Supplementary Notes and Figures for

Subclonal reconstruction of tumors using machine learning and population genetics

Giulio Caravagna,Timon Heide, Marc Williams, Luis Zapata, Daniel Nichol, Katevan Chkhaidze, William Cross, George A. Cresswell, Benjamin Werner, Ahmet Acar, Louis Chesler, Christopher C. Barnes, Guido Sanguinetti, Trevor A. Graham, Andrea Sottoriva

Correspondence: Andrea Sottoriva and Trevor A. Graham.
E-mail: andrea.sottoriva@icr.ac.uk, t.graham@qmul.ac.uk

This PDF file includes:

Supplementary Figures 1 to 23
Captions for Databases 1 to 2
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Databases 1 to 2

Contents

List of Figures

1. Note 1: Analysis of single sample simulated data

Required softwares.

- DPpackage (CRAN version 1.1-7.4);
- pyClone (Conda version 0.13);
- DPclust (GitHub version 2.2.5);
- SciClone (GitHub version 1.10).

Latest versions of the tools that we have developed and use here:

- MOBSTER (https://caravagn.github.io/mobster/), model-based tumour subclonal deconvolution;
- BMix (https://caravagn.github.io/BMix/), Maximum Likelihood Binomial and Beta-Binomial univariate mixtures;
- VIBER (https://caravagn.github.io/VIBER/), variational Binomial multi-univariate mixtures;
- CNAqc (https://caravagn.github.io/CNAqc/), data QC, processing and visualisation;
- TEMULATOR (https://github.com/T-Heide/TEMULATOR), non-spatial tumour growth simulator;
- CHESS (https://github.com/sottorivalab/CHESS.cpp), Cancer Heterogeneity with Spatial Simulations.

A. Simulated datasets. We used a non-spatial simulator for a stochastic branching process of tumour growth, which is implemented in the TEMULATOR non-spatial tumour growth simulator. This tool has been previously described(1), the data are discussed in vignette "2. Simulated single-sample data analysis" available as Supplementary Data.

The non-spatial simulator is based on the Gillespie algorithm(2), and was used to generate the data we used to test MOBSTER with single biopsies. The method allows the simulation of variant allelic frequencies (over time) of mutations accruing during the growth of a tumor with a known clonal structure. The following small modifications were made. Instead of a Poisson distributed coverage C_i of a mutant allele x_i, an over-dispersed Beta-Binomial distribution was used. Given the averaging sequencing depth c and a constant small dispersion parameter $\rho = 0.08$, per-allele coverage C_i values were determined as

$$C_i \sim \mathsf{Bin}(n = c/\mu, p_i)$$

with $\mu = 0.6$ and

$$p_i \sim \mathsf{Beta}(\alpha = \mu/(\rho - 1), \beta = (\mu - 1)(\rho - 1))$$

Variant allele frequencies values were assumed to be Binomial distributed, with the known fraction of mutated tumor cells in the population (x), given normal contamination ($1 - a$) and constant ploidy ($\pi = 2$)

$$VAF \sim \mathsf{Bin}(n = C_i, p = ax/(a\pi + 2 - 2a)) * 1/C_i$$

For the non-spatial simulations we simulated, similar to previous work (1), one ancestral population and just a single mutant subclone. Each subclonal driver has been inserted at a fixed time-step, selected so that the consequent stochastic dynamics where ending with subclones of sizes that were reasonably detectable with our simulated sequencing depth. This simplifies the synthetic data generation with respect to introducing a mutant stochastically (e.g. with Gillespie rates) as the majority of mutant clones would not be detected at all in the data and hence would not be useful to test MOBSTER (see Kessler and Levine for the expected frequency of selected subclones in a Luria-Delbruck model (3)). In the future, we might perform more extensive tests to relate the growth parameters of a tumor to the quality of the data it generates to perform subclonal deconvolution. The evolutionary parameters were set as follows, and kept constant through simulations: the tumor mutation rate was $m = 16$ (in mutations per cell doubling), the death rate was $\mu = 0.2$, the total number of reactions was $t_{end} = 179782830$. Data had average sequencing depth of 120 (120x simulated coverage) and the tumor had $N_c = 500$ clonal mutations. Nine random simulations for various subclone birth rates λ_s were perfomed with

$$\{\lambda_s = 1 + 0.1i \mid i = 1, \dots, 13\}$$

and number reactions (prior to initiation of a subclonal expansion)

$$\{t_s = 2^i \mid i = 4, \dots, 14\}$$

were simulated (the clock t being expressed in reaction numbers). All simulations in which the subclone accumulated less than 50 mutations prior to its transformation (i.e. less than 4-5 divisions) were removed and three datasets with specific fraction of mutated cells in the population (x_s, the CCF of the subclone) were generated by randomly selecting from the remaining simulations as follows:

- 20 effectively neutral cases where $x_s < 5\%$, and 20 with $x_s > 90\%$;
- 110 cases with a detectable subclone, with $20\% < x_s < 80\%$.

These cases represent tumors with small subclones (almost undetectable), tumors where the subclone has swept through the overall population and cases where the subclone is detectable within the VAF spectrum. We also remark that the genomes that we simulate are diploid (hence CCF$= 2*$VAF), which makes it simpler to assess the performance of the method regardless our ability to call copy number. In the future, we might also improve on this respect the type of simulated cohorts.

To analyze the influence of sequencing depth as well as sample purity, two additional datasets with variable purity ($a = \{0.3, 0.6, 0.7, 0.9\}$) or variable average depth ($\{40, 60, 80, 100, 200\}$) and otherwise identical parameters were created.

B. Example fits of high-quality WGS data. We used synthetic data for $n = 150$ non-spatial tumors to measure how tails confound subclonal deconvolution (30 cases of neutral tumors with 0 subclones, 120 with one subclone). To run MOBSTER we adjust the observed VAF, knowing that simulated mutations have no coy number associated (balanced diploid). We remark that copy number events and mutation multiplicities (i.e., the number of copies of the mutation in the cancer genome) are another confounder that shape the adjusted VAF distribution. The formula for the standard adjustment of allelic frequencies reads as

$$C = 1/2 * (V * [(m + M - 2) * p + 2])/(c * p)\,,$$

where V is the raw VAF (ratio between depth of reads with the mutant allele, over overall depth at the mutant locus), m and M the minor and major copy number of the mutation (so $m + M$ is the segment ploidy), p is sample purity, and c are the copies of the mutant allele. This is half the value of the Cancer Cell Fractions, and allows using Beta distributions that range in $[0, 1]$ for the peaks. In our simulated tumors $p = m = M = c = 1$, and thus $C = V/p$ as expected (adjustment for purity).

Unless otherwise specified, we have run MOBSTER (Supplementary Figure 1) with these parameters: number of clones $k = 1, \ldots, 5$ (fit with and without a tail, for every k), initialisation via peak detection and 10 independent runs for each parameters configuration, fit via iterative method of moments, with a cutoff at 6000 iterations, and stopping when the variation in the estimate of $\boldsymbol{\pi}$ is smaller than $\epsilon = 10^{-8}$. Thus, for each tumour, $5 \times 2 \times 10 = 100$ independent fits are compute; the best model is the one that minimises the score used for model selection (usually ICL or reICL) across all runs. MOBSTER fits of tumours with and without subclones are in Supplementary Figure 2, for simulated whole-genome sequencing (WGS) data (120x coverage, 100% sample purity). Errors in subclonal deconvolution happen when MOBSTER's fits do not match the generative model.

Typical mismatches are summarised in the same figure, and can be classified as follows:

1. *when there is an almost total, but not complete, subclonal sweep, and one mixture component fits multiple clones.* This happens because MOBSTER clusters the frequency spectrum with Beta distributions, which have variance large enough to cover multiple, close one another, Binomial peaks. This limitation can be solved in a downstream analysis of non-tail read counts, where the Binomial variance will depend explicitly on coverage, and model-selection of the mixtures size can split a large-variance Beta into multiple Binomials. We note that this might not hold using Beta-Binomial distributions in downstream anlaysis;

2. *when a subclone is "hidden" by the tail.* This is a genuine error of MOBSTER, and is a general problem of clustering groups of observation that, by definition, overlap (the tail and the subclone). This type of error seem is likely when the subclones are small compared to tail mutations, and therefore the signal of tails overloads the subclones.

3. *when there is a statistical competition between a model with low-frequency subclones, and a tail.* In some circumstances, the ICL score favours a Beta fit to the leftmost part of the frequency spectrum, outputting a model without tail. In this case, MOBSTER will assign genuine tail mutations to a subclone, a minor inconvenient if we still identify the subclone. By design, the reICL score uses a reduced entropy to penalise for the overlap of subclones, achieving better intermix of tails and subclones. Such cases might be identified by alternative selection-based metrics.

C. Tail detection and fit precision with high-quality WGS data. For 150 tumours (Supplementary Figure 3) we measured:

- **Identification of the true** k**.** We have measured k, the number of clusters fit (neutral tumours $k = 1$, non-neutral tumours $k = 2$, one subclone) and other statistics to conclude that MOBSTER is highly accurate. We have paid particular attention to 59 simulated cases where we fail to fit the true k, and checked if the problems stem from the errors discussed in Supplementary Figure 2, or not. We have found that: 1) in 17% of cases the peak of the subclone hides below the tail (undetectable frequency), 2) in 70% of cases the subclonal sweep is almost complete, and MOBSTER fits one Beta component to jointly fit the clonal cluster and the subclone, and 3) in the remaining 13% of cases MOBSTER mistakes the true model.

- **Odds ratio**. To compare models with and without tails, we have computed ICL (or reICL) odds ratios between the highest-scoring models with, and without tail (e.g. a model with one tail and one clonal cluster, versus a model with one clonal cluster and a subclone).

- **Precision of the fit.** We have measured the fit precision as (i) the rate of true positives and false negatives, and (ii) as the euclidean distance between each predicted peak, and its closest true peak. To match peaks we have used a 5% tolerance (i.e., we have centred an interval at the true peak value, and accepted a fit within $\pm\delta_{\text{peaks}}/2$ with $\delta_{\text{peaks}} = 0.05$). Measurements show that both error rates and distances decrease when we fit a tail.

D. Confounding factors for tail detection: coverage and tumour purity. Sequencing coverage and purity impact on the detectability of low frequency subclonal mutations, and can make it virtually impossible to distinguish tails from genuine subclones. To assess their effect we have analysed 80 of the 150 generated tumour, in 5 configurations of purity and Poisson-distributed coverage ($n = 480$ tumours, Extended Data Figure 1). We have tested the two confounders separately, generating for each tumour WGS data (i) with mean coverage $40\times$, $60\times$, $80\times$, $100\times$, $120\times$ and $200\times$, and fixed 100%, purity and (ii) tumour purity spanning 0.3, 0.5, 0.7 and 0.9, at fixed coverage $120\times$. For every simulated dataset we have removed mutations with < 6 reads harbouring the alternative allele, matching a 5% VAF cutoff at $120\times$ (Supplementary Figure 3).

Simulations show that it can be difficult to fit tails with coverage below $80/100\times$, even with perfect tumour purity. A similar trend has been observed for purity below 0.7 at $120\times$. These considerations derive from the analysis of both the size of the tail (i.e., number of mutations assigned), and the percentage of tumours fit with a tail. Overall, they suggest why many studies with coverage below $100\times$ and variable levels of purity might have failed reporting tails before our work. In the future, these measurements could be used to aid experimental design, ensuring that suitable configurations of purity and coverage are selected to study cancer evolution.

E. Identification of subclones. One crucial feature of a subclonal deconvolution algorithm is its ability to identify subclones. Besides the tests discussed above, we have carried out specialised tests to measure this aspect of MOBSTER in complex VAF profiles.To speed up the process of generating these tumours, instead of simulating the underlying branching process and rejecting undesired simulations, we have sampled datasets directly from MOBSTER's generative model.

E.1. Subclones with different size and peak. In these tests we have used a clonal cluster with $n = 400$ mutations, and a tail with $n = 500$ mutations and randomly sampled Pareto shape. We have then simulated a subclone with peak ranging in $[0.05, 0.23]$, and with a number of mutations that spans from 50 to 1250. We have measured the probability of fitting a tail with ICL, as well as the probability of fitting the subclone with 10 independent runs per configuration of parameters.

Results in Supplementary Figure 6 show that MOBSTER can fit both the subclone and the tail for a wide range of parameter values, but that the overlap of tails and subclones complicates the inference. In some configurations the subclone turns out to be difficult to detect, as suggested from the average number of Beta components fit by MOBSTER. For instance, it is extremely difficult to call a subclone that is small (50% of the tail size), or very low frequency (< 0.1 VAF peak). In these cases, the data harbours a very weak signal of subclonal selection, i.e., the subclone hides perfectly under the tail and is almost impossible to detect it from the VAF profile. The subclonal peak is successfully identified in all remaining cases.

E.2. Fitting tumours that do not have a tail. Through a direct simulation of the VAF distribution we have assessed that MOBSTER is not biased towards "calling" tails with ICL (Supplementary Figure 6). This has been verified simulating a tumour with 2 clones (Beta components), and without tail. The VAF distribution has only two "bumps" for to the clonal cluster and the subclone. We have analysed only mutations with VAF above 5%, and simulated a subclonal peak spanning 0.05 to 0.2, with increases of 0.02 VAF. This mimics the positioning of a subclone that is growing out of the tail: we want MOBSTER to fit it with a Beta, not with a tail. Results suggest that the subclone is always detectable when its peak is above 7% (Beta fit). When the subclonal peak hits at the detectability limit, instead, the observed distribution of the subclonal mutations splits into two symmetrical shapes that decay like a power law. In this case subclones are undistinguishable from a tail, and MOBSTER mistakes them. These general results have shown that the method is largely reliable, and that it can correctly disentangle tails from subclones with suitable data in a variety of settings.

E.3. Complex mixtures with several subclones. We have simulated datasets with a very large number of subclones (Supplementary Figures 8 and 9), random Beta peaks and different variance values. The Beta variance reflects how wide the subclonal peaks are, a feature that captures sequencing over-dispersion (intuitively, like in a Beta-Binomial hierarchical formulation): the lower the variance, the lower the dispersion. In these tests we have used Beta components with the same variance v randomly sampled in the parametric interval $v \in [10^{-4}, y]$, where $y = 10^{-4}$ (low), $y = 10^{-3}$ (mid) and $y = 10^{-2}$ (high).

The density for some example Beta components is shown in Supplementary Figures 8. As in the other tests, we have measured fit precision matching simulated peaks with 3% tolerance. In panel B of Supplementary Figures 8 we report a mixture with 4 simulated peaks, which we underfit with 3 Beta components. At 3% tolerance, with 1 out of 3 matched peaks, the rate of false positives is $1/3$, and therefore the rate of true positives is $1 - 1/3 = 2/3$. The rate of false negatives is $1/4$, because one simulated component is not matched by any of the fit peaks. Concerning mixture size, the ratio between the simulated and the fit k is $4/3 = 1.3$, suggesting underfit. Examining simulated datasets and fits from Supplementary Figure 8, where each colour represents a component (simulated or fit), we see the effect of the Beta variance (low, mid or high) on the data, and its effect on the fit. Intuitively, we see that the signal with high variance (0.01) and many clones ($k = 7$) represent a very difficult setup because the mixture components largely overlap; in these cases it seems impossible to retrieve the true k.

The results from $n = 288$ simulated datasets are reported in Supplementary Figure 9. We have simulated mixtures with 100, 500, 1000 or 5000 points, generated from $k = 3, 5$ or 7 components. In these tests the tail's shape is randomly sampled between 0.5 and 3. The proportion of each component is randomly sampled (from a Dirichlet distribution with parameter 1), and the tests have been carried out for each configuration of Beta variance. The observed mean rates of true positives, false positives and false negatives suggest that MOBSTER is precise even with large mixtures or few samples, even for mid and low Beta variance. The ratio of true to fit Beta components as a function of the simulated mixture size, for the different input variances, has shown that the expected result (red dashed line) is missed only for cases that have many components with high variance. These are the type of cases pictured also in Supplementary Figure 8. Concerning the simulated tail, instead, we have fond agreement between the fit and simulated values of the shape. In these tests the diagonal red line gives the expected values for the tail; note that when a low-frequency subclone largely overlaps with the tail – i.e., the subclonal peak is close to 0 – the error in the tail parameter fit grows because the signals are hard to separate.

E.4. Model selection strategies. From the overall set of simulations analysed in Supplementary Figure 3, we performed model selection by using the BIC, ICL and reICL scores (Supplementary Figures 4). Results confirm the intuitions underlying these methods: (i) BIC does not minimise overlapping mixture components, and consistently calls two clones, (ii) ICL is the most stringent method to call tails, which are often dropped in favour of a Beta subclone to reduce the entropy among the tail and the clones, (iii) in our simulated tumours, reICL always calls a tail and subclones, since its reduced entropy term penalises only the overlap among Beta components, disregarding the tail, and (iv) ICL and reICL are similar and more precise than BIC. reICL calls tails that overlap with subclones, which is reasonable when we want to "clean" the signal from neutral mutations, and call subclonal peaks.

F. Comparison with popular competing methods. Our analysis shows that if we directly cluster data of tumours with a tail, then we overfit the number of clones regardless what method we use to approach subclonal deconvolution.

Preamble: Bayesian methods for Binomial mixtures. At the core of widely used tools like pyClone, DPclust, PhyloWGS, and SciClone (4–7) there are two Bayesian statistical methodologies for mixture modeling (see (8, 9) for a thorough introduction):

- non-parametric Dirichlet Processes that describe distributions over distributions (here over mixtures of any size); posterior estimates for those models are computed via Markov Chain Monte Carlo (MCMC) sampling;

- semi-parametric Dirichlet Finite Mixture Models for a mixture of a finite number of Binomial distributions; posterior estimates for these models are approximated via Variational Inference (VI).

We focus on how these methods determine the number of output clusters k. Both models use a parameter α to model the propensity at which a new cluster is created during the fit; this parameter can be set either to a predefined value (point estimate), or learnt from the data via Bayesian computations. The interpretation of α is according to adopted model, the intuitions follow.

- with Dirichlet Processes (DP), α is termed concentration or scaling. A draw from a DP is a full mixture distribution, and is usually obtained through well-known generative models (e.g., the stick-breaking construction) that exploits exchangeability theorems for random variables (9). If we analyse the input observations in a specific order, the behaviour for a new observation x_n given the previous $n-1$ is to draw a new sample from a baseline measure G_0 with probability

  $$x_n \mid x_{n-1}, \ldots, x_1 \sim \frac{\alpha}{n-1+\alpha}\,.$$

  A draw generates a new cluster and fixes its parameters according to some prior distribution $p(\boldsymbol{\theta})$; with Binomial components the sample is a new success probability p from a conjugate Beta prior. Thus, α drives the fit of k: in the limit behaviour $\alpha \to 0$, the realisations are all concentrated at a single value and $k = 1$. Instead, if $\alpha \to \infty$ the realisations become continuous, and infinite clusters are created. Note that in practice the models are semi-parametric because the number of observations is an upper bound to k; the theoretical construction, however, is truly non-parametric as k can grow to infinite for $n \to \infty$.

  We want to highlight that in many implementations of DPs, posterior estimates of α are computed using a conjugate Gamma prior $\alpha \sim \Gamma(a, b)$ with hyper-parameters a (shape) and b (rate). Notice that sometimes (e.g., in pyClone), the Gamma density is equivalently expressed through a scale parameter which is the inverse of the rate.

- with Finite Mixture Models fit via variational inference, the statistical model uses a Dirichlet prior on the mixing proportions. Assuming k components

  $$\boldsymbol{\pi} \mid \boldsymbol{\alpha} \sim \mathsf{Dir}(\alpha_i, \ldots, \alpha_k)$$

  where $\alpha_i \in \mathbb{R}$ are real values that determine prior beliefs on the mixing proportions. We usually set a constant across all dimension to have flat prior (i.e., $\forall i.\, \alpha_i = \alpha_*$) on the mixture weights.

  A variational method is semi-parametric, as it estimates k by considering all mixtures with less then an upper bound of components $k < K$. As for DPs we can take $k = n$ for n input points, but we suggest a much lower value to speed up convergence. Thus, α controls for the final number of clusters returned by the method. In a mean-field variational formulation, posterior estimates are approximated as

  $$p(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}) = q(\mathbf{Z})\, q(\boldsymbol{\pi}) \prod_{w=1}^{k} q(\theta_w)\,.$$

  where the q are suitable variational distributions. Here $\mathbf{Z}$ are latent variables, $\boldsymbol{\pi}$ mixing proportions and $\boldsymbol{\theta}$ parameters; the posterior distributions q are assumed to factorise. This formulation is an optimisation problem in which we minimise a Kullback-Leibler divergence, and we measure convergence by monitoring the *Evidence Lower Bound* (ELBO).

Synthetic tests. It is intuitive to expect that a direct Binomial clustering of all input read counts inflates k. Here, we have confirmed that this happens (i) with baseline implementations of the two statistical methodologies described above, and (ii) with popular tools for subclonal deconvolution that are based on the above methodologies. As baseline implementation of DPs we have used the R package DPpackage which provides a high-performance Fortran MCMC sampler (function DPbetabinom) that supports both scalar values for α, and the Gamma prior (10). As baseline implementation of variational mixtures we have used a new in-house multivariate Binomial mixture model, where we have full control over the parameters of the fit, and we can we monitor the ELBO to assess convergence. This new model is called VIBER (Variational Binomial Mixtures), and is detailed in Note 3.

The two tests have assessed different aspects of the problem, with the second test also assessing the impact of tool-specific post-fit heuristics(e.g., filtering for cluster size or strength of clustering assignments, clone merging and density smoothing). In both tests we have observed similar trends and systematic errors, with clusters originating from polyphyletic tail mutations. To assess the effect of α on the fit, we have scanned pointwise estimates of $\alpha = 1, 10^{-2}, 10^{-4}, 10^{-8}$ for both methodologies, and a prior $\alpha \sim \Gamma(0.01, 0.01)$ for the DPs. In this case we refer to the scale parametrisation of the Gamma prior which is adopted by the DPpackage, which corresponds to a parametrisation with rate $1/0.01 = 100$.

Comparison against baseline implementations of the methodologies. We have analysed all the high-quality simulated $n = 150$ tumours following the protocols discussed in the Main Text. For each tumour we have fit (1) the full set of read counts and (2) non-tail mutations identified with MOBSTER. To determine non-tail mutations we have used *hard clustering assignments* from the latent variables $z_{i,k}$ (responsibilities): $\text{cluster}(x_i) = \arg\max_{c=1,\ldots,k} z_{c,k}$. Notice that since it can happen that MOBSTER fits the data without a tail, then this means that in our tests we are also considering the ability of MOBSTER to fit tails to the data.

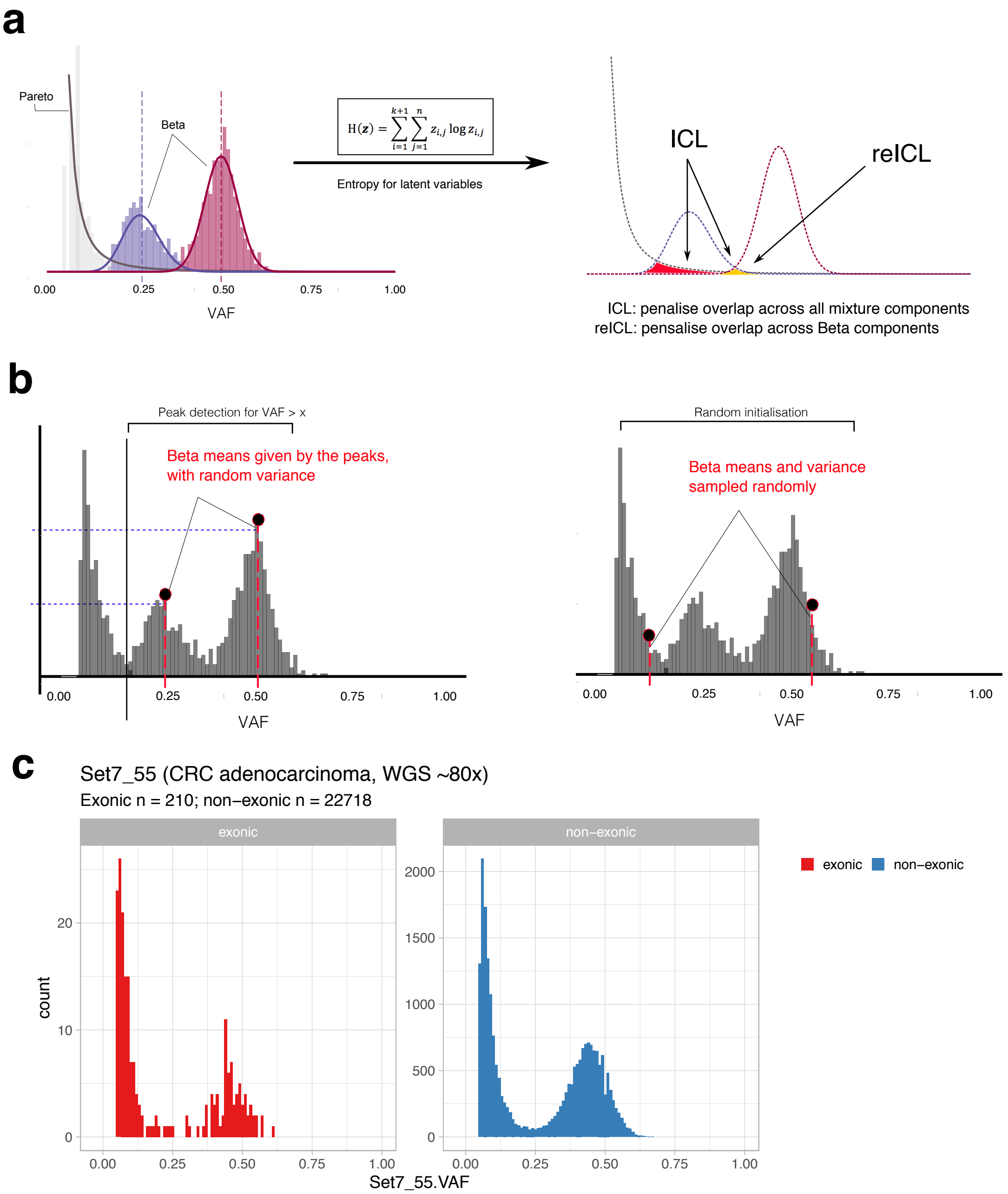
The fit parameters of the methods are as follows. With the DPpackage, we have used 10,000 MCMC steps with a burn-in of 5,000 steps and a thinning of 3, which gives over $3,000$ samples to estimate the posterior distributions of the parameters. With VIBER's variational method we have set the upper bound on the number of clusters $k = 5$, for the Beta priors of the Binomial mixture components a standard flat prior with $a_0 = b_0 = 1$, and we have monitor the ELBO's change until they are below $\epsilon = 10^{-10}$. To avoid local optima of this method we have also sampled 10 independent initial conditions for the fit, without observing substantial variations across results. Key results are in the Main Text, and a full set of results is available as Supplementary Figure **??**. From the ratio $r = k_{\text{fit}}/k$ between the estimated number of clusters k_{fit} and the true k we observe large errors ($r > 1$, overfit) without MOBSTER. For tumours without subclones the tail is large and the median error for high α can reach $r = 4$: i.e., we use $k = 4$ Binomial clusters (1 clonal cluster, plus three subclones), where there are 0 simulated subclones. The error diminishes with lower values of α, but persists. This demonstrates how input parameters influence the outputs. For tumours with subclones, tail and subclones overlap and the error is reduced but remains inflated. The trend observed for variations of α is the same across all tumours, regardless of the number of subclones. These errors reduce almost to zero if one uses MOBSTER to "clean" the signal from tail mutations. We still observe a trend due to the variation of α, but it is evident that for $\alpha \leq 10^{-4}$ the true model is retrieved. With $\alpha = 10^{-4}$ we obtain the same performance of a Gamma prior on α, in the Dirichlet Process fit. Both a prior and a pointwise estimate of α cannot retrieve the true model, without MOBSTER. In general, the combination with lowest variance and highest precision of the fit, is obtained by combining MOBSTER with the Dirichlet Process fit, and stringent $\alpha = 10^{-4}$.

Comparison against well-known tools. In the Main Text we have shown a comparison carried out using DPclust, pyClone – with both Binomial and Beta-Binomial distributions – and sciClone on the same test set described above ($n = 150$ tumours, $120\times$ median coverage and purity 100%). DPclust and pyClone are based on Dirichlet Processes, sciClone on variational Dirichlet finite mixtures. The tests above has been used to identify optimal values for α, a parameter which is hardcoded in the implementation of DPclust and sciClone. Parameters and comments on these simulations are reported in the Main Text.

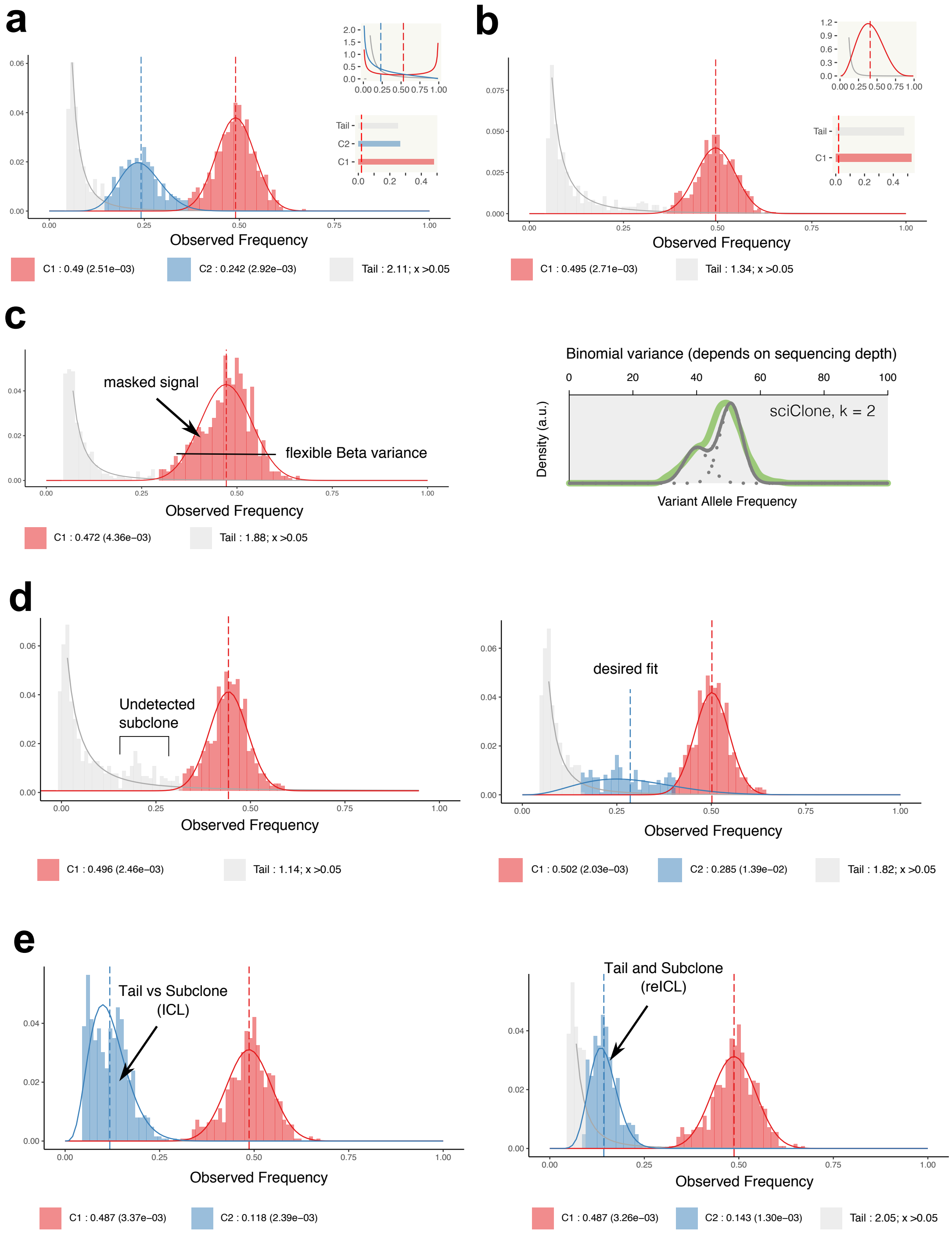
G. Evolutionary parameters from fits. We used MOBSTER fits to measure quantitative evolutionary parameters of tumor growth (Supplementary Figure 5). We could do that by adapting the computations originally carried out in Williams et al. (11). A model-derived distribution for these values could be created implementing a parametric bootstrap strategy from MOBSTER fits, re-sampling the data to fit new models and determine other measures of the evolutionary parameters. If we fit a tail with MOBSTER, we can derive the tumor mutation rate μ scaled by the probability of lineage survival β

$$\frac{\mu}{\beta} = \frac{M}{(1/f_m - 1/f_M)}$$

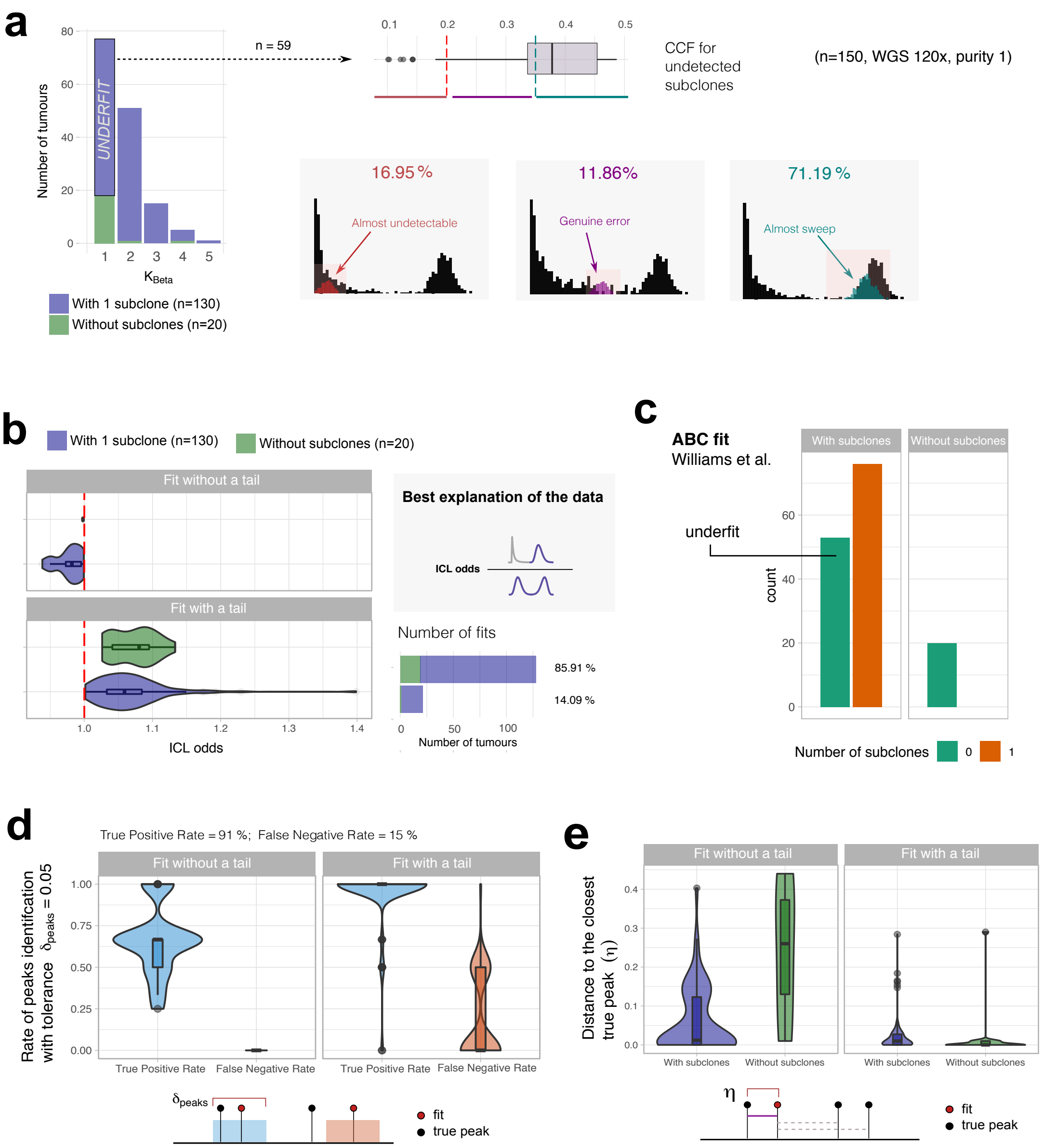
where f_m is the minimum VAF and f_M is the maximum, and M is the number of mutations between f_m and f_M. The ranges of the VAF distribution can be taken from the posterior assignments of the latent variables that map mutations to the tail (if any), to avoid outliers we use the mutations within certain empirical quantiles (for instance, 2% and 98%). The time unit of this rate are tumor-doubling times, and the conversion to rate per base pairs can be computed dividing this value by number of sequenced nucleotides; for a whole diploid genome this conversion factor would be roughly 3×10^9. With the mutation rate and the fit parameters of each subclone we can calculate the time the subclone emerges, and its selection intensity s. Selection $s > 0$ is defined as the relative growth rates of host tumor cell populations versus the subclone. See our earlier work for a detailed explanation of the formulas to derive the selection coefficient and emergence time for a subclones(11).



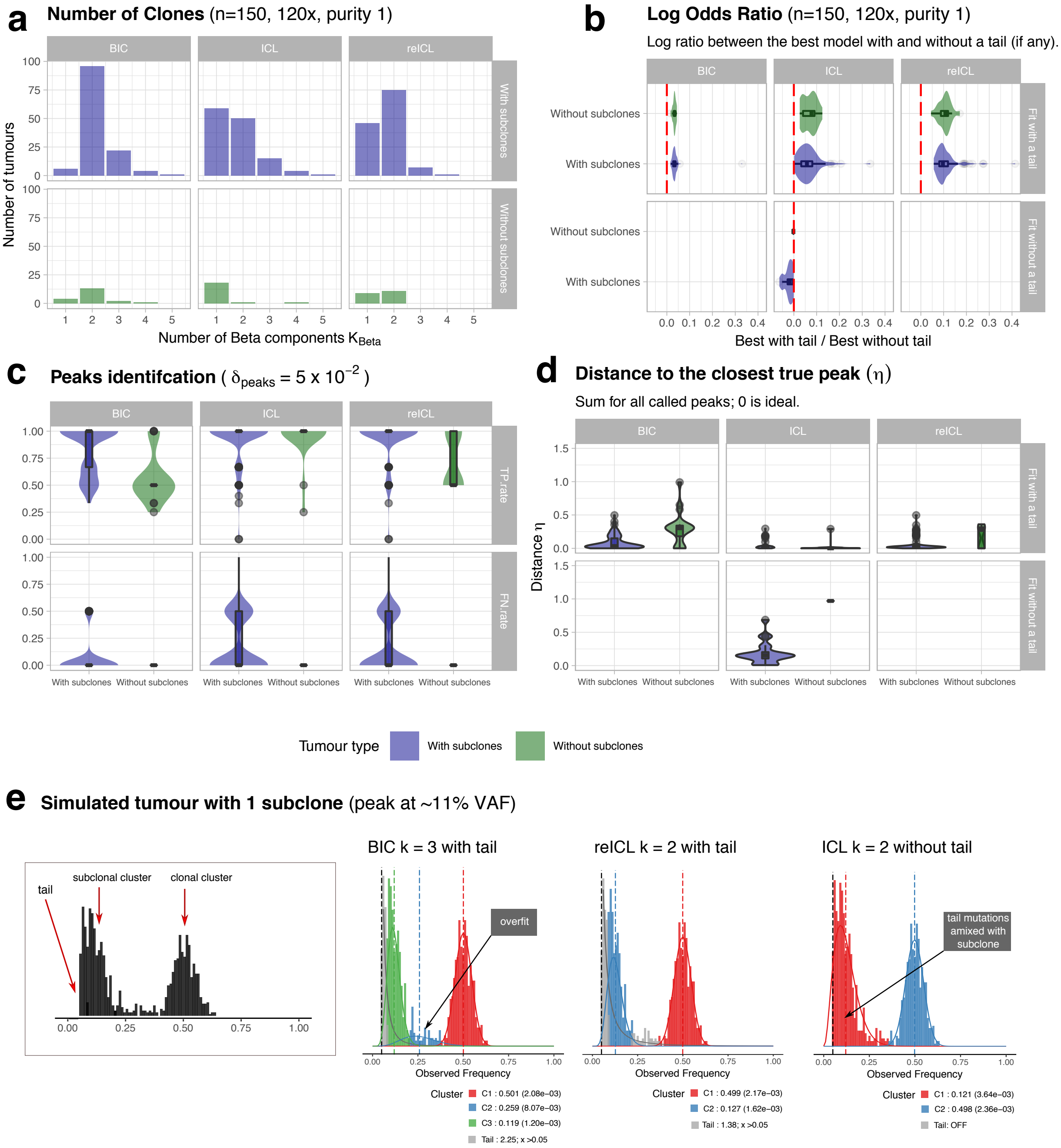
Supplementary Figure 1. MOBSTER**: model selection and initialisation. a.** MOBSTER model with $k = 2$ Beta for the clones, and a Pareto power-law tail. When we fit a model, we use the entropy $H(\mathbf{z})$ of the latent variables $\mathbf{z}$ to reduce the overlap between the components of the mixture. There are two ways to introduce the entropy penalisation, called ICL and reICL. ICL uses the entropy of all $\mathbf{z}$, penalising the overlap of all mixture components (Beta and tail). This can lead to excessive penalisations when $K > 1$, since a subclone by definition must overlap with the tail (red and yellow areas). This is a peculiar characteristic of this clustering problem, where the data points (subclone versus tail) are not separated. reICL is a variation to the ICL which uses a reduced entropy (yellow area) computed subsetting $\mathbf{z}$ to the mutations whose hard clustering assignments differ from the tail. To compute reICL the reduced set of $\mathbf{z}$ is renormalised, and a standard entropy is computed. **b.** MOBSTER's two heuristics to determine the fit's initial conditions. In left, peak detection is used to determine the two peaks and set the Beta means, while the variance is randomly sampled. In right, totally random initial condition (mean and Beta variance). **c.** VAF distributions for diploid mutations of a colorectal carcinoma (CRC) sequenced at median $80\times$ (whole genome, WGS). The two distributions show ~ 200 mutations that map to the exonic part of the genome, and ~ 22000 that map outside. In a case like this, the exome distribution seems to preserve most features of the whole-genome counterparts. Analyses of this data are described in Supplementary Note 4.



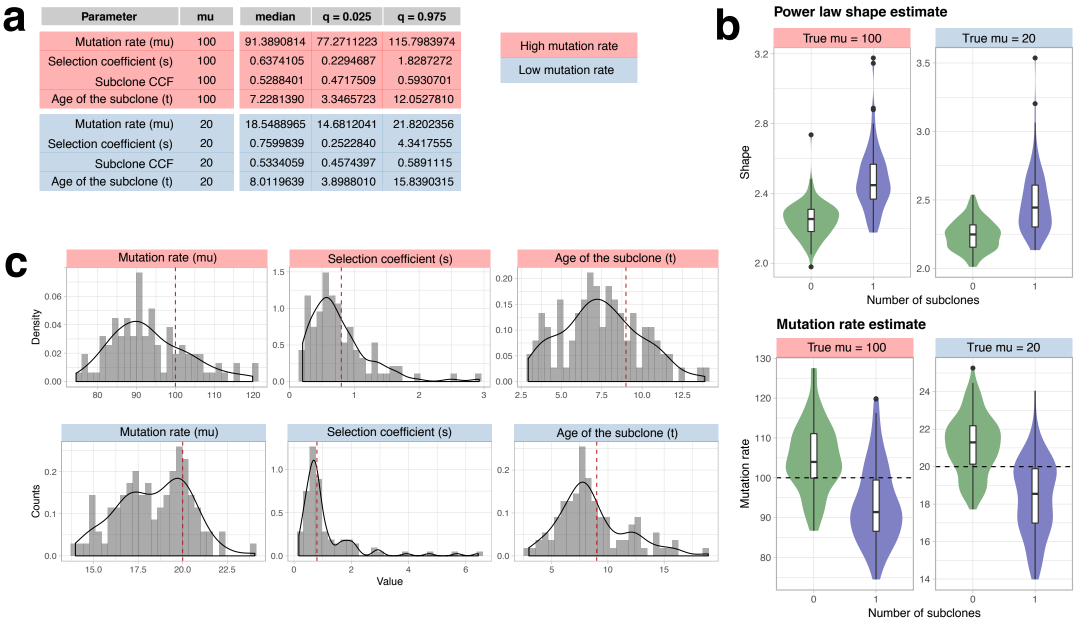
Supplementary Figure 2. Examples MOBSTER **fits with simulated tumours: WGS coverage mean $120\times$, purity 100%, diploid mutations.** **a.** Perfect fit for a polyclonal tumour with 1 subclone. The top histogram is the VAF data coloured according to hard clustering assignments; the tail is in grey and the components' parameters are reported in the caption (mean and variance). Smaller panels show the initial condition of the fit computed by MOBSTER's peak detection routine, and final mixing proportions. **b.** As in panel a), perfect fit for a monoclonal tumour. **b.** As in panel a), perfect fit for a monoclonal tumour. **c.** With ICL, MOBSTER fits $k = 1$ Beta components to a tumour where the subclone has almost totally swept through. This can happen because Beta distributions can have flexible variance. This problem can often be fixed in downstream analysis of read counts for non-tail mutations via Binomial distributions. Here this is shown with sciClone, which correctly retrieves $k = 2$ components from the mutations in C1. **d.** Example of a false negative with MOBSTER, which missesa small subclone with peak at 0.25. In the right part of the panel we show the fit that would have matched the true model (ranked 3rd with ICL). **e.** Fit for a tumour with a very low-frequency large subclone. With ICL, this induces a competition due to the entropy term computed on all the latent variables (therefore penalising mutations at the crossing of the subclone and the tail). With ICL, the best model contains a Beta cluster which, however, contains extra mutations that are clearly part of the tumour tail. The alternative model selections strategy that uses reICL can mitigate these situations and call both a tail and a subclone (see also Supplementary Figure 4).



Supplementary Figure 3. MOBSTER **fit of $n = 150$ synthetic single-sample tumours: general statistics.** All boxplots and violins show mean and inter quartile range (IQR), upper whisker is 3rd quartile +1.5 * IQR and lower whisker is 1st quartile - 1.5 * IQR. **a.** Identification of the true number of clones in the data, per tumour type. In 59 cases we do not detect the subclone in the data. We investigate this and find that in those cases input data harbours a weak signal of selection. Of 59 cases, in $\sim 17\%$ (10) a very low frequency subclone "hides" under a tail – this could be due to lo a small selective advantage compared to its ancestor. In $\sim 70\%$ (40) cases a subclonal sweep is observed, MOBSTER uses one Beta component to model overlapping clusters – these cases might be resolved in a subsequent analysis (Supplementary Figure 2A). In the remaining cases (9) MOBSTER genuinely misleads the subclone for genetic drift. **b.** We measured the confidence of the method's prediction by computing the ICL odds between the top model with, and without tail. Odds above 1 support tails. **c.** We also run the analysis with the ABC method described by (11), which uses a branching process simulator to fit the data. This method by construction will always fit a tail to the VAF distribution; in this tests we find underfit in 60 cases of tumours with a subclone where ABC fits a model without subclonal clusters. **d, e.** From the $n = 150$ simulated cases we measured the precision of the fit as the rate of calling or missing a true peak (true positive, false negative). Peaks are matched with tolerance $\delta_{\text{peaks}} = 0.05$, and MOBSTER is very precise. We measure also the distance between the predicted fits and their closest true peak. Results show that tails reduce the error of the fit, consistently.

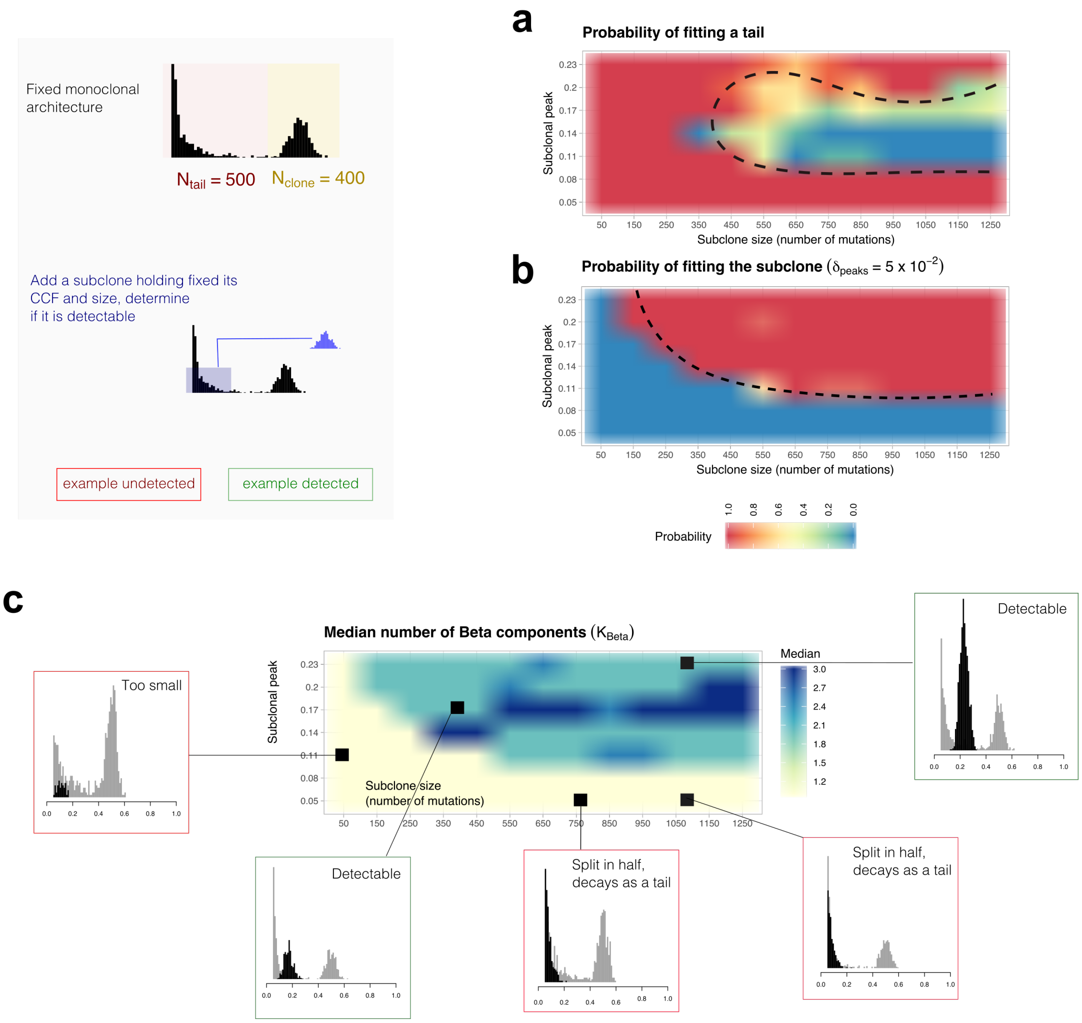


Supplementary Figure 4. Model selection with BIC, ICL and reICL in MOBSTER**.** All boxplots and violins show mean and inter quartile range (IQR), upper whisker is 3rd quartile +1.5 * IQR and lower whisker is 1st quartile - 1.5 * IQR. **a.** For the same $n = 150$ synthetic tests described in the Main Text, we show the number of clones detected by performing model selection with BIC, ICL and reICL. **b,c,d.** Similarly, we plot the same measures of precision and fit confidence for the best models obtained with those different criteria of model selection. **e.** Example different fits for a tumour with 1 subclone. We can see how BIC calls multiple clones, plus a tail. reICL retrieves the true model and ICL drops a tail in favour of a subclonal component. This is due to the different entropy terms used in ICL, against reICL; BIC does not use entropy at all.

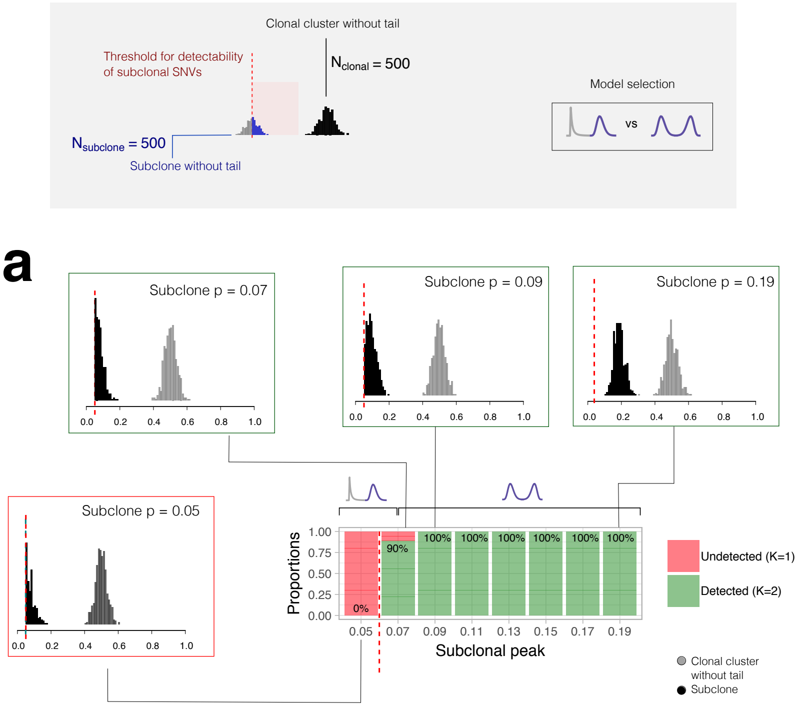


Parameter	mu	median	q = 0.025	q = 0.975
Mutation rate (mu)	100	91.3890814	77.2711223	115.7983974
Selection coefficient (s)	100	0.6374105	0.2294687	1.8287272
Subclone CCF	100	0.5288401	0.4717509	0.5930701
Age of the subclone (t)	100	7.2281390	3.3465723	12.0527810
Mutation rate (mu)	20	18.5488965	14.6812041	21.8202356
Selection coefficient (s)	20	0.7599839	0.2522840	4.3417555
Subclone CCF	20	0.5334059	0.4574397	0.5891115
Age of the subclone (t)	20	8.0119639	3.8988010	15.8390315

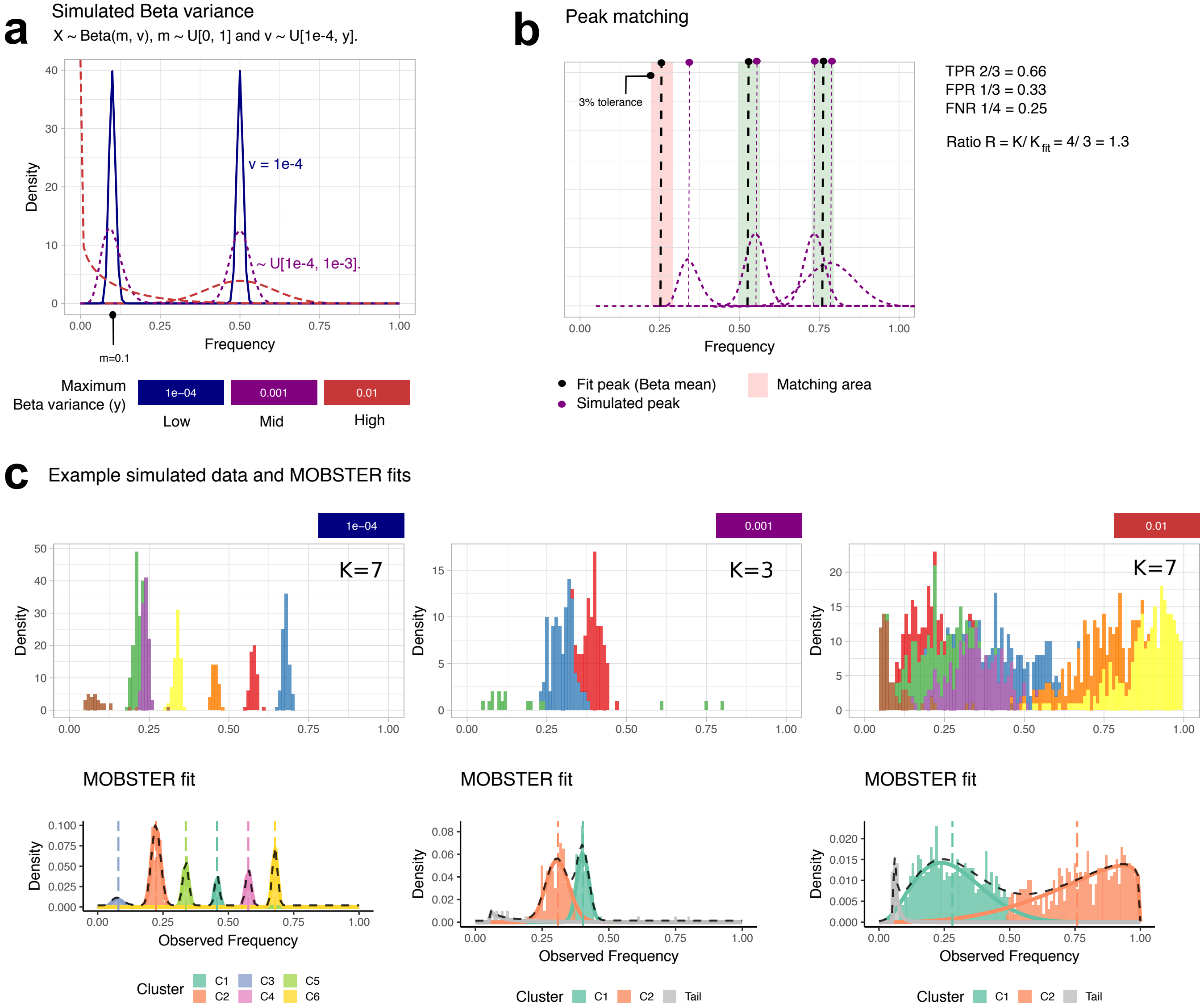
Supplementary Figure 5. Extra tests for MOBSTER**: determining evolutionary parameters from fits.** All boxplots and violins show mean and inter quartile range (IQR), upper whisker is 3rd quartile +1.5 * IQR and lower whisker is 1st quartile - 1.5 * IQR. **a.** Summary statistics of a batch of tumours ($n = 100$, with 0 or 1 subclone) simulated with low/ high mutation rate ($\mu = 20, 100$), from which we measured the evolutionary parameters from MOBSTER's fits. **b.** For the tests in panel a, boxplots of the power law shape estimate and the mutation rate (dashed line, ground truth), in tumour cell doublings. **c.** Histogram of the estimates of mutation rate, selection coefficient (s) and age of the subclone estimated; the simulated value (dashed line) is the line.



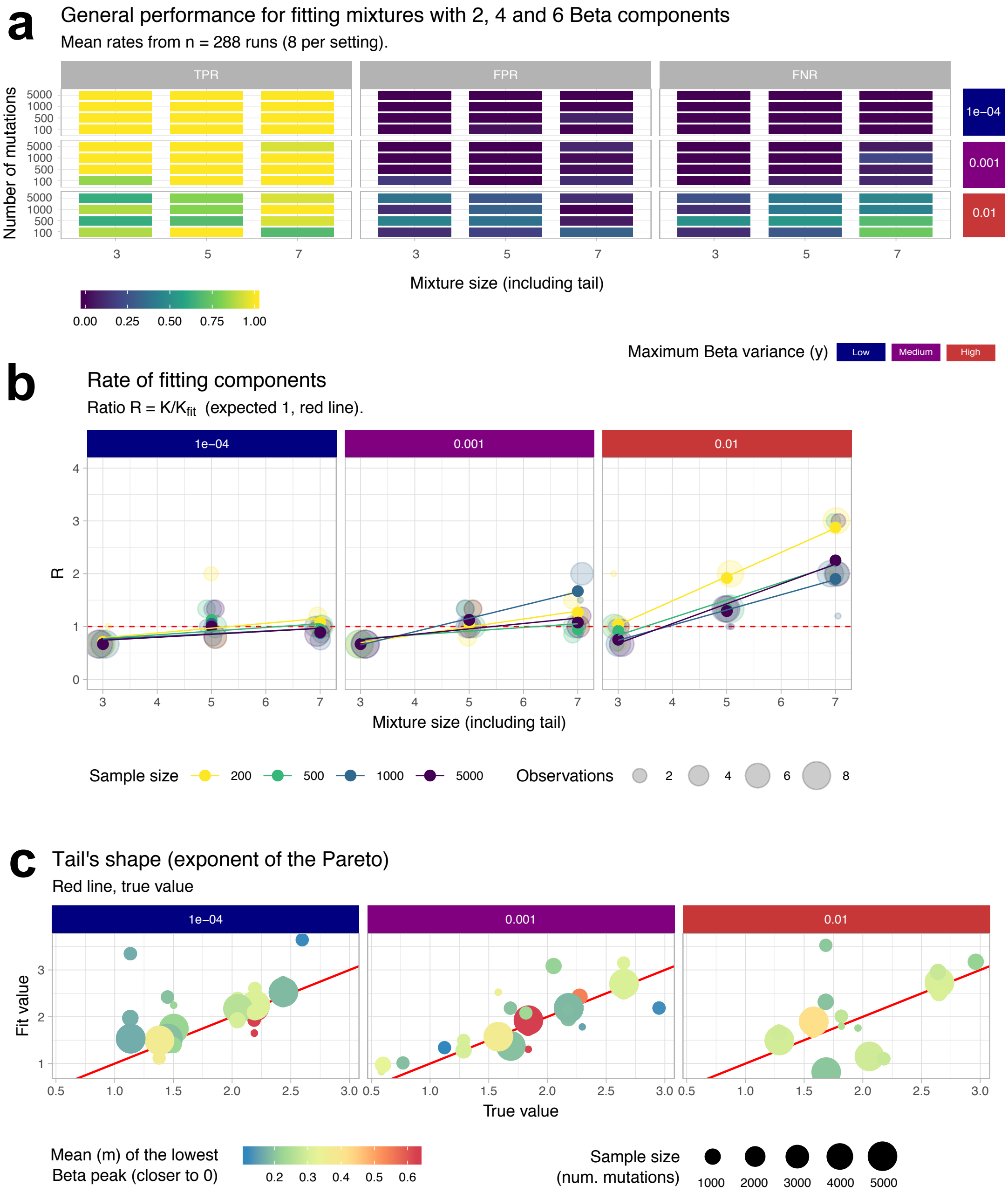
Supplementary Figure 6. Extra tests for MOBSTER**: calling small subclones.. a,b.** Through a direct simulation of the frequency spectrum we assessed the ability of the method to call subclones, as a function of their size (number of mutations), and peak (i.e., VAF/ CCF). We simulated a clonal cluster with 400 mutations (Beta $a = b = 100$, so $\mu = 0.5$) , and a tail with 500 SNVs (Pareto scale $x_* = 0.05$, shape randomly sampled in $[0.1, 3]$). We simulated a subclone with peak spanning from 0.05 to 0.23 (increases by 0.03), and with a number of mutations that span from 50 to 1250 (increases by 100). We measure the probability of fitting a tail, and compare it to the probability of fitting the subclone with 10 independent runs per parameters configuration. Results show when subclonal detection is difficult because the subclone is either small, or at very low frequency. **c.** We measure the median number of Beta components fit by MOBSTER. In the boxes we plot examples from this test.



Supplementary Figure 7. Extra tests for MOBSTER**: fitting tumours that do not have a tail. a.** We used the frequency spectrum to assess the bias of the method towards calling tails. We simulated a low-frequency subclone and a clonal cluster without a tail, both with 500 mutations each. We considered a detectability limit on allelic frequency of 5% (0.05) and simulated the subclonal peak spanning from 0.05 to 0.2, with increases of a factor 0.02. This mimics a subclone growing out of the frequency spectrum of the tail. In the panel we report the probability of detecting the subclone with a Beta (true model) against the probability of mistaking the subclone for a tail. Results show that the method is not biased, meaning that as soon as the subclonal distribution is evident (i.e., peak greater than 7%) the methods fits the subclone with a Beta, and the true model is correctly retrieved. The subclone is mistaken when its distribution is peaked exactly at the detectability limit of allelic frequency. In boxes some examples plot from this test.



Supplementary Figure 8. Tumours with large number of subclones: parameters, example data and fits. a. Density of the simulated Beta distributions shows 3 types of components (low, mid and high variance). In this test we want to assess MOBSTER's performance as a function of the variance of the Beta distributions used. **b.** Peak matching strategy with some fixed tolerance to measure the number of true positives, false positives and negatives comparing peaks in the fit to the generative model. **c.** Example simulated data for MOBSTER's generative models with $K-1$ Beta clusters, and a tail. The plots shows that for large K and large variance of the Beta distributions, many complonents overlap and is difficult to retrieve the true model. In the bottom row we show some example MOBSTER's fits of the simulated data.



Supplementary Figure 9. Tumours with large number of subclones: performance. With reference to the tests shown in Supplementary Figure 8, we have simulated data from $4 \times 3 \times 3$ configurations, repeated 8 times each for a total of $n = 288$ simulations. **a.** For models with 2, 4 or 6 Beta clusters (i.e., subclones co-existing), and mutations spanning from 100 to $5,000$ (low to high somatic load), we measured the rates of true positives, false positives and negatives comparing peaks in the fit to the generative model, as explained in Supplementary Figure 8. **b.** For the same tests, we report the ratio $(r = k/k_{fit})$ between the number of components in the generative model (k) and in the fit. The expect match is at $r = 1$; we plot this measure split by the variance used in the test to show that for large variance there is a systematic under-fit because mixing components are very much overlapping. **c.** For the same tests, we report the fit exponent of the Pareto power law tail (α) matched to the true value. All points here should be on the diagonal. This plot shows that when the esitimate is off, it is because there is a very low frequency subclone fit in the spectrum of the tail. The effect of this is to reduce the number of mutations that should be assigned to the tail, and therefore estimate a steeper decay than the actual one.

2. Note 2: Subclonal deconvolution from multi-region datasets

In this Note we investigate the influence of neutral tails on subclonal deconvolution from *multi-region sequencing data*. In the single sample case, we have theoretical understanding of the clone size distribution under neutral evolution ($\sim 1/f^2$) and so subclonal selection is detected as a deviation from this expectation. Uncertainty over the spatial structure of a tumour (e.g. the spatial spread of clones during tumour growth), coupled with the constraints and potential sampling-bias inherent to multi-region biopsies (e.g. variable number and physical sample size, non-uniform spatial sampling, heterogeneity in tumour content across tumour tissue, etc.) preclude an analogous simple theoretical understanding of the clone size distribution in spatial multi-region data. For these reasons, MOBSTER fits a general power law tail– rather than assuming a fixed exponent equal to 2 – which gives flexibility to consider deviations from strict exponential growth due to spatial structure (see (12)), or sampling bias.

In a sequencing dataset of multiple spatially-distinct[*] samples from a tumour, clone identification is performed by identifying groups of mutations that are at the same VAF/ CCF in one sample (e.g. represent a cluster in single-region sequencing data) and which remain clustered together in other samples from the same patient (13). Differences in the position of the cluster between samples represents differences in the abundance of the clone between samples (13, 14). We generated synthetic data from spatial simulations of tumour growth with and without subclones, and observed that same phenomenon that we see with single-sample data. Namely the fact that standard methods identify clusters from subsets of alleles that are neutrally evolving, giving the misleading illusion of numerous selected clones if we attempt to translate clone trees into clonal evolution models. Strikingly, this error grows dramatically with the number of biopsies collected (number of regions). This means that, contrary to intuition, more samples generate proportionally more noise than signal for subclonal reconstruction that seek to determine genuine clonal expansions related to positive selection forces. This unpleasant effect is generated by different sources of spatial sampling bias, and the complex way in which passenger mutations from neutral tails spread in physical space within a growing neoplasm.

In this Note we present (i) the simulation system that we have used to simulate tumour evolution in space, (ii) three theoretical concept of spatial confounders for tumour evolution estimation from multi-region data and (iii) the result of our spatial analysis. In Supplementary Note 2 ?Subclonal deconvolution from multi-region datasets? we describe the spatial data generation mechanism, as well as a system developed to perform virtual staining of simulated tissues which can be used to visually understand the diffusion of alleles in space, and the analysis of multi-region data with MOBSTER.

Simulation system for 2D tumours and virtual tissue staining. To observe the confounders we have carried out multivariate inference from synthetic data generated with CHESS, a spatial 2D simulator that we have recently published (15).

In brief, CHESS could generate tumors with multiple clones and 2 to 9 spatially-separated biopsies (10,000 tumour cells per bulk). iI this simulator birth (α) and death (ω) reaction rates of cells on 2D square lattice were simulated via the Gillespie algorithm (2). A cell selected to die was removed from the lattice. Daughter cell created during birth were either placed on a random empty neighbouring grid point (as of Moore neighbourhood) or if no empty adjacent grid point existed, on a neighbouring grid point that was freed by pushing adjacent cells into a random direction up to a given distance d. The number of additional mutations introduced into the genome of a daughter cell was drawn from a Poisson distribution with mean given by the mutation rate μ, a parameter of the simulation. New subpopulations were inserted at given time point t_j by modification of a random member of a selected subpopulation. This simplifies the synthetic data generation with respect to introducing a mutant stochastically (e.g. with Gillespie again) as the majority of mutant clones would not be detected at all in the data (too small or too late – see Kessler and Levine for the expected frequency of selected subclones in a Luria-Delbruck model (3)) and hence would not be useful to test MOBSTER. Simulation of sequencing data for each bulk sample (squares on the lattice) was done as for the non-spatial simulator.

To aid the explanation of the confounders discussed in this Note, we have implemented a "virtual staining" method that color all cells that have one or more mutations that we want to stain for. Each cell is coloured by its true clone's color, thus clonal mutations provide a representation of the simulated diffusion of clones in space, while private subclonal mutations represent wedges of such plot. With these plots, it is also straightforward to assemble the clone tree for complex scenarios by nesting staining plots, in a kind of Russian doll matryoshka effect.

A. Theoretical definition of spatial confounders affecting multi-region sequencing studies.

We first give a theoretical characterisation of the confounders that affect multi-region sequencing studies that do not account of the evolutionary properties of tumours growing in space. The confounders for simplicity are always represented in 2D models of tumour growth. In summary they can be described as:

- the *hitchhikers mirage*, which stems from breaking cell phylogenies into nested clusters that do not represent linear evolution;
- the *ancestor effect*, which refers to clusters of high-frequency mutations that do not imply positively selected clones;
- the *admixing deception*, from spatially-close genetically-distant neutral lineages that generate extra peaks in the VAF spectrum.

A.1. The hitchhikers mirage. In Extended Data Figure 2 we show the schematic of a tumour with a founder clone (blue) that gives rise to a new positively selected subclone (green). Two spatially distinct samples are collected from the tumour, respectively containing only the original clone (S1) and the new subclone with selective advantage (S2). Therefore, sequencing data from sample S2 contains the subclonal driver event – which is clonal in S2 – while both samples contain neutral passengers.

[*] We note that here we refer to samples collected from the same anatomical location (e.g., a primary tumour, or a metastasis). The case of a matched primary/ metastasis requires a slightly different interpretation of the idea of clone as a cellular population that manages to move and seed in a distant tissue de facto originates a new clonal expansion that fixates.

The marginal VAF distribution of each sample shows the expected neutral tail due to passenger mutations, and the joint tumour distribution shows multiple additional clusters on top of the two clones. The cluster(s) at low frequency (red) are due to neutral tails, as in the single sample case. The higher-frequency cluster (yellow) is instead much subtler, as it originates from passenger mutations that accrue in the lineage that goes onto forming the selected subclone (orange shaded mutations). When the subclone expands, in fact, the passenger mutations that accrue prior the subclonal driver are already present in every cell of the subclone, and therefore appear clonal in S2 (*hitchhiker mutations*). However, some of these hitchhikers (orange) are also present in branching lineages that segregate into other parts of the tumour, at varying frequencies.

For example, in a series of cell doublings (i.e., cell division 1, 2, 3 etc.), orange mutations reach frequency 50%, 25%, 12.5%, 6.25% etc. in the rest of the tumour. Since these passengers are detected also in sample S1, a clustering algorithm can "break" the group of subclonal mutations (orange plus green) into two clusters with different Binomial parameters. It is important to observe that splitting orange and green mutations is statistically consistent with the observation that these alleles "move" differently across biopsies. However, from an evolution point of view, the separate clusters have little to do with the actual selective forces shaping this tumour. In particular, the cutting point in the cell division tree where the clusters are split (between the blue and green clones), is totally determined by the anatomical location of the sampled biopsies, which means that a different sampling might have given a very different cutting point and clustering outcome. Moreover, the orange cluster does not even correspond to a single clone (in a traditional sense); rather, that clusters is just a mixture of neutrally evolving ancestors that happen to be found in S1, and is completely determined by spatial sampling. From a selection-centred evolutionary perspective, this type of "clone" does not seem interesting or useful to trace the evolutionary history of a tumour, as it gives the illusion of a linear sweep that has never happened in the tumour history. Note that a direct consequence is that the annotation of driver mutations into these type of clusters is of course wrong. Due to the rather subtle nature of this confounder, we refer to it as the "hitchhikers mirage".

An example of this confounder as it can observed in a simulated tumour is Supplementary Figure 10, where we show the staining that supports exactly the theoretical intuition of this confounder. Fortunately, this confounder can be – in principle – resolved if we remove neutral tails. Conceptually, if we run MOBSTER separately on each sample we can find that orange mutations belong to the tail in sample S2. Then, we can safely remove them from the subsequent read-counts clustering, which identifies only one cluster in the data.

A.2. The ancestor effect. Mutations found at high frequency in a small sample of a larger population can be erroneously associated with selection, following the fallacious argument that if a mutation has high frequency then it must have been selected. This is wrong because it does not consider the fact that, in multivariate comparative analysis (biopsy versus biopsy), we always expect to find most recent common ancestors that by definition are high-frequency, but have nothing special in terms of selection forces. We refer to this confounder as the "ancestor effect", and show it's theoretical concept in Extended Data Figure 3.

This confounder can be easily explained thinking of a monoclonal tumour (or analogously as rephrasing this confounder inside a subclonal expansion of a polyclonal tumour), where we sample two biopsies S1 and S2. By definition of most recent common ancestor (MRCA), we expect in each sample to detected private clonal mutations corresponding to each sample, which lead to the definition of the MRCA for each biopsy and the MRCA of both biopsies. The former MRCAs are defined as the cells that contained all mutations minus the private ones, the latter is the cell that contains the intersection of the mutations of both private MRCAs. We remark that in this monoclonal tumour no differentially selected clones exist. Multi-region sampling inevitably resolves the clonal ancestry of the samples (e.g. finds clusters) from these arbitrary MRCAs, which by construction are neutrally evolving ancestors picked up by the specific spatial sampling. This can be seen showing the clone tree associated with this data, where the two "clones" private to each biopsy (S1: green; S2: orange) are used to reconstruct the tree, and the cells in the simulation have been virtually stained for the mutations in each identified cluster (including the clonal cluster). This is bound to be ubiquitous in spatial sampling, as inevitably a localised biopsy will contain genetically more related cells. An example of the pattern that we expect to see in the data distribution as due to this confounder is sketched in Extended Data Figure 3.

A.3. The admixing deception. Sampling bias is particularly problematic when there is some level of admixing of distant lineages within a sample. By definition, we can divide the tumour in two ancestral lineages (notionally "left" and "right") that are the decedents of the first branch point of the phylogenetic tree (i.e. the first cancer cell division to produce two surviving lineages).

This is illustrated for the simplest possible neutral case in Extended Data Figure 4, and we note the scenario is further complicated when positively selected subclones are present. We analyse two spatially distinct tumour samples in this simple neutral case. Sample S1 contains cells entirely from the left lineage, whereas sample S2 contains, by chance, cells from both left and right lineages. This is bound to happen as somewhere in the tumour two distant parts of a phylogeny must meet. Sample S1 contains a large proportion of the lineages coming from the right tree (80%, blue), and a minor proportion of the lineages coming from the left part of the tree (20%, orange). These two groups of lineages have very distant MRCAs despite being spatially close, which causes the detection of subclones that are however entirely due to neutral evolution. Both samples contain the clonal mutations (black), but only sample S1 contains the green mutations (ancestor effect). In S2, however, the observed adjusted VAF can split into two subclonal clusters at CCF approximately 80% (adjusted VAF 0.4, blue) and 20% (adjusted VAF 0.1, orange). There are strong implications for our ability to identify the true generative model: the same subclone distribution can also be obtained from a tumour with subclones under selection, and distinguishing between admixing and positive selection is therefore extremely challenging.

This confounder needs to be considered because we expect clonal admixture within samples. The admixing skews the frequency of the dominant lineage in a sample by a small gap, enough to separate it from the clonal mutations (blue cluster is skewed to the left of the black cluster). An example of the pattern that we expect to see in the data distribution as due to this confounder is sketched in Extended Data Figure 4.

B. Reconstruction of tumours with multi-region sequencing data. In these tests we have crated synthetic datasets in which 2 to 9 biopsy samples are resected from a number of baseline synthetic tumours with one ($n = 50$), two ($n = 10$) or three ($n = 10$) subclones with increasing fitness (birth rates). Simulations were ended when any of its cells reached the edge of the 800×800 2D lattice, which accounts to roughly $\approx 5 \times 10^5$ cells. Time points at which a new subpopulation was introduced ($t_{\text{clone}} = \{0, 4, 6.7\}$) and corresponding birth rates ($\alpha = \{1, 1.6, 2.4\}$) were chosen to allow coexistence of each subpopulation at approximately equal abundance at the end of the simulation. The remaining parameters were kept constant: $\mu = 10$, $N_{\text{clonal}} = 100$, $\bar{C} = 100$, $\omega = 0$ and $d = 100$ (see the Online Methods for the relevant equations for simulating tumour dynamics). Bulk samples of about $10,000$ cells (100×100 pixels) were taken along the outer perimeter with an equal angular distance relative to the centre between them.

In order to focus on the effect of spatial confounders, in these tests we haved fixed the mean sequencing coverage to $100\times$, with purity 100% for all bulk samples. In the test we have compared the fits from a multivariate variational Binomial clustering run on read counts of all data, compared to the same analysis run on non-tail mutations detected with MOBSTER. In the context of multivariate analysis, tail mutations are computed running MOBSTER on each simulated biopsy, and computing the union of all mutations that are part of a tail.

B.1. Example tumours. In the case of a neutral tumour (Supplementary Figure 11) with 2 bulk biopsies, MOBSTER can filter out the tails from both samples. Analysis of the joint VAF distribution shows the relationship between the mutations in the samples, which cluster into $k = 3$ groups: one clonal cluster (green), and two subclonal clusters due to the ancestor fallacy (one private to S1, and one private to S2). Due to the admixing deception, the mean VAFs of these two clusters are slightly shifted below the expected clonal peak (about 0.5 adjusted VAF). The contamination of mixing lineages is minimal, and the second lower-frequency VAF cluster cannot be detected at the simulated resolution of $100\times$ WGS. The effects of ancestor sampling can be clearly appreciated in the virtual staining of the spatial distribution of the mutations, where we can color the cells based on mutational cluster assignments. In the staining we see that a tumour split by sampling into two subpopulations misleads subclonal estimation. Clustering without MOBSTER suffers from a much worse overfit effect, and detects twice as many clones ($k = 6$) with two false positive clusters of private tail mutations (clusters C1 and C2).

In Supplementary Figure 12 we use these principles to show that similar errors happen when a new subclone emerges within the background (blue staining). Three biopsies of this tumour are sampled, and MOBSTER is used to identify and filter out tail mutations. The VAF distribution for the boundary sample S1, which has a clear bump as predicted by theory, is a clear mixture of background and selected subclone. Sample S2 is entirely composed of the background clone, and shows intra-sample neutral dynamics. Sample S3, instead, shows a small cluster due to sampling bias and genetic drift, which is misleading because the biopsy is monoclonal and contains only the selected subclone. Multivariate subclonal reconstruction run after MOBSTER detects $k = 5$ clusters. Virtual staining of each cluster clearly shows that multiple clusters are the result of neutral drift (C1), ancestor fallacy (C4, C5) and admixing deception (C1, C4 and C5, which are shifted). This is where informed interpretation of the data needs to take place after subclonal reconstruction. A similar scenario can be identified if one analyses a tumour where two subclones are present, such as the one described in Supplementary Figure 13. Even in these case most of the clusters observed with a standard analysis can be explained with the proposed confounders.

B.2. Guidance for subclonal reconstruction with* MOBSTER *using multiple biopsies. We seek to find a conservative approach to reason on the output of subclonal reconstruction with multiple samples. Ideally, we would like to achieve a situation in which we can translate the output clone tree into a clonal evolution model, in a straightforward way.

In the presented neutral tumour the true model is a single monoclonal population. The subclonal reconstruction identifies $k = 3$ clusters, but two of them are just due to sampling two bulks. Being private to each sample they cannot be discriminated from the ancestor fallacy. A conservative analysis could then eliminate them, recovering the true model. Importantly, the exact same phylogeny could be obtained with two monoclonal samples from two distinct clones (e.g. equivalent to a sampling for a tumour with one subclone). In that case, however, it is reasonable to expect the length of the two branches outgoing C1 to be significantly different. A longer branch can indicate possible selection, under the assumption that the size of the biopsy samples is similar; if this were not the case, biopsy size would act as an extra confounder for branches length (larger biopsies have more cells, and therefore less private mutations as their ancestor is further back in time). In the polyclonal tumour with one subclone, we recover a more complex clone tree because we work with three biopsy samples. The true model here is a founder clone which gives rise to one positively-selected subclone. Cluster C2 corresponds to the founder clone and contains only truncal mutations. Cluster C3 contains mutations that belong to the common ancestors of both C5 and C1, but not C4. Importantly, all the leaves are clusters that are private to each sample. The most conservative clonal tree is then “linear evolution” of C2 (clonal) to subclone C3, the true tumour evolution. We note that the additional power in calling C3 a true subclone comes from the fact that the subclone has been observed in multiple biopsies.

Based on these observations, the strategy that we propose in order to minimise the effect of spatial confounders is *heuristic*. When for each Binomial cluster we have computed its parameter (peak) for each sample, we retain only those that have value above a threshold x in at least y biopsies (for instance $x = 0.05$ in $y \geq 2$ biopsies); usually we can test larger values of y according to the number of available samples. This is an intuitive heuristic which imposes a certain amount of “empirical evidence” to call a subclone; the parameters x and y should always be set based on the study and the data under scrutiny. Following this strategy we have tested the performance of MOBSTER against standard subclonal reconstruction with $n = 559$ spatially-simulated tumours, with different subclonal architectures and multi-region sampling as explained above. In each run we have measured statistics from the clone trees that can be reconstruct processing the output clusters, following this heuristic. We note that for read counts clustering we have used a in-house variational Binomial mixture model (Supplementary Note 3 of this document), run with and without MOBSTER.

Results are summarised in Supplementary Figure 14. A first observation is that, with more biopsies available, more clusters are observed. This is intuitive in light of the effect of the confounders that we have just explained, but the number of clusters is particularly high with the standard analysis that does not account for neutral tails. Concerning the proposed heuristics, MOBSTER results are more confident and less noisy than the ones that we can obtain without. The median error reaches 0% (top score) as measured in violations of the pigeonhole principle according to the clone trees generated our previously published method for clone tree inference from cancer cell fractions and subclonal deconvolution outputs (16). Briefly, from the output clusters the method computes all possible clone trees that fit each sample independently, and assemble the detected edges in a graph. Then, it scans the graph to detect all possible spanning trees rooted in the clonal cluster, and ranks trees using a scoring method for the number of violations of the pigeonhole principle[†] (perfect score 1 has 0 violations).

We are here interested in solving confounders that augment the number of trees, and measure their effect on the goodness of fit (as measured by the violations of the pigeonhole principle). Out tests suggest that, with our analysis, much fewer clone trees ($p < 0.001$, one-sided Binomial test) constructed via the standard pigeonhole principle can be fit to data if we run MOBSTER. This trend is observed in more than 400 cases, and in a large number of datasets we find one tree only if we use MOBSTER. As an alternative, the standard analysis finds a number of trees that spans from 5 to about 100 (max cutoff of the tree sampler). Precisely, in 38% ($n = 212$) of simulated cases we identify only one tree with MOBSTER, while in 16% ($n = 92$) of cases both analyses find only one tree. We have also found that trees derived after MOBSTER have median 0 violations (perfect fit), while "standard" trees have systematic violations of the principle. The number of violations with MOBSTER is also significantly lower ($p < 0.001$, one-sided Kolmogorov-Smirnov test), and results expanded by number of biopsies and subclones in the simulated cohort confirm this trend, and show that the error rate without MOBSTER increases with the number of samples taken.

Interestingly, we observe that the strategy that we propose to identify clones is less effective without MOBSTER because tail mutations that spread across samples tend to form clusters that co-occur in all samples, and thus no reasonable value of y (minimum number of samples where a subclone is detected) can identify them.

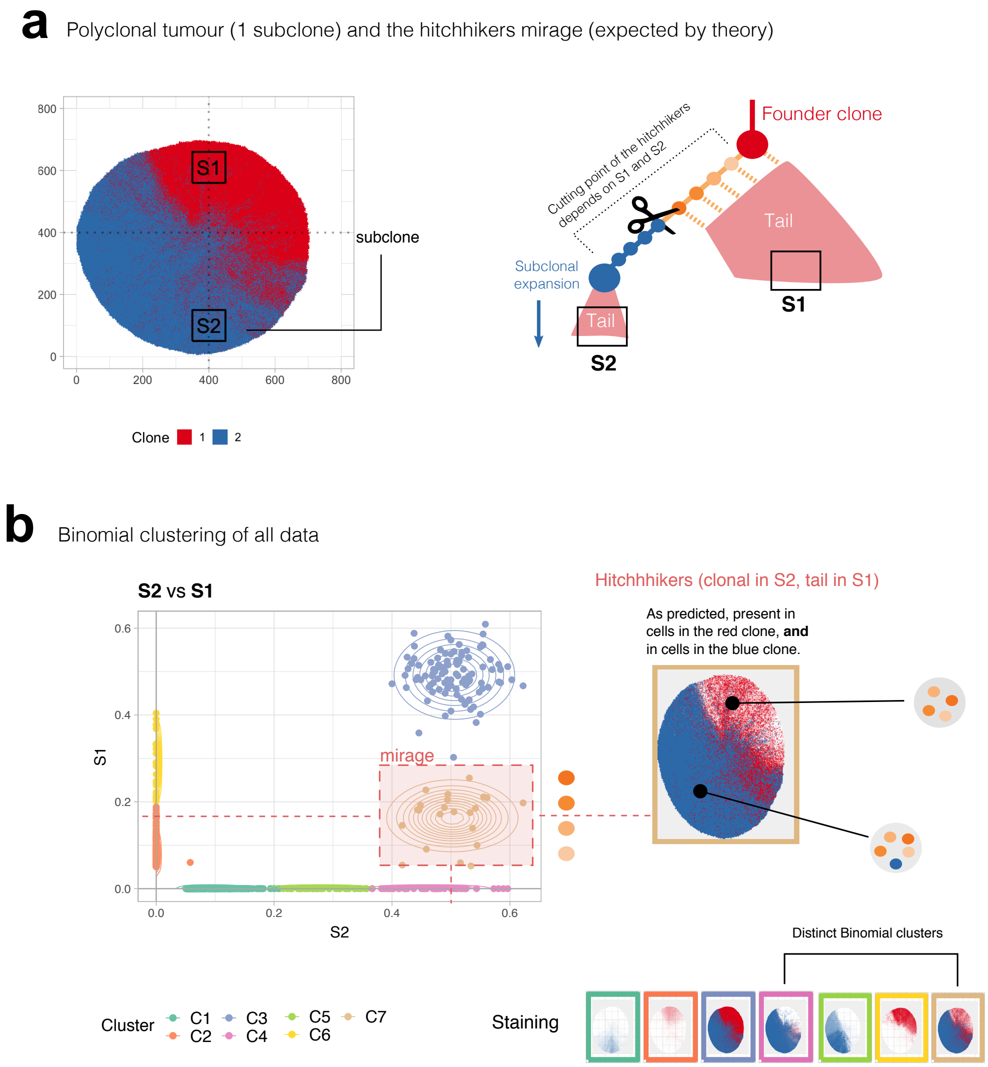
C. Summary. These analyses suggest that evolutionary interpretations derived from the structure of clone trees should take into account the effects of tumour spatial sampling bias. In general, these analyses also highlight that mutations that derive from neutral processes should be accounted for with MOBSTER, is we seek to determine clones that grow due to forces of positive selection. Importantly, these set of tests also show that the spatial confounders that we have identified are driving the output of standard subclonal deconvolution from multiple tumour biopsies. This problem can be partially ameliorated leveraging on MOBSTER, combined with a posteriori inspection and curation based on a heuristic that we have discussed.

It is fundamental to keep this ideas in mind when attempting to draw conclusions about the evolution of a tumour from the results of these analyses. In this process, we often feel legitimate to measure covariates or other parameters of interest from the output clone trees, as if they were representative of the actual clonal evolution of the tumour. *This is not safe,* because the confounders induce features in the clone trees that do not necessarily resemble the tumour's clonal evolution.

In Extended Data Figure 5 we show a hypothetical tree of putative clones recovered after filtering tail mutations with MOBSTER. The tree contains only two positively selected clones (A and D). The other clones (B, C and E) are just due to the confounding effects discussed above and represent arbitrary ancestors that evolved neutrally. There is nothing "special" about those ancestors: they are not phenotypically distinct subpopulations, and they have not experienced subclonal selection. Most importantly, resampling the same tumour would lead to different neutral clones to be sampled, thus a different tree where only the selected subclones would be the same. Drawing a Muller plot of the evolutionary history of the tumour from this clone tree would give a misleading picture of the history of subclonal selection in this tumour. This can be more clearly appreciated from the phylogenetic tree of the individual cells in the tumour, also represented in Extended Data Figure 5, where the true subclone is annotated in orange. When we map the "clones" from the clone tree into the true cell phylogeny, we can see that clones C, B and E are just random ancestors in the phylogeny. If we resampled the tumour again, we would have picked different cells, and therefore ancestors every time, obtaining a different clone tree. This is also true for the branching structure of the tree, which depends on cluster C. For this reason, drawing conclusions from the structure (e.g., linear, branching) or the size (e.g., number of clones) of the clone tree without accounting for these confounding factors, can be misleading.

Summarising, in this Supplementary Note we have proposed some guidelines on how to perform subclonal reconstruction on multi-region sequencing data using MOBSTER, integrating careful data interpretation using evolutionary principles. It is important to note however that depth and sequencing noise will inevitably limit the detectability of small subclones with any method, as most mutants are either too weakly selected or arrive too late to grow to a size that is detectable in the data (17). Hence, as the expected number of detectable subclones under positive selection in the data is low, in this study we assumed at most k possible subclonal clusters in the data as "prior" for our analysis. However, downstream analysis can be performed agnostically to the number of possible subclones (e.g. with non-parametric Dirichlet clustering).

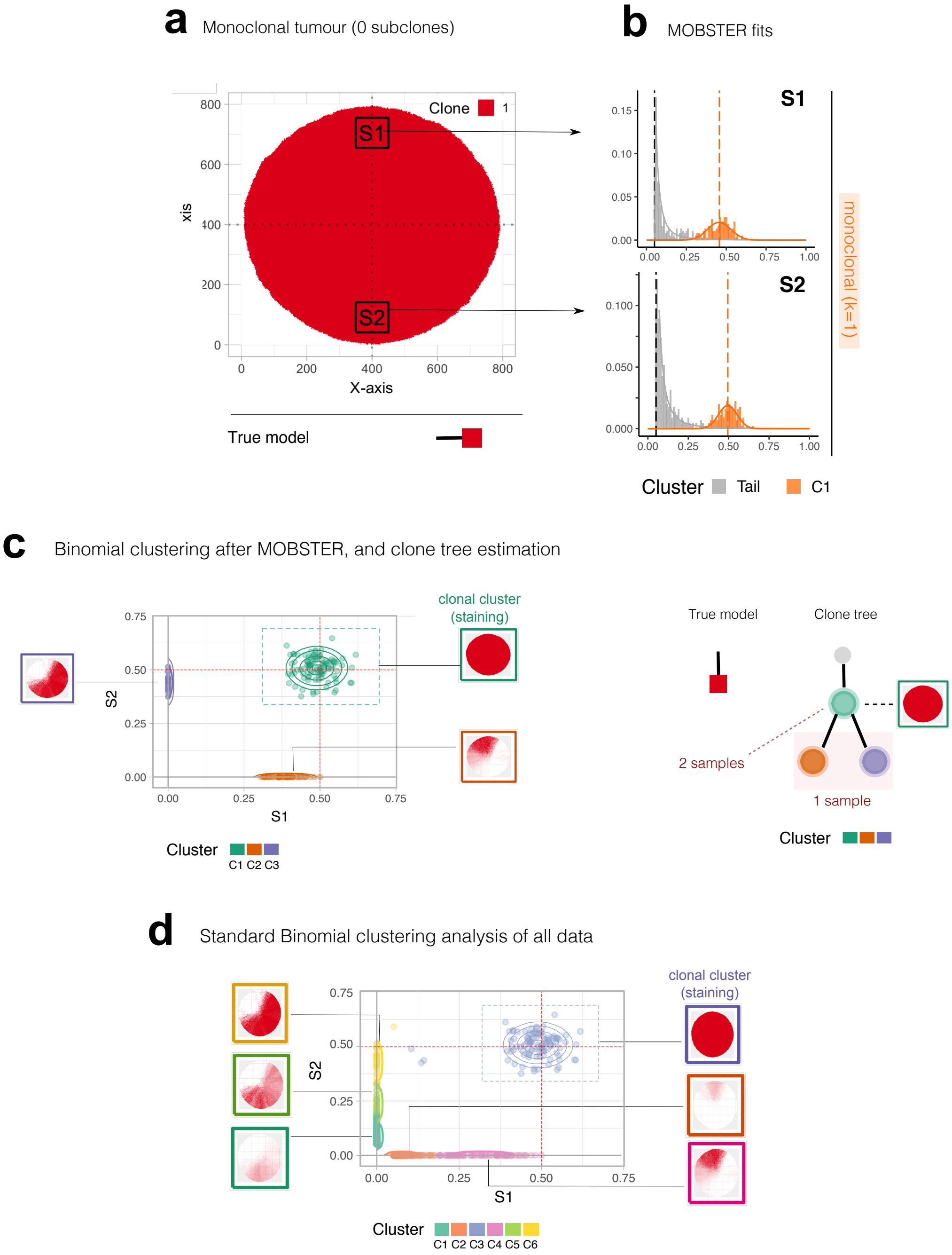
[†] A violation for a sample is when, in a branch x towards y and z, the observed CCF (here adjusted VAF) of x is lower than that of y plus z.



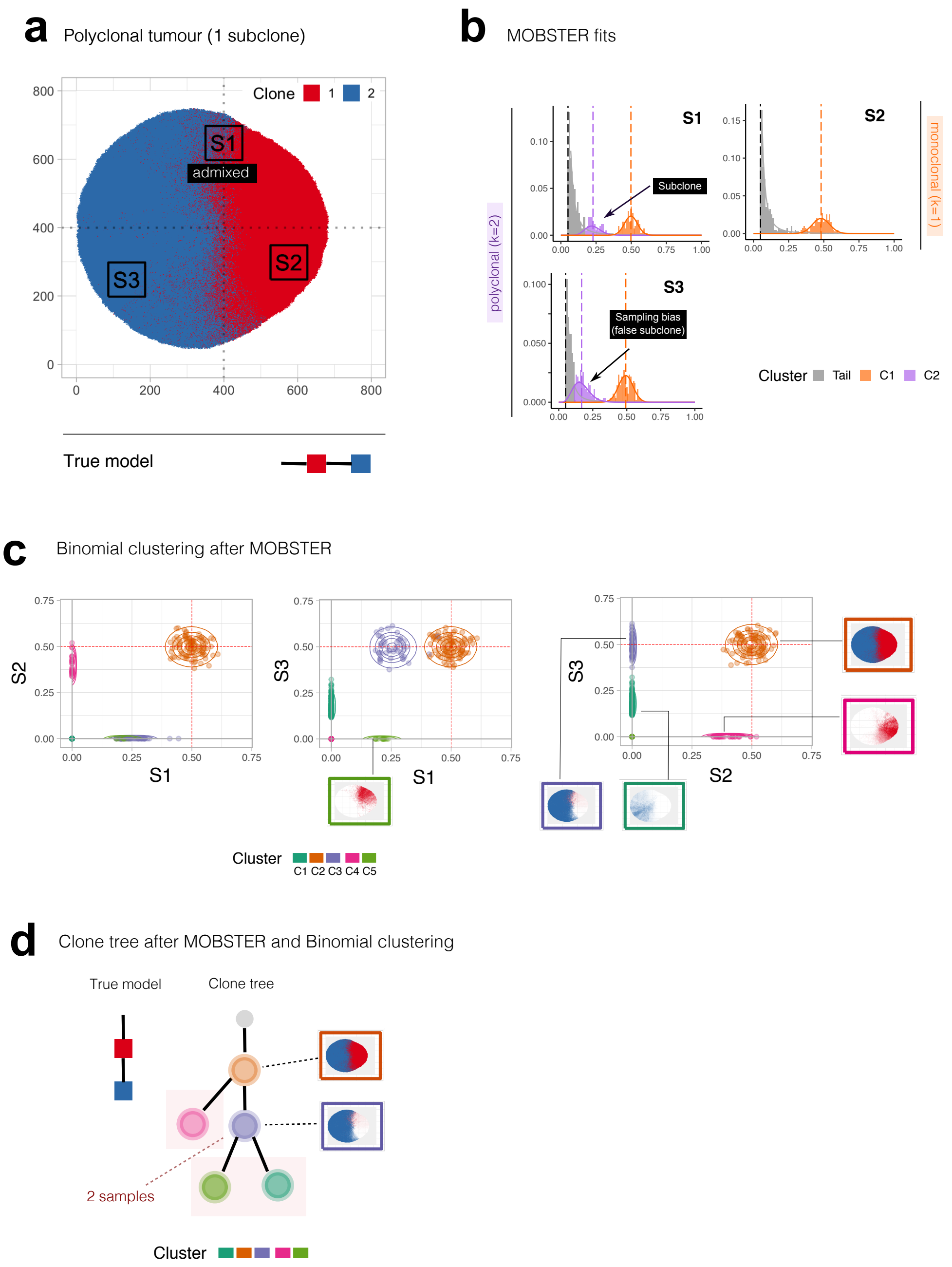
a Polyclonal tumour (1 subclone) and the hitchhikers mirage (expected by theory)

b Binomial clustering of all data

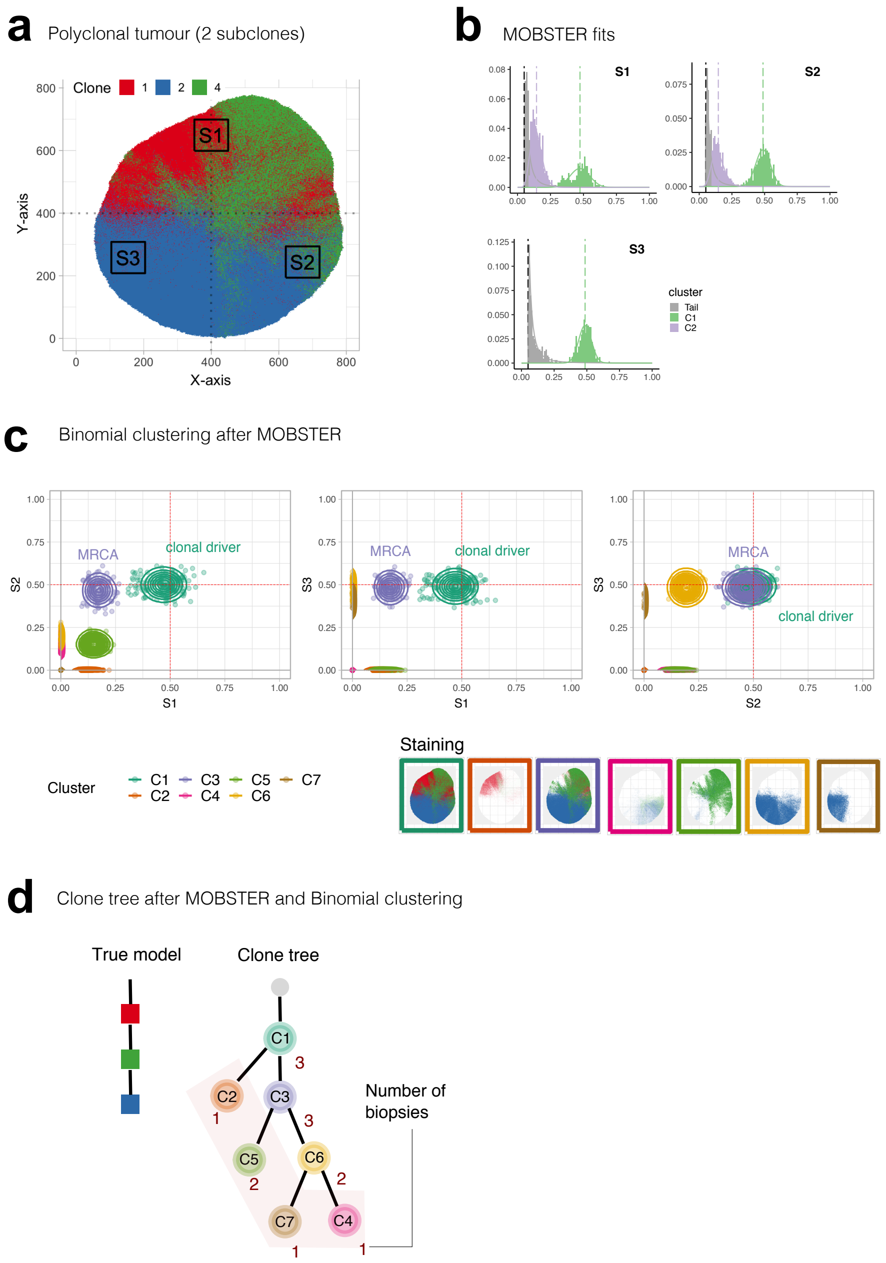
Supplementary Figure 10. Example hitchhiker mirage in multi-region sequencing data. a. A tumour with one subclone, as required to observe the hitchhiker mirage when we collect two monoclonal biopsies (here simulated at 100× WGS). The effect of the mirage as predicted by theory is sketched no the right. **b.** When we display the data and compute Binomial clusters without controlling for tails, we see the effect of the hitchhiker mirage. Cluster C7 contains mutations that in sample S1 are tail, while in S2 are clonal because they hitchhiked to the subclonal driver which originated the blue subclone. The Binomial mixture breaks therefore the hitchhikers into multiple clusters. While this is conceptually correct as those alleles do move differently across biopsies, the portrayed clonal architecture contains at least one extra node that might generate more trees that can fit the data, and more complex clonal evolution models than required.



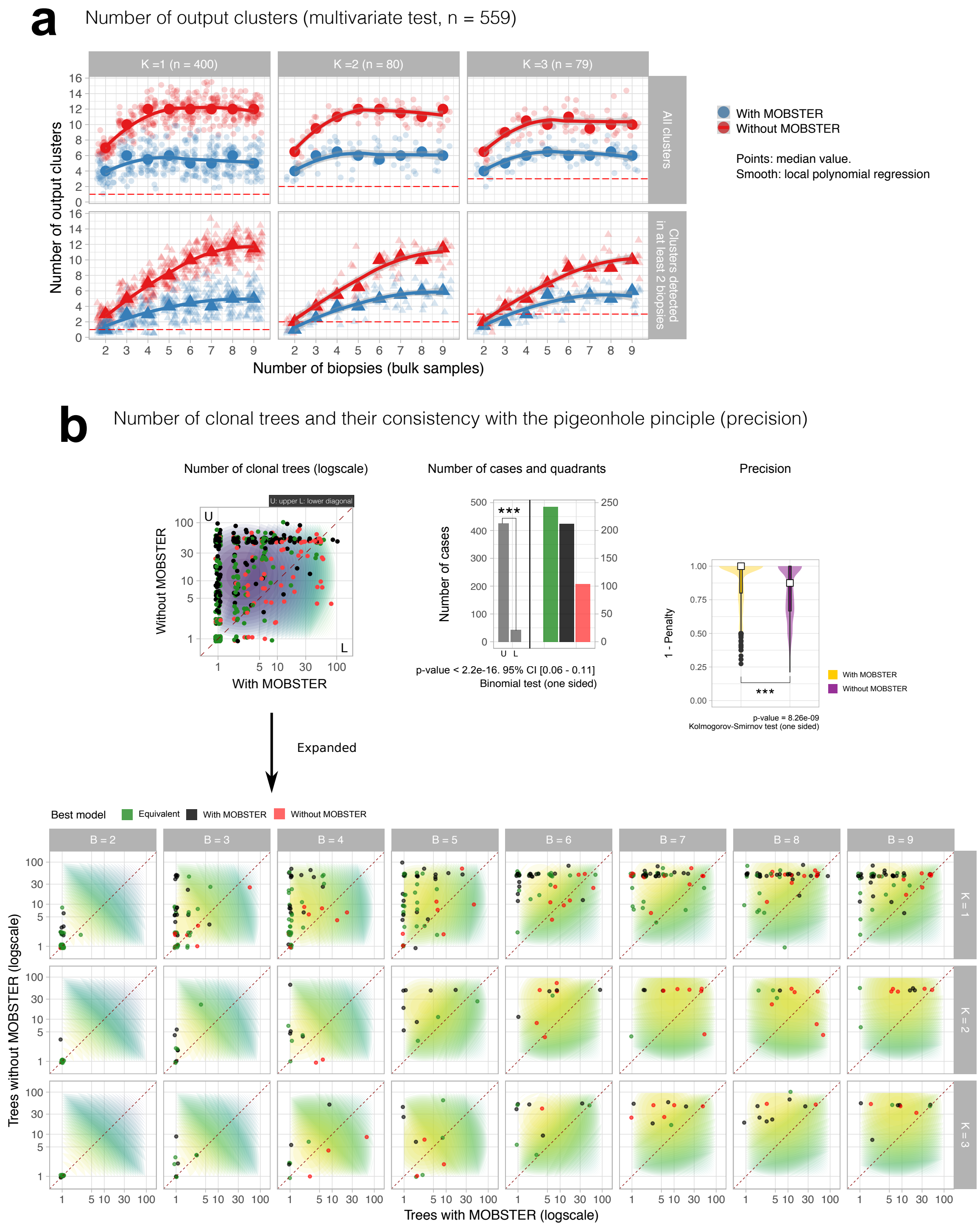
Supplementary Figure 11. Example spatial analysis of a monoclonal tumour. a, b. Spatial simulation of a tumor without subclones (i.e., monoclonal) where two samples were collected and sequenced at $100\times$ WGS. MOBSTER fit of the two monoclonal biopsies shows the tails in the data, that can be identified and removed. **c.** A multivariate analysis of non-tail mutations with a variational Binomial mixture after MOBSTER identifies $k = 3$ clusters (one truncal and two private to each biopsy), showing the admixing deception (shift of both private clusters from the 0.5 clonal expectation). The staining of these clusters shows the ancestor effect as well. The estimated clone tree is also reported. When we try to detect the actual clones under positive selection and their clonal architecture, a conservative heuristic is to consider only clusters detected in a minimum number of biopsies. We note that removing private subclones can underfit, as in the case of a genuine subclone detected in a single biopsy. In this tumor, removing leaf clusters/ clones leads to the identification of the true model. **d.** When we do not control for tails and carry out a standard analysis of all data, we detect $k = 6$ clusters and a much more complex evolutionary history; staining of the mutations identified with this clustering shows the error.



Supplementary Figure 12. Example spatial analysis of a tumour with one subclone. a. Spatial simulation of a tumor with one subclone under positive selection (blue), and collection of 3 biopsies (one boundary, and two monoclonal) sequenced at $100\times$ WGS. **b.** The fit of MOBSTER to S1 and S2 is perfect, but in S3 MOBSTER calls an extra subclone likely due to genetic drift (the true tail deviate from a power law, resulting in a distribution resembling a genuine subclonal cluster). **c.** When we analyze read counts for non-tail mutations after MOBSTER we detect a clear subclone in S1 against S3. This is correct because the subclone has swept through only S3, while it is admixed to its ancestor in S1. In all samples the ancestor fallacy and admixing deception are clearly observable. **d.** As in Supplementary Figure 11, removing leaf clusters/ clones leads to the identification of the true model.



Supplementary Figure 13. Example spatial analysis of a tumour with two subclones. a. Spatial simulation of a tumor with two subclone under positive selection (blue and green, evolving linearly from a red ancestral tumour-initiating clone), and collection of 3 biopsies (two boundary, and one monoclonal) sequenced at $100\times$ WGS. **b.** This sampling is tricky, as boundary biopsies (S1 and S2) contain an admixed signal. However, the fit with MOBSTER is perfect for all biopsies, with the clones are properly detected using 2, 2 and 1 Beta components. **c.** We run a variational fit for Binomial mixtures on the read counts of non-tail mutations. In bottom we show the virtual staining of the simulated tumour, which shows the presence of clear most recent common ancestors among the clusters. **d.** We manually curate a clonal tree from the virtual staining in panel c), and annotate the number of times each cluster occur in the 3 biopsy samples.



Supplementary Figure 14. Multivariate analysis of 2D tumours with multiple biopsies and subclones. All boxplots and violins show mean and inter quartile range (IQR), upper whisker is 3rd quartile +1.5 * IQR and lower whisker is 1st quartile - 1.5 * IQR. **a.** Performance of MOBSTER versus standard methods for multi-region sequencing analysis, with $n = 559$ simulated 2D tumors with a number of biopsies that ranges from 2 to 9. Each tumor contains either 0, 1 or 2 subclones, and we measure statistics from the reconstructed clone trees after variational Binomial clustering of read counts after MOBSTER. In this panel for variable number of overall clones k ($k-1$ subclones) we report the number of clones called with and without MOBSTER, estimated via variational Binomial mixtures. The red dashed line shows the actual number of true clones under positive selection in the tumour; we compare the overall number of clusters, and the reduced count where we remove those that occur only in a single biopsy. **b.** Plots summarising all tests, with each point coloured to show if the best model is obtained with MOBSTER (black), without (red) or equivalently (green). The barplot shows the number of points in the upper and lower diagonals of the plot; the p-value is for a null where the point has equal chance of being above or below the diagonal (no improvement across analyses). In the bottom plot we show the same counts for the number of detected clonal trees with and without MOBSTER, divided by number of subclones in the simulation, and number of collected biopsies (2 to 9). Results show that the detection of more points in the upper diagonal, which means fewer trees available with MOBSTER, is systematic across all simulated configurations of tumour and sampling. The trees are constructed following the pigeonhole principle, whose violations are measured and used to rank trees(16). The distribution of the scores for pigeonhole violations is shown as one minus the penalty score assigned to each tree, according to the scoring system presented in (16); values closer to 1 (0 penalty) reflect fewer violations of the principle, and therefore trees that are better fits to the data. We test for the difference among scores with a one-sided Kolmogorov-Smirnov test. These plots show that with MOBSTER we systematically retrieve fewer clonal trees with consistent higher quality, at a rate that is significantly higher ($p < 0.001$ in both tests).

3. Note 3: Variational inference for Dirichlet mixtures of multivariate Binomials

There are several tools to cluster read counts from tumour mutations with/ without MOBSTER's preliminary analysis. Besides testing sciClone, pyClone and DPClust, we have start working on the development of a generalised framework for Cancer Evolution analyses (future work). For this reason, we have implemented two R packages:

- a new variational method for Dirichlet mixtures of multivariate Binomial distributions which is now available as VIBER (https://caravagn.github.io/VIBER/);
- a new implementation of maximum-likelihood mixtures with both Binomial and Beta-Binomial components, which is now available as BMix (https://caravagn.github.io/bmix/).

The statistical model underneath VIBER is analogous to SciClone (7); our implementation however contains the (i) post-fit heuristic to filter output clusters that we discuss in this manuscript, and (ii) exposes all parameters for fit and convergence (e.g., the concentration, priors, etc.), which in SciClone are hardcoded inside the tool. For some single-sample tests and real-data application of this paper, we often use BMix, which allows to measure dispersion of the sequencing data..

In this Note we describe VIBER; we first introduce its variational method for single-sample (i.e., univariate) data, and then we extend it to multi-region (i.e., multivariate) data.

Data. The input data $\mathbf{X}$ consists of N pairs of values describing independent experiments $\mathbf{X} = [\boldsymbol{s}\ \boldsymbol{t}]$ where $\boldsymbol{s}$ and $\boldsymbol{t}$ are the N-dimensional vectors with the counts of successful and total trials in each experiment. In this context $\boldsymbol{s}$ represents reads with the mutated allele, while $\boldsymbol{t}$ the total number of reads at the locus (the per-locus coverage). For the n-th datapoint, we denote its values as scalars s_n and t_n; this notation is used to explicit the Binomial likelihood – e.g., $\mathsf{Bin}(s_n \mid t_n, \theta_k)$.

Bayesian model. We consider a *Dirichlet Finite-Mixture Model with K Binomial mixtures*, where each mixture's parameter – the Binomial probability θ – has a conjugate Beta prior

$$\begin{aligned} \boldsymbol{\pi} \mid \boldsymbol{\alpha} &\sim \mathsf{Dir}(\alpha_0, \ldots, \alpha_0) && \text{[mixture proportions]} \\ \mathbf{Z} \mid \boldsymbol{\pi} &\sim \mathsf{Cat}(\boldsymbol{\pi}) && \text{[latent variables]} \\ \theta_k \mid a, b &\sim \mathsf{Beta}(a_0, b_0) && \text{[parameter of a component]} \\ f_k \mid \theta_k &\sim \mathsf{Bin}(\boldsymbol{s} \mid \boldsymbol{t}, \theta_k) && \text{[likelihood of a component]} \end{aligned}$$

so the fixed-value *hyperparameters* are the scalars α_0, a_0 and b_0.

Joint and variational distributions. We want to infer the following factorized distribution

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}) &= p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta})\, p(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}) \\ &= p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta})\, p(\mathbf{Z} \mid \boldsymbol{\pi}, \boldsymbol{\theta})\, p(\boldsymbol{\pi}, \boldsymbol{\theta}) \\ &= p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta})\, p(\mathbf{Z} \mid \boldsymbol{\pi}, \boldsymbol{\theta})\, p(\boldsymbol{\pi})\, p(\boldsymbol{\pi}) = \underbrace{p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})}_{\text{mix. likelihood}}\, p(\mathbf{Z} \mid \boldsymbol{\pi})\, p(\boldsymbol{\pi}) \prod_k p(\theta_k) \end{aligned} \qquad [1]$$

where the term that depends on the Binomial mixtures are $p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})$ and the prior $p(\theta_k)$. Notice that we write product terms over k without the upper index K; we use the same notation for the samples, and omit N. As usual, the variational distribution is written as

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}) = q(\mathbf{Z})\, q(\boldsymbol{\pi}, \boldsymbol{\theta})\,. \qquad [2]$$

Key equations. We will use these equations to compute expectations.

- the multinomial distribution of the latent variables $\mathbf{Z} \mid \boldsymbol{\pi} \sim \mathsf{Cat}(\mathbf{Z} \mid \boldsymbol{\pi})$ is

$$\ln p(\mathbf{Z} \mid \boldsymbol{\pi}) = \ln \mathsf{Cat}(\mathbf{Z} \mid \boldsymbol{\pi}) = \ln \prod_{n,k} \pi_k^{z_{nk}} = \sum_{n,k} z_{nk}\ \ln \pi_k\,. \qquad [3]$$

- the Binomial likelihood of the data given the latent variables with iid samples is

$$\begin{aligned} \ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}) &= \ln \prod_{n,k} \mathsf{Bin}(s_n \mid t_n, \theta_k)^{z_{nk}} = \sum_{n,k} z_{nk} \ln \mathsf{Bin}(s_n \mid t_n, \theta_k) \\ &= \sum_{n,k} z_{nk} \Big\{ s_n \ln \theta_k + (t_n - s_n) \ln(1-\theta_k) \Big\} + \mathit{const} \\ \ln \mathsf{Bin}(s_n \mid t_n, \theta_k) &= \ln \binom{t_n}{s_n} \theta_k^{s_n} (1-\theta_k)^{t_n - s_n} = s_n \ln \theta_k + (t_n - s_n) \ln(1-\theta_k) + \mathit{const}\,. \end{aligned} \qquad [4]$$

Here and in what follows *const* is the normalization constant.

- the prior for the mixture components is the Dirichlet likelihood with hyperparameter $\boldsymbol{\alpha}_0 = [\alpha_0 \cdots \alpha_0]$ (notice that the components are independent)

$$\ln p(\boldsymbol{\pi}) = \ln \mathsf{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}_0) = \ln \prod_k \pi_k^{\alpha_0 - 1} + \mathit{const} = (\alpha_0 - 1) \sum_k \ln \pi_k + \mathit{const}\,. \qquad [5]$$

- The conjugate prior for the parameters of a mixture component $\theta_k \sim \mathsf{Beta}(a_0, b_0)$ has hyperparameters a_0 and b_0 and log-likelihood

$$\ln p(\theta_k) = \ln \mathsf{Beta}(\theta_k \mid a_0, b_0) = (a_0 - 1)\ln\theta_k + (b_0 - 1)\ln(1-\theta_k) + \mathit{const}\,. \qquad [6]$$

A. Update factors $q(\mathbf{Z})$, $q(\boldsymbol{\pi})$ and $q(\theta_k)$.

Derivation of $q(\mathbf{Z})$ This is the expectation with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ of the logarithm of the joint distribution that we want to infer

$$\begin{aligned} \ln q(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\theta}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta})] \\ &= \mathbb{E}_{\boldsymbol{\theta}}[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\pi}}[\ln p(\mathbf{Z} \mid \boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\pi}}[\ln p(\boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\theta}}[\ln p(\boldsymbol{\theta})] \qquad [\text{ln of } (1)] \\ &= \mathbb{E}_{\boldsymbol{\theta}}[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\pi}}[\ln p(\mathbf{Z} \mid \boldsymbol{\pi})] + \mathit{const}\,, \end{aligned}$$

where with const we group all terms that do not depend on $\mathbf{Z}$ (the parameter of q).

These are expectations of factors computed above: for the first term the expectation is over $\boldsymbol{\theta}$, and distributes over the Binomial terms, for the latter over π_k. We obtain

$$\mathbb{E}_{\boldsymbol{\pi}}[\ln p(\mathbf{Z} \mid \boldsymbol{\pi})] = \sum_{n,k} z_{nk}\, \mathbb{E}_{\pi_k}\left[\ln \pi_k\right]\,, \qquad [\text{by } (3)]$$

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}[\ln p(\mathbf{X} \mid \mathbf{Z}, \theta)] &= \sum_{n,k} z_{nk}\, \mathbb{E}_{\theta_k}\left[s_n \ln\theta_k + (t_n - s_n)\ln(1-\theta_k)\right] + \mathit{const} \qquad [\text{by } (4)] \\ &= \sum_{n,k} z_{nk} \Big\{ s_n \mathbb{E}_{\theta_k}\left[\ln\theta_k\right] + (t_n - s_n)\mathbb{E}_{\theta_k}\left[\ln(1-\theta_k)\right] \Big\} + \mathit{const}\,. \end{aligned}$$

Notice that the normalization constants are absorbed in const (they do not depend on θ_k). We then combine everything to derive $q(\mathbf{Z})$, and group the terms by z_{nk}

$$\begin{aligned} \ln q(\mathbf{Z}) &= \sum_{n,k} z_{nk} \Big\{ \mathbb{E}_{\pi_k}\left[\ln\pi_k\right] + s_n \mathbb{E}_{\theta_k}\left[\ln\theta_k\right] + (t_n - s_n)\mathbb{E}_{\theta_k}\left[\ln(1-\theta_k)\right] \Big\} + \mathit{const} \\ &= \sum_{n,k} z_{nk}\, \ln\lambda_{nk} + \mathit{const} \qquad [7] \end{aligned}$$

where we defined $N \times K$ terms

$$\ln\lambda_{nk} = \mathbb{E}_{\pi_k}\left[\ln\pi_k\right] + s_n \mathbb{E}_{\theta_k}\left[\ln\theta_k\right] + (t_n - s_n)\mathbb{E}_{\theta_k}\left[\ln(1-\theta_k)\right]\,. \qquad [8]$$

The *responsibilities* r_{nk} are computed taking logarithms out, and exponentiating

$$q(\mathbf{Z}) \propto \prod_{n,k} \lambda_{nk}^{z_{nk}}$$

so we can finally normalize this distribution

$$q(\mathbf{Z}) = \prod_{n,k} r_{nk}^{z_{nk}}\,, \qquad r_{nk} = \frac{\lambda_{nk}}{\sum_j \lambda_{jk}}\,. \qquad [9]$$

Derivation of $q(\boldsymbol{\pi}, \boldsymbol{\theta})$ First we show that $q(\boldsymbol{\pi}, \boldsymbol{\theta})$ can be factorized, and then proceed further on its factors

$$\begin{aligned} \ln q(\boldsymbol{\pi}, \boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta})] = \\ &= \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})] + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z} \mid \boldsymbol{\pi})] + \mathbb{E}_{\mathbf{Z}}[\ln p(\boldsymbol{\pi})] + \mathbb{E}_{\mathbf{Z}}[\ln p(\boldsymbol{\theta})]\,. \qquad [\text{ln of } (1)] \end{aligned}$$

We consider all terms because they involve the parameters of $q(\boldsymbol{\pi}, \boldsymbol{\theta})$. Expectations over $\mathbf{Z}$ are trivial for terms that do not depend on latent variables, while the mixture likelihood has now an expectation over $\mathbf{Z}$

$$\begin{aligned} \mathbb{E}_{\mathbf{Z}}[\ln p(\boldsymbol{\pi})] &= \ln p(\boldsymbol{\pi}) \\ \mathbb{E}_{\mathbf{Z}}[\ln p(\boldsymbol{\theta})] &= \sum_k \ln p(\theta_k) \\ \mathbb{E}_{\mathbf{Z}}[p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})] &= \sum_{n,k} \mathbb{E}_{\mathbf{Z}}[z_{nk}] \ln \mathsf{Bin}(s_n \mid t_n, \theta_k)\,. \qquad [\text{by } (4)] \end{aligned}$$

When we put all together we obtain

$$\ln q(\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_k \underbrace{\sum_n \Big\{\mathbb{E}_{\mathbf{Z}}[z_{nk}]\ \ln \mathsf{Bin}(s_n \mid t_n,\theta_k) + \ln p(\theta_k)\Big\}}_{\ln q(\theta_k)} + \underbrace{\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z}\mid\boldsymbol{\pi})] + \ln p(\boldsymbol{\pi})}_{\ln q(\boldsymbol{\pi})}$$

and by taking logarithms out and exponentiating we obtain the factorisation

$$q(\boldsymbol{\pi},\boldsymbol{\theta}) = q(\boldsymbol{\pi}) \prod_k q(\theta_k)\,. \tag*{[10]}$$

Derivation of $q(\boldsymbol{\pi})$ This involves only the Dirichlet distribution, which is the conjugate prior of the categorical (8). The expectation for the latent variables are the responsibilities ($\mathbb{E}_{\mathbf{Z}}[z_{nk}] = r_{nk}$)

$$\begin{aligned}
\ln q(\boldsymbol{\pi}) &= \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z}\mid\boldsymbol{\pi})] + \ln p(\boldsymbol{\pi}) + const \\
&= \sum_{n,k} \underbrace{\mathbb{E}_{\mathbf{Z}}[z_{nk}]}_{r_{nk}} \ln \pi_k + \sum_k \ln \pi_k^{\alpha_0 - 1} + const && \text{[by (3) and (5)]} \\
&= \sum_{n,k} \ln \pi_k^{r_{nk}} + \ln \pi_k^{\alpha_0 - 1} + const && [\mathbb{E}_{\mathbf{Z}}[z_{nk}] = r_{nk}]
\end{aligned}$$

Then by taking out the logarithms and exponentiating one gets

$$\begin{aligned}
q(\boldsymbol{\pi}) &\propto \prod_k \pi_k^{\alpha_0 - 1} \cdot \pi_k^{\sum_n r_{nk}} = \prod_k \pi_k^{\alpha_0 + N_k - 1} && [N_k = \textstyle\sum_n r_{nk}] \\
&= \mathsf{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}_0 + N_k) && [11]
\end{aligned}$$

so the posterior is Dirichlet with components $\alpha_k = \alpha_0 + N_k$

***Derivation of** $q(\theta_k)$**.* The posterior for a Binomial mixture component with Beta prior is Beta by conjugacy, as we show here.

$$\begin{aligned}
\ln q(\theta_k) &= \sum_n \mathbb{E}_{\mathbf{Z}}[z_{nk}]\ \ln \mathsf{Bin}(s_n \mid t_n,\theta_k) + \ln p(\theta_k) + const \\
&= \sum_n \mathbb{E}_{\mathbf{Z}}[z_{nk}] \Big\{ s_n \ln \theta_k + (t_n - s_n) \ln(1-\theta_k) \Big\} + (a_0 - 1)\ln\theta_k + (b_0 - 1)\ln(1-\theta_k) + const
\end{aligned}$$

Considering $\mathbb{E}_{\mathbf{Z}}[z_{nk}] = r_{nk}$ as in the derivation of $q(\boldsymbol{\pi})$ we have

$$\begin{aligned}
\ln q(\theta_k) &= \sum_n \ln \theta_k^{s_n r_{nk}} + \ln(1-\theta_k)^{(t_n - s_n) r_{nk}} + \ln \theta_k^{a_0 - 1} + \ln(1-\theta_k)^{b_0 - 1} + const \\
&= \ln \theta_k^{a_0 - 1} \prod_n \theta_k^{s_n r_{nk}} \cdot (1-\theta_k)^{b_0 - 1} \prod_n (1-\theta_k)^{(t_n - s_n) r_{nk}} + const
\end{aligned}$$

which has the general form $\ln \theta_k^x \cdot (1-\theta_k)^y + const$ with

$$x = \sum_n s_n r_{nk} + a_0 - 1 \qquad\qquad y = \sum_n (t_n - s_n) r_{nk} + b_0 - 1\,.$$

Now if we define $s_k^\star = \sum_n s_n r_{nk}$ and $t_k^\star = \sum_n t_n r_{nk}$, we find

$$q(\theta_k) \propto \theta_k^{a_0 + s_k^\star - 1} \cdot (1-\theta_k)^{b_0 + t_k^\star - s_k^\star - 1} = \mathsf{Beta}(\theta_k \mid a_0 + s_k^\star, b_0 + t_k^\star - s_k^\star)\,, \tag*{[12]}$$

which suggests the posterior update rules $a_k = a_0 + s_k^\star$ and $b_k = a_0 + t_k^\star - s_k^\star$.

B. Update equations (variational E and M-steps). By the previous equations we obtain the E-variational and M-variational steps.

Variational E-step This step consists in computing the responsibilities r_{nk}, which depend on our approximation to the posterior over the parameters. To compute r_{nk} we have to compute the λ_{nk} terms, then the normalization is just empirical to get each of the r_{nk} values. The formula for the λ_{nk} terms is

$$\ln \lambda_{nk} = \mathbb{E}_{\pi_k}\left[\ln \pi_k\right] + s_n \mathbb{E}_{\theta_k}\left[\ln \theta_k\right] + (t_n - s_n)\mathbb{E}_{\theta_k}\left[\ln(1-\theta_k)\right] \ , \qquad [13]$$

which can be split in the computation of three expectations (under q).

Term $\mathbb{E}_{\pi_k}\left[\ln \pi_k\right]$ is known (8) to be $\mathbb{E}_{\pi_k}\left[\ln \pi_k\right] = \Psi(\alpha_k) - \Psi(\boldsymbol{\alpha}_*)$ with $\boldsymbol{\alpha}_* = \sum_k \alpha_k$, where Ψ is the *digamma* function. For term $s_n \mathbb{E}_{\theta_k}\left[\ln \theta_k\right]$it is possible to see that[‡]

$$\begin{aligned}
\mathbb{E}_{\theta_k}\left[\ln \theta_k\right] &= \int_0^1 \ln \theta_k \cdot q(\theta_k)\, d\theta_k = \frac{1}{B(a_k, b_k)} \int_0^1 \ln \theta_k \cdot \theta_k^{a_k - 1} \left(1-\theta_k\right)^{b_k - 1} d\theta_k && \text{[by def.]}\\
&= \frac{1}{B(a_k, b_k)} \frac{\Gamma(a_k)\,\Gamma(b_k)}{\Gamma(a_k+b_k)} \Big[\Psi(a_k) - \Psi(a_k+b_k)\Big] \\
&= \Psi(a_k) - \Psi(a_k+b_k) && \text{[since } B(a_k,b_k) = \Gamma(a_k)\Gamma(b_k)\Gamma(a_k+b_k)^{-1}\text{]}
\end{aligned}$$

which means that $s_n \mathbb{E}_{\theta_k}\left[\ln \theta_k\right] = s_n \Big[\Psi(a_k) - \Psi(a_k+b_k)\Big]$.

Term $(t_n - s_n)\mathbb{E}_{\theta_k}\left[\ln 1 - \theta_k\right]$. This term is similar to above; the expectation is again over the same Beta distribution. To solve it, we can apply a direct substitution for $1 - \theta_k$

$$\begin{aligned}
\mathbb{E}_{\theta_k}\left[\ln 1 - \theta_k\right] &= \int_0^1 \ln(1-\theta_k) \cdot q(\theta_k)\, d\theta_k && \text{[by def.]}\\
&= \frac{1}{B(a_k, b_k)} \int_0^1 \ln(1-\theta_k) \cdot \theta_k^{a_k - 1} \left(1-\theta_k\right)^{b_k - 1} d\theta_k \\
&= \frac{-1}{B(b_k, a_k)} \int_{0=y}^{1=y} \ln y \cdot y^{b_k - 1} \left(1-y\right)^{a_k - 1} dy && [y = 1-\theta_k,\, d\theta_k = -dy]\\
&= \Psi(b_k) - \Psi(a_k+b_k) && \text{[as above, with reversed sign]}
\end{aligned}$$

where we used $B(b_k, a_k) = B(a_k, b_k)$ and obtained a Beta with swapped parameters, overall

$$(t_n - s_n)\mathbb{E}_{\theta_k}\left[\ln 1 - \theta_k\right] = (t_n - s_n)\Big[\Psi(b_k) - \Psi(a_k+b_k)\Big] \ .$$

Computing the λ_{nk} terms. The formula for λ_{nk} is just

$$\begin{aligned}
\lambda_{nk} &\propto \exp\left\{\Psi(\alpha_k) - \Psi(\boldsymbol{\alpha}_*) + s_n\Big[\Psi(a_k) - \Psi(a_k+b_k)\Big] + (t_n - s_n)\Big[\Psi(b_k) - \Psi(a_k+b_k)\Big]\right\}\\
&\propto \exp\left\{\Psi(\alpha_k) - \Psi(\boldsymbol{\alpha}_*) + s_n\Big[\Psi(a_k) - \Psi(b_k)\Big] + t_n\Big[\Psi(b_k) - \Psi(a_k+b_k)\Big]\right\} \ , \qquad [14]
\end{aligned}$$

and values for r_{nk} can be computed by normalizing these terms.

Variational M-step The other quantities of interest are the rules to compute the posterior approximations to the parameters of the Dirichlet and Beta distributions, and they have been derived earlier.

$$\alpha_k = \alpha_0 + N_k \qquad a_k = a_0 + s_k^\star \qquad b_k = b_0 + t_k^\star - s_k^\star \qquad [15]$$

where

$$N_k = \sum_n r_{nk} \qquad s_k^\star = \sum_n s_n r_{nk} \qquad t_k^\star = \sum_n t_n r_{nk}\ .$$

C. Variational evidence lower bound $\mathcal{L}(q)$. The *evidence lower bound* (ELBO) $\mathcal{L}(q)$ is computed at each iteration of our variational inference to assess convergency. For our mixture of Binomials the ELBO is

$$\begin{aligned}
\mathcal{L}(q) &= \sum_{\mathbf{Z}} \int\int q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}) \ln\left\{\frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta})}{q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta})}\right\} d\boldsymbol{\pi} d\boldsymbol{\theta} = \\
&= \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta}}\left[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta})\right] - \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta}}\left[\ln q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta})\right]
\end{aligned}$$

[‡] There is an explicit derivation at https://stats.stackexchange.com/questions/241993/pdf-of-y-logx-when-x-is-beta-distributed-the-expected-value-of-y/242020

Since p and q factorize we can rewrite the bound as

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta}}\left[\ln p(\mathbf{X} \mid \mathbf{Z},\boldsymbol{\theta})\, p(\mathbf{Z} \mid \boldsymbol{\pi})\, p(\boldsymbol{\pi}) \prod_k p(\theta_k)\right] - \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\theta}}\left[\ln q(\mathbf{Z}) q(\boldsymbol{\pi}) \prod_k q(\theta_k)\right] \\ &= \mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\ln p(\mathbf{X} \mid \mathbf{Z},\boldsymbol{\theta})\right] + \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi}}\left[\ln p(\mathbf{Z} \mid \boldsymbol{\pi})\right] + \mathbb{E}_{\boldsymbol{\pi}}\left[\ln p(\boldsymbol{\pi})\right] + \sum_k \mathbb{E}_{\theta_k}\left[\ln p(\theta_k)\right] \\ &\quad - \mathbb{E}_{\mathbf{Z}}\left[\ln q(\mathbf{Z})\right] - \mathbb{E}_{\boldsymbol{\pi}}\left[\ln q(\boldsymbol{\pi})\right] - \sum_k \mathbb{E}_{\theta_k}\left[\ln q(\theta_k)\right] .\end{aligned}$$

We rewrite this form of the ELBO by using the KL-divergence among the involved distributions (Section E)

$$\mathcal{L}(q) = \mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\ln p(\mathbf{X} \mid \mathbf{Z},\boldsymbol{\theta})\right] + w[q(\mathbf{Z}), p(\mathbf{Z} \mid \boldsymbol{\pi})] + w[q(\boldsymbol{\pi}) \parallel p(\boldsymbol{\pi})] + \sum_{k=1}^{K} w[q(\theta_k) \parallel p(\theta_k)]$$

where $w[q,p] = -\mathsf{KL}[q,p]$. The only terms that we need to compute are the first, and the last, as the other are independent on the mixture and derived in Section E.

Term $\mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\ln p(\mathbf{X} \mid \mathbf{Z},\boldsymbol{\theta})\right]$. Once expanded, this is an expectation of independent random variables (z_{nk}, θ_k and $1-\theta_k$), thus is the product of their expectations; here we include also the normalization constant $\ln C_n = \ln t_n! - \ln s_n! - \ln(t_n - s_n)!$ for a Binomial random variable

$$\begin{aligned}\mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\ln p(\mathbf{X} \mid \mathbf{Z},\boldsymbol{\theta})\right] &= \mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\sum_{n,k}\left\{s_n \ln\theta_k + (t_n - s_n)\ln(1-\theta_k) - \ln C_n\right\}\right] \\ &= \sum_{n,k} \mathbb{E}_{\mathbf{Z}}\left[z_{nk}\right] \mathbb{E}_{\theta_k}\left[s_n \ln\theta_k + (t_n - s_n)\ln(1-\theta_k) - \ln C_n\right] .\end{aligned}$$

These terms have been computed for the variational E-step, we have

$$\mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\ln p(\mathbf{X} \mid \mathbf{Z},\boldsymbol{\theta})\right] = \sum_{n,k} r_{nk}\left\{s_n\Big[\Psi(a_k) - \Psi(b_k)\Big] + t_n\Big[\Psi(b_k) - \Psi(a_k + b_k)\Big] - \ln C_n\right\} .$$

Term $\sum_k w[q(\theta_k) \parallel p(\theta_k)]$. A term of this summation is the negative KL-divergence among two Beta random variables, one given by the prior p, the other by the posterior q (8)

$$\begin{aligned}w[q(\theta_k), p(\theta_k)] &= -\mathsf{KL}[q(\theta_k) \parallel p(\theta_k)] \\ &= -\left\{\ln\frac{B(a_0,b_0)}{B(a_k,b_k)} + (a_k - a_0)\Psi(a_k) + (b_k - b_0)\Psi(b_k) + (a_0 - a_k + b_0 - b_k)\Psi(a_k + b_k)\right\} \\ &= \ln\frac{B(a_k,b_k)}{B(a_0,b_0)} + (a_0 - a_k)\Psi(a_k) + (b_0 - b_k)\Psi(b_k) + (a_k - a_0 + b_k - b_0)\Psi(a_k + b_k) .\end{aligned}$$

The ELBO $\mathcal{L}(q)$. Define the following quantities

$$\begin{aligned}\hat{\Psi}_{a_k,b_k} &= \Psi(a_k) - \Psi(b_k) & \hat{\Psi}_{b_k,a_k} &= \Psi(b_k) - \Psi(a_k + b_k) \\ \ln\rho_k &= \Psi(\alpha_k) - \Psi(\boldsymbol{\alpha}_*) & w[q(\theta_k) \parallel p(\theta_k)] &= \omega_k\end{aligned}$$

to see that the ELBO has analytical form

$$\begin{aligned}\mathcal{L}(q) = \sum_{n,k} & r_{nk}\left\{s_n\hat{\Psi}_{a_k,b_k} + t_n\hat{\Psi}_{b_k,a_k} - \ln C_n + \ln\rho_k - \ln r_{nk}\right\} \\ &+ \ln\frac{C(\boldsymbol{\alpha}_0)}{C(\boldsymbol{\alpha})} + \sum_k\Big[(\alpha_0 - \alpha_k)\ln\rho_k + \omega_k\Big] .\end{aligned} \qquad [16]$$

D. Multivariate extension. Each data point is now a vector with dimension W, we denote with an extra superscript the input data when we refer to specific component of the data. As in other papers, we assume that the W dimensions are independent; the correlation structure is given by the latent variables which map each point (a vector) to one of K clusters.

Statistical model. With respect to the univariate case we now have a vector of parameters $\boldsymbol{\theta}_k$ for each mixture component, and we use multivariate distributions (written in bold)

$$\begin{aligned}\boldsymbol{\theta}_k \mid \boldsymbol{a},\boldsymbol{b} &\sim \textbf{Beta}(\boldsymbol{a}_0,\boldsymbol{b}_0) && \text{[parameter of a mixture]} \\ f_k \mid \boldsymbol{\theta}_k &\sim \textbf{Bin}(\boldsymbol{s} \mid \boldsymbol{t},\boldsymbol{\theta}_k) && \text{[likelihood of a mixture]}\end{aligned}$$

and our hyperparameters for the Beta are now vectors as well. The joint and the variational distributions are unchanged, provided that we assume a further factorization of the likelihood function and the prior. The multinomial distribution (latent variables), and the prior for the components are as in the univariate case.

Binomial likelihood of the data. The log-likelihood of a multivariate Binomial random variable with independent dimensions is

$$\ln \mathbf{Bin}(\boldsymbol{s}_n \mid \boldsymbol{t}_n, \boldsymbol{\theta}_k) = \ln \prod_w \mathsf{Bin}(s_{n,w} \mid t_{n,w}, \theta_{k,w}) = \sum_w \left[s_{n,w} \ln \theta_{k,w} + (t_{n,w} - s_{n,w}) \ln(1 - \theta_{k,w}) \right] + const\,.$$

We omit to write that w ranges from one to W to reduce the notation. The Binomial complete likelihood reads as

$$\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}) = \ln \prod_{n,k} \mathbf{Bin}(\boldsymbol{s}_n \mid \boldsymbol{t}_n, \boldsymbol{\theta}_k)^{z_{nk}} = \sum_{n,k} z_{nk} \ln \mathbf{Bin}(\boldsymbol{s}_n \mid \boldsymbol{t}_n, \boldsymbol{\theta}_k)\,. \qquad [17]$$

Prior for the mixture parameters. The prior is a W-dimensional multivariate Beta random variable $\boldsymbol{\theta}_k \sim \mathbf{Beta}(\boldsymbol{a}_0, \boldsymbol{b}_0)$ with vector hyperparameters and independent dimensions

$$\begin{aligned} \ln p(\boldsymbol{\theta}_k) &= \ln \mathbf{Beta}(\boldsymbol{\theta}_k \mid \boldsymbol{a}_0, \boldsymbol{b}_0) = \ln \prod_w \mathsf{Beta}(\theta_{k,w} \mid a_{0,w}, b_{0,w}) \\ &= \sum_w \Big[(a_{0,w} - 1) \ln \theta_{k,w} + (b_{0,w} - 1) \ln(1 - \theta_{k,w}) \Big] + const\,. \qquad [18] \end{aligned}$$

Update factors for variational distributions, and update steps. We need to derive $q(\mathbf{Z})$ and $q(\boldsymbol{\pi}, \boldsymbol{\theta}) = q(\boldsymbol{\pi}) \prod_k q(\boldsymbol{\theta}_k)$. Note that term $q(\boldsymbol{\pi}) \propto \mathsf{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}_0 + N_k)$ is still a Dirichlet posterior as in the univariate case.

Term $q(\mathbf{Z})$. For term $\ln q(\mathbf{Z})$ the only change is in the expectation of the complete likelihood

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}[\ln p(\mathbf{X} \mid \mathbf{Z}, \theta)] &= \mathbb{E}_{\boldsymbol{\theta}} \left[\sum_{n,k} z_{nk} \ln \mathbf{Bin}(\boldsymbol{s}_n \mid \boldsymbol{t}_n, \boldsymbol{\theta}_k) \right] = \\ &= \sum_{n,k} z_{nk} \left\{ \sum_w \Big[s_{n,w} \ln \mathbb{E}_{\theta_{k,w}}[\theta_{k,w}] + (t_{n,w} - s_{n,w}) \mathbb{E}_{\theta_{k,w}}[\ln(1 - \theta_{k,w})] \Big] \right\} + const \end{aligned}$$

We derive as usual $\ln q(\mathbf{Z}) = \sum_{n,k} z_{nk}\ \ln \lambda_{nk} + const$, but now we define $N \times K$ terms

$$\ln \lambda_{nk} = \mathbb{E}_{\pi_k}\left[\ln \pi_k\right] + \sum_w \Big[s_{n,w} \ln \mathbb{E}_{\theta_{k,w}}[\theta_{k,w}] + (t_{n,w} - s_{n,w}) \mathbb{E}_{\theta_{k,w}}[\ln(1 - \theta_{k,w})] \Big]\,. \qquad [19]$$

The responsibilities are defined as usual as $r_{nk} \propto \lambda_{nk}$.

Term $q(\boldsymbol{\theta}_k)$. We derive the posterior parameter for a multivariate Binomial mixture component similarly to the univariate case, but with a vector of parameters per component

$$\begin{aligned} \ln q(\boldsymbol{\theta}_k) &= \sum_n \mathbb{E}_{\mathbf{Z}}[z_{nk}]\ \ln \mathbf{Bin}(\boldsymbol{s}_n \mid \boldsymbol{t}_n, \boldsymbol{\theta}_k) + \ln p(\boldsymbol{\theta}_k) + const \\ &= \sum_n \mathbb{E}_{\mathbf{Z}}[z_{nk}] \left\{ \sum_w \ln \mathsf{Bin}(s_{n,w} \mid t_{n,w}, \theta_{k,w}) \right\} + \sum_w \ln \mathsf{Beta}(\theta_{k,w} \mid a_{0,w}, b_{0,w}) + const \end{aligned}$$

Considering $\mathbb{E}_{\mathbf{Z}}[z_{nk}] = r_{nk}$ we have

$$\ln q(\boldsymbol{\theta}_k) = \ln \prod_w \theta_{k,w}^x \cdot \prod_w (1 - \theta_{k,w})^y + const\,.$$

with the following substitutions

$$x = \sum_n s_{n,w} r_{nk} + (a_{0,w} - 1) \qquad\qquad y = \sum_n (t_{n,w} - s_{n,w}) r_{nk} + (b_{0,w} - 1)$$

Now if one defines

$$s^\star_{k,w} = \sum_n s_{n,w} r_{nk} \qquad\qquad t^\star_{k,w} = \sum_n t_{n,w} r_{nk}$$

then

$$\begin{aligned} q(\boldsymbol{\theta}_k) &\propto \prod_w \theta_{k,w}^{a_{0,w} + s^\star_{k,w} - 1} \cdot (1 - \theta_{k,w})^{b_{0,w} + t^\star_{k,w} - s^\star_{k,w} - 1} \\ &= \prod_w \mathsf{Beta}(\theta_{k,w} \mid a_{0,w} + s^\star_{k,w}, b_{0,w} + t^\star_{k,w} - s^\star_{k,w}) = \mathbf{Beta}(\boldsymbol{\theta}_k \mid \boldsymbol{a}_0 + S, \boldsymbol{b}_0 + T - S)\,, \qquad [20] \end{aligned}$$

which suggests the posterior update rules (per component)

$$a_{k,w} = a_{0,w} + s^\star_{k,w} \qquad [21]$$

$$b_{k,w} = a_{0,w} + t^\star_{k,w} - s^\star_{k,w}\,. \qquad [22]$$

Update equations. By the previous equations we obtain the E-variational and M-variational steps. The formula for the λ_{nk} terms is

$$\ln \lambda_{nk} = \mathbb{E}_{\pi_k}\left[\ln \pi_k\right] + \sum_w \Big[s_{n,w} \ln \mathbb{E}_{\theta_{k,w}}[\theta_{k,w}] + (t_{n,w} - s_{n,w}) \mathbb{E}_{\theta_{k,w}}[\ln(1-\theta_{k,w})] \Big]$$

which is split again as three expectations under q. The first term is analogous to the univariate case, the other terms are just adjustment to the multivariate case of the original terms (since they are expectations per-component)

$$\begin{aligned} s_{n,w} \mathbb{E}_{\theta_{k,w}}\left[\ln \theta_{k,w}\right] &= s_{n,w} \Big[\Psi(a_{k,w}) - \Psi(a_{k,w} + b_{k,w}) \Big] \\ (t_{n,w} - s_{n,w}) \mathbb{E}_{\theta_{k,w}}\left[\ln 1 - \theta_{k,w}\right] &= (t_{n,w} - s_{n,w}) \Big[\Psi(b_{k,w}) - \Psi(a_{k,w} + b_{k,w}) \Big] \,. \end{aligned}$$

The formula for λ_{nk} (and normalized r_{nk} values), are obtained as

$$\lambda_{nk} \propto \exp\left\{ \Psi(\alpha_k) - \Psi(\boldsymbol{\alpha}_*) + \sum_w s_{n,w} \Big[\Psi(a_{k,w}) - \Psi(b_{k,w}) \Big] + t_{n,w} \Big[\Psi(b_{k,w}) - \Psi(a_{k,w} + b_{k,w}) \Big] \right\} .$$

The variational M-step uses the posterior approximations to the parameters derived earlier.

Variational evidence lower bound $\mathcal{L}(q)$. Consider the ELBO

$$\mathcal{L}(q) = \mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})\right] + \omega[q(\mathbf{Z}), p(\mathbf{Z} \mid \boldsymbol{\pi})] + \omega[q(\boldsymbol{\pi}) \parallel p(\boldsymbol{\pi})] + \sum_k \omega[q(\boldsymbol{\theta}_k) \parallel p(\boldsymbol{\theta}_k)]$$

where $\omega[q,p] = -\mathsf{KL}[q,p]$. The only terms that we need to compute are the first, and the last, as the other are independent on the mixture and derived in Section E.

Term $\mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})\right]$. For a multivariate Binomial random variable

$$\begin{aligned} \mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})\right] &= \mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\sum_{n,k} z_{nk} \Big\{ \sum_w \ln \mathsf{Bin}(s_{n,w} \mid t_{n,w}, \theta_{k,w}) \Big\} \right] \\ &= \sum_{n,k} \mathbb{E}_{\mathbf{Z}}\left[z_{nk}\right] \mathbb{E}_{\boldsymbol{\theta}_k}\left[\sum_w \ln \mathsf{Bin}(s_{n,w} \mid t_{n,w}, \theta_{k,w}) \right] \\ &= \sum_{n,k} \mathbb{E}_{\mathbf{Z}}\left[z_{nk}\right] \sum_w \mathbb{E}_{\theta_{k,w}}\left[\ln \mathsf{Bin}(s_{n,w} \mid t_{n,w}, \theta_{k,w})\right] \\ &= \sum_{n,k} r_{nk} \sum_w \mathbb{E}_{\theta_{k,w}}\left[\ln \mathsf{Bin}(s_{n,w} \mid t_{n,w}, \theta_{k,w})\right] \,. \end{aligned}$$

Now the inner expectation rewrites as

$$\mathbb{E}_{\theta_{k,w}}\left[\ln \mathsf{Bin}(s_{n,w} \mid t_{n,w}, \theta_{k,w})\right] = \mathbb{E}_{\theta_{k,w}}\left[s_{n,w} \ln \theta_{k,w} + (t_{n,w} - s_{n,w}) \ln(1 - \theta_{k,w}) \right] \,.$$

which are the terms derived for $\ln \lambda_{nk}$ in the univariate case. Here the normalisation constant depends on the dimension $\ln C_{n,w} = \ln t_{n,w}! - \ln s_{n,w}! - \ln(t_{n,w} - s_{n,w})!$ and we have

$$\mathbb{E}_{\mathbf{Z},\boldsymbol{\theta}}\left[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})\right] = \sum_{n,k} r_{nk} \left\{ \sum_w s_{n,w} \Big[\Psi(a_{k,w}) - \Psi(b_{k,w}) \Big] + t_{n,w} \Big[\Psi(b_{k,w}) - \Psi(a_{k,w} + b_{k,w}) \Big] - \ln C_{n,w} \right\} .$$

Term $\sum_k \omega[q(\boldsymbol{\theta}_k) \parallel p(\boldsymbol{\theta}_k)]$. If we rewrite

$$\sum_k \omega[q(\boldsymbol{\theta}_k) \parallel p(\boldsymbol{\theta}_k)] = \sum_k \left\{ \sum_w \omega[q(\theta_{k,w}) \parallel p(\theta_{k,w})] \right\}$$

we retrieve a negative KL-divergence among Beta random variables (as in the univariate case),

$$\omega[q(\boldsymbol{\theta}_k) \parallel p(\boldsymbol{\theta}_k)] = -\sum_w \mathsf{KL}[q(\theta_{k,w}) \parallel p(\theta_{k,w})] = \sum_w \left[\ln \frac{B(a_{k,w}, b_{k,w})}{B(a_{0,w}, b_{0,w})} + \phi_w \right] .$$

where

$$\phi_w = (a_{0,w} - a_{k,w})\Psi(a_{k,w}) + (b_{0,w} - b_{k,w})\Psi(b_{k,w}) + (a_{k,w} - a_{0,w} + b_{k,w} - b_{0,w})\Psi(a_{k,w} + b_{k,w})$$

The ELBO $\mathcal{L}(q)$. Define the following quantities

$$\hat{\Psi}^{w}_{a_k,b_k} = \Psi(a_{k,w}) - \Psi(b_{k,w}) \qquad\qquad \hat{\Psi}^{w}_{b_k,a_k} = \Psi(b_{k,w}) - \Psi(a_{k,w} + b_{k,w})$$

$$\ln \rho_k = \Psi(\alpha_k) - \Psi(\boldsymbol{\alpha_*}) \qquad\qquad \omega[q(\boldsymbol{\theta}_k) \parallel p(\boldsymbol{\theta}_k)] = \eta_k$$

to see that the ELBO has analytical form

$$\mathcal{L}(q) = \sum_{n,k} r_{nk} \Big\{ \sum_{w} s_{n,w} \hat{\Psi}^{w}_{a_k,b_k} + t_{n,w} \hat{\Psi}^{w}_{b_k,a_k} - \ln C_{n,w} + \ln \rho_k - \ln r_{nk} \Big\}$$
$$+ \ln \frac{C(\boldsymbol{\alpha}_0)}{C(\boldsymbol{\alpha})} + \sum_k \Big[(\alpha_0 - \alpha_k) \ln \rho_k + \eta_k \Big] .$$

E. Deriving an ELBO.

Relation to the Kullback–Leibler divergence. When we derive the ELBO among p and q (distributions) we have

$$w(X, Y) = \mathbb{E}_{X\sim q}\left[\ln Y \sim p\right] - \mathbb{E}_{X\sim q}\left[\ln X \sim q\right]$$

where X is the random variable associated to the posterior approximation, and Y to the prior. Here we show that it is possible to derive a compact form for $w(X, Y)$ without computing explicitly the two expectations. First, both expectations are under q, while the expectation term comes from either p or q. Thus, the general ELBO term is a negative cross-entropy

$$\mathbb{E}_{X\sim q}\left[\ln Y \sim f_Y\right] = \int q(x) \ln f_Y(x)\, dx = -\Big\{ \mathcal{H}[X \sim q] + \mathsf{KL}(q \parallel f_Y) \Big\} = -\underbrace{\mathcal{H}[X \sim q, Y \sim f_Y]}_{\text{cross-entropy}},$$

with $f_Y = \{p, q\}$. This is an actual cross-entropy when $X \neq Y$ $(f_Y = p)$, and reduces to a negative entropy when $X = Y$ $(f_Y = q)$ because the KL-divergence from self is 0

$$\mathbb{E}_{X\sim q}\left[\ln X \sim q\right] = \int q(x) \ln q(x)\, dx = -\mathcal{H}[X \sim q] .$$

This means that we can find a form for the ELBO which does not require the expectations with respect to q and p. In practice

$$\begin{aligned} w(X, Y) &= \mathbb{E}_{X\sim q}\left[\ln Y \sim f_Y\right] - \mathbb{E}_{X\sim q}\left[\ln X \sim q\right] \\ &= -\mathcal{H}[X \sim q, Y \sim p] - (-\mathcal{H}[X \sim q]) \\ &= \mathcal{H}[X \sim q] - \mathcal{H}[X \sim q, Y \sim p] \\ &= \mathcal{H}[X \sim q] - (\mathcal{H}[X \sim q] + \mathsf{KL}(q \parallel p)) = -\mathsf{KL}(q \parallel p) \end{aligned}$$

This alternative formulation becomes handy when the KL-divergence among q and p is known.

Terms independent of the mixture likelihoods (see (8)).

Proportions of the mixture $\boldsymbol{\pi}$. The KL-divergence for Dirichlet distributions with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is

$$\begin{aligned} \mathsf{KL}[q \sim \mathsf{Dir}(\boldsymbol{\alpha}) \parallel p \sim \mathsf{Dir}(\boldsymbol{\beta})] &= \ln \underbrace{\frac{\Gamma(\boldsymbol{\alpha}_*)}{\prod_k \Gamma(\alpha_k)}}_{C(\boldsymbol{\alpha})} + \ln \underbrace{\frac{\prod_k \Gamma(\beta_k)}{\Gamma(\boldsymbol{\beta}_*)}}_{1/C(\boldsymbol{\beta})} + \sum_k (\alpha_k - \beta_k) \underbrace{\Big[\Psi(\alpha_k) - \Psi(\boldsymbol{\alpha}_*) \Big]}_{\ln \rho_k} \\ &= \ln \frac{C(\boldsymbol{\alpha})}{C(\boldsymbol{\beta})} + \sum_k (\alpha_k - \beta_k) \ln \rho_k \end{aligned}$$

where $\boldsymbol{\alpha}_* = \sum_k \alpha_k$. In our case we have a scalar hyperparameter α_0, which leads to $p \sim \mathsf{Dir}(\boldsymbol{\alpha}_0 = [\cdots \alpha_0 \cdots])$

$$\begin{aligned} w(q(\boldsymbol{\pi}), p(\boldsymbol{\pi})) &= -\mathsf{KL}[q \sim \mathsf{Dir}(\boldsymbol{\alpha}) \parallel p \sim \mathsf{Dir}(\boldsymbol{\alpha}_0)] \\ &= -\Big\{ \ln \frac{C(\boldsymbol{\alpha})}{C(\boldsymbol{\alpha}_0)} + \sum_k (\alpha_k - \alpha_0) \Big[\Psi(\alpha_k) - \Psi(\boldsymbol{\alpha}_*) \Big] \Big\} \\ &= \ln \frac{C(\boldsymbol{\alpha}_0)}{C(\boldsymbol{\alpha})} + \sum_k (\alpha_0 - \alpha_k) \ln \rho_k . \end{aligned}$$

To avoid numerical issues in the implementation it is convenient to use the transform

$$\ln \frac{C(\boldsymbol{\alpha}_0)}{C(\boldsymbol{\alpha})} = \ln C(\boldsymbol{\alpha}_0) - \ln C(\boldsymbol{\alpha}) = \ln \Gamma\left(K\alpha_0\right) - K \ln \Gamma(\alpha_0) - \ln \Gamma \left(\sum_k \alpha_i \right) + \sum_k \ln \Gamma(\alpha_i) .$$

Latent variables. For this term we have

$$\begin{aligned}w(q(\mathbf{Z}), p(\mathbf{Z} \mid \boldsymbol{\pi})) &= \mathbb{E}_{\mathbf{Z},\boldsymbol{\pi}}\left[\ln p(\mathbf{Z} \mid \boldsymbol{\pi})\right] - \mathbb{E}_{\boldsymbol{Z}}\left[\ln q(\mathbf{Z})\right] \\ &= \sum_{n,k} r_{nk} \ln \rho_k - \sum_{n,k} r_{nk} \ln r_{nk} \\ &= \sum_{n,k} r_{nk} (\ln \rho_k - \ln r_{nk})\,.\end{aligned}$$

with as above $\ln \rho_k = \Psi(\alpha_k) - \Psi(\boldsymbol{\alpha}_*)$. It is easy to verify why this is correct: if one changes sign this becomes an entropy minus a cross-entropy, which is a KL-divergence.

4. Note 4: Analysis of patient derived data

All summary statistics for all fit samples of this paper are available in the Excel Supplementary Table 1.

A. Single sample datasets. We obtained WGS mutation calls for the breast cancer sample PD4120a and the AML Platinum sample from our earlier manuscript(11); originally these two samples have been discussed in two separate publications (5, 18). We also downloaded publicly-available WGS molecular data from high-purity lung samples stored at the Comprehensive Omics Archive of Lung Adenocarcinoma (19) (COALA, reachable at http://genome.kaist.ac.kr/). The analysis of lung cancers and the data downloaded from COALA are described in vignette "4. Single-sample cross-sectional lung cases", available as Source Data.

MOBSTER analyses these tumors in a few minutes on a standard laptop. As in previous work (11), we analyzed all single-sample cases focusing on SNVs from diploid regions of the tumour genome. For PD4120a, we used data from chromosome 3, a single largely diploid chromosome (which has$\approx$ 70% tumor purity and $n = 4,643$ SNVs in chromosome 3). In this fit the tail is as large as the largest subclone – i.e., $\approx 1,000$ SNVs ($\approx 20\%$ of the tumor mutational burden), compared to the largest subclone ($\approx 1,100$ SNVs). For AML Platinum we used multiple chromosomes, consistently with previous works (11). With this sample we obtained a similar result re-analyzing $n = 1,332$ SNVs in diploid regions of the tumour; this hematological sample has very high purity ($> 90\%$), and the subclones have 103 and 116 SNVs each. Compared to the breast sample, the tail that we fit to this tumor is much smaller (66 SNVs), suggestive of a lower mutation rate for this type of tumour. For the lung samples we inspected the data available on the COALA portal (i.e., the circos plot reporting copy number and structural variations) to determine chromosome arms with diploid karyotypes without loss of heterozygosity, and lack of structural rearrangements.

All these datasets are among the highest-resolution single-sample WGS studies available so far in the public domain, with a coverage spanning from 100x (lung samples) up to 320x (AML sample). Comparatively, samples from large-scale studies often do not hit the minimum coverage limit ($\approx$100x) that we suggest to detect reliable subclonal structures. All datasets have been analyzed with multiple setups of parameters and tools, and all analyses have given consistent results. To show a wide set of analyses, in the Main Text we report the breast and AML analysis carried out using the VAF adjusted by tumor purity, while for the lung cases we report the analysis of raw VAF. A cutoff of 0.05 (5%) has been used to filter low frequency mutations, adjusted for purity if the VAF was adjusted too. In all cases we clustered read counts of non-tail mutations with both sciClone and our in-house Binomial mixture model BMix (default parameters).

For the breast and AML samples, we computed manually the clonal trees from the output clusters, following the pigeonhole principle, and we measured the evolutionary parameters of both tumors. We found evolutionary estimates that are in agreement with the previous analysis which used Approximate Bayesian Computation to fit a stochastic branching process model (11) of tumor growth to data. The parameters identified are reported in the Main Text.

We used other single-sample datasets, as discussed in the Main Text. We report for instance cases from the COALA portal where we analysed diploid mutations that have clearly low-quality data (Supplementary Figure 19). In practice, either due to low purity, coverage or both, the signal in these data is hard to interpret and we can barely see a cluster of clonal mutations.Therefore, in cases like these we cannot really conclude much about the evolution of these tumours.

B. Multi-region datasets. For the two colorectal carcinoma Set07 and Set06 we generated whole-genome profiles sequenced at median coverage 100x, and adopted an analysis pipeline similar to what described in Cross et al. (20). The original analysis referred to these two patients as "carcinoma 6" and "carcinoma 7", but sequenced the DNA with 40x median coverage.

Files in CSV format with somatic mutations and copy number segments, as well as the scripts to run this analysis are in vignette "5. Multi-region cross-sectional colorectal carcinomas", in Source Data; the analysis took about 20 minutes on a laptop.

Mutation and CNA data. We used Mutect and CloneHD to call somatic mutations and copy number segments across all samples (copy number calls were also double checked with another caller, data not shown); sample purity was estimated from clonal diploid mutations. The purity of most samples is around 80%; for Set07 values per sample are 0.88, 0.88, 0.88 and 0.80, for Set06 they are 0.66, 0.72, 0.80, 0.80, 0.80, and 0.80 (Supplementary Figure 15).

Because these tumours are largely diploid, we used somatic SNVs in diploid regions to assess the tumour architecture. Reliable diploid regions are identified from segments with minor and major copy number equal 1, excluding areas around the centromere of each chromosome. To reduce false positive calls and contamination from germline mutations, we retained only somatic mutations called privately to each sample (to implement this filter, we also used data from a third adenoma (20)). Then, we imposed other filters on the data, removing all SNVs with adjusted VAF below a 5% cutoff (further adjusted for sample purity), and removed those with VAF above 0.7 (70%) as they originate from likely miscalled diploid segments. These filters altogether identify good quality somatic calls; after the filter we retain a large number of SNVs (approx. 80×10^3 in total), split as $\approx 50,000$ for Set07 (Supplementary Figure 15) and $\approx 30,000$ for Set06 (Supplementary Figure 17).

Deconvolution. From the available SNVs, we run MOBSTER with ICL for model selection, to be more stringent in calling tails and subclones. Each run of MOBSTER converged to a solution with 0 subclones (i.e., $k = 1$) and subclonal mutations assigned to a very evident power law tail. The tails fit for these data suggest that, at this resolution of the analysis, we do not have evidence of ongoing positive subclonal selection in both patients. This is also in line with the preliminary analysis (Figure 2 in Cross et al. (20)) which observed this from sample trees, a type of phylogenetic trees with biopsy samples as leaves.

To compute the final clusters from read counts data, we used VIBER, our in-house variational model for multivariate Binomial mixtures (Supplementary Note 3). We run VIBER using parameters similar to those used in the multi-region simulations. Output clusters have been filtered retaining those observed with inferred cluster VAF > 0.05 in at least 2 samples, and also measuring the proportion of mutations in each cluster that are private to a biopsy, and the overall proportion of mutations with respect to the tumour burden (tail and non-tail mutations). Using these strategies we filtered out clusters (i.e., ancestors) with mostly private mutations, as well as very small clusters. The overall analysis shows us two large mono-clonal tumours, with no evidence of subclones expanding (Supplementary Figures 15 and 17).

dN/dS analysis. We used a method that does not use VAF to detect subclones under positive selection hiding under the tails of these tumors. Our point is to use alternative methods to assess whether MOBSTER's tail mutations seem likely neutral, or not, with a different measurement. We used the dndscv tool by Martincorena et al. (21) to estimate the dN/dS ratio of non-synonymous to synonymous substitutions for patients (separately and pooled).

To achieve statistical power for this analysis and achieve a minimum number of coding substitutions from these two patients, we computed the estimates across all genes(default dndscv gene_list = NULL). We split mutations into two groups, comparing non-tail mutations identified as clonal by our analysis, and pooling together all other mutations (i.e., tail mutations, and those removed by our heuristics). The presence of possible positive selection would be supported by estimated values strictly greater than 1 (dN/dS > 1); neutral mutations should have dN/dS ≈ 1, and mutations under negative (or purifying) selection would show dN/dS < 1. For each estimate, the method returns both a point estimate of the dN/dS value, as well as a 95% confidence interval (CI), ideally we would like the CI not to contain 1(21).

For pooled data and clonal mutations we find a value above one (dN/dS $= 1.55$; CI $[0.91, 2.65]$) as expected, which breaks down for Set07 (dN/dS $= 1.78$; CI $[0.87, 3.6]$) and Set06 (dN/dS $= 1.29$; CI $[0.58, 2.8]$). This analysis of the remaining mutations confirms lack of evidence for positive subclonal selection, with pooled dN/dS values below or almost equal to 1, and narrow confidence intervals (pooled dN/dS $= 0.85$; CI $[0.70, 1.02]$; Set07 dN/dS $= 0.95$; CI $[0.76, 1.20]$ and Set06 dN/dS $= 0.68$; CI $[0.50, 0.92]$). Interestingly, the observation of negative selection forces, hereby suggested by dN/dS values below 1, has been recently linked to power-law neutral dynamics for the clone size distribution (22).

Summarizing, this analysis provides further evidence to our conclusions, screening off the possibility that we have missed detectable subclones (false negatives). Of course, the strength of this conclusions is limited by the number of coding substitutions in the input data for dN/dS, which here is generated from only two patients; for consistency we report that we obtained similar dN/dS estimates using an alternative method(23) (data not shown).

Clone trees. From the clusters we computed the possible clone trees (data not shown), measuring violations of the pigeonhole principle with our previous method (16), as done in multivariate simulations. Trees have been computed from the clusters obtained with and without MOBSTER (Supplementary Figures 16 and 18). In both cases, we find a much more complex clonal history without MOBSTER, because the confounders due to neutral evolution create small subclonal clusters that lead to the construction of complicated clone trees, exactly as we observed in simulations (Main Text). Since these clusters can be attached to different internal tree nodes, violations of the pigeonhole principle during tree-assembly can happen. Without MOBSTER even the best tree has at least one such violation, otherwise the output trees are trivial (monoclonal).

Beyond tree structure, we are interested in how the SNVs organise in clusters. This is fundamental, as clustering assignments are a proxy for the putative genotype of each clone. We mapped the clusters across analyses: if the two analyses were perfectly concordant, every cluster should map to only one other cluster. This is not the case, and from the mappings we see that MOBSTER's tail mutations spread evenly across clusters (Supplementary Figures 16 and 18). This confirms that neutral mutations confound subclonal deconvolution. From these plots we also see that the clonal cluster determined without MOBSTER contains tail SNVs that are clonal in some biopsy, but subclonal in others. This suggests a complex scenario of intra-tumor spatial heterogeneity for these samples, and the effect of spatial sampling bias when we attempt a reconstruction without MOBSTER.

Driver mutations. We also checked for the presence of somatic driver mutations in the VAF spectrum, and annotated the same set of driver events already identified by Cross et al. (20). These are somatic mutations (SNVs and indels) in well-known colorectal driver genes in COSMIC: for patient Set07 we find mutations in APC (p.R787X and p.R1432X), KRAS (p.G12D), SMAD3 (p.Y42X) and TP53 (p.E159X), while for patient Set06 we find mutations in APC (p.R216X), KRAS (p.G12V), PIK3CA (p.C420R), ARID1A (p.W1453X) and TCF7L2 (indel). Some of these mutations are not SNVs, or happen to be found in non-diploid regions (20); we used the model fits to map, a posteriori, such mutations to the clusters detected by our analysis.

A remark is due for the presence of a PIK3CA mutation, which is annotated as driver in the trunk of the tree of Set06. Cross et al have annotated that mutation originally as clonal (20); in our data we find it in a clonal cluster (5 out of 6 samples), and tail in sample Set6_42. For this reason, the SNV is removed from successive Binomial clustering. Nonetheless, that mutation in Set6_42 has adjusted VAF about 30%, which places the mutation slightly below the point of crossing of the tail and the clonal cluster. This is also reflected by the posterior latent variables (data not shown) that assign the mutation to the tail (≈80% probability) and the clonal cluster (≈20% probability).

The mutation is clearly clonal in all the other biopsies (adjusted VAF well above 40%). However, the purity of Set6_42 is well below the purity of the other five samples (66% versus 80%), which makes it possible that this reduced VAF is also due to excessive contamination of normal tissue in this sample, and consequent higher noise in the sequencing data. For this reason, we preferred to be consistent and annotated this mutation as clonal in Set06. This discrepancy does not affect our analysis, or our conclusions. In particular, in the dN/dS analysis – which is the only one that might be affected by this clonal/ subclonal classification – we have assigned the PIK3CA mutation to the set of tail mutations as reflected by our analysis. This avoids us to bias the reported statistics in favor of neutrality, which is a key point of our analysis.

C. PCAWG cohort. We obtained somatic mutations and copy number calls from the PCAWG cohort. In our analyses we have used consensus purity and ploidy estimates, consensus copy number calls and consensus somatic (snv_mnv) mutations with flag passonly. We have not implemented additional quality check steps on the consensus calls from PCAWG, but filtering input VAF values consistently with other analyses.

For each available patient we have mapped all reported SNVs to the available consensus copy number segments. Then, we have split mutation data into up to five "karyotypes", according to the copy state of each segment:

- loss of heterozygosity (LOH, "1:0" or genotype A);
- copy-neutral LOH ("2:0"or genotype AA);
- diploid ("1:1" or genotype AB);
- triploid ("2:1" or genotype AAB);
- tetraploid ("2:2" or genotype AABB).

We use "A:B" to represent the copies of the major (A) and minor (B) alleles. We have then fit raw VAF values for each karyotype, if the number of available mutations was above a minimum cutoff. In Source Data we provide a list of the samples we used. By considering mutations in five possible copy-states for all the cohort we obtain inputs for about ≈ 8000 fits to run with MOBSTER.

Deconvolution. The sequencing coverage of this cohort is below our observed minimum coverage requirements (≈ 100x) to detect reliable subclonal dynamics. Some cases are therefore difficult to assess because at PCAWG's depth and purity it seems often statistically impossible to distinguish a genuine subclonal cluster due to selection from the leftover of a down-sampled "degenerate" neutral tail. As an example diploid tumours where we cannot fit a tail, but the data resolution seems to low draw a conclusion about the subclones, are shown in Supplementary Figure 20.

Therefore we first opted for a conservative analysis using the ICL model-selection criterion, and then carried out a less stringent analysis with reICL (Supplementary Figure 21). In both analyses, we identified the best fit and assigned mutations to clusters; and then performed Binomial clustering of non-tail mutations using BMix, our in-house Binomial model that score mixtures by ICL (this time, with a Binomial likelihood function). In both VAF and read counts fits, we scanned for a model with maximum 3 clones (2 subclones). The ICL analysis is more stringent and requires a stronger signal in the data to determine tails, especially when they co-exist with subclones. This means that we are obtaining a lower bound for the number of subclones in the data, when reporting summary statistics with ICL.

In order to assess the effect of ICL on our reported "3% of cases with a subclone" (Supplementary Figure 21), we have also re-computed the same statistics with reICL (less conservative). We report a scatter plot of the full-cohort with both scores (analogous to Figure 4). We also plot, in the same figure for the diploid cases determined in PCAWG, the number of Beta mixture components fit by MOBSTER. This is not the same as the number of Binomial clusters, but still it allows to compare MOBSTER's mixture complexity among analyses. We note that in most cases each MOBSTER Beta component is fit by a single Binomial distribution, which saved us from re-computing Binomial fits from non-tail mutations determined with reICL. The two analyses change the number of cases classified as "monoclonal" versus "polyclonal", but the large majority of cases is classified consistently (approx. 3% of polyclonal with ICL, 7% with reICL). In both cases, our estimates of the prevalence of subclonal positive selection are very much lower than those reported by the PCAWG consortium (24). Example cases with a possible subclone in our analysis are in Supplementary Figures 22.

D. Longitudinal datasets. Upon request to the authors, we have obtained somatic mutation calls and copy number segments for $n = 16$ matched primary-relapse glioblastoma samples recently published (25). These are patients that have been diagnosed with a primary IDH-wildtype glioblastoma, and that have presented a relapse tumor after therapy; the available data consist in high-depth whole-genome sequencing at about 100x median coverage. Given the size of this cohort and the sequencing setup, this dataset provides the ideal information to study clonal evolution under therapy over time.

To analyze these data, we split samples based on collection time-points, shortly called pre and post. In this dataset, mutations have been annotated in the corresponding VCF files only if they had non-zero VAF. Without coverage for mutations that are private to a biopsy, we could not run a canonical multivariate Binomial analysis, and rather split the analysis of pre and post sample for every patient, processing the VAF of major karyotypes, adjusted for tumor purity estimated from the peaks in the VAF distribution of diploid mutations. We analyzed:

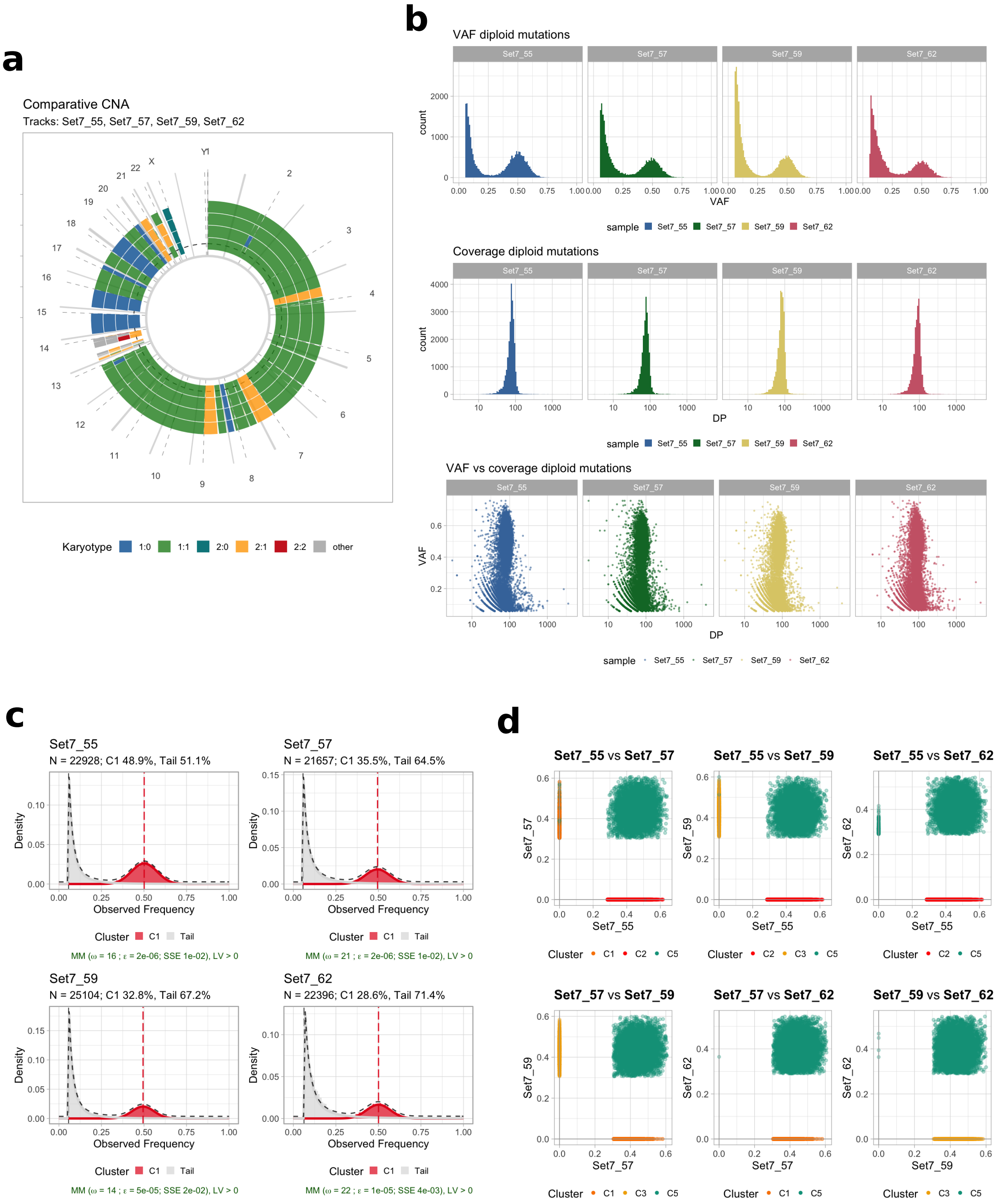
- diploid (“1:1”);
- triploid (“2:1”);
- and tetraploid (“2:2”)

clonal copy number segments, where the numbers follow the notation used in the PCAWG analysis. Then, for every sample we used the fit of the mutations that match the tumor ploidy to classify a sample; for a tumor that is overall diploid we used mutations from karyotype “1:1”, while for a tumor that underwent genome doubling we used mutations from karyotype “2:2”.

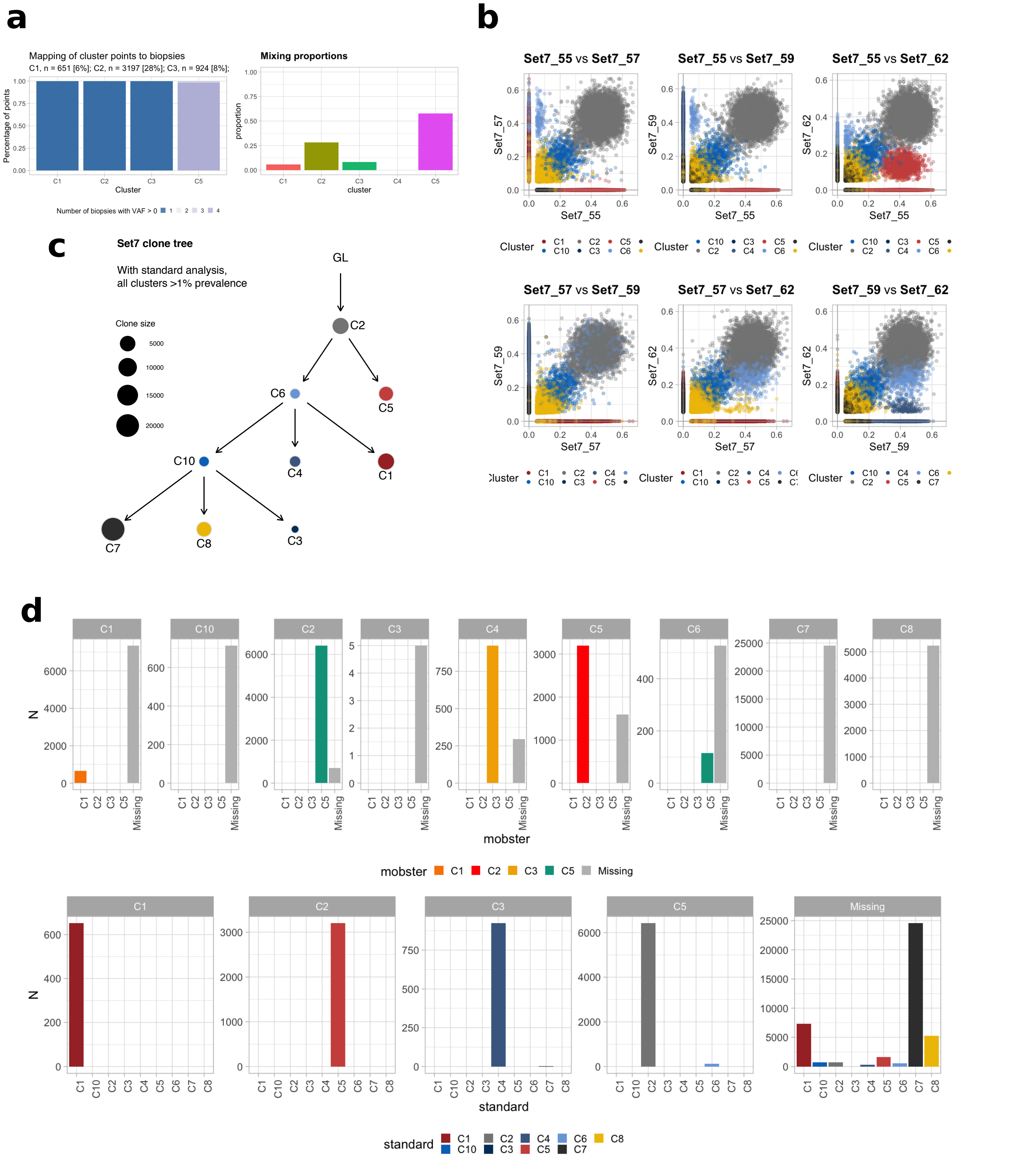
Deconvolution and dN/dS statistics. Our reanalysis did identify a significant number of subclones likely under positive selection, consistently with the original analysis of these data (25). The original analysis employed a standard clustering (Binomial mixtures) of the input data, and the resulting clone trees reflected complex polyclonal structures. With our analysis, in most cases, the predicted clonal architectures are much simpler because we get rid of neutral mutations. All the analyzed cases are available in Supplementary Figure 23.

In each plot we report the fit from the most prevalent karyotype in each biopsy, as well as a two-dimensional plot (pre versus post), colored by mutation status. In the figure, we also report analysis of non-tail mutations with BMix’s Binomial mixture model, as well as the VAF change over time for a pool of genes that have been originally annotated as potential drivers by Korber et al (25). In the Main Text we have included the cases that we deem more representative of the observed patterns of evolution. These subclones were observed only in the relapse tumor in $n = 3$ cases (with one relapse sample containing multiple selected subclones) or only in the primary tumor in $n = 2$ cases. In $n = 1$ case, different subclones were detectable in the primary and relapse samples. In a few cases we also observe the effect of treatment (temozolomide) in increasing the tumor mutation rate. In a number of cases which include tumors with overall ploidy 4 (genome doubled cases, often found at relapse), we do not find evidence of subclones growing in the primary tumor. Notably, in all the analyzed cases we were able to fit a tail to the data, and the ancestor effect induced by spatial sampling of the tumor at diagnosis and relapse was also very evident, with clonal mutations private to each individual biopsy (Supplementary Figure 23).

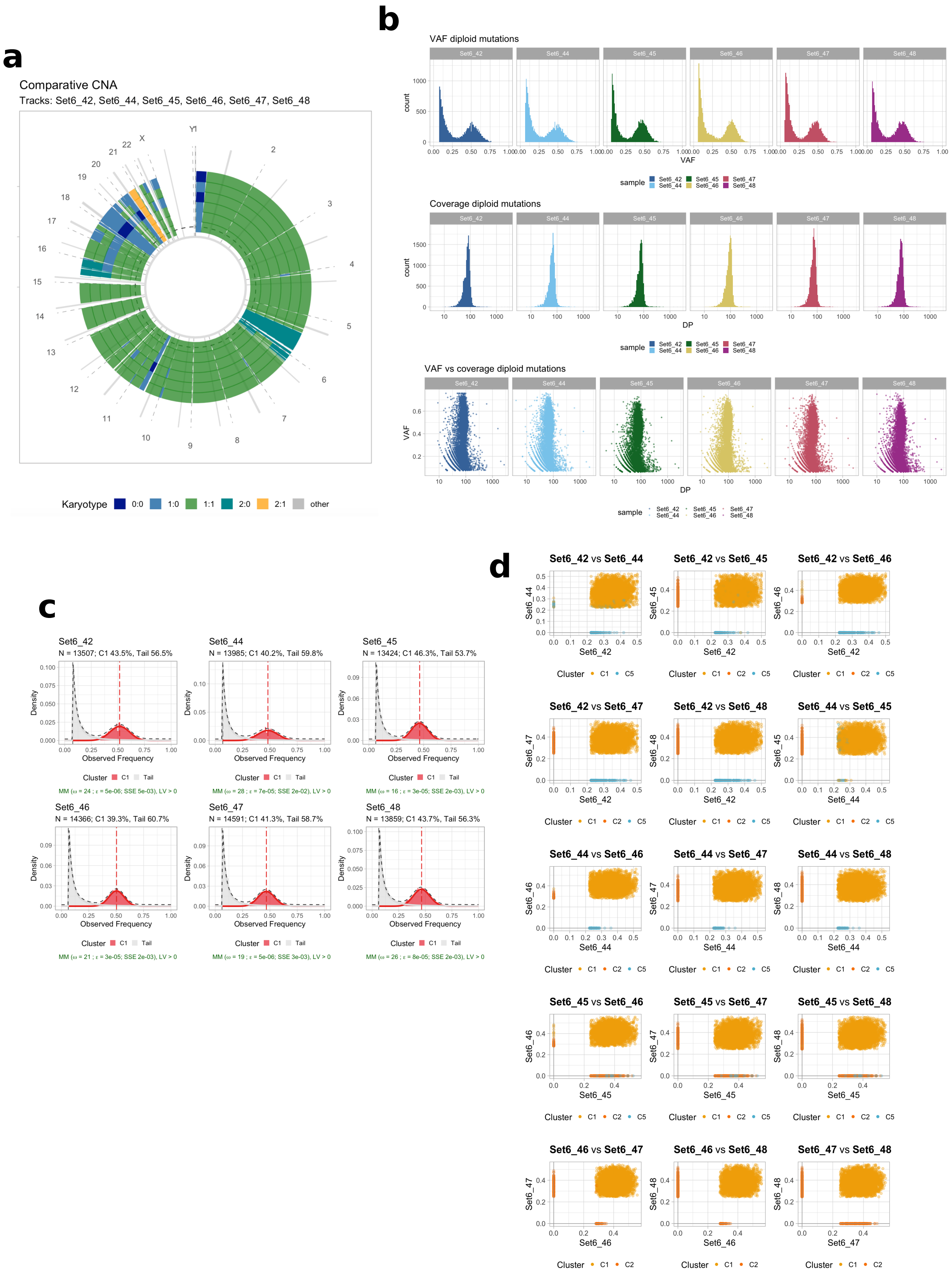
As for other samples, we also assessed the dN/dS statistics with dndscv (21). To run this analysis, we used 74 genes from the Intogen database (https://www.intogen.org) that are likely glioblastoma drivers, and compared tail versus non-tail mutations across all the analysed karyotypes. Results are consistent and dN/dS values of tail mutations approach 1.0, suggesting that tails are unlikely to hide small subclones, at least driven by nonsynonymous somatic mutations.



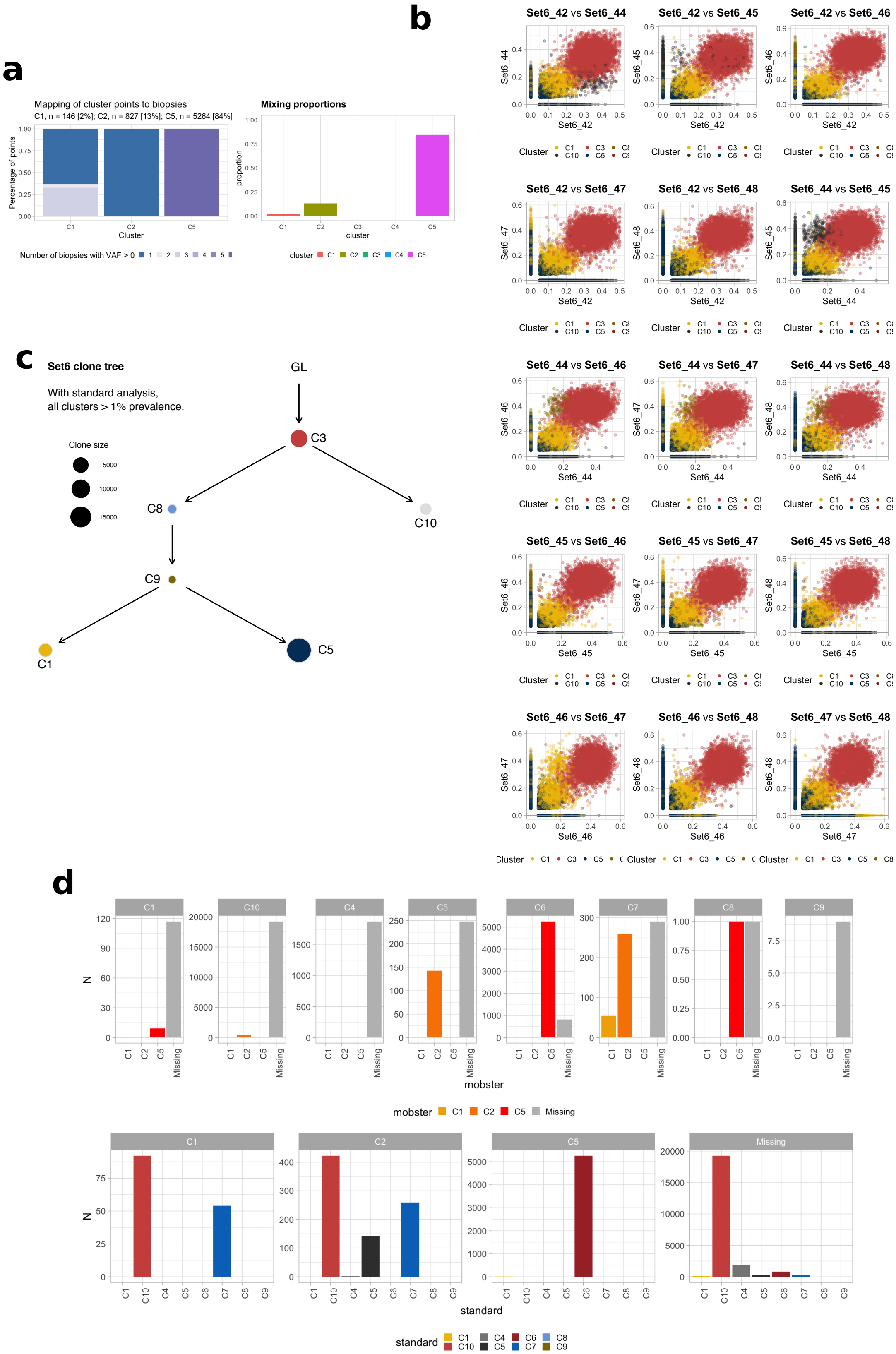
Supplementary Figure 15. Analysis of `Set07` with MOBSTER **a.** `CloneHD` Copy Number calls for the `Set07` colorectal carcinoma, visualised with `CNAqc`. Each line in the circular layout is a whole-genome segment for one of the input biopsies. In the plot we colour the karyotypes defined as number of copies of the Major and minor alleles. **b.** For each sample, we show the VAF of diploid mutations, their coverage and one measurement against the other. **c,d.** MOBSTER fits of the input samples (c), and downstream clustering of non-tail mutations with VIBER.



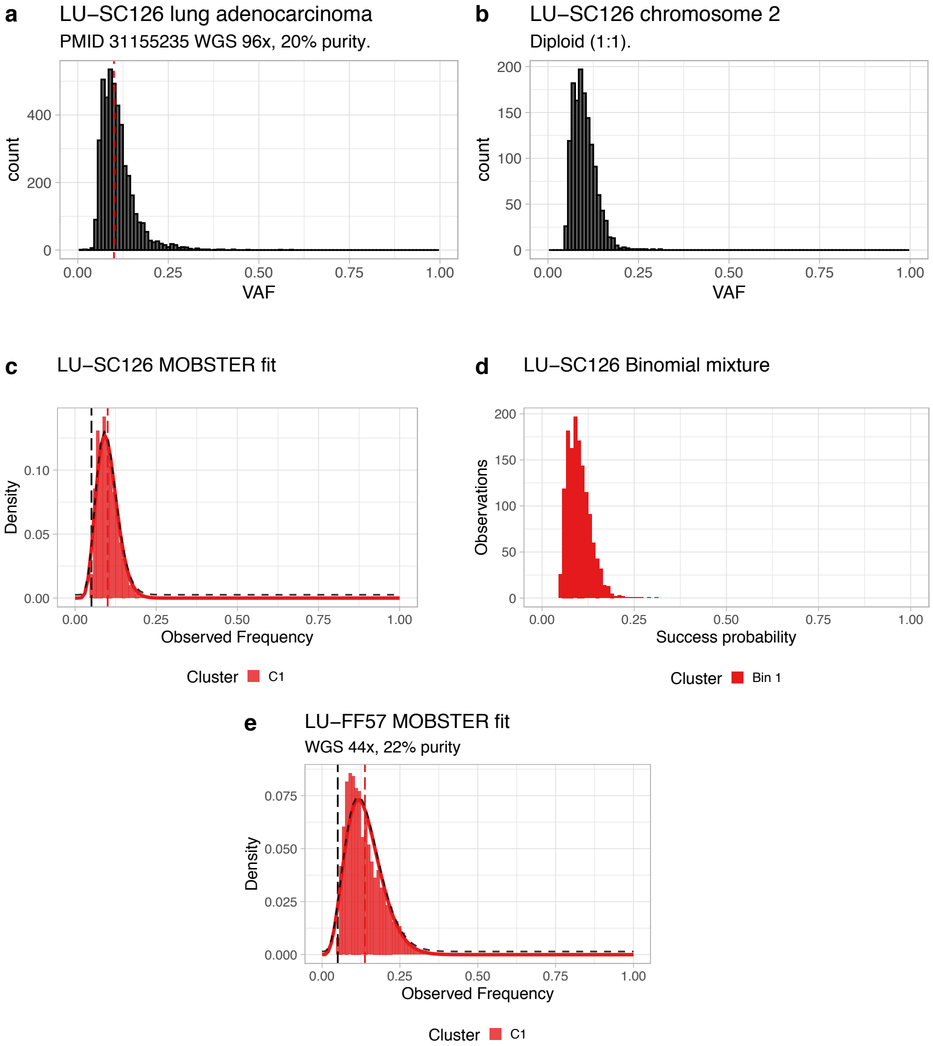
Supplementary Figure 16. Analysis of `Set07`**, more details. a.** In left, for every MOBSTER clusters we report the proportion of cluster points with non-zero VAF in x of the input biopsies. In right, we show the model's mixing proportions. These plots suggest that most clusters contain private mutations, or are very small. **b,c.** Standard analysis obtained by clustering all diploid mutations with VIBER. Many clusters generate a complex clone tree (panel c), whereas the analysis with MOBSTER suggest that this tumour is mono-clonal. **d.** Mapping among MOBSTER's clusters, and the clusters found by the standard analysis (panel b).



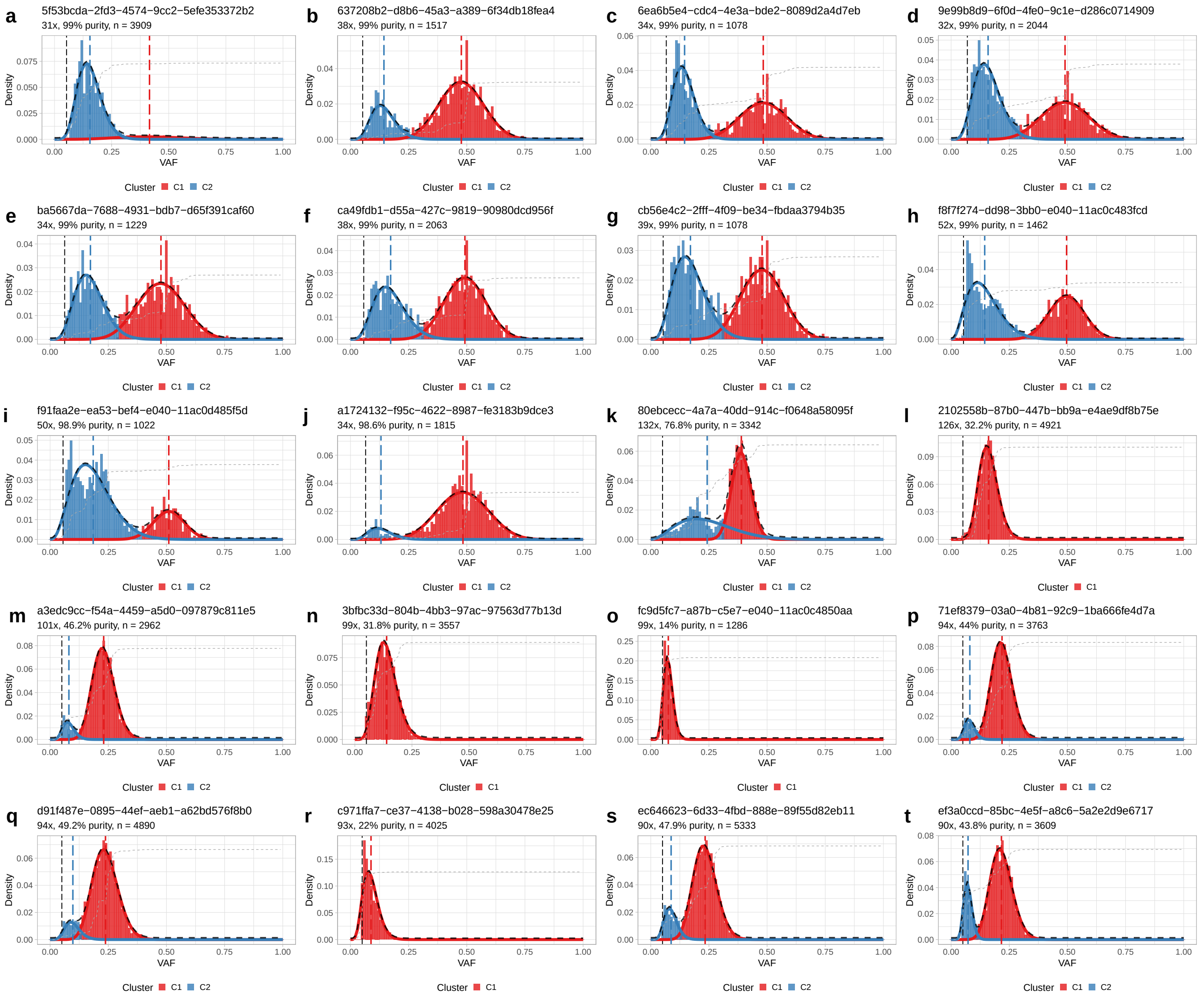
Supplementary Figure 17. Analysis of `Set06` **with** MOBSTER **a.** `CloneHD` Copy Number calls for the `Set06` colorectal carcinoma, visualised with `CNAqc`. Each line in the circular layout is a whole-genome segment for one of the input biopsies. In the plot we colour the karyotypes defined as number of copies of the Major and minor alleles. **b.** For each sample, we show the VAF of diploid mutations, their coverage and one measurement against the other. **c,d.** MOBSTER fits of the input samples (c), and downstream clustering of non-tail mutations with VIBER.



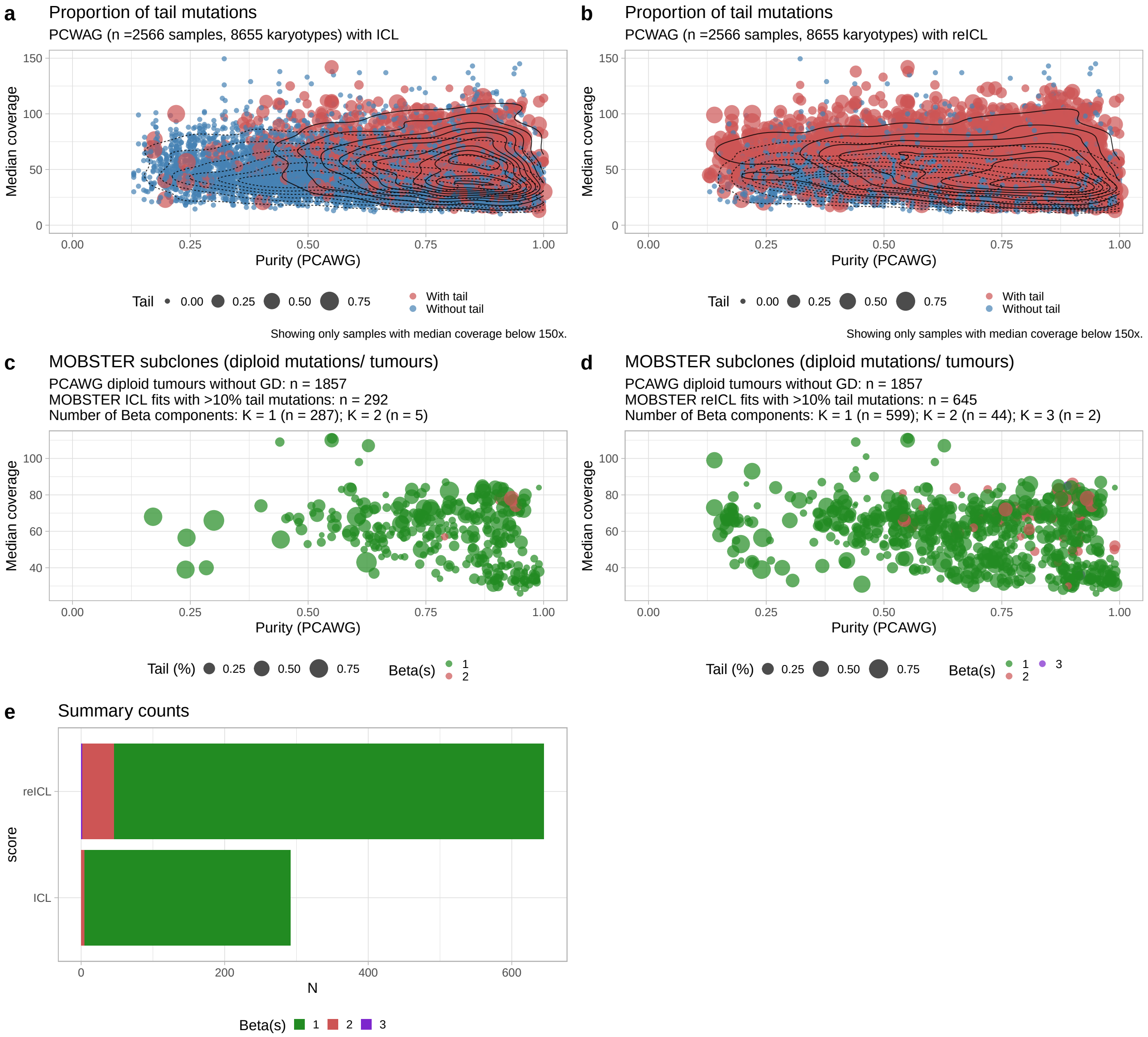
Supplementary Figure 18. Analysis of `Set06`**, more details. a.** In left, for every MOBSTER clusters we report the proportion of cluster points with non-zero VAF in x of the input biopsies. In right, we show the model's mixing proportions. These plots suggest that most clusters contain private mutations, or are very small. **b,c.** Standard analysis obtained by clustering all diploid mutations with VIBER. Many clusters generate a complex clone tree (panel c), whereas the analysis with MOBSTER suggest that this tumour is mono-clonal. **d.** Mapping among MOBSTER's clusters, and the clusters found by the standard analysis (panel b).



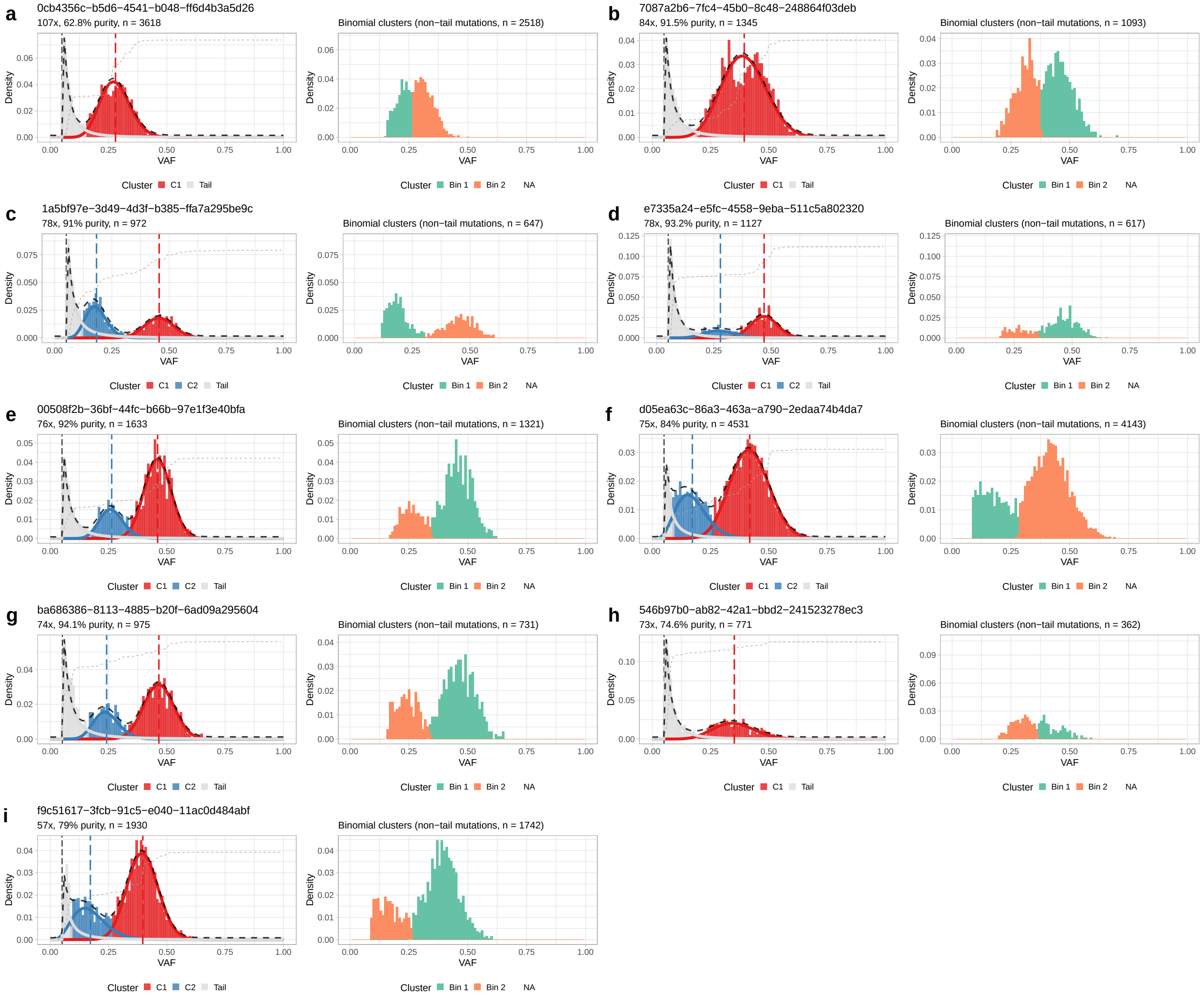
Supplementary Figure 19. Low quality WGS data that challenges the reconstruction. Cases of patient-derived WGS samples that are difficult to analyse because the resolution of the data is low. In practice, either due to low purity, low coverage or both, the signal in these data is hard to interpret and we can barely see a cluster of clonal mutations.Therefore, we cannot really conclude much about the evolution of these tumours; for instance, we cannot determine if the data contains all tumour clonal mutations (and therefore, drivers), if there is any pattern of subclonal evolution, or if there are tails. All samples are lung adenocarcinomas available from the COALA portal (http://genome.kaist.ac.kr/), and we analyse diploid mutations **a,b.** Sample `LU−SC126`. A case with quite high coverage (WGS $96\times$), but low purity (20%). All mutations in left panel, and mutations in genomic regions with diploid karyotype (without LOH) in the right panel. **c,d.** MOBSTER fits and Binomial clustering o non-tail mutations. The only signal detectable is that of the clonal population, no tails or no subclones can be detected with these data. **e.** Sample LU-FF57. This sample has low coverage (WGS $44\times$) and low purity (22%). The fit with MOBSTER again is a monoclonal tumour without tail.



Supplementary Figure 20. Diploid tumours where we could not fit a tail in `PCAWG`**.** With the same setup described in the Main Text and summarised in Supplementary Figure 21 , we have $n = 919$ cases of diploid tumours with at least 1000 SNVs where we cannot fit a tail using `ICL`. Here we have ranked them by purity (a-j) and median coverage (k-t), and plot the top 10 for each category. We report here the MOBSTER fit of each one of these cases. There is a number of different possibilities why these cases are more difficult to fit a tail of neutral mutations. **(l, n, o, r).** Some low-purity cases where there is not enough information to fit any subclonal evolutionary dynamics in these data – which means that we cannot assess whether there is any expanding subclone in these tumours, and therefore we should not use these samples to report cohort-level statistics. **(m, p, q, s, t).** Cases with a similar pattern of data quality that might be too low to determine reliable evolutionary dynamics. In these cases it is difficult to establish if C2 is a real subclone just because it seems that we see a little portion of its data – we note that in this situation we should be equally skeptical if that cluster was fit to a tail, which is the reason we have imposed a minimum tail size in Supplementary Figure 22 as a proxy for data quality. **(a).** Cases where the presence of a very small clonal cluster (compared to the subclone), raises suspicion on the possibility that C1 is just the result of miscalled loss of heterozygosity. Was that the case, the real tumour purity would be well below 100%. **(b, c, d, e, f, g, j, j, k).** Cases where the left-most part of the tail VAF spectrum might be affected by other sources of noise. This is suggested by the fact that the data histogram does not decay sharply when we impose a minimum cutoff on the VAF – 5%, which we have used for this and other analyses. In these distributions this characteristic decay might be due to real subclones sitting on top of tails that have huge shape value (low mutation rates), or might be due to VAF-dependent noise originating from consensus calling strategies based on tools which have different rates of false negatives with low-frequency VAF values. A consequence of this type of distribution is that fitting a power law tail with scale parameter as the minimum of the data – the maximum-likelihood estimator of this distribution – is not statistically convenient. To mitigate this effect, one could increase the VAF cutoff to produce sharper decays; however, that would easily lead to the inflation of fabricated neutral tails, which we want to avoid. Given this difficulties and peculiar shape of the VAF distribution, we remain agnostic to the possibility that in any of these tumours there are one or more subclone. **(h, i).** Cases where `ICL` penalises excessively the fit, because of its stringent entropy-derived penalty as explained in Supplementary Figure 4. In these cases we would be better off using `reICL`, which would have fit to data multiple Betas plus a tail (data not shown), classifying these cases as the ones in Supplementary Figure 22.



Supplementary Figure 21. Summary of `PCAWG` fits with `ICL` and `reICL`. a,b. The same type of scatter plot as shown in the Main Text for the `PCAWG` cohort. Here we report fits by `ICL` in panel (a), and fits by `reICL` in panel (b). The latter type of scoring system is less stringent in calling tails; in fact the number of red points (fits with tail) is clearly larger in panel (b). **c,d.** The `PCAWG` cases for which we report summary statistics in the Main Text are diploid (assessed by `PCAWG`), without Genome Doubling (GD, assessed by `PCAWG`); these are $n = 1857$ cases. For these cases we examine the fits of the main diploid karyotype (i.e., mutations mapping to copy number segments with single copy of both the major and minor alleles, assessed by `PCAWG`), obtained with `ICL` and `reICL`. We seek to examine fits that have a clear signal of a neutral mutations, as we expect following the theory of clonal evolution; to determine this set we take fits with at least 10% of the mutational load assigned to the tail component (fits with smaller tails might have fit tails instead of genuine subclones). `ICL` is more conservative in calling tails, and in fact we find more tails among cases fit with `reICL` (roughly three times more, numbers in then panel). In panel (c) and (d) we report the scatter of these points, coloured by the number of Beta components used in MOBSTER fits, with size proportional to the fit tail. We remark that the actual number of subclones is given by fitting Binomial mixture from non-tail mutations, which is the statistics that we report in the Main Text. Here we report Beta components from MOBSTER, because we do not have Binomial fits for non-tail mutations identified by `reICL` (we have them for fits by `ICL`, Main Text). **e.** Summary counts of the coloured points in the scatters of the previous panels shows that the amount of tumours with a possible subclone is much smaller than the proposed number reported by the original evolutionary-unaware `PCAWG` analysis.

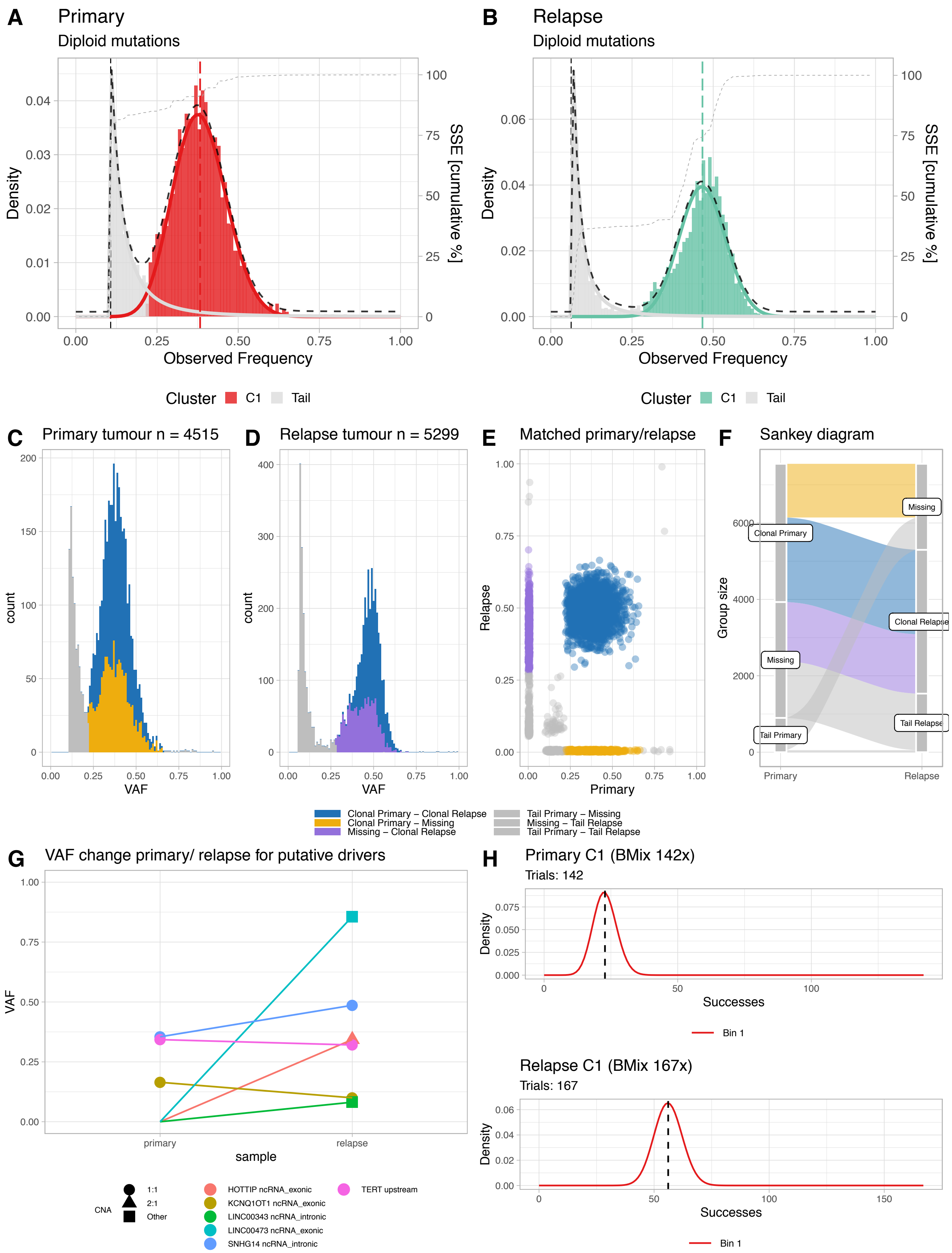


Supplementary Figure 22. Diploid tumours with a tail and a subclone in PCAWG**.** We have analysed the PCAWG cohort as described in the Main Text. Out of the PCAWG tumours (see Supplementary Figure 21) that have ploidy 2 (rounded to integer), that are not genome doubled (no_wgd) and where we can fit a tail that contains at least 10% of the mutational load with with ICL, we identify $n = 9$ cases that contain one subclone (determined from Binomial fits of non-tail mutations). The cases are reported here, with the annotated purity and median coverage for the mutations that map to the clonal diploid copy number segments estimated by the consortium. In this plot there are different types of cases, some with very low frequency subclones (f, i), others with subclones that have very high cancer cell fractions (a, b, h).

Supplementary Figure 23. Analysis of longitudinal IDH-wt glioblastoma patients. Multi-page figure with the analysis of $n = 16$ IDH-wt glioblastoma patients patients available from (25). This legend applies to all figures from page this page to the References; each page correspond to one analysed patient, with the patient identifier reported on top of the page. Each figure has the same panel labels across all patients, and its description is in this legend. In the Main Text we report the description of the analysis, which for short consists in analysing separately primary and relapse samples of each patient, splitting somatic SNVs based on the copy number segments they map to (diploid, triploid and tetraploid karyotypes). As filtering, we report the analysis from output clusters that contain at least 50 SNVs. **a.** MOBSTER analysis of the primary biopsy, for the largest karyotype (i.e., the karyotype the higher number of mutations). **b.** MOBSTER analysis of the relapse biopsy, for the largest karyotype (i.e., the karyotype the higher number of mutations) as in panel a). **c,d, e, f.** Colouring of the data based on the pattern of occurrence of the mutations, and MOBSTER's analysis. For instance, “Clonal Primary - Clonal Relapse“ means that the mutations are present in the putative clonal cluster both at point of care and relapse, “Clonal Primary - Missing“ are mutations that are clonal private to the primary biopsy (i.e., they have VAF 0 in the matched relapse sample). Panel e) reports the joint VAF values, and panel f) the Sankey diagram that shows how the clusters spread across the biopsies. **g.** Difference in VAF for a pool of events annotated in the original paper as putative drivers (25). These are either SNVs in coding and non-coding regions of the genome (note mutations in the promoter of `TERT`). **h.** Binomial mixture analysis of mutations mapped to each one of the clusters computed by MOBSTER; this panel reports the discrete Binomial density fit to data at the median coverage observed in the reference cluster.

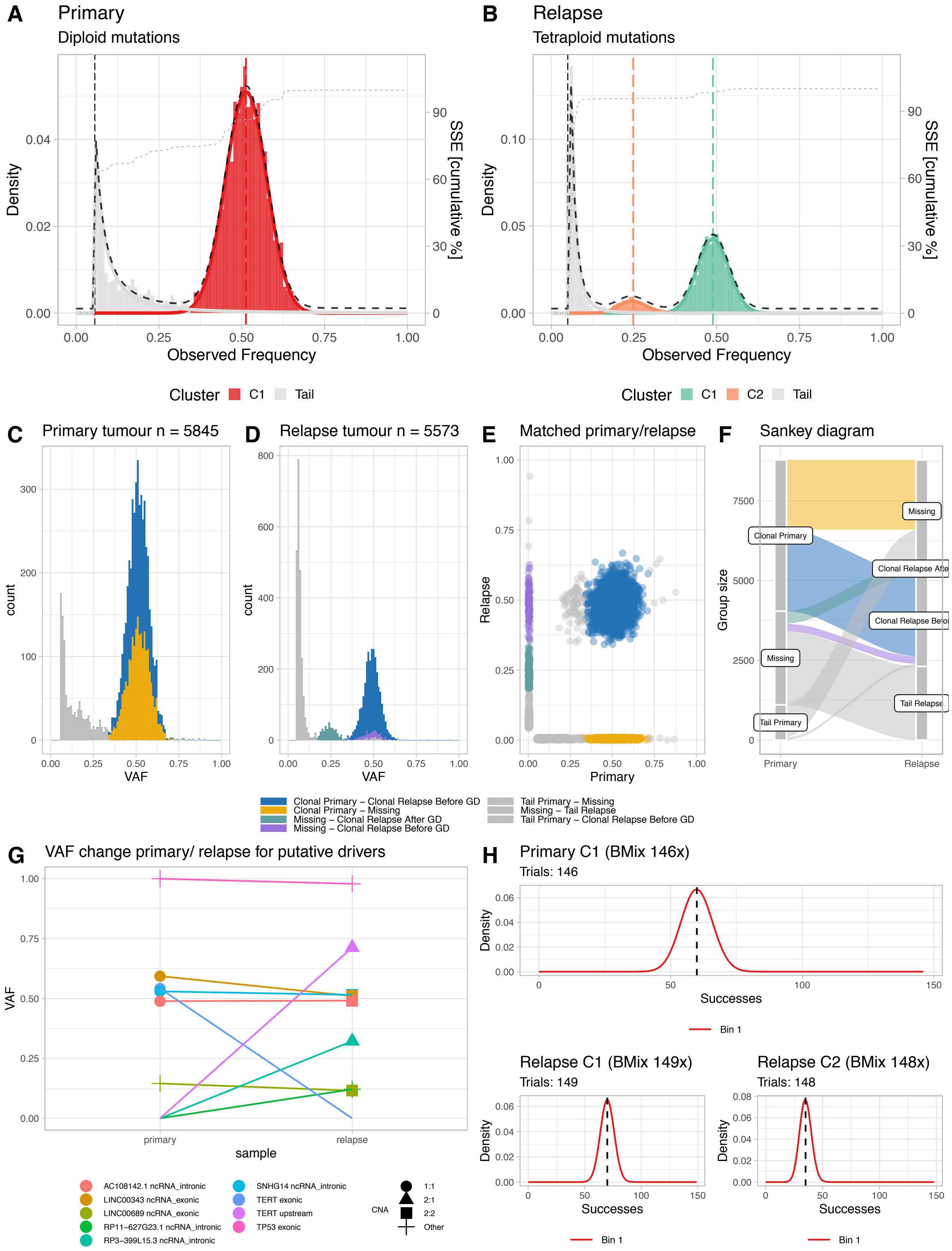
H043–5VWP

Groups with at least 50 SNVs.



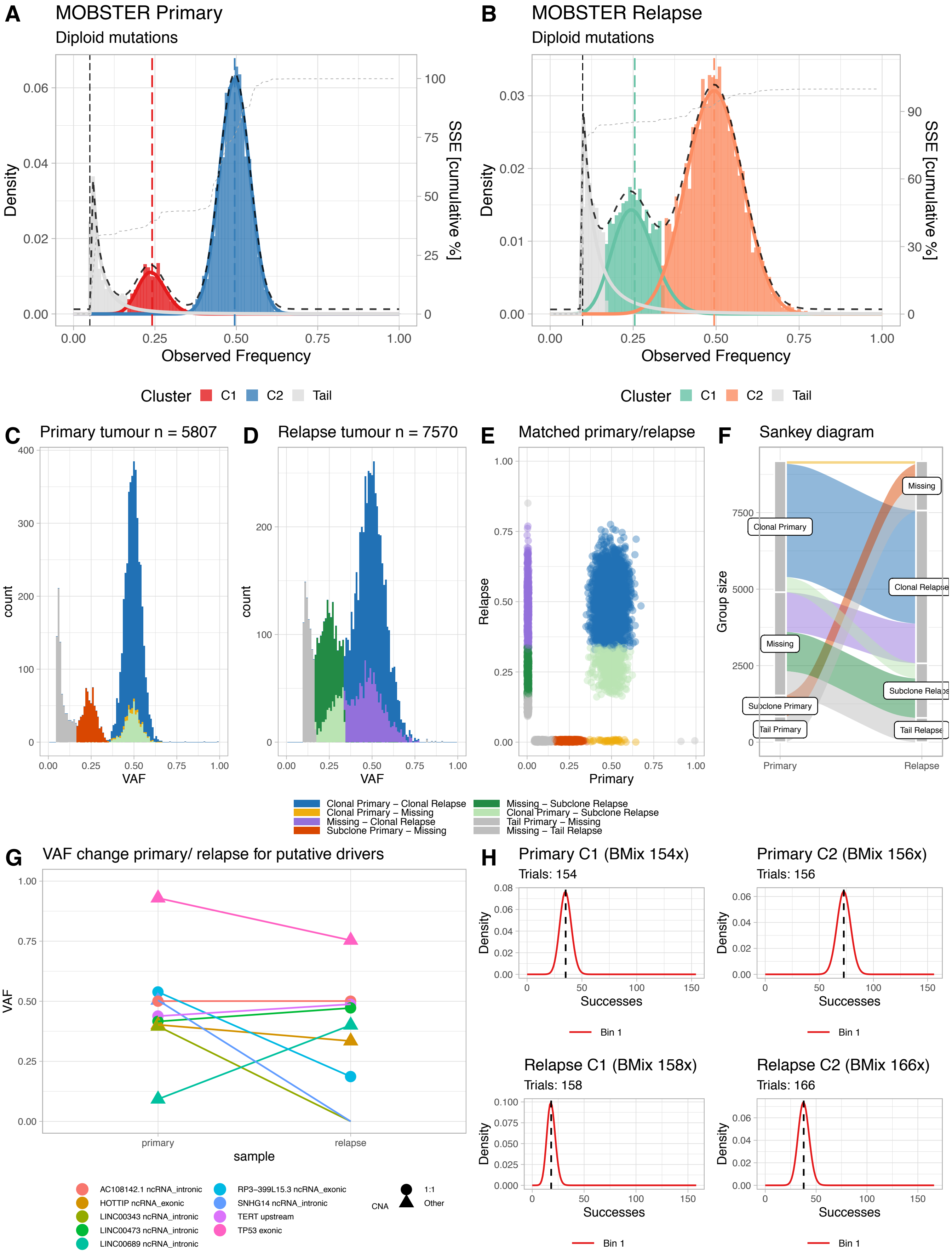
H043–6F91

Groups with at least 50 SNVs.



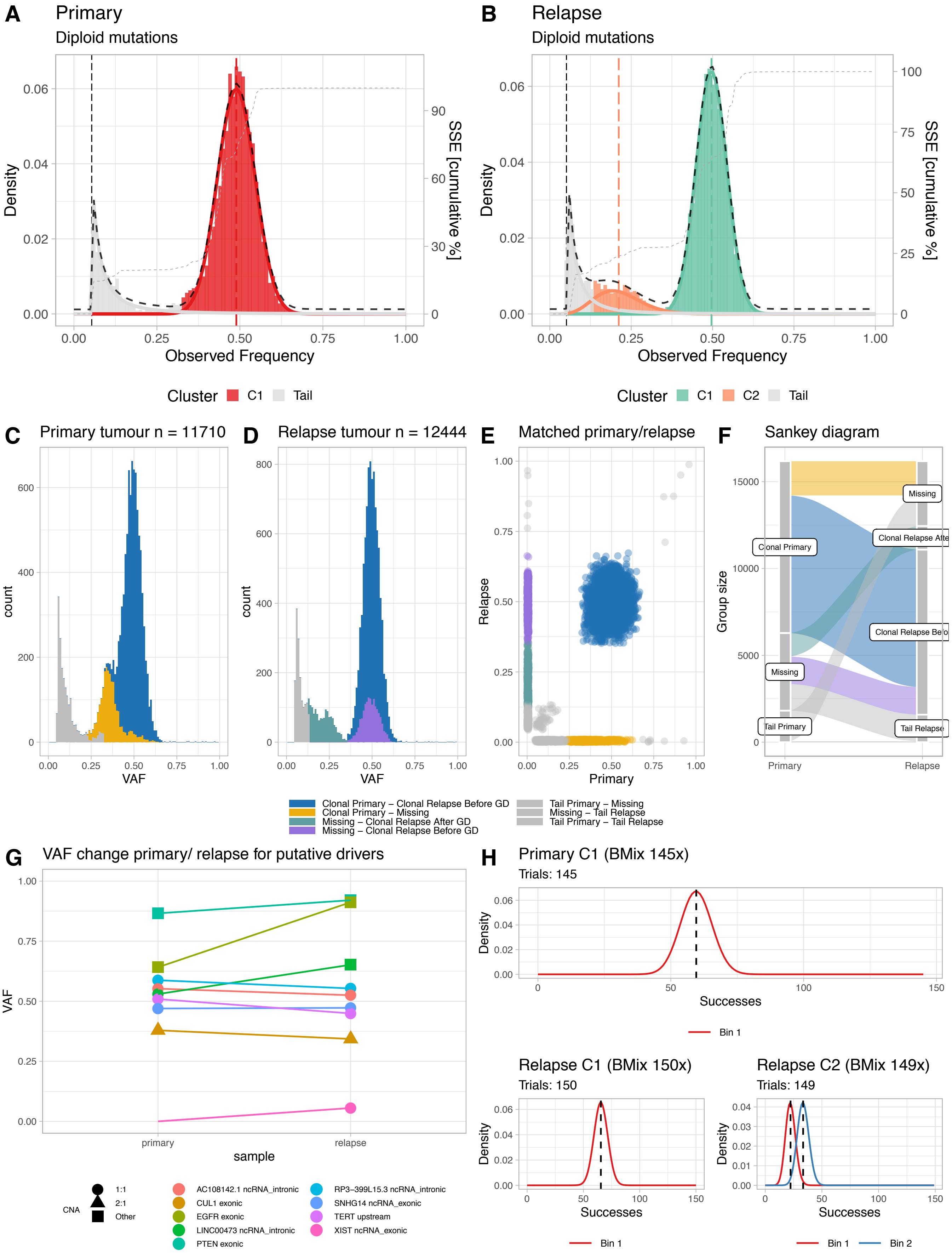
H043–28GK

Groups with at least 50 SNVs.



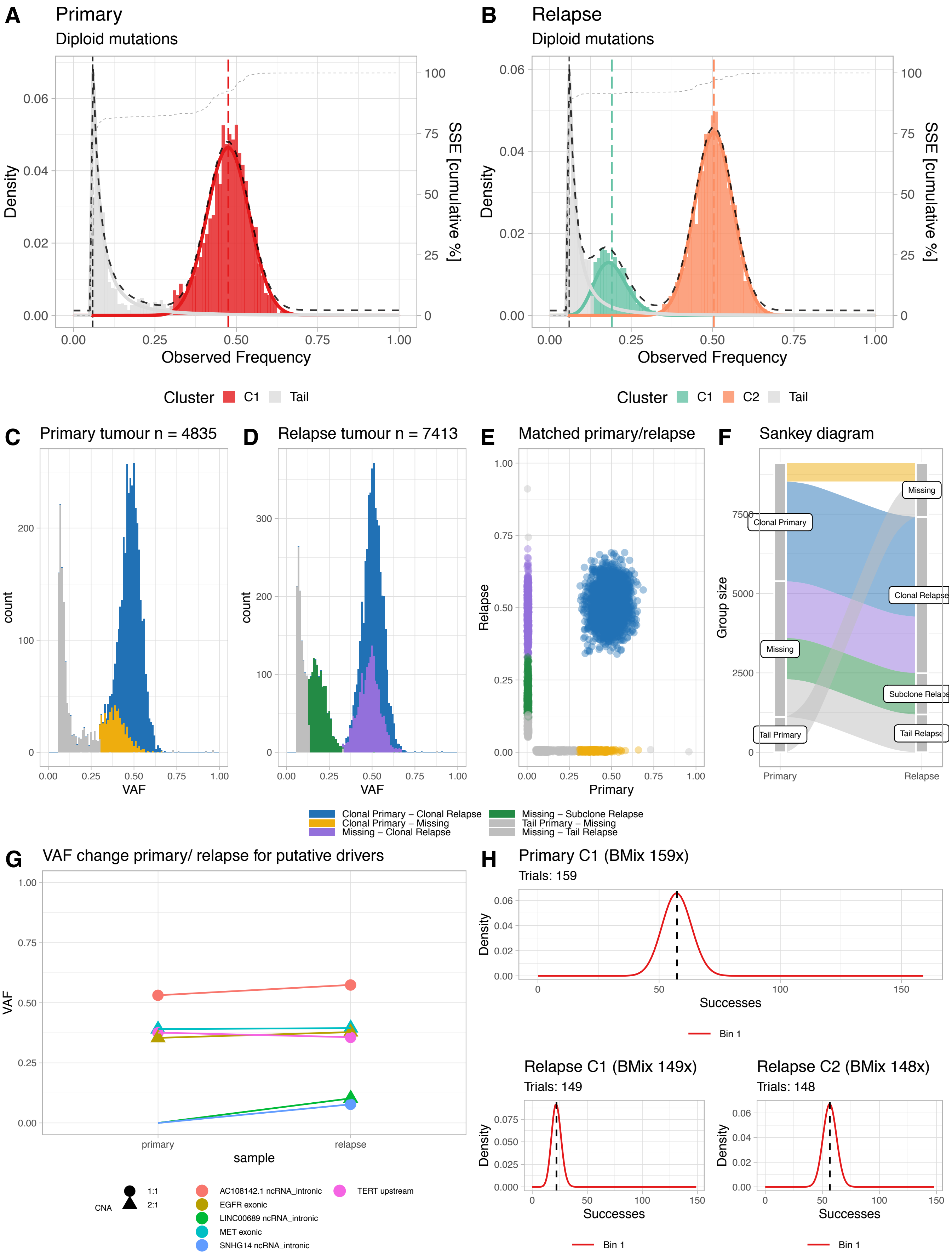
H043–B7R7

Groups with at least 50 SNVs.



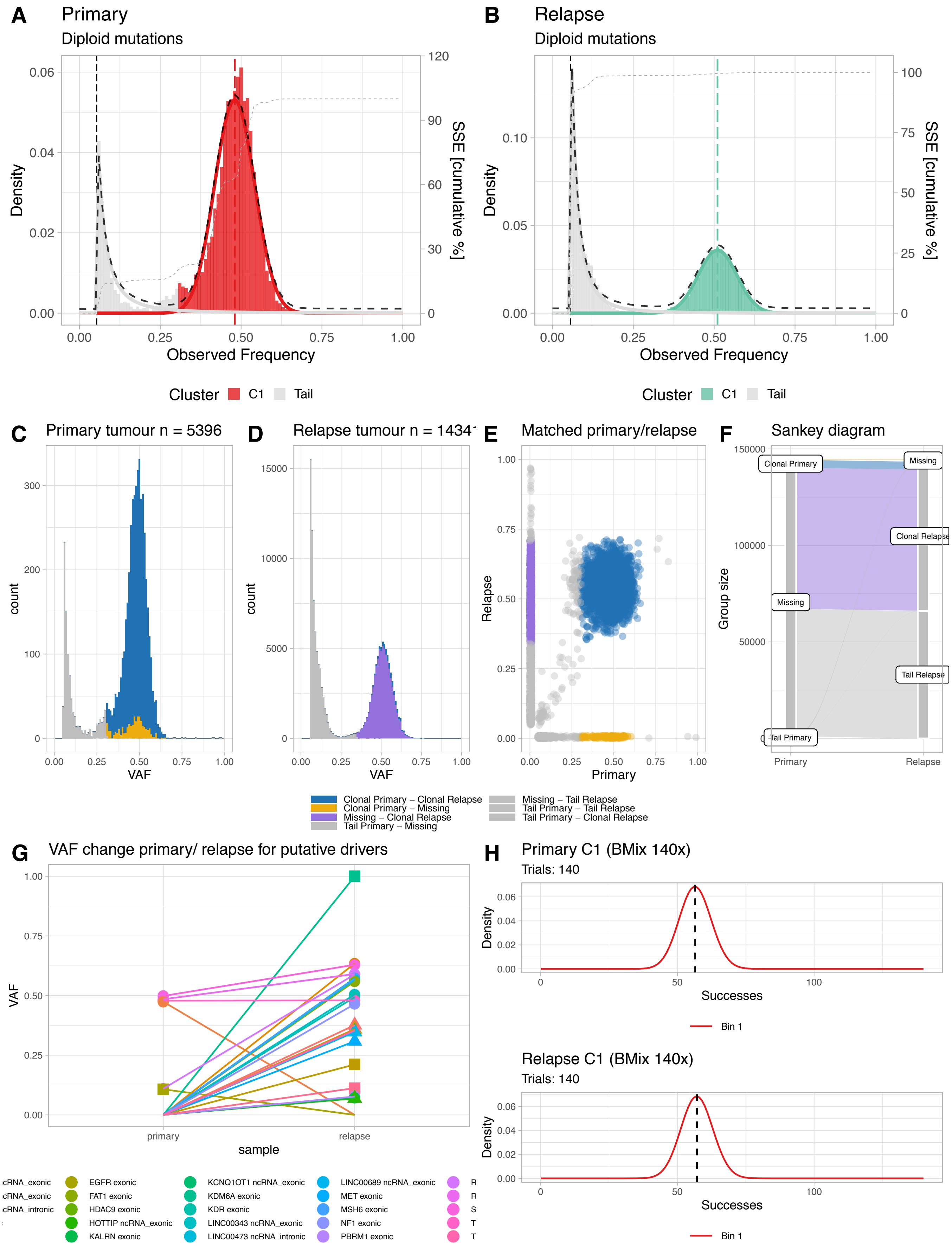
H043–BU96

Groups with at least 50 SNVs.



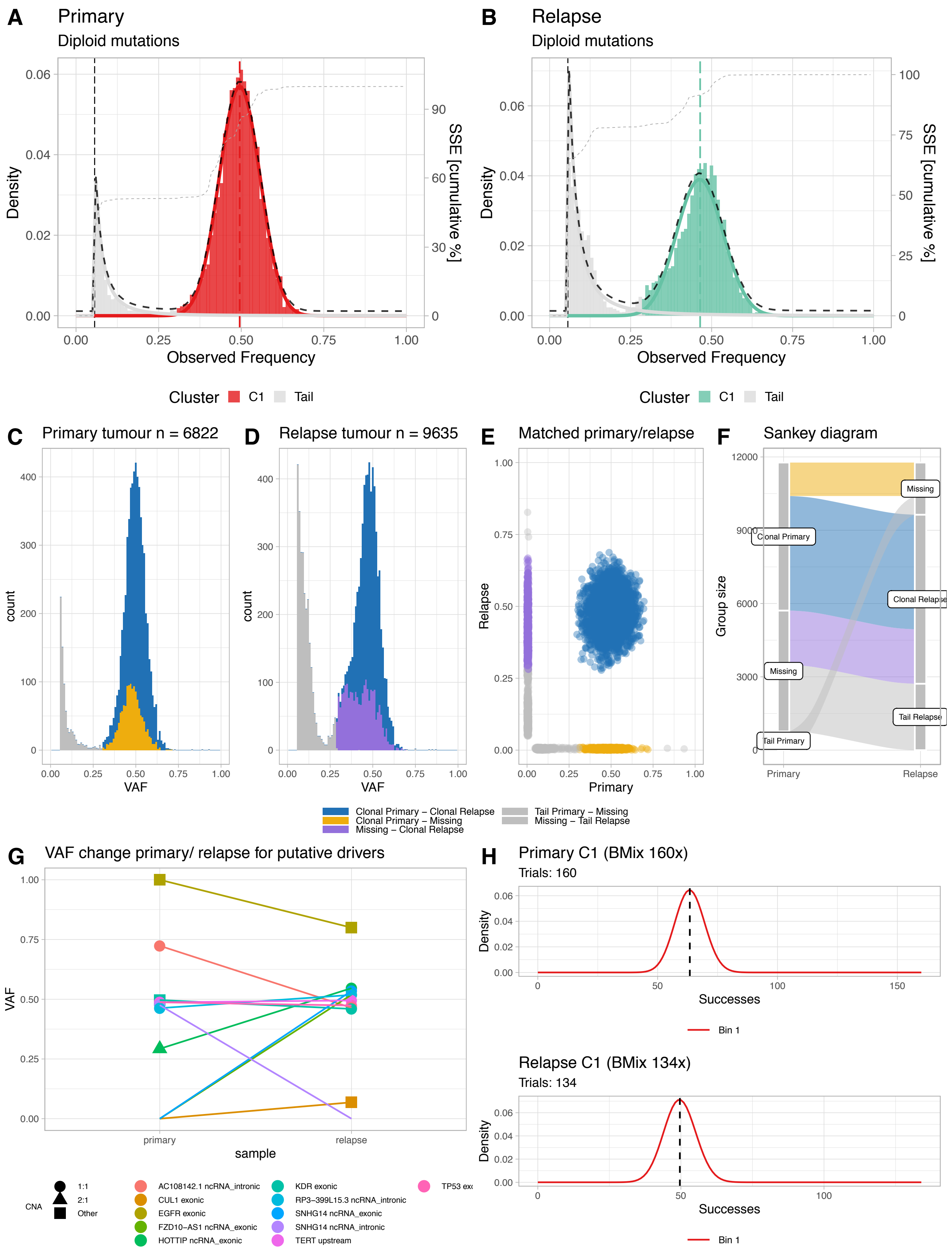
H043–D9MRCY

Groups with at least 50 SNVs.



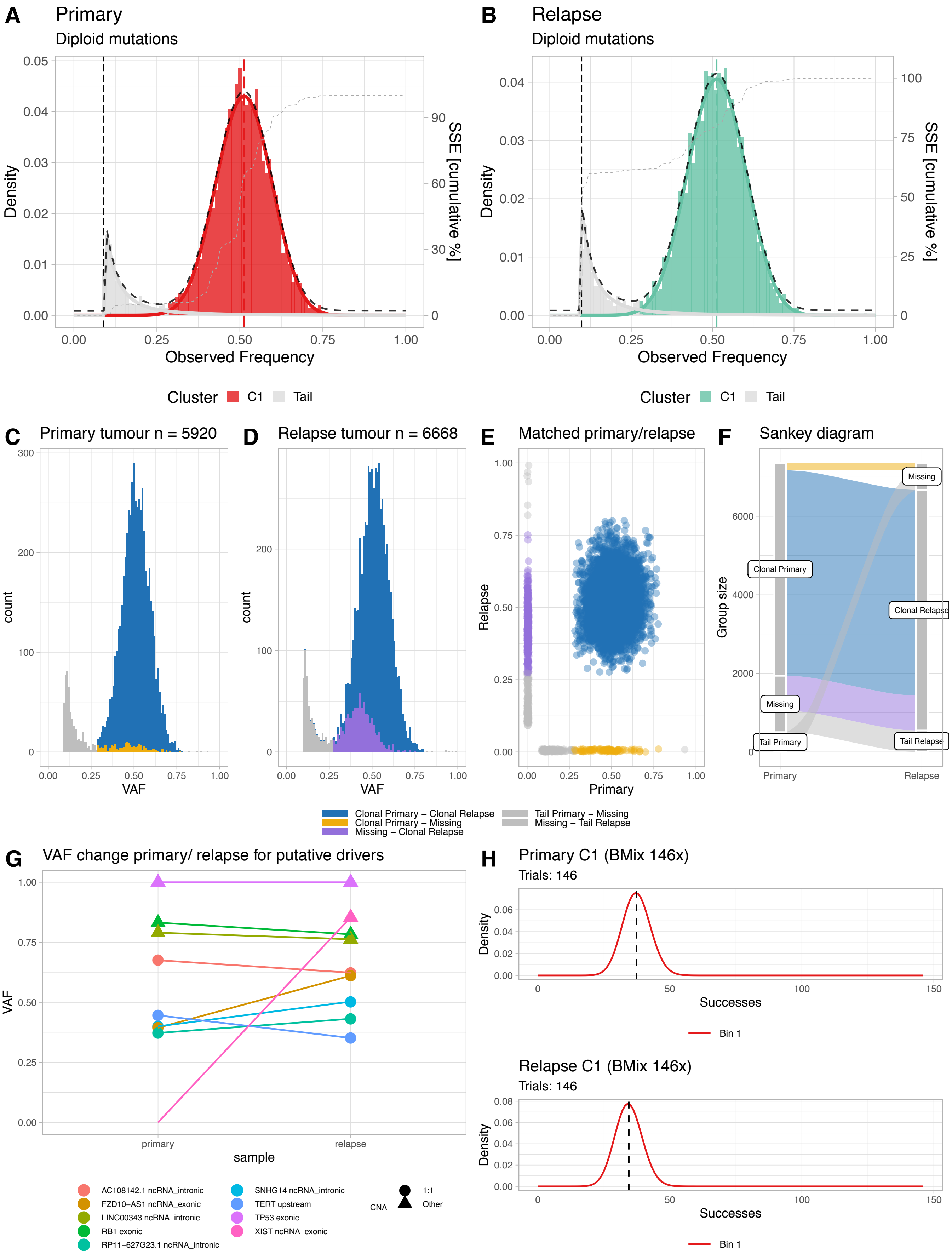
H043–DSX2

Groups with at least 50 SNVs.



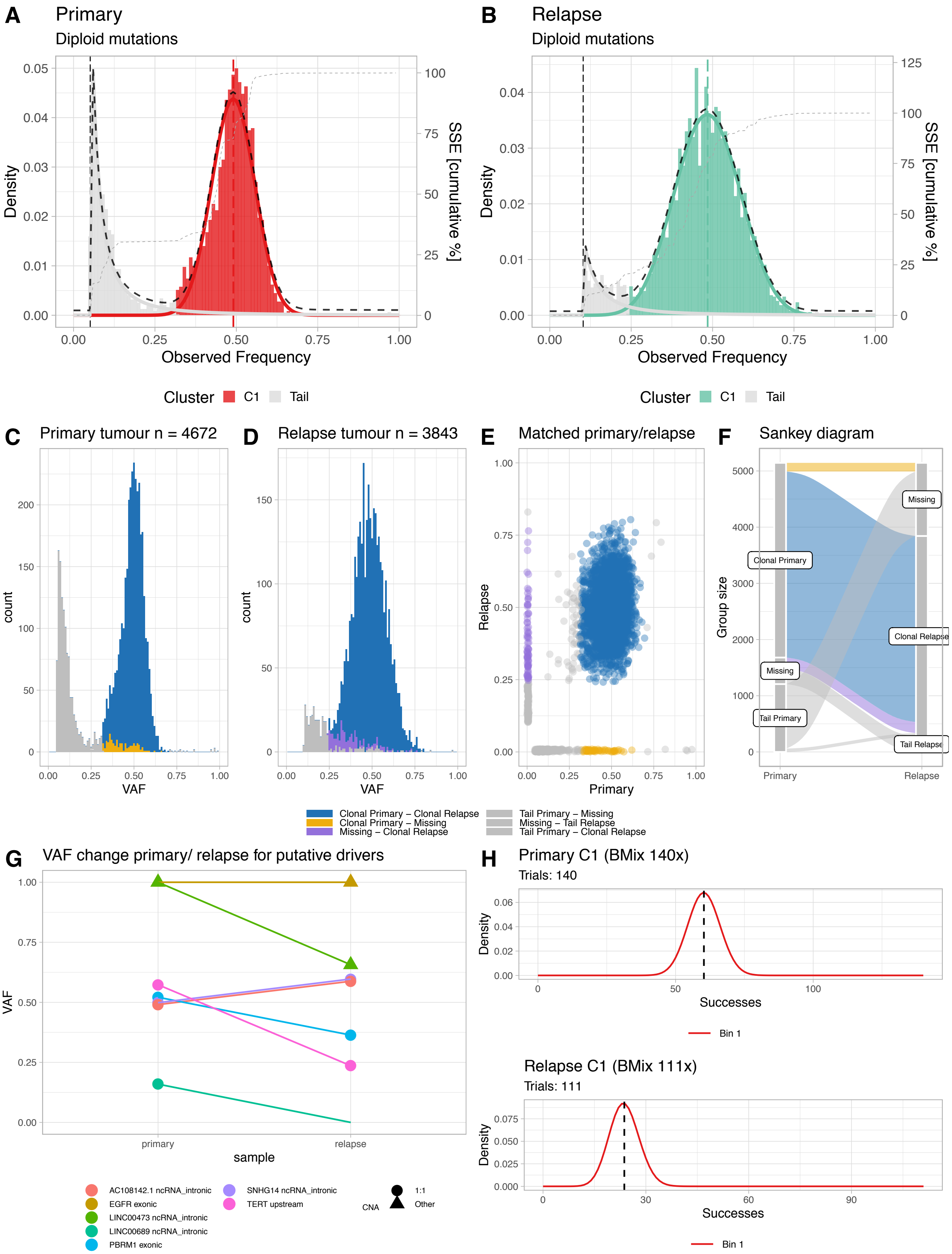
H043–GESMJV

Groups with at least 50 SNVs.



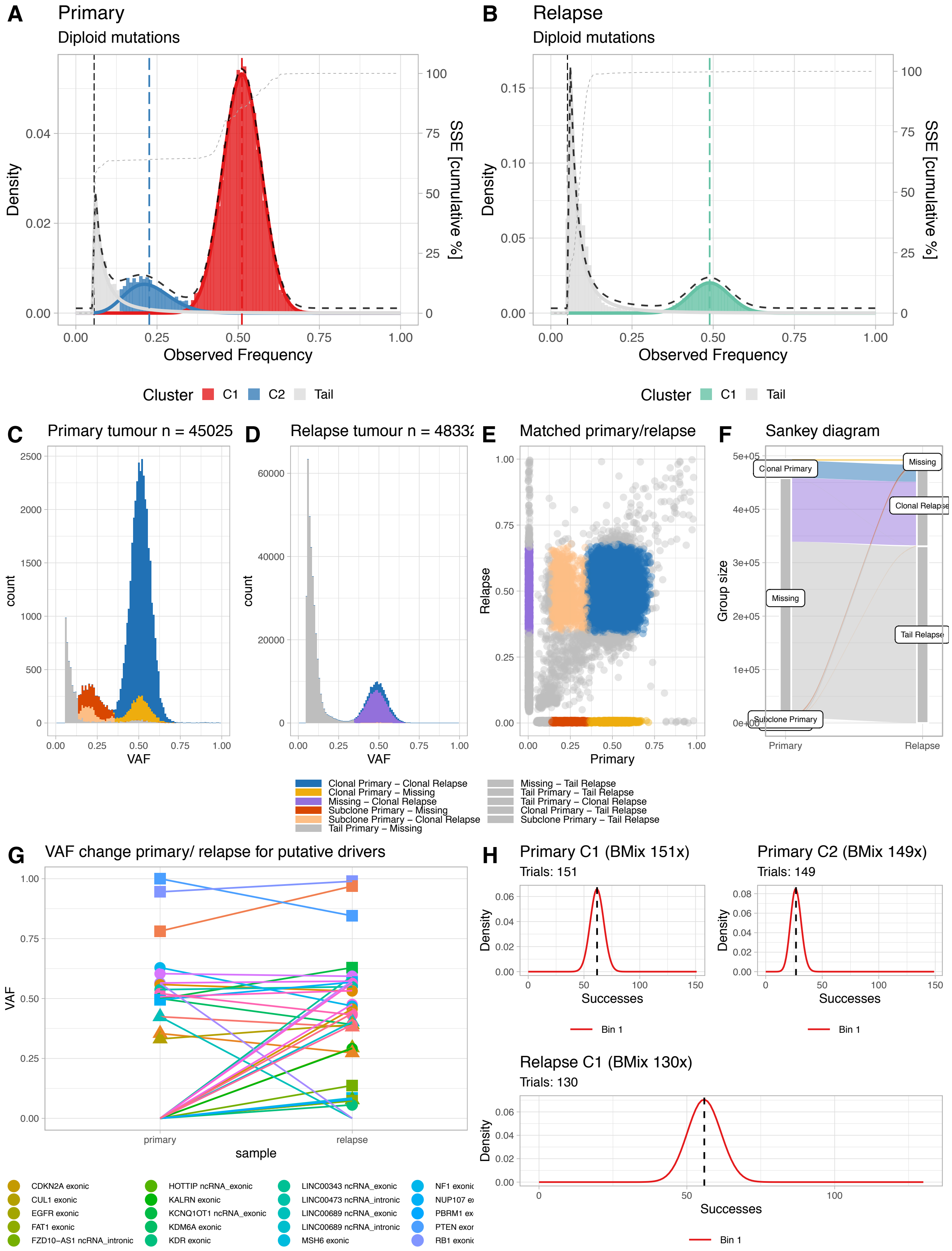
H043–GKS176

Groups with at least 50 SNVs.



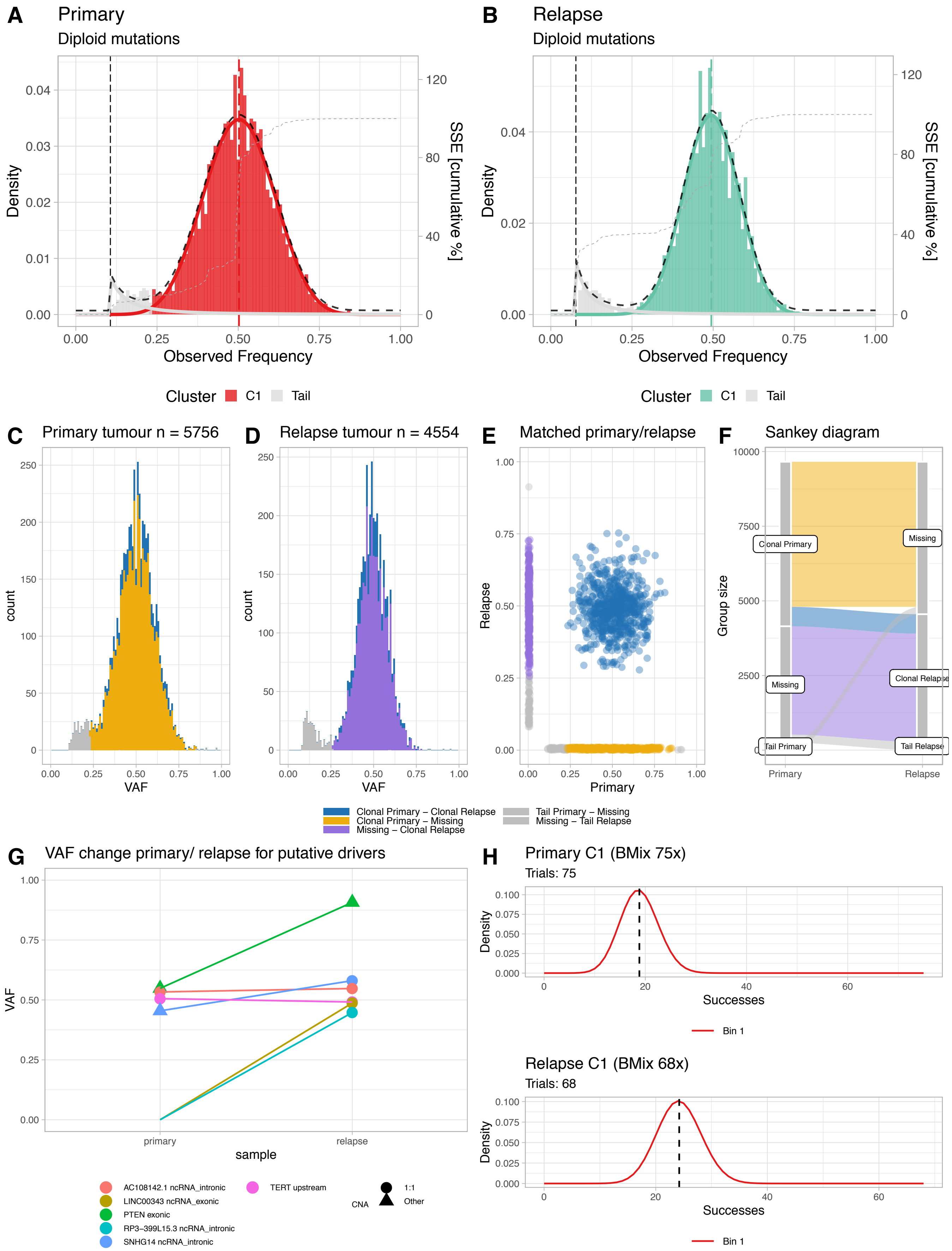
H043–KZWS

Groups with at least 50 SNVs.



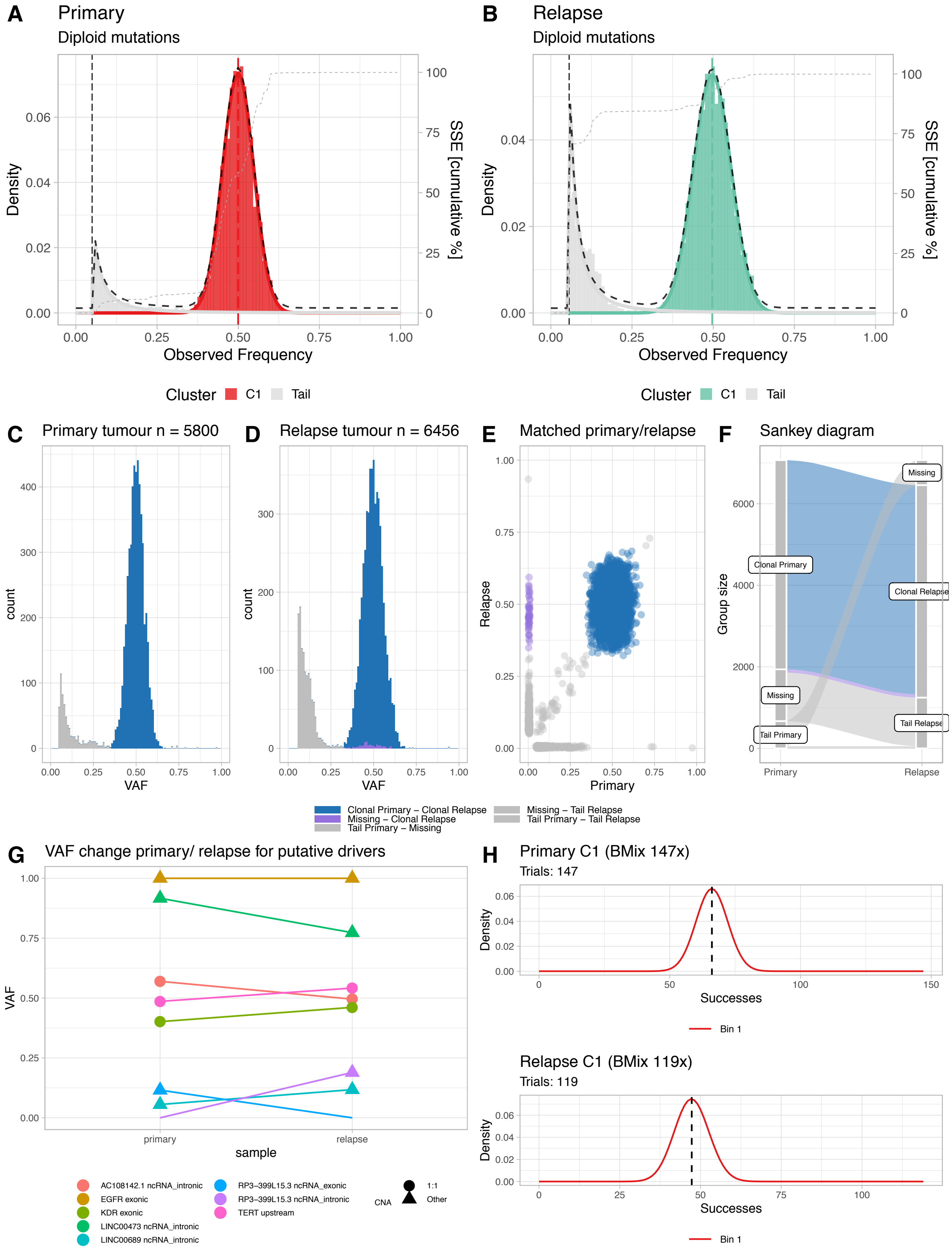
H043–LNWEGT

Groups with at least 50 SNVs.



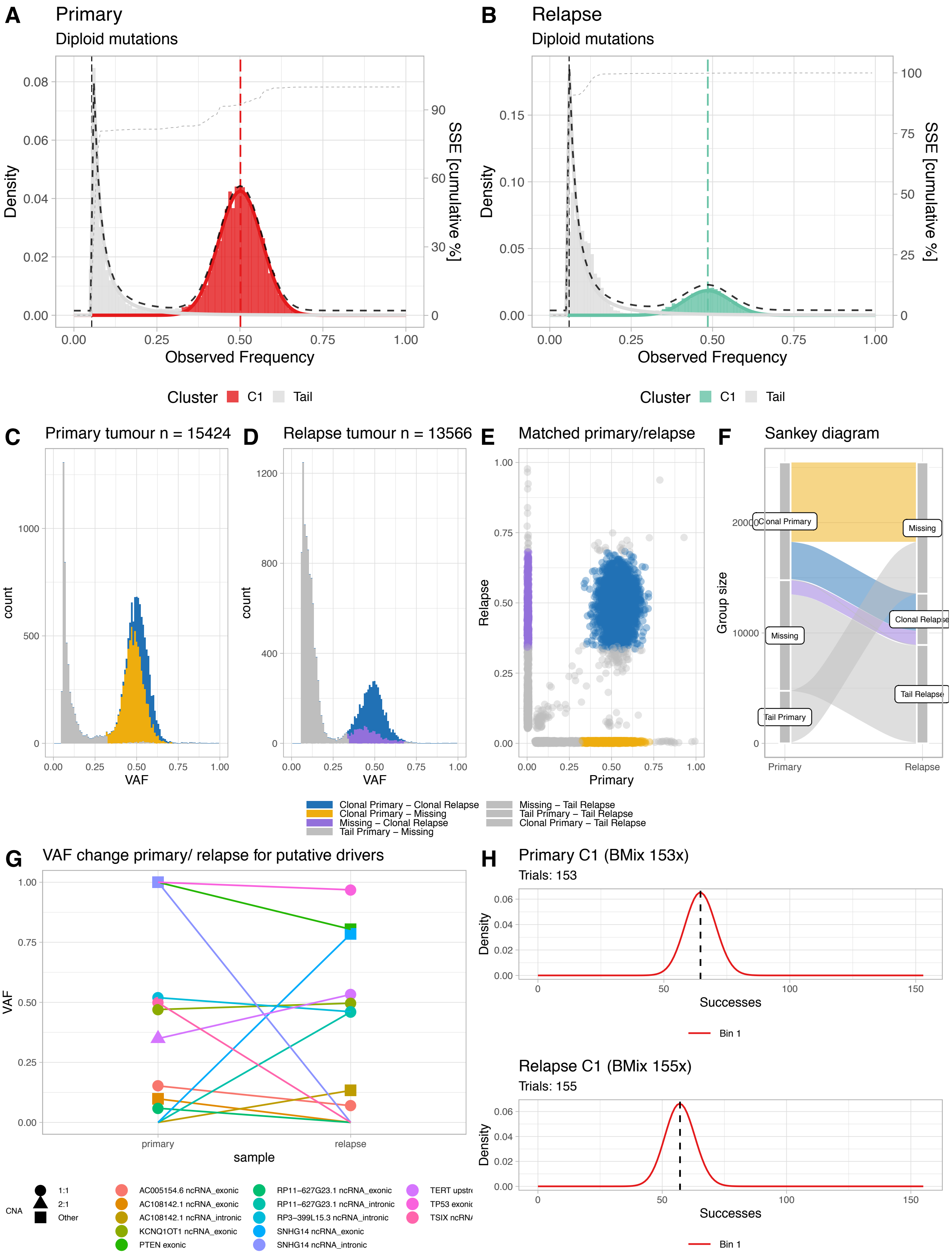
H043–N7LCPV

Groups with at least 50 SNVs.



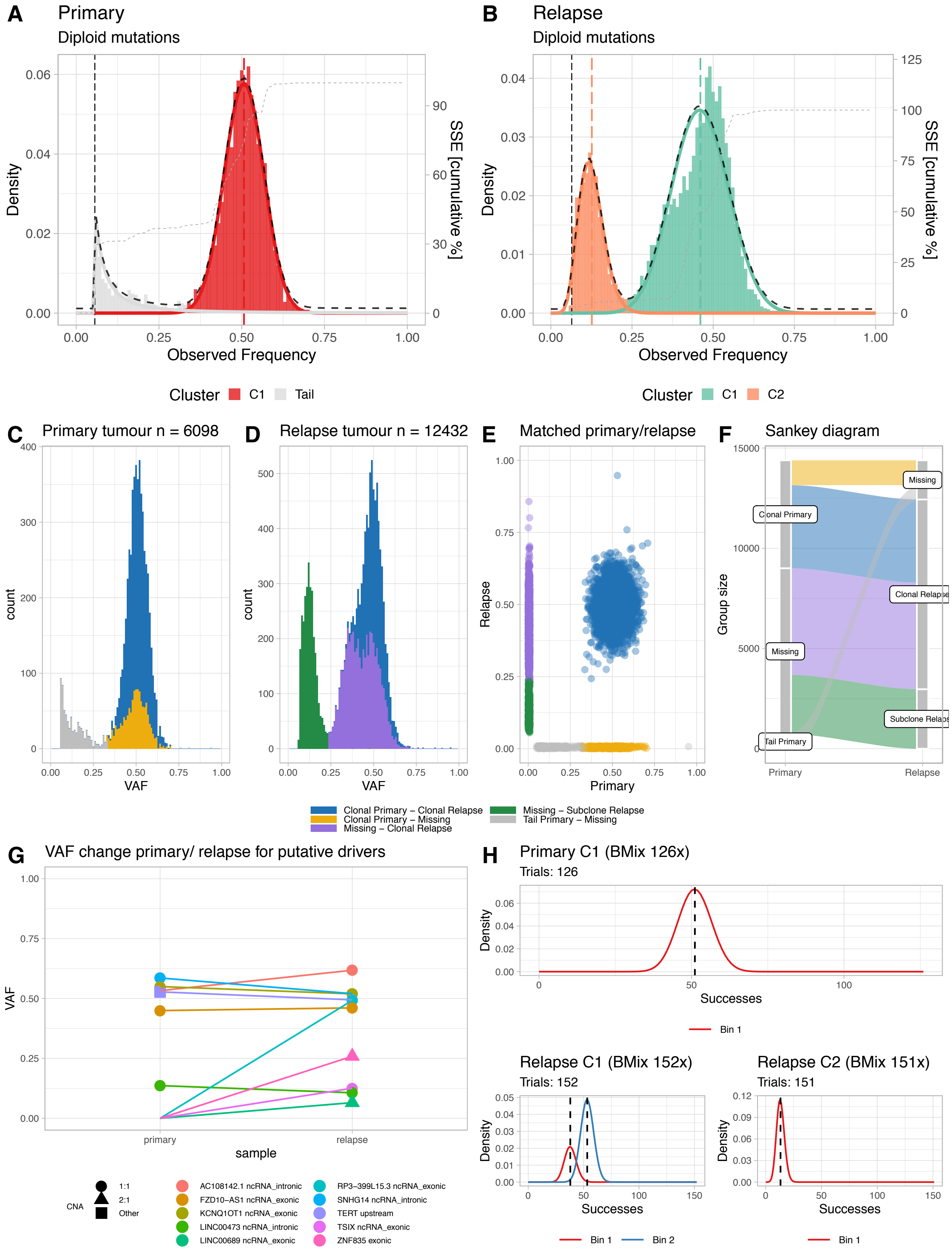
H043–NAFCCV

Groups with at least 50 SNVs.



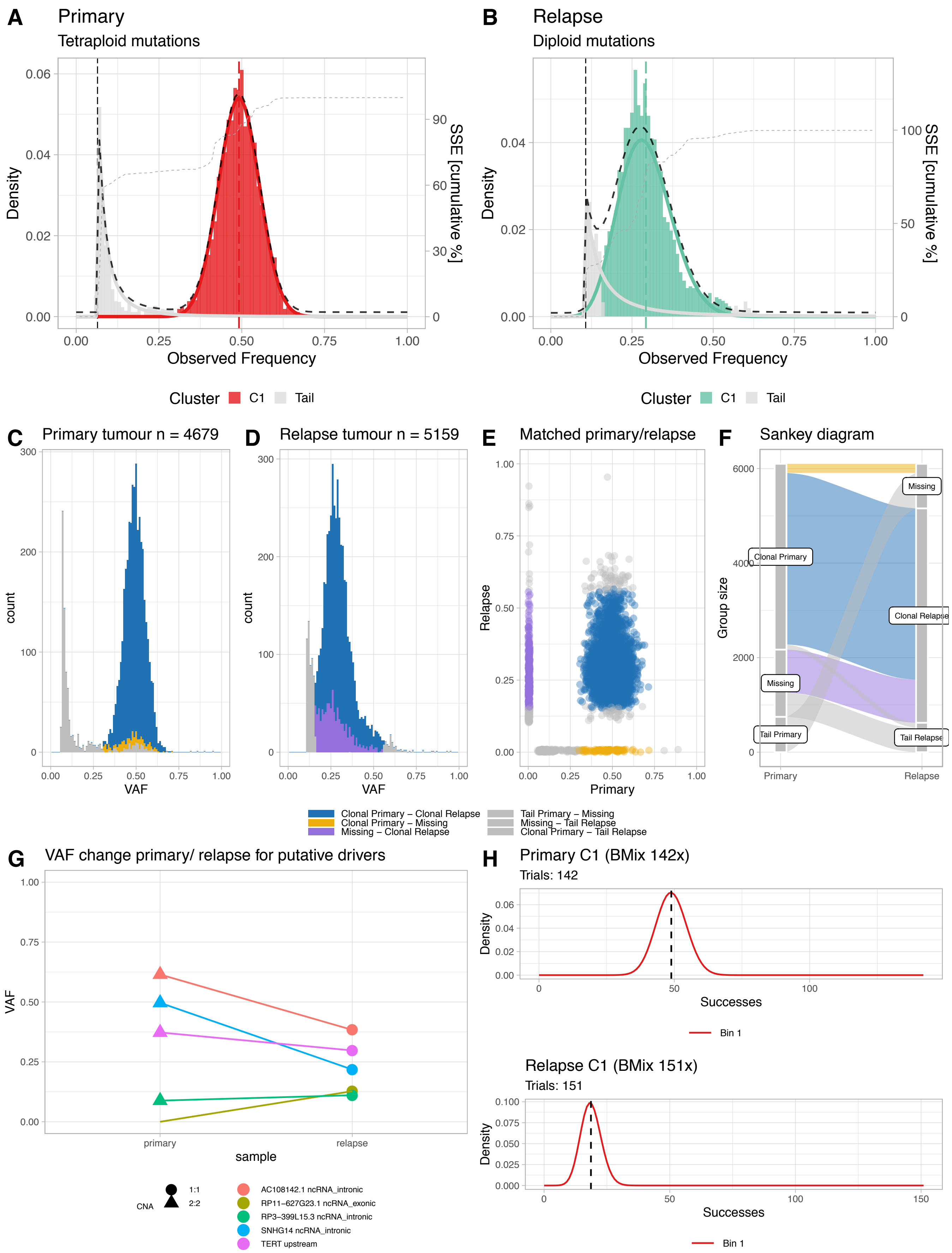
H043–PWC258

Groups with at least 50 SNVs.



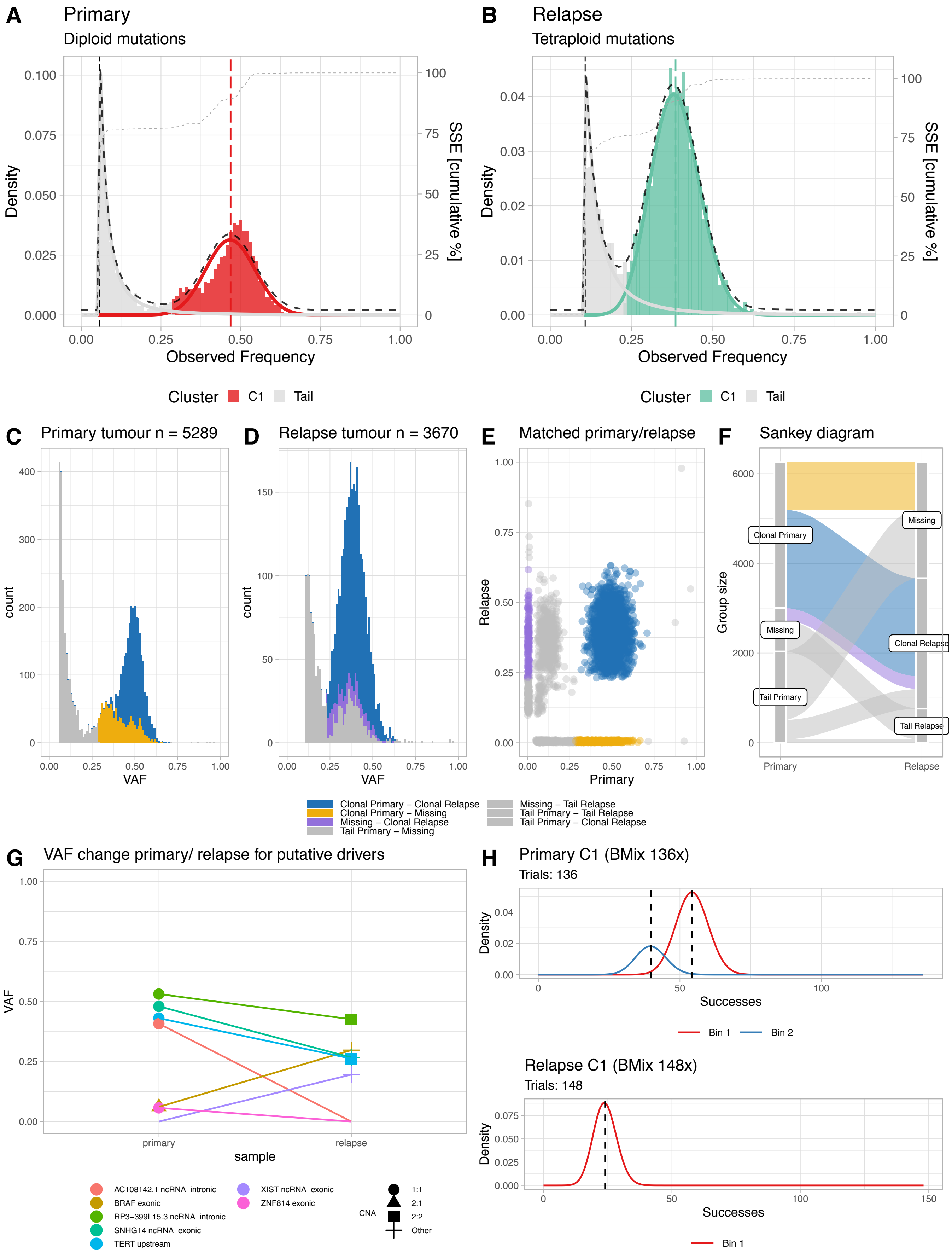
H043–QHGXQQ

Groups with at least 50 SNVs.



H043–4PGF

Groups with at least 50 SNVs.



Additional data table 1 (Supplementary_Table_1)

Summary fits for all the samples analysed in this manuscript (multi-sheet Excel).

Additional data table 2 (Source_Data)

MOBSTER preliminary package version (updated version at the GitHub repository https://caravagn.github.io/mobster/); vignettes to replicate the analyses of the paper, and WGS data of colorectal multi-region samples.

References

1. Heide T, et al. (2018) Reply to 'neutral tumor evolution?'. *Nature genetics* 50(12):1633–1637.
2. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* 81(25):2340–2361.
3. Kessler DA, Levine H (2013) Large population solution of the stochastic luria–delbrück evolution model. *Proceedings of the National Academy of Sciences* 110(29):11682–11687.
4. Roth A, et al. (2014) Pyclone: statistical inference of clonal population structure in cancer. *Nature Methods* 11(4):396–398.
5. Nik-Zainal S, et al. (2012) The life history of 21 breast cancers. *Cell* 149(5):994–1007.
6. Deshwar AG, et al. (2015) Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology* 16(1):1.
7. Miller CA, et al. (2014) Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS computational biology* 10(8):e1003665.
8. Bishop CM (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. (Springer-Verlag, Berlin, Heidelberg).
9. Teh YW (2011) Dirichlet process in *Encyclopedia of machine learning*. (Springer), pp. 280–287.
10. Jara A, Hanson TE, Quintana FA, Müller P, Rosner GL (2011) Dppackage: Bayesian semi-and nonparametric modeling in r. *Journal of statistical software* 40(5):1.
11. Williams M, et al. (2018) Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics* 50(6):895.
12. Fusco D, Gralka M, Kayser J, Anderson A, Hallatschek O (2016) Excess of mutational jackpot events in expanding populations revealed by spatial luria–delbrück experiments. *Nature communications* 7:12760.
13. Gundem G, et al. (2015) The evolutionary history of lethal metastatic prostate cancer. *Nature* 520(7547):353.
14. Yates LR, et al. (2015) Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine* 21(7):751–759.
15. Chkhaidze K, et al. (2019) Spatially constrained tumor growth affects the patterns of clonal selection and neutral drift in cancer genomic data.
16. Caravagna G, et al. (2018) Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature Methods* 15(9):707.
17. Kessler DA, Levine H (2015) Scaling solution in the large population limit of the general asymmetric stochastic luria–delbrück evolution process. *Journal of statistical physics* 158(4):783–805.
18. Griffith M, et al. (2015) Optimizing cancer genome sequencing and analysis. *Cell systems* 1(3):210–223.
19. Lee JJK, et al. (2019) Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. *Cell* 177(7):1842–1857.
20. Cross W, et al. (2018) The evolutionary landscape of colorectal tumorigenesis. *Submitted*.
21. Martincorena I, et al. (2017) Universal patterns of selection in cancer and somatic tissues. *Cell* 171(5):1029–1041.
22. Lakatos E, et al. (2019) Evolutionary dynamics of neoantigens in growing tumours. *BioRxiv* p. 536433.
23. Zapata L, et al. (2018) Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome biology* 19(1):67.
24. Dentro SC, et al. (2018) Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *BioRxiv* p. 312041.
25. Körber V, et al. (2019) Evolutionary trajectories of idhwt glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer cell* 35(4):692–704.