with all drivers in the trunk. Analysis without MOBSTER would have inflated the number of subclones (right panel; Supplementary Figures 20-23). **(d)** WGS data at 100x from 6 biopsies of cancer Set06 also showed neutral subclonal dynamics. Without MOBSTER we would have inflated the number of selected subclones (right panel; Supplementary Figures 24-27). **(e)** dN/dS analysis for Set06 and Set07 comparing truncal vs subclonal mutations confirmed lack of evidence for positive selection at the subclonal level, corroborating our conclusions. **(f)** Three lung cancer cases from ref[24] sequenced at 100x WGS were consistent with neutral subclonal dynamics.

Figure 4. Analysis of 2,566 whole-genomes from PCAWG with MOBSTER. (a) Fit of a PCAWG[25] tumor with 55x coverage and 66% purity using standard methods. **(b)** At this data resolution, neutral tails are under-sampled (Figure 1j,k) and cannot be distinguished from selected subclones. **(c)** In PCAWG cases with higher coverage (67x) and purity (74%), neutral tails can be clearly detected using MOBSTER. **(d)** Analysis of the same tumor with standard methods would have identified multiple subclonal clusters, including a cluster of neutral tail mutations. **(e)** We analyzed n=2,566 PCAWG samples, plotted here for purity vs coverage. Blue dots are tumors where MOBSTER cannot fit a tail. Red cases have a neutral tail. Percentage of tail mutations determines dot size. The marginal histograms report the normalized number of cases with tail. **(f)** We focused on the 902 diploid cases with coverage >30x and purity >65% (median of the cohort) where we could fit a tail. Using a panel of 191 pan-cancer driver genes, we show that tail mutations have dN/dS~1, providing no evidence of positive selection (point estimate and Confidence Intervals from dndscv). Clonal and subclonal non-tail mutations show dN/dS>1, consistent with positive selection. **(g)** If we take the 298 diploid cases with a tail containing at least 10% of the total mutational burden, we find evidence of a selected subclone only in 9 cases (3% of tumors). Similar proportions are obtained if we impose a 5% or 2% cutoff on the size of the tail. See Supplementary Figures 29-31.

Figure 5. Analysis of longitudinal glioblastoma samples with MOBSTER. (a). Patient H043−BU96 is one of $n = 16$ IDH-wildtype glioblastomas for which we analyzed WGS data (~100x) from pre-treatment and post-treatment longitudinal samples previously generated[28]. **(b)** Analysis following MOBSTER identified subclones private to the primary (yellow) and relapse (green) tumor respectively, the latter containing a putative driver mutation in LINC00689. **(c)** Patient H043−KZWs MOBSTER fits. **(d)** Here a subclone detected in the primary went on to sweep through the relapse, which was hypermutant after temozolomide treatment (zoom-in logscale panel). **(e)** Patient H043−PWC258 MOBSTER fits. **(f)** Here the primary sample showed neutral evolutionary dynamics, whereas the relapse contained detectable subclones possibly mixing with the neutral tail. An additional high-frequency subclone was detected from a downstream analysis using Binomial clustering of read counts (purple cluster, split into 2 Binomial components). **(g)** MOBSTER can also be used to identify and assign clusters that are produced by whole-genome duplications, or more general aneuploid states. In such contexts, we expect to see peaks in the VAF distribution that distinguish mutations that happened before and after genome doubling. In the case of patient H043−6F91, a diploid primary tumor (neutral) became whole-genome duplicated at relapse. **(h)** Orthogonal dN/dS analysis (point estimate and Confidence Intervals from dndscv) of mutations in 74 putative GBM driver genes assigned to neutral tails versus non-tail provided evidence of selection only in non-tail mtuations. The full list of analyzed cases is available in Supplementary Figure 32.

Online Methods

Model-based clustering of cancer subclonal populations with MOBSTER

The subclonal deconvolution problem is popular in the cancer literature[35]. Given read counts for a list of mutations detected from bulk sequencing of multiple tumor samples, we want to detect clusters of mutations that represent cancer subpopulations admixed in our samples. The problem can be framed to include any type of somatic mutation for which we can estimate the frequency, in the data, of the somatic (i.e., alternative) allele. Usually, the mutations that are easier to call are Single Nucleotide Variants (SNVs); more complex structural variations or insertion-deletions are more challenging to determine accurate allelic frequencies. Regardless mutation types, our aim is to use determine mutations clusters that suggest cancer subpopulations (i.e., clones) under positive selection.

MOBSTER is a mixed method that combines two types of random variables to approach this problem.

The frequency spectrum and the observational process. Kessler and Levin[10] have shown that, in the large population solution of the stochastic Luria-Delbrück model, the probability of having m mutants follows a fat-tail Landau distribution

$$p(m) = \frac{1}{\mu N} f_{\text{Landau}}\left(\frac{m}{\mu N} - \log \mu N + \gamma - 1\right) .$$

Here N is population size, μ the average fraction of birth events and γ the Euler constant. The asymptotic behavior of f_{Landau} can be approximated as $f_{\text{Landau}}(x) = 1/x^2$, which leads to the power-law approximation that has also been derived by others[12-14] as $p(m) \approx 1/m^2$.

A generative model for this power law can be constructed with a standard Markovian stochastic birth-death process of cell division – sometimes called *branching process*[16]. The existence of patterns of neutral evolution is thus a consolidated result from Population Genetics arguments that describe the spread of alleles in growing populations without recombination, such as cancer[17]. In other words, the *progeny of each clone* accumulates neutral passenger mutations until any of their daughter cells acquires a new mutation that undergoes selection because it triggers a new clonal expansion with increased fitness: the power-law spectrum emerges therefore by the frequencies of passengers. When a daughter cell enjoys a clonal expansion, however, the frequency of the variant alleles that accrued from the ancestor cell to the actual cell that acquired the driver, will grow. Eventually, this new subclonal expansion will become detectable if selection forces are strong compared to background (which is the clone within this cell was born). In a recursive fashion, the progeny of this new cell/ subclone will start dividing, giving rise to another power-law distributed tail of within-clone neutral dynamics. Example subclonal evolutionary dynamics are shown in the vignette "1. Example subclonal dynamics" (Supplementary Data), where we animate a subclonal expansion which shows how subclones emerge from low frequency up until they sweep, and how the allele frequency distribution changes over time.

Importantly, we want to make it clear that the power-law part of the spectrum – i.e., the *tail* – results from the accumulation of passenger mutations in the progeny of each clone. We note that this result – in particular the exponent 2 (shape) – refers to the total population structure of the tumor, which is accessible only in the theoretical scenario in which we can sequence all the cancer cells. Therefore, any specific finite sample that we collect and sequence, which is also contaminated by normal cells, might exhibit deviations from this theoretical distribution[16]. Deviations from strict exponential growth – e.g., due to spatial constrains – can also cause theoretical deviations from the exponent two[13,42]. However, we use this result to create a parametric model-based approach to analyze cancer data (i.e., we fix the type of distribution, but not its parameters).

Input data and conceptualization. We work with sequencing data for the variant alleles of n somatic mutations, which we can pre-process in different ways. One option is to adjust Variant Allele Frequency (VAF) values for copy number and purity, retrieving the so-called Cancer Cell Fractions (CCF) and re-scaling them into [0, 1] by halving the CCF. With these adjusted VAF values we expect a clonal peak at roughly 50% VAF, with outliers spreading around 0.5 but well below 1; compared to CCF, these values avoid the truncation of values above 1[3]. Another similar option is to adjust VAF values only by copy number, obtaining the so-called Cellular Prevalence (CPs). A third option is using directly the raw VAF data; in this last scenario we can further split mutations by karyotype – i.e., the absolute copy number segments where they map to – and account for the fact that different aneuploidy states have different expected distributions (e.g., a triploid tumor is expected to have two peaks of mutations, plus a tail and possibly subclonal clusters).

On real data, we suggest to use mutations that map to copy number segments with common karyotypes (i.e., copy states), such as diploid regions (with or without loss of heterozygosity), and triploid and tetraploid segments. Mutations mapping to more complex karyotypes (e.g., highly amplified oncogenes) can always be mapped post hoc, after clustering, and should account for a small subset of the tumor's mutational burden. We stress to use mutations in high-confident copy-number regions to carry out subclonal deconvolution; miscalled copy number states confound the inference creating artifact clusters of mutations. As a best practice, we usually attempt a first fit using diploid genomes without losses of heterozygosity (i.e., regions with one copy of the major and minor alleles), where we can identify high-confidence diploid SNVs.

Regardless the representations, a model for the *frequency spectrum* ρ of the observed mutations with $k \geq 1$ detectable clones is a random variable that follows

$$\rho \sim \sum_{i=1}^{k} (Y_i + B_i) \ ,$$

where

- $Y_i \propto x^{-\alpha}$ is a power-law random variable for frequencies of neutral mutations in the progeny of clone i. The generic exponent $\alpha > 0$ gives flexibility to accommodate all the confounders described above;
- $B_i \in [0,1]$ is a Beta random variable modelling the signal of clone i. In layman terms, B_i models the "peak" in the VAF distribution due to the hitchhikers of the clone. These distributions range in $[0,1]$, rendering them suitable to describe allelic frequencies (and also motivating why we scale CCF values to fit this range). For the sake of simplification, we assume here to work with adjusted VAF values, so that aneuploidy states (amplified, unamplified) are adjusted to form a single peak in the distribution (i.e., exactly as with CCF).

This model looks simple, and further observations are required to turn it into a mixture of standard random variables. In this formulation, the random variables for the tail and the bump of a clone are coupled to capture a joint signal. While the overall mixing proportions can be assumed to be independent, this compound random variable requires an extra level of mixing within each clone – i.e., another mixing weight to properly capture the proportions of the clone tail, and bump. We can however simplify this model accepting to track at finer detail only the clusters of each clone, which we use to identify subpopulations in the frequency spectrum (i.e., we use the clone's peak, obtained from the cluster's mean, to assess the phylogenetic history of the tumor).

We therefore simplify the model by noting that all tails have the same exponent $\alpha > 0$, which holds if all clones have the same mutation rate. If the mutation rate does not change among subclones – i.e., when there are no hypermutant subclones – all tails are described by the same theoretical distribution, and can be represented as multiple instances of the same random variable. Thus, we group them together in a single power-law tail

$$\rho \sim \left(Y + \sum_{i=1}^{k} B_i \right) .$$

Here the random variables have the same meaning as above, but the clone is no longer indexed by i. This model has a key advantage over the one where each clone "emits" its own tail: the random variables are decoupled and allow a simple mixture-model formulation which we will present below.

Before concluding, we observe that given ρ, the *observational model* for read counts collected from NGS sequencing, is a standard binomial process $n|\rho, m \sim \mathrm{Bin}(n|m, \rho)$, where m is the coverage (total number of reads), and w the number of reads harboring the variant allele; ρ is then the success probability for m iid Bernoulli trials. It is important to observe that the frequency spectrum and the observational process look at the data from different perspectives: the former is a distribution on allelic frequencies, while the latter on read counts. In this observational model we can in principle use Beta-Binomial distributions to account for coverage overdispersion.

Relation to other models in the literature. The literature is rich with models that describe the above observational process and variation thereof, either with Binomial or Beta-Binomial distributions. We briefly discuss those that are more related to our framework.

Bayesian methods that employ Dirichlet Processes for infinite Binomial mixture models are a popular generalization of the observational process. These non-parametric methods can fit an unspecified number of clusters k to data, simplifying model selection procedures. pyClone[5], DPclust[3] and PhyloWGS[6] are three popular tools for clonal deconvolution that in different ways use this framework. pyClone and DPclust implement Binomial mixtures, with the former also supporting Beta-Binomial distributions; in both cases a stick-breaking construction for Dirichlet Process priors is adopted[43]. PhyloWGS, instead, combines Binomial distributions with a tree stick-breaking construction for the Dirichlet Process priors[44], which allows PhyloWGS to cluster jointly the input SNVs, and construct a phylogenetic tree for the detected clones.

An alternative popular approach based on finite mixture models is SciClone[7], which supports Binomial, Beta and Gaussian mixtures. SciClone fits the models to data via Variational Inference, an information-theoretic approach to approximate the posterior distribution over the model's parameters. SciClone is a hybrid tool, as it can cluster allelic frequencies via Beta/ Gaussian mixtures, and read counts via Binomial mixtures. We want to note that, with Beta distributions, canonical Bayesian modeling leads to intractable priors, even if the conjugate prior distribution of the Beta distribution can be found by following the principles of conjugate priors for the exponential family. For this reason, Variational Inference of Beta mixtures exploits a Gamma approximation to the prior and posterior distributions, originally derived by Mao and Li[45]. In this approximation we cannot derive the so-called evidence lower bound, a standard measure to monitor convergence of a variational fitting algorithm.

These models are related to MOBSTER's framework: they assume that ρ can be approximated by a point-process (e.g. a Dirac distribution) centered at the Beta means. The potential pitfall is clear: by applying the observational process to neutral mutations, the number of clones is overestimated. Clusters will be called from tail mutations (polyphyletic lineages), which is wrong when we look for clones under selection. We note that SciClone with Beta distributions models the allele frequency spectrum as well, however, they do not account for power-law tails of neutrally-evolving mutations.

Distributions and likelihood. MOBSTER implements a statistical model to fit n VAF values to Y, the tail, and to any one of the B_i Betas, the clones (predefined in number). From a fit, tail mutations can be removed inspecting clustering assignments, and other methods can be used to fit the observational process on the read counts of the remaining data. For this reason, MOBSTER is complementary to the tools mentioned above, as it works upstream the observational process. Nonetheless, our method provides also a preliminary indication on the possible number of subclones in the tumor: with high-quality data with low dispersions, one can expect the same number of clones to be confirmed by downstream analysis of non-tail mutations.

The fit uses a pre-specified number of $k+1$ components, where Y is a Pareto Type-I distribution as the power-law tail. For a *scale* x_* and *shape* $\alpha > 0$, its density is

$$g(x|\, x_*, \alpha) = \alpha x_*^{\alpha} \frac{1}{x^{\alpha+1}}$$

for $x > x_*$, and 0 otherwise. Notice that the density is 0 for values below the scale parameter, which requires a sharp cutoff on the input VAF, and that its support is $[0, +\infty)$. The model also uses k Beta distributions $B_1, \dots, B_k$ to model clonal and subclonal clusters. For a *shape* $a > 0$ and $b > 0$ the density of a Beta random variable is

$$h(x|a, b) \;=\; \frac{x^{a-1}(1-x)^{b-1}}{\mathrm{B}(a, b)}$$

where $\mathrm{B}(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$ is the beta-function. The support of this distribution is $[0, 1]$, the full frequency spectrum.

The overall model uses a Dirichlet prior on the abundance of each clone; thus MOBSTER is a Finite Dirichlet Mixture Model with both Beta and Pareto distributions. The model likelihood for a dataset $X = \{x_i | i = 1, \dots, n\}$ where we assume each x_i to be iid, is a combination of two types of densities

$$p(D|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \left[\pi_1 g(x_i|x_*, \alpha) + \sum_{w=2}^{k} \pi_w h(x_i|a_{w-1}, b_{w-1}) \right].$$

We use $\boldsymbol{\theta}$ as a shorthand to the model parameters, and $\boldsymbol{\pi} = [\pi_1 \dots \pi_{k+1}]$ for the mixing proportions – a standard Dirichlet variable on the $(k+1)$-dimensional probability simplex. Notice that, just for notational convenience, we are assuming that the first model component is the Pareto random variable (the tail); we hold this setup fixed even if the model does not fit a tail (in that case we force $\pi_1 = 0$). Because of this, we use the index $w-1$ for the parameters of the Beta distributions just to reflex that their index start from one.

Fitting MOBSTER. The formulation uses $n \times (k+1)$ latent variables $\boldsymbol{z}$. A variational approach to fit this mixture is theoretically possible: we could use conjugate Gamma priors for the Pareto, and we would approximate the posteriors for the Beta components as in sciClone. However, we could only approximate a criterion for convergence of the fit, as mentioned above.

We fit the model parameters via Maximum Likelihood Estimation (MLE) through an adaptation of a standard Expectation-Maximization approach (EM). This alternative is faster than a Bayesian Monte Carlo strategy, at the drawback of inferring a point estimate of the parameters. The lack of an explicit measure of uncertainty in the prediction (confidence) can be mitigated using the bootstrap.

We perform these steps to fit a MOBSTER model. In the E-step, we compute the posterior estimates of the latent variables as usual, once we account for the two different distributions involved

$$z_{w,1}\,|\boldsymbol{\theta} \;\propto\; \pi_1 g(x_i|x_*,\alpha) \qquad\qquad z_{w,i}\,|\boldsymbol{\theta} \;\propto\; \pi_i h(x_w|a_i,b_i)$$

In both cases the normalisation constant C_w is the overall density mass for point x_w

$$C_w = \pi_1 g(x_w|x_*,\alpha) + \sum_{i=2}^{k} \pi_i h(x_w|a_i,b_i)\,.$$

In the M-step, for the Pareto tail, we begin by noting that the scale x_* of the distribution can be set to its MLE[46], which is known to be the smallest observed frequency $x_* = \min X$. This is a constant of the data, so we have one less parameter to fit. We fit the Pareto shape α, given x_*; switching to the log-likelihood and including latent variables its MLE estimator is

$$\alpha_{\mathrm{MLE}} = -\frac{\sum_{i=1}^{n} z_{i,1}}{\sum_{i=1}^{n} z_{i,1}\log(x_*/x_i)}$$

For the Beta clones, in the M-step, the MLE estimator for the distributions has no closed form; we can resort to approximate it numerically, increasing the computational burden. We can also rely on a recent analytical result on the Moment-Matching (MM) estimator of mixtures of Betas by Schröder and Rahmann[47]. MM consists in matching t empirical moments of the data X to the theoretical moments of the distribution, and solving for them. Here $t = 2$ (mean and variance); a Beta distribution has mean μ and variance σ given by

$$\mu = \frac{a}{a+b} \qquad\qquad \sigma = \frac{ab}{(a+b)^2(1+a+b)}\,.$$

For a Beta, conditioned on the latent variables, the MM estimator is

$$\mu_{i_{\mathrm{MM}}} = \frac{\sum_{w=1}^{n} z_{w,i}x_w}{n\pi_i} \qquad\qquad \sigma_{i_{\mathrm{MM}}} = \frac{\sum_{w=1}^{n} z_{w,i}(x_w-\mu)^2}{n\pi_i}\,.$$

Given estimates for μ_i and σ_i, we can re-parametrize the Beta as

$$a_{i_{\mathrm{MM}}} = \left(\frac{1-\mu_i}{\sigma_i} - {\mu_i}^{-1}\right)\,{\mu_i}^2 \qquad\qquad b_{i_{\mathrm{MM}}} = \mu_i({\mu_i}^{-1}-1)\,.$$

We remark that MM is not the same as computing the MLE, which computes the zeroes of the derivative of the likelihood with respect to the parameters $\boldsymbol{\theta}$, $\partial h/\partial\boldsymbol{\theta}$. Thus, the properties of standard EM do not hold when we compute updates via MM: we cannot guarantee that the likelihood increases monotonically, because we cannot employ Jensen's inequality. It is however shown[47] that the differences between the estimators are negligible in most cases. For the sake of precision, Schröder and Rahmann propose to call a fit through the MM for Beta distributions the "iterative method of moments", rather than EM.

In MOBSTER's implementation we provide both a standard EM fit with numerical solution for the MLE of Beta distributions, and the faster iterative method of moments. In the former case we monitor convergence of the likelihood, as standard. In the latter we use the posterior estimates of $\boldsymbol{\pi}$ since the likelihood is not monotonically increasing. A theoretical property of this MM approach is that, in each step, before updating the component weights, the expectation of the estimated density equates the sample mean. In particular, this is true at a stationary point; a proof of this is in Lemma 1 of Schröder and Rahmann[47].

Initial conditions. As standard in EM approaches, we compute the fit with several random initial conditions. We provide two heuristics to compute the initial condition of the fit (Supplementary Figure 1). One is based on a peak detection heuristic applied in the frequency range $[0.1, 1]$ to VAF values binned with size 0.01. To detect k initial peaks we perform kmeans clustering of each peak's x-coordinate, and store their centres. If there are $w < k$ peaks to cluster, we sample $k - w$ random values in $(0, 1)$ for the remaining peaks. We use the centers of these clusters as the mean of k Beta distributions with randomized variance sampled in $[10^{-3}, 0.25]$; we do sample variance values until the corresponding Beta parameters a and b are positive. For the tail, α is randomly sampled in the interval $[0.01, 5]$. These values provide wide ranges of different initial distributions. An alternative method to select the initial condition of the fit is totally randomized.

Experimental results show that peak detection is a more robust initialization method; the random counterpart sometimes leads to Beta distributions with mean approaching one, a region of parameter values where the likelihood becomes less stable, leading to numerical difficulties. In many cases, we test fits with both initial conditions and retain the best one.

Clustering assignments and model selection. We do not want the fit to be biased towards tails, as we would miss low-frequency subclones that hide in the tail. Besides, simulations suggest limits to the detectability of tails, and therefore we shall not assume tail to be always present in the data. For this reason, MOBSTER can "turn off" the Pareto component of the mixture (i.e., setting $\pi_1 = 0$) and fit just k Beta. Hence, we can perform model selection for $1 \leq k \leq K$ considering both models with and without a tail. This induces a statistical competition and allows us to select the model that best explains the data, with or without a tail.

In MOBSTER we compute the negative log-likelihood $\mathsf{NLL} = -\log f(X \mid \boldsymbol{\theta}, \boldsymbol{\pi})$ of the data, which we use to derive the usual AIC and BIC scores $\mathrm{BIC} = 2\mathrm{NLL} + |\boldsymbol{\theta}| \log \mathrm{n}$, and $\mathrm{AIC} = 2\mathrm{NLL} + 2|\boldsymbol{\theta}|$.

These criteria favor simpler fits by penalizing a model for the number of its parameters $|\boldsymbol{\theta}|$. A model with k Beta distributions and one tail has $|\boldsymbol{\theta}| = 3k + 2$ parameters ($k + 1$ for the Dirichlet mixture $\boldsymbol{\pi}$, $2k$ for the Beta(s) and 1 for the Pareto tail). The fit without tail model has $|\boldsymbol{\theta}| = 3k - 1$ parameters; fewer parameters reduce less the penalty, thus favoring fits without a tail.

In MOBSTER we want to drive the fit to select separate clusters, i.e., fits with few overlapping components, which we do not achieve using BIC or AIC. We achieve these separations by using instead two types of entropy terms. In one case we compute, from the latent variables, the usual entropy $\mathrm{H}(\boldsymbol{z})$

$$\mathrm{H}(\boldsymbol{z}) = \sum_{i=1}^{k+1} \sum_{j=1}^{n} z_{i,j} \log z_{i,j}$$

and obtain the standard Integrative Classification Likelihood (ICL) $\mathrm{ICL} = \mathrm{BIC} + \mathrm{H}(\boldsymbol{z})$, approximated through the BIC[48]. In this paper we also introduce a heuristic variation to the ICL, which we call reICL, a reduced-entropy criterion where we use the entropy of mutations that are not assigned to a tail (Supplementary Figure 1). This is defined as $\mathrm{reICL} = \mathrm{BIC} + \mathrm{H}(\hat{\boldsymbol{z}})$, where $\hat{\boldsymbol{z}}$ are the latent variables for the set of mutations $\{\mathrm{x} \mid 1 \neq \operatorname{argmax} \boldsymbol{z}_{x,.}\}$, re-normalized. Notice that in practice $\hat{\boldsymbol{z}}$ is defined from the hard clustering assignments that we use to assign mutations to clusters; cluster "1" is the label to identify tail mutations.

Entropy terms in ICL and reICL help to fit separate clusters because overlapping mixture components have higher entropy, and therefore penalty. The maximum entropy distribution is the uniform one, which is when we cannot confidently assign mutations to clusters (a point seems to be equally-well explained by multiple components). By definition, ICL will push towards fits with a clear separation among tail *and* Beta components, while reICL will only require separation of the Beta ones. This modification to the ICL seems reasonable because the Pareto tail overlaps - by definition - to all subclonal clusters, and this leads to strong entropy penalizations with

ICL. For this reason, ICL will be more stringent in calling tails than reICL, which drops a part of the entropy penalty restricting its computation to $\hat{\boldsymbol{z}}$. See also Supplementary Figure 1 for a graphical explanation.

Notice that, because we are using NLL, we seek to *minimize* these scores. In the tests, we investigate different model-selection strategies, and choose as default score for model selection in MOBSTER reICL, which seems to provide a nice tradeoff. Between the ability to identify the Beta components, while retaining the tail structure.

Analysis of synthetic data

In the Supplementary Note and in the Supplementary Data (vignettes "Simulated single-sample data analysis" and "Simulated multi-sample data analysis") we explain how we used branching processes to generate tumors without and with space, and present output metrics to assess precision and sensitivity of our analyses (number of clusters, confidence in the predictions, rates of false/true positives/negatives, the effect of coverage and purity and the ability to identify subclones). In the tests we used MOBSTER and other tools for subclonal deconvolution.

We found MOBSTER and the analyses built around it to be accurate, across all simulated tumors. In all cases tails improve fit quality, from a statistical point of view. This clustering problem is challenging because tails and clones overlap, confounding weak signals of subclonal selection at the low-frequency VAF. We used our performance and combinations of coverage and purity to identify minimum requirements for reliable deconvolution in non-spatial data. In general, we assessed that we can fit subclones and tails for a wide range of parameter values, but overlapping distributions complicate the inference. MOBSTER does not show biases and can identifies subclones, even when they have low VAF (Supplementary Note).

From multi-region data (Supplementary Note) of polyclonal tumors we identified three confounders that inflate the number of clones reported by a "standard" analysis. The confounders contribute Binomial clusters that cannot be directly linked to clonal evolution patterns originating from positive selection. Branching structures originating from the confounders are also misleading, and do not reflect selection-driven branched evolution. One of the confounders can be solved by MOBSTER; two require extra heuristics discussed in the Supplementary Note.

Analysis of patient derived data

The description of all the data analyzed is in the Supplementary Note, as well as in the Supplementary Data. All summary statistics for all fit samples of this paper are available in Supplementary Table 1.

Data Availability

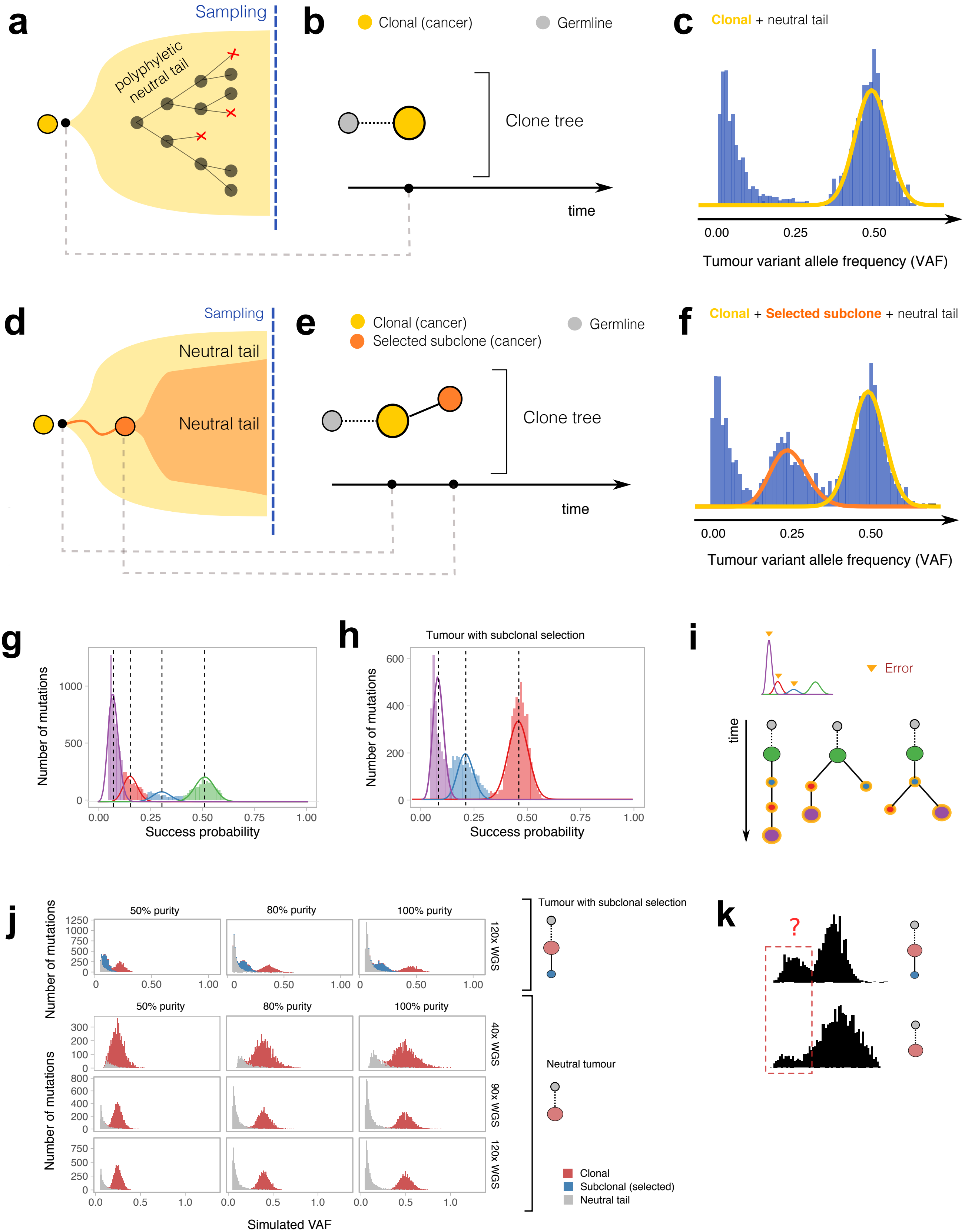
Data in Figure 3a were from Nik-Zainal *et al.* 2012[3]. Data in Figure 3b were from Griffith *et al.* 2015[20]. Data in Figure 3c-e were cases from Cross et al. 2018[21], here re-sequenced at higher sequencing depth. Sequence data from those colorectal cancer cases have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001003066. Further information about EGA can be found on https://ega-archive.org. Diploid SNVs and copy number calls are available in the Supplementary Data in vignette "5. Multi-region cross-sectional colorectal carcinomas". Data in Figure 3f were from Lee *et al.* 2019[24]. Data in Figure 4 are available through the PCAWG consortium[25]. Whole-genome variant call data in Figure 5 that were not available from the original publication, were provided upon email request by Korber *et al.* 2019[28].

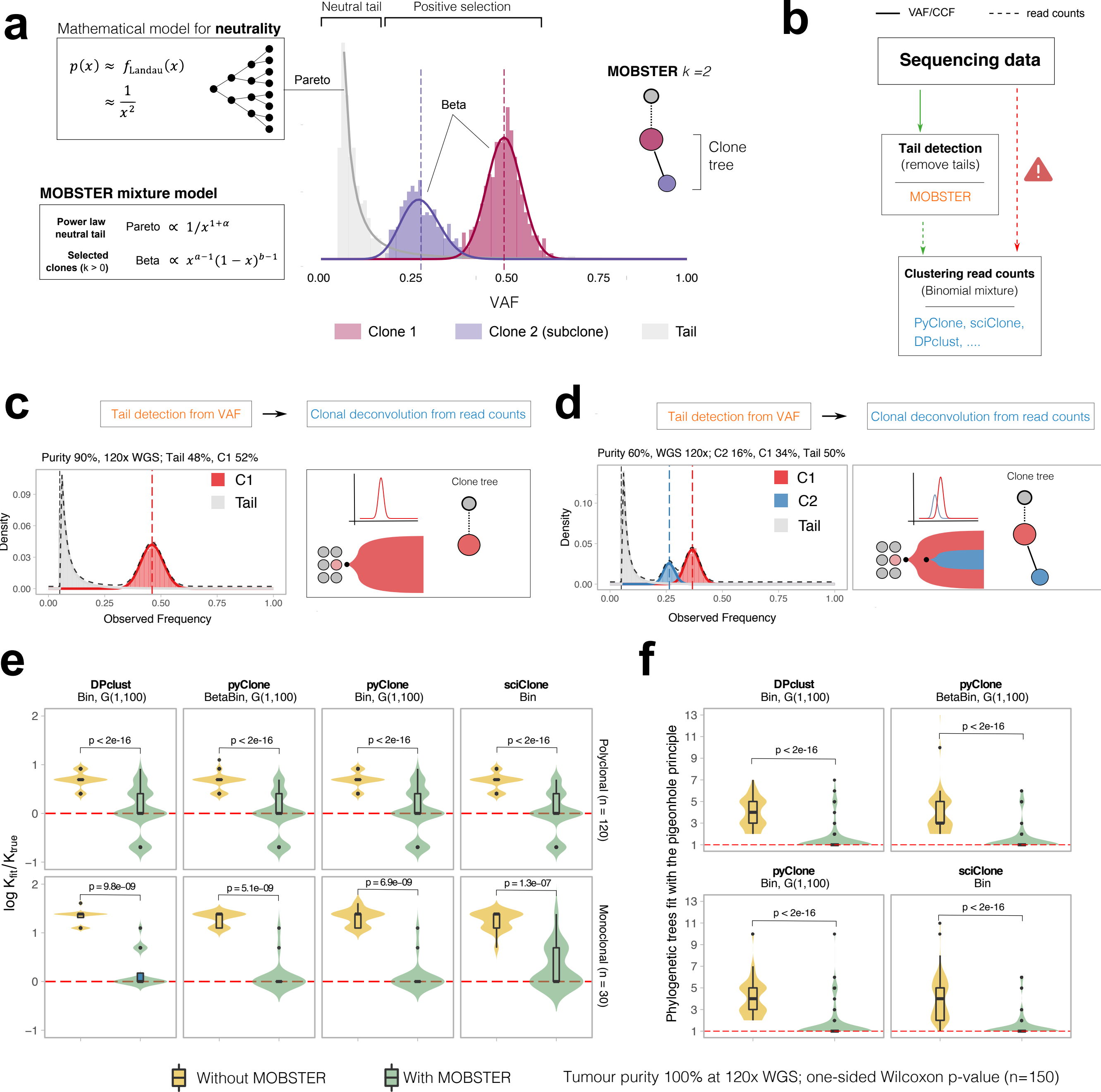
Code Availability

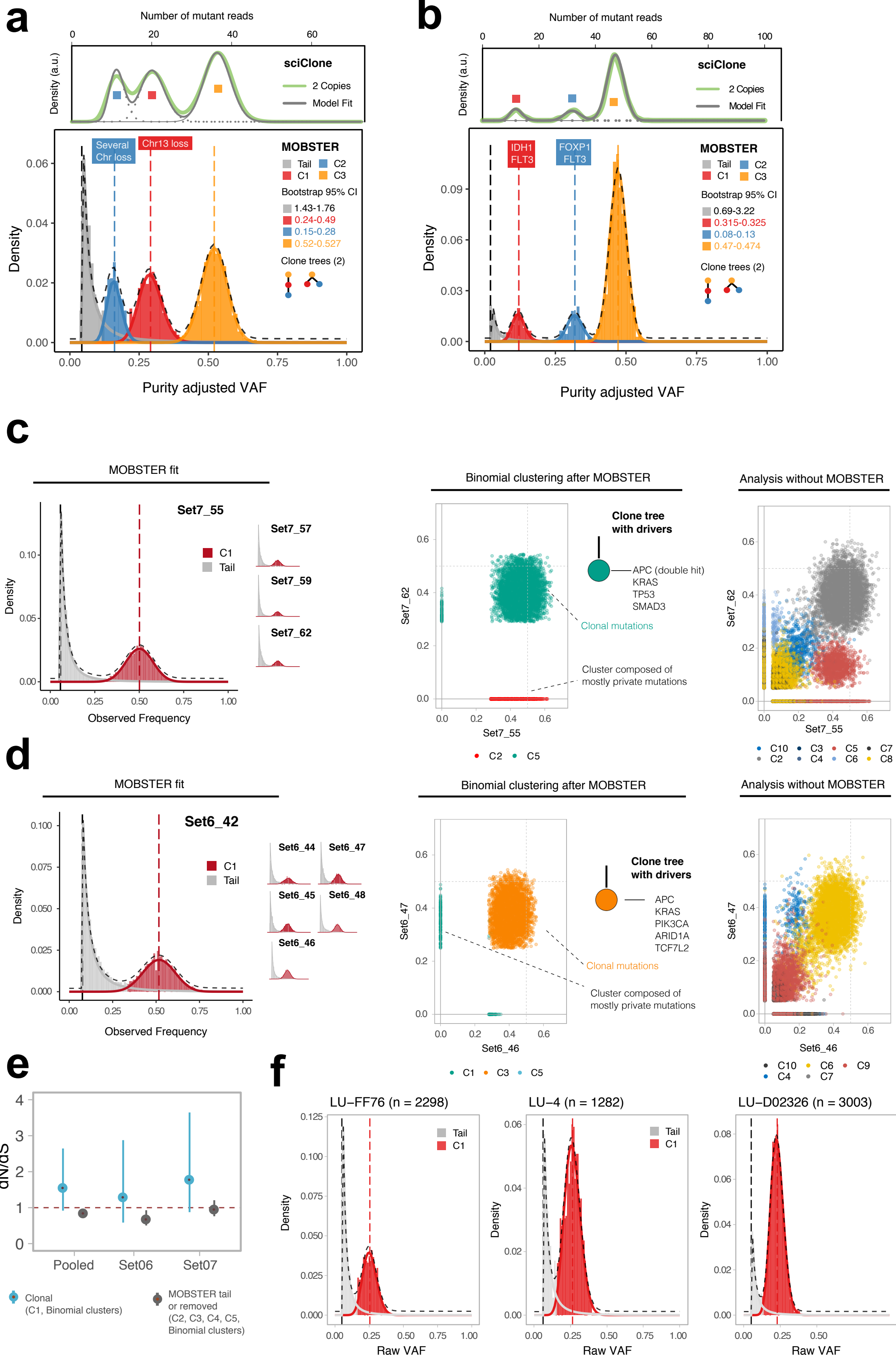
MOBSTER is available as an R package at https://github.com/sottorivalab/mobster/; future updates, as well as all vignettes and manuals are maintained at https://caravagn.github.io/mobster/. A repository with all Supplementary Data is available at https://github.com/sottorivalab/mobster_supp_data. Supplementary Data contain vignettes that show the analysis of single-sample and multi-region simulated tumors, the whole analysis of multi-region colorectal samples and single-sample lung cancers, and summary results from the PCAWG and GBM cohorts. Somatic single nucleotide variants and copy number calls used for the analysis of multi-region colorectal samples are also available as Supplementary Data. The implementation of all other R packages that we have developed are available at https://caravagn.github.io/ .

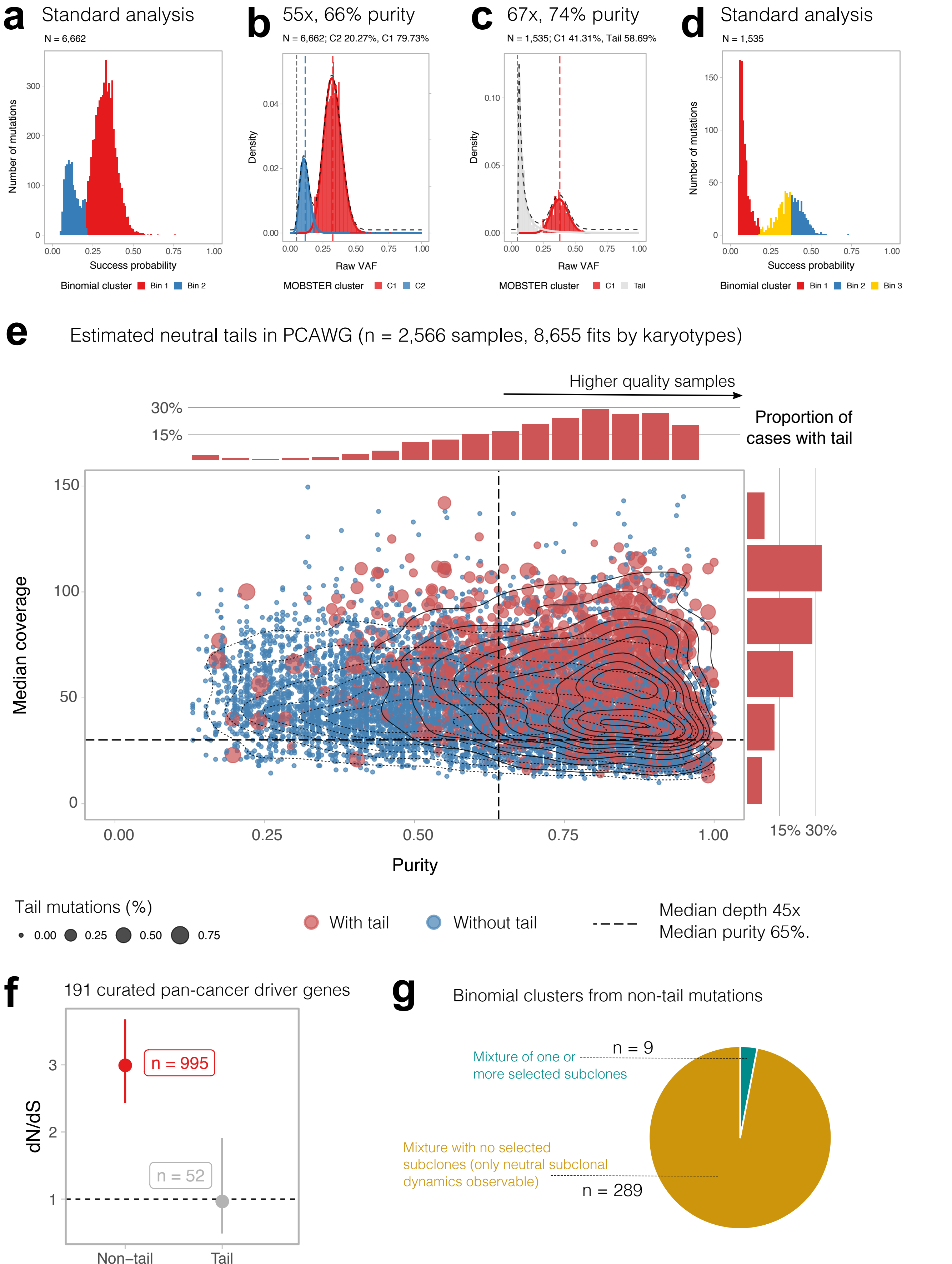
Methods-only References

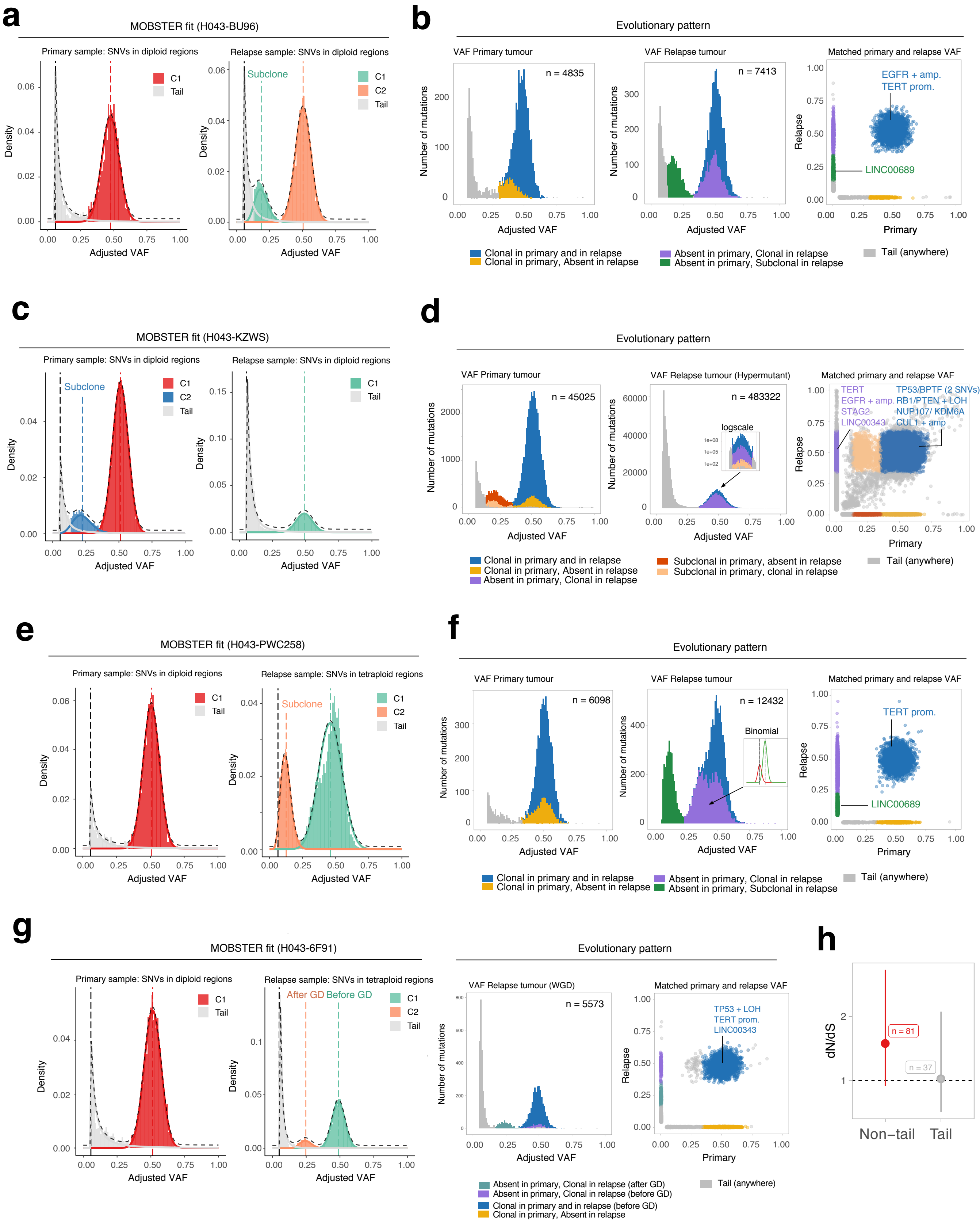
42. Fusco, D., Gralka, M., Kayser, J., Anderson, A. & Hallatschek, O. Excess of mutational jackpot events in expanding populations revealed by spatial Luria–Delbrück experiments. *Nat Comms* **7,** 12760 (2016).
43. Teh, Y. W. in *Encyclopedia of Machine Learning* 280–287 (Springer, Boston, MA, 2011).
44. Ghahramani, Z., Jordan, M. I. & Adams, R. P. Tree-Structured Stick Breaking for Hierarchical Data. 19–27 (2010).
45. Ma, Z. & Leijon, A. Bayesian Estimation of Beta Mixture Models with Variational Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33,** 2160–2173 (2011).
46. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-Law Distributions in Empirical Data. *SIAM Review* **51,** 661–703 (2009).
47. Schröder, C. & Rahmann, S. A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification. *Algorithms for Molecular Biology 2017 12:1* **12,** 21 (2017).
48. Biernacki, C., Celeux, G. & Govaert, G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22,** 719–725 (2000).

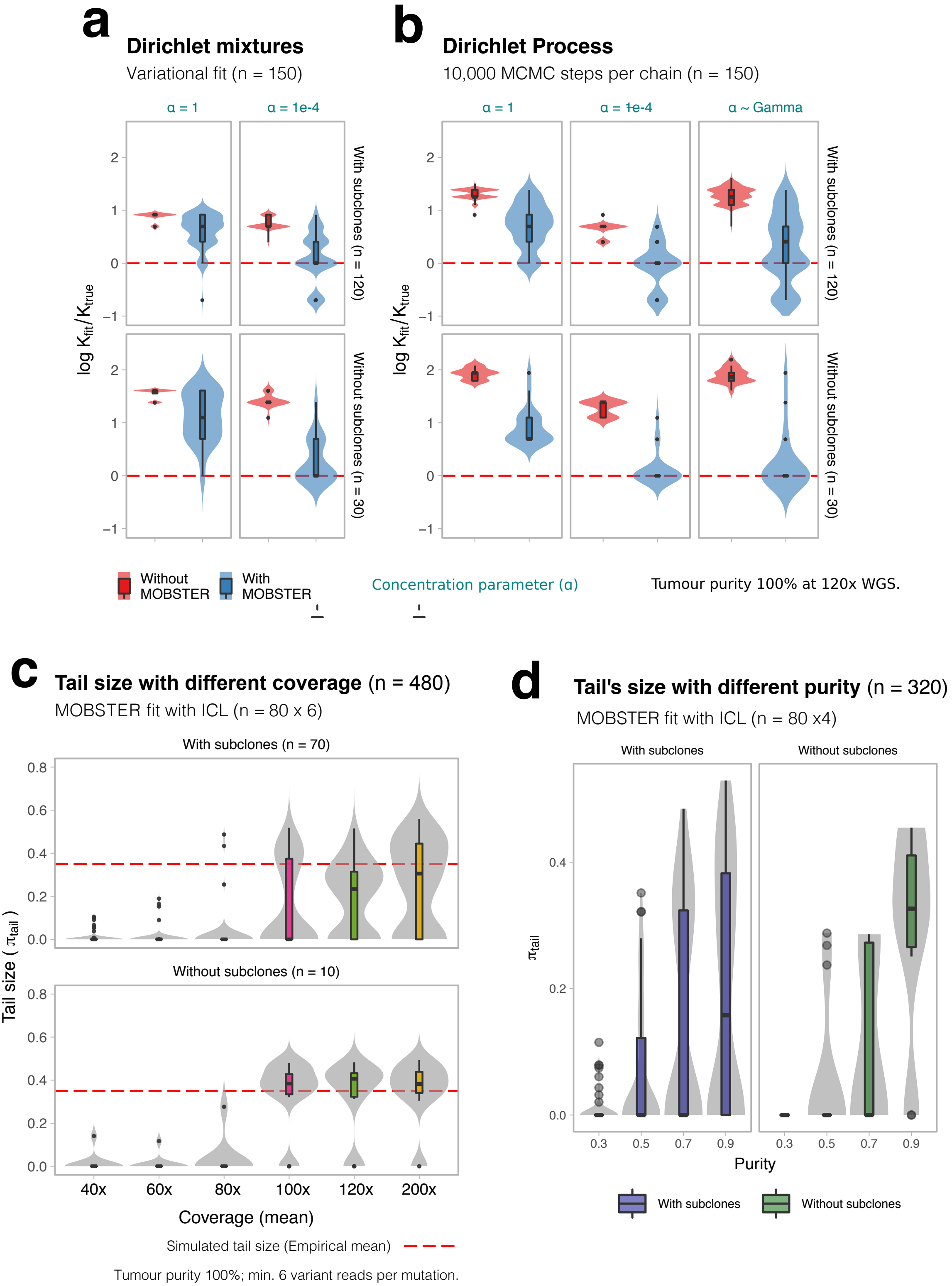

a
Sampling
polyphyletic neutral tail
b
Clonal (cancer)
Germline
Clone tree
time
c
Clonal + neutral tail
0.00
0.25
0.50
Tumour variant allele frequency (VAF)
d
Sampling
Neutral tail
Neutral tail
e
Clonal (cancer)
Selected subclone (cancer)
Germline
Clone tree
time
f
Clonal + Selected subclone + neutral tail
0.00
0.25
0.50
Tumour variant allele frequency (VAF)
g
Number of mutations
1000
500
0
0.00
0.25
0.50
0.75
1.00
Success probability
h
Tumour with subclonal selection
Number of mutations
600
400
200
0
0.00
0.25
0.50
0.75
1.00
Success probability
i
Error
time
j
Number of mutations
50% purity
80% purity
100% purity
1250
1000
750
500
250
0
0
0.50
1.00
120x WGS
Tumour with subclonal selection
50% purity
80% purity
100% purity
300
200
100
0
40x WGS
800
600
400
200
0
90x WGS
750
500
250
0
120x WGS
0.0
0.5
1.0
Simulated VAF
Number of mutations
Neutral tumour
Clonal
Subclonal (selected)
Neutral tail
k
?


a
Mathematical model for neutrality
$p(x) \approx f_{\text{Landau}}(x)$
$\approx \frac{1}{x^2}$
MOBSTER mixture model
Power law neutral tail
Pareto $\propto 1/x^{1+\alpha}$
Selected clones (k > 0)
Beta $\propto x^{a-1}(1-x)^{b-1}$
Neutral tail
Positive selection
Pareto
Beta
0.00
0.25
0.50
0.75
1.00
VAF
Clone 1
Clone 2 (subclone)
Tail
MOBSTER k =2
Clone tree
b
VAF/CCF
read counts
Sequencing data
Tail detection
(remove tails)
MOBSTER
Clustering read counts
(Binomial mixture)
PyClone, sciClone, DPclust, ...
c
Tail detection from VAF
Clonal deconvolution from read counts
Purity 90%, 120x WGS; Tail 48%, C1 52%
C1
Tail
Density
0.00
0.03
0.06
0.09
0.00
0.25
0.50
0.75
1.00
Observed Frequency
Clone tree
d
Tail detection from VAF
Clonal deconvolution from read counts
Purity 60%, WGS 120x; C2 16%, C1 34%, Tail 50%
C1
C2
Tail
Density
0.00
0.05
0.10
0.00
0.25
0.50
0.75
1.00
Observed Frequency
Clone tree
e
DPclust
Bin, G(1,100)
pyClone
BetaBin, G(1,100)
pyClone
Bin, G(1,100)
sciClone
Bin
p < 2e-16
p < 2e-16
p < 2e-16
p < 2e-16
p = 9.8e−09
p = 5.1e−09
p = 6.9e−09
p = 1.3e−07
Polyclonal (n = 120)
Monoclonal (n = 30)
log K_{fit}/K_{true}
2
1
0
−1
f
DPclust
Bin, G(1,100)
pyClone
BetaBin, G(1,100)
pyClone
Bin, G(1,100)
sciClone
Bin
p < 2e-16
p < 2e-16
p < 2e-16
p < 2e-16
Phylogenetic trees fit with the pigeonhole principle
1
3
5
7
9
11
13
Without MOBSTER
With MOBSTER
Tumour purity 100% at 120x WGS; one-sided Wilcoxon p-value (n=150)


a
Number of mutant reads
0
20
40
60
Density (a.u.)
sciClone
2 Copies
Model Fit
MOBSTER
Tail
C1
C2
C3
Bootstrap 95% CI
1.43-1.76
0.24-0.49
0.15-0.28
0.52-0.527
Clone trees (2)
Several Chr loss
Chr13 loss
Density
0.06
0.04
0.02
0.00
0.00
0.25
0.50
0.75
1.00
Purity adjusted VAF
b
Number of mutant reads
0
20
40
60
80
100
Density (a.u.)
sciClone
2 Copies
Model Fit
MOBSTER
Tail
C1
C2
C3
Bootstrap 95% CI
0.69-3.22
0.315-0.325
0.08-0.13
0.47-0.474
Clone trees (2)
IDH1 FLT3
FOXP1 FLT3
Density
0.09
0.06
0.03
0.00
0.00
0.25
0.50
0.75
1.00
Purity adjusted VAF
c
MOBSTER fit
Set7_55
C1
Tail
Density
0.10
0.05
0.00
0.00
0.25
0.50
0.75
1.00
Observed Frequency
Set7_57
Set7_59
Set7_62
Binomial clustering after MOBSTER
Set7_62
Set7_55
0.0
0.2
0.4
0.6
Clone tree with drivers
APC (double hit)
KRAS
TP53
SMAD3
Clonal mutations
Cluster composed of mostly private mutations
C2
C5
Analysis without MOBSTER
Set7_62
Set7_55
0.0
0.2
0.4
0.6
C10
C2
C3
C4
C5
C6
C7
C8
d
MOBSTER fit
Set6_42
C1
Tail
Density
0.100
0.075
0.050
0.025
0.000
0.00
0.25
0.50
0.75
1.00
Observed Frequency
Set6_44
Set6_47
Set6_45
Set6_48
Set6_46
Binomial clustering after MOBSTER
Set6_47
Set6_46
0.0
0.2
0.4
0.6
Clone tree with drivers
APC
KRAS
PIK3CA
ARID1A
TCF7L2
Clonal mutations
Cluster composed of mostly private mutations
C1
C3
C5
Analysis without MOBSTER
Set6_47
Set6_46
0.0
0.2
0.4
0.6
C10
C4
C6
C7
C9
e
dN/dS
4
3
2
1
0
Pooled
Set06
Set07
Clonal (C1, Binomial clusters)
MOBSTER tail or removed (C2, C3, C4, C5, Binomial clusters)
f
LU–FF76 (n = 2298)
Tail
C1
Density
0.125
0.100
0.075
0.050
0.025
0.000
0.00
0.25
0.50
0.75
1.00
Raw VAF
LU–4 (n = 1282)
Tail
C1
Density
0.06
0.04
0.02
0.00
0.00
0.25
0.50
0.75
1.00
Raw VAF
LU–D02326 (n = 3003)
Density
0.08
0.06
0.04
0.02
0.00
0.00
0.25
0.50
Raw VAF


a
Standard analysis
N = 6,662
Number of mutations
Success probability
Binomial cluster Bin 1 Bin 2
b
55x, 66% purity
N = 6,662; C2 20.27%, C1 79.73%
Density
Raw VAF
MOBSTER cluster C1 C2
c
67x, 74% purity
N = 1,535; C1 41.31%, Tail 58.69%
Density
Raw VAF
MOBSTER cluster C1 Tail
d
Standard analysis
N = 1,535
Number of mutations
Success probability
Binomial cluster Bin 1 Bin 2 Bin 3
e
Estimated neutral tails in PCAWG (n = 2,566 samples, 8,655 fits by karyotypes)
Higher quality samples
30%
15%
Proportion of cases with tail
Median coverage
Purity
15% 30%
Tail mutations (%)
0.00 0.25 0.50 0.75
With tail
Without tail
Median depth 45x
Median purity 65%.
f
191 curated pan-cancer driver genes
dN/dS
n = 995
n = 52
Non–tail
Tail
g
Binomial clusters from non-tail mutations
Mixture of one or more selected subclones
n = 9
Mixture with no selected subclones (only neutral subclonal dynamics observable)
n = 289


a
MOBSTER fit (H043-BU96)
Primary sample: SNVs in diploid regions
Density
0.00
0.02
0.04
0.06
C1
Tail
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
Relapse sample: SNVs in diploid regions
Subclone
C1
C2
Tail
Density
0.00
0.02
0.04
0.06
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
b
Evolutionary pattern
VAF Primary tumour
n = 4835
Number of mutations
0
100
200
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
VAF Relapse tumour
n = 7413
Number of mutations
0
100
200
300
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
Matched primary and relapse VAF
EGFR + amp.
TERT prom.
LINC00689
Relapse
0.00
0.25
0.50
0.75
1.00
0.00 0.25 0.50 0.75 1.00
Primary
Clonal in primary and in relapse
Clonal in primary, Absent in relapse
Absent in primary, Clonal in relapse
Absent in primary, Subclonal in relapse
Tail (anywhere)
c
MOBSTER fit (H043-KZWS)
Primary sample: SNVs in diploid regions
Subclone
C1
C2
Tail
Density
0.00
0.02
0.04
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
Relapse sample: SNVs in diploid regions
C1
Tail
Density
0.00
0.05
0.10
0.15
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
d
Evolutionary pattern
VAF Primary tumour
n = 45025
Number of mutations
0
1000
2000
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
VAF Relapse tumour (Hypermutant)
n = 483322
logscale
1e+08
1e+05
1e+02
Number of mutations
0
20000
40000
60000
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
Matched primary and relapse VAF
TERT
EGFR + amp.
STAG2
LINC00343
TP53/BPTF (2 SNVs)
RB1/PTEN + LOH
NUP107/ KDM6A
CUL1 + amp
Relapse
0.00
0.25
0.50
0.75
1.00
0.00 0.25 0.50 0.75 1.00
Primary
Clonal in primary and in relapse
Clonal in primary, Absent in relapse
Absent in primary, Clonal in relapse
Subclonal in primary, absent in relapse
Subclonal in primary, clonal in relapse
Tail (anywhere)
e
MOBSTER fit (H043-PWC258)
Primary sample: SNVs in diploid regions
C1
Tail
Density
0.00
0.02
0.04
0.06
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
Relapse sample: SNVs in tetraploid regions
Subclone
C1
C2
Tail
Density
0.00
0.02
0.04
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
f
Evolutionary pattern
VAF Primary tumour
n = 6098
Number of mutations
0
100
200
300
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
VAF Relapse tumour
n = 12432
Binomial
Number of mutations
0
200
400
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
Matched primary and relapse VAF
TERT prom.
LINC00689
Relapse
0.00
0.25
0.50
0.75
1.00
0.00 0.25 0.50 0.75 1.00
Primary
Clonal in primary and in relapse
Clonal in primary, Absent in relapse
Absent in primary, Clonal in relapse
Absent in primary, Subclonal in relapse
Tail (anywhere)
g
MOBSTER fit (H043-6F91)
Primary sample: SNVs in diploid regions
C1
Tail
Density
0.00
0.02
0.04
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
Relapse sample: SNVs in tetraploid regions
After GD
Before GD
C1
C2
Tail
Density
0.00
0.05
0.1
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
Evolutionary pattern
VAF Relapse tumour (WGD)
n = 5573
Number of mutations
0
200
400
600
800
0.00 0.25 0.50 0.75 1.00
Adjusted VAF
Matched primary and relapse VAF
TP53 + LOH
TERT prom.
LINC00343
Relapse
0.00
0.25
0.50
0.75
1.00
0.00 0.25 0.50 0.75 1.00
Primary
Absent in primary, Clonal in relapse (after GD)
Absent in primary, Clonal in relapse (before GD)
Clonal in primary and in relapse (before GD)
Clonal in primary, Absent in relapse
Tail (anywhere)
h
dN/dS
1
2
n = 81
n = 37
Non−tail
Tail

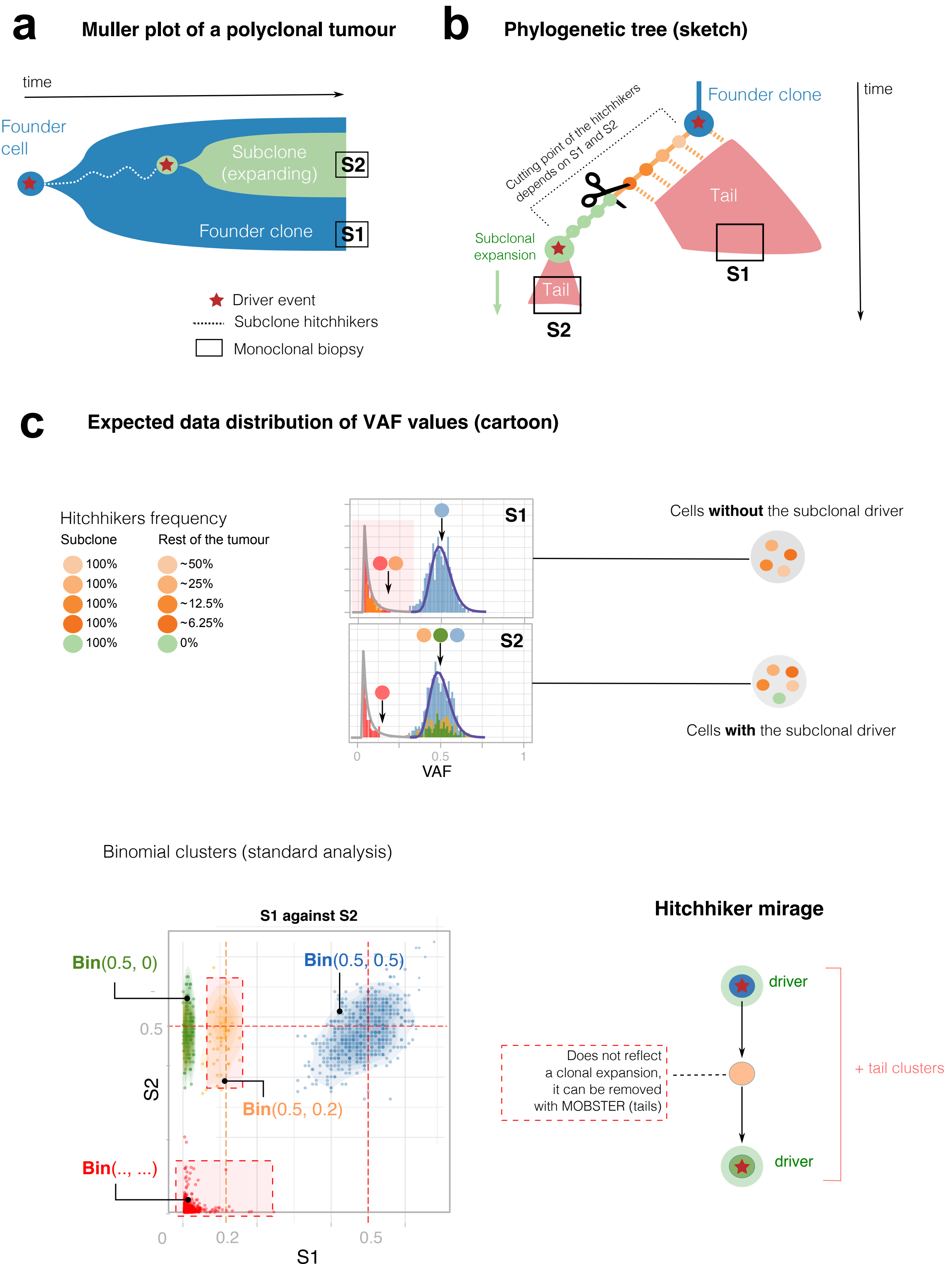


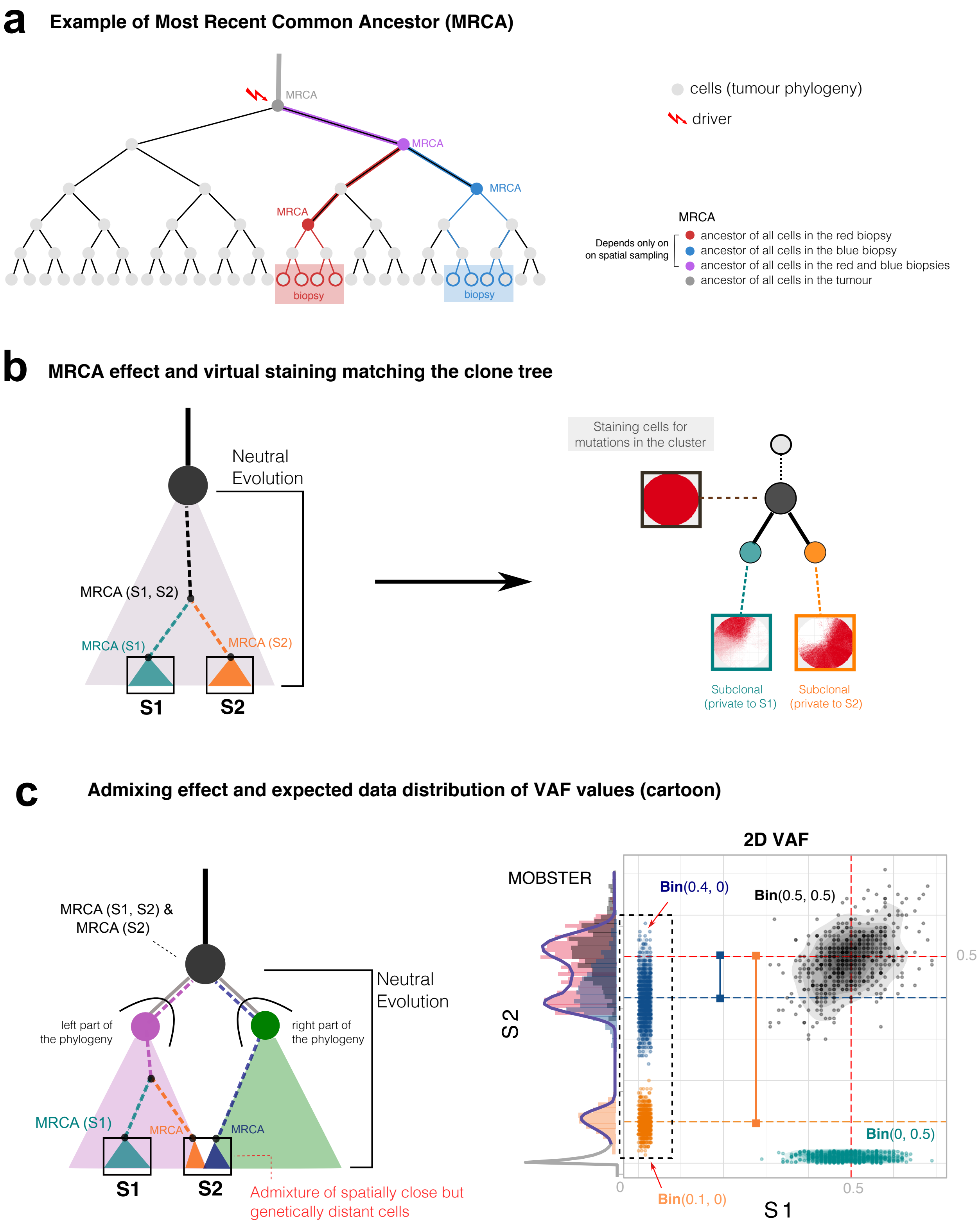
a **Dirichlet mixtures**
Variational fit (n = 150)

b **Dirichlet Process**
10,000 MCMC steps per chain (n = 150)

c **Tail size with different coverage** (n = 480)
MOBSTER fit with ICL (n = 80 x 6)

d **Tail's size with different purity** (n = 320)
MOBSTER fit with ICL (n = 80 x4)


a
Muller plot of a polyclonal tumour
time
Founder cell
Subclone (expanding)
S2
Founder clone
S1
★ Driver event
Subclone hitchhikers
Monoclonal biopsy
b
Phylogenetic tree (sketch)
Founder clone
time
Cutting point of the hitchhikers depends on S1 and S2
Tail
S1
Subclonal expansion
Tail
S2
c
Expected data distribution of VAF values (cartoon)
Hitchhikers frequency
Subclone
Rest of the tumour
100%
100%
100%
100%
100%
~50%
~25%
~12.5%
~6.25%
0%
S1
S2
0
0.5
1
VAF
Cells without the subclonal driver
Cells with the subclonal driver
Binomial clusters (standard analysis)
S1 against S2
Bin(0.5, 0)
Bin(0.5, 0.5)
Bin(0.5, 0.2)
Bin(.., ...)
0.5
S2
0
0.2
0.5
S1
Hitchhiker mirage
driver
Does not reflect a clonal expansion, it can be removed with MOBSTER (tails)
driver
+ tail clusters

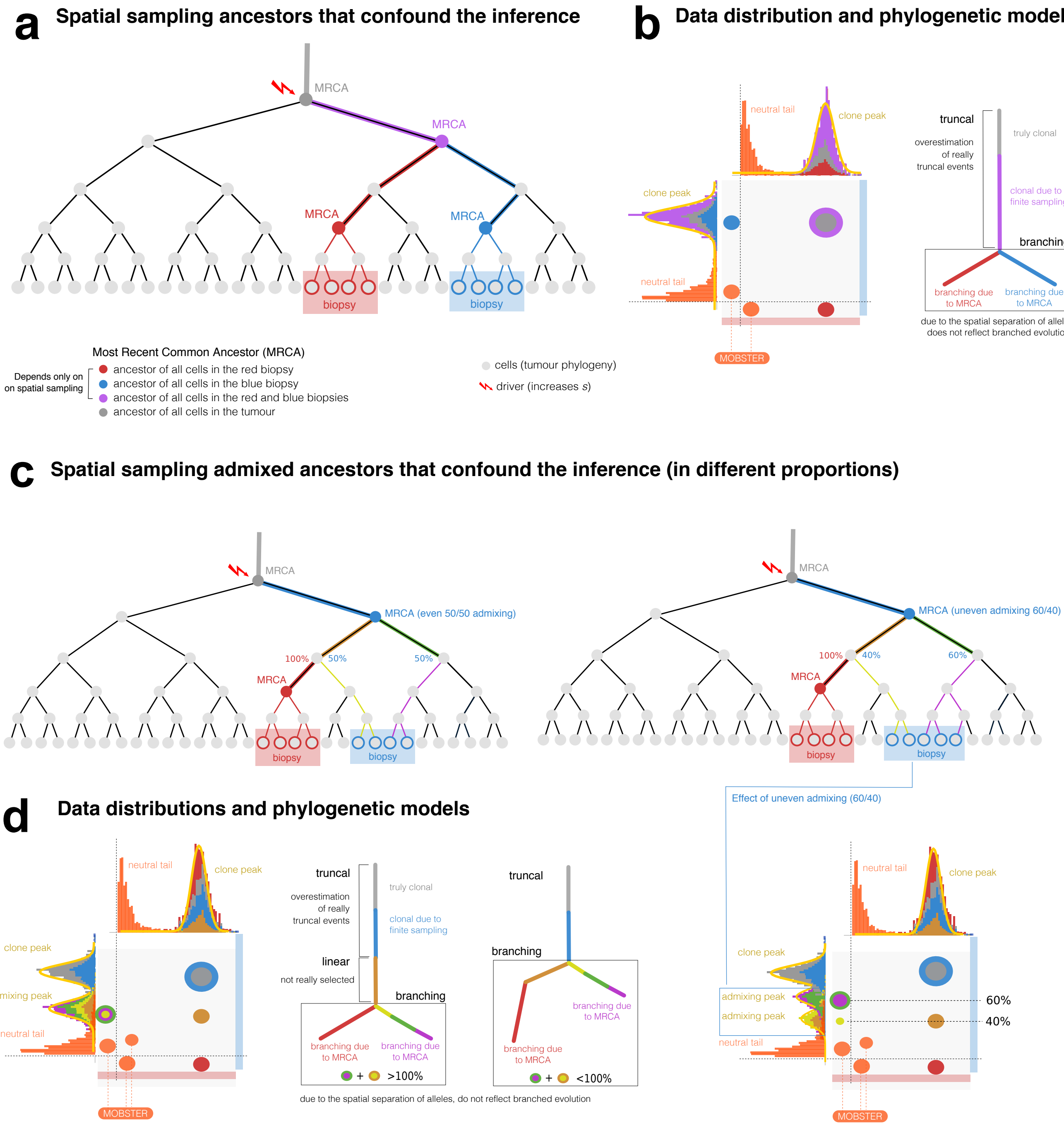


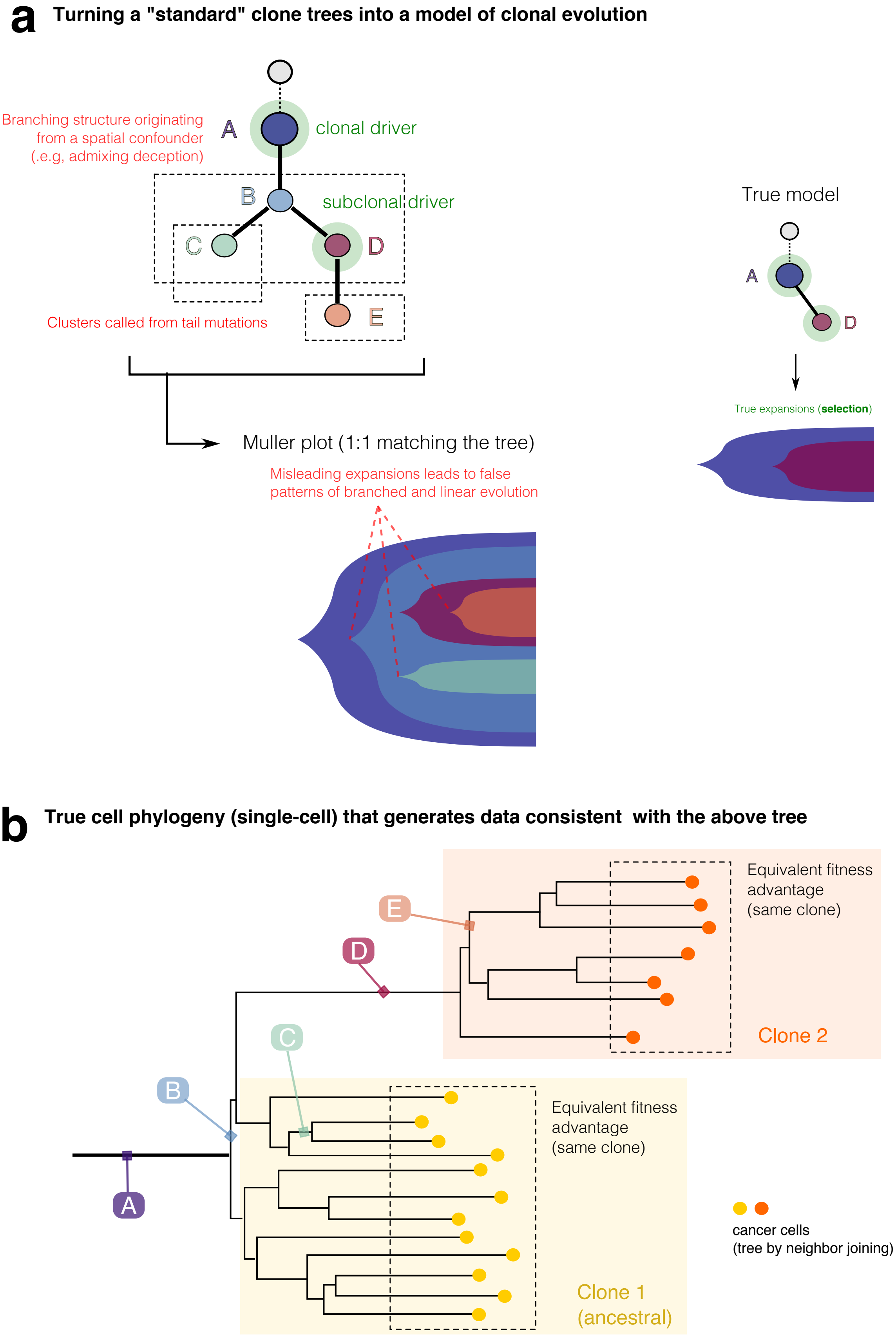
a Example of Most Recent Common Ancestor (MRCA)

b MRCA effect and virtual staining matching the clone tree

c Admixing effect and expected data distribution of VAF values (cartoon)

a
MRCA
MRCA
MRCA
MRCA
biopsy
biopsy
● cells (tumour phylogeny)
driver
MRCA
Depends only on on spatial sampling
● ancestor of all cells in the red biopsy
● ancestor of all cells in the blue biopsy
● ancestor of all cells in the red and blue biopsies
● ancestor of all cells in the tumour
b
Neutral Evolution
MRCA (S1, S2)
MRCA (S1)
MRCA (S2)
S1
S2
Staining cells for mutations in the cluster
Subclonal (private to S1)
Subclonal (private to S2)
c
MRCA (S1, S2) & MRCA (S2)
left part of the phylogeny
right part of the phylogeny
Neutral Evolution
MRCA (S1)
MRCA
MRCA
S1
S2
Admixture of spatially close but genetically distant cells
2D VAF
MOBSTER
Bin(0.4, 0)
Bin(0.5, 0.5)
Bin(0, 0.5)
Bin(0.1, 0)
S2
S1
0
0.5
0.5


a
Spatial sampling ancestors that confound the inference
MRCA
MRCA
MRCA
MRCA
biopsy
biopsy
Most Recent Common Ancestor (MRCA)
Depends only on
on spatial sampling
ancestor of all cells in the red biopsy
ancestor of all cells in the blue biopsy
ancestor of all cells in the red and blue biopsies
ancestor of all cells in the tumour
cells (tumour phylogeny)
driver (increases s)
b
Data distribution and phylogenetic model
neutral tail
clone peak
clone peak
neutral tail
MOBSTER
truncal
overestimation
of really
truncal events
truly clonal
clonal due to
finite sampling
branching
branching due
to MRCA
branching due
to MRCA
due to the spatial separation of allel
does not reflect branched evolutio
c
Spatial sampling admixed ancestors that confound the inference (in different proportions)
MRCA
MRCA (even 50/50 admixing)
100%
50%
50%
MRCA
biopsy
biopsy
MRCA
MRCA (uneven admixing 60/40)
100%
40%
60%
MRCA
biopsy
biopsy
Effect of uneven admixing (60/40)
d
Data distributions and phylogenetic models
neutral tail
clone peak
clone peak
mixing peak
neutral tail
MOBSTER
truncal
overestimation
of really
truncal events
truly clonal
clonal due to
finite sampling
linear
not really selected
branching
branching due
to MRCA
branching due
to MRCA
+ >100%
truncal
branching
branching due
to MRCA
branching due
to MRCA
+ <100%
due to the spatial separation of alleles, do not reflect branched evolution
neutral tail
clone peak
clone peak
admixing peak
admixing peak
neutral tail
60%
40%
MOBSTER



a Turning a "standard" clone trees into a model of clonal evolution

b True cell phylogeny (single-cell) that generates data consistent with the above tree

A
clonal driver
Branching structure originating from a spatial confounder (.e.g, admixing deception)
B
subclonal driver
C
D
E
Clusters called from tail mutations
Muller plot (1:1 matching the tree)
Misleading expansions leads to false patterns of branched and linear evolution
True model
A
D
True expansions (selection)
E
D
C
B
A
Equivalent fitness advantage (same clone)
Clone 2
Equivalent fitness advantage (same clone)
Clone 1 (ancestral)
cancer cells
(tree by neighbor joining)