PERSPECTIVE
https://doi.org/10.1038/s41592-020-01013-2

nature methods

Check for updates

A practical guide to cancer subclonal reconstruction from DNA sequencing

Maxime Tarabichi[1,2,19], Adriana Salcedo[3,4,5,6,7,19], Amit G. Deshwar[8,19], Máire Ni Leathlobhair[9,10,19], Jeff Wintersinger[11], David C. Wedge[9,12,13,20], Peter Van Loo[1,20], Quaid D. Morris[7,11,14,15,16,20] and Paul C. Boutros[3,4,5,6,14,17,18,20] ✉

Subclonal reconstruction from bulk tumor DNA sequencing has become a pillar of cancer evolution studies, providing insight into the clonality and relative ordering of mutations and mutational processes. We provide an outline of the complex computational approaches used for subclonal reconstruction from single and multiple tumor samples. We identify the underlying assumptions and uncertainties in each step and suggest best practices for analysis and quality assessment. This guide provides a pragmatic resource for the growing user community of subclonal reconstruction methods.

Cancers evolve from a single cell through the sequential acquisition of somatic mutations, some of which enable the hallmark traits of cancer[1,2]. The descendants of this cell, which share its genotype, form the initial cancer clone. Selection, mutation, drift and spatial separation of clonal populations may then give rise to related but genetically distinguishable descendant subpopulations within a single tumor. These subclones can be evaluated from DNA sequencing studies, which have started to quantify key aspects of tumor development, such as metastatic seeding patterns[3,4] and mutations present in all tumor cells (that is, clonal mutations) that may be targets for treatment and early intervention[5,6]. Tumor heterogeneity has important clinical consequences: tumors with complex subclonal structures can be more aggressive[7,8] and are more likely to develop drug resistance and metastases[9].

The process of subclonal reconstruction involves three key aspects. First, it characterizes the major populations of cells in a given tumor by identifying the somatic mutations present in each one. Second, it quantifies the proportion of cells from each clone in the tumor (its cellular prevalence; Box 1). Third, it reconstructs the phylogenetic path by which the different clones evolved from their common ancestor, and ultimately from a normal host cell. Subclonal reconstruction can be performed from DNA sequencing data for a single tumor sample or from multiple samples collected over time and/or space. The DNA sequencing data themselves can be generated via a variety of sequencing strategies[10]. Thus, the accuracy and resolution of each feature of the subclonal reconstruction is shaped by the experimental design and the mutational characteristics of the specific tumor being reconstructed.

We focus here on computational methods for subclonal reconstruction using bulk DNA sequencing data, which remains the most widely used approach, although single-cell techniques continue to rapidly improve in quality and cost. We first outline the fundamental principles of subclonal reconstruction from a single heterogeneous tumor sample and then extend them to subclonal reconstruction from multiple samples. We next review the key approaches used for subclonal reconstruction, along with their limitations. Finally, we close with some perspectives on how the field may move and summarize our recommendations for subclonal reconstruction in practice (Table 1), providing a step-by-step practical guide with an example case (Supplementary Note).

Overview of subclonal reconstruction

Mutations belonging to the initiating cell of the most recent clonal sweep are expected to occur in every cell in the tumor. We refer to these as 'clonal' and distinguish them from 'subclonal' mutations, which arise in descendant subpopulations. We assume familiarity with some key technical terms in cancer genomics (Box 1).

In the standard workflow for single sample subclonal reconstruction, most subclonal reconstruction methods consider single nucleotide variants (SNVs), small indels, and larger copy number alterations (CNAs; Fig. 1a and Box 1). They use the variant allele frequency (VAF) of SNVs to infer the proportion of sampled cells bearing the SNV (cellular prevalence, or CP; Box 1). They do so by using the read structure (Fig. 1b) to first reconstruct copy number state, learning regions of clonal and subclonal copy number change (Fig. 1b,c). Algorithms then group SNVs with similar cellular prevalences, assuming that these occurred within a single distinct clone.

[1]The Francis Crick Institute, London, UK. [2]Wellcome Sanger Institute, Hinxton, UK. [3]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [4]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA. [5]Jonsson Comprehensive Cancer Center, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [6]Institute for Precision Health, University of California, Los Angeles, Los Angeles, CA, USA. [7]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [8]The Edward S. Rogers Sr. Department of Electrical & Computer Engineering, University of Toronto, Toronto, Ontario, Canada. [9]Big Data Institute, University of Oxford, Oxford, UK. [10]Ludwig Institute for Cancer Research, University of Oxford, Oxford, UK. [11]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [12]Oxford NIHR Biomedical Research Centre, Oxford, UK. [13]Manchester Cancer Research Centre, University of Manchester, Manchester, UK. [14]Vector Institute, Toronto, Ontario, Canada. [15]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [16]Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [17]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada. [18]Department of Urology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [19]These authors contributed equally: Maxime Tarabichi, Adriana Salcedo, Amit G. Deshwar, Máire Ni Leathlobhair. [20]These authors jointly directed this work: David C. Wedge, Peter Van Loo, Quaid D. Morris, Paul C. Boutros. ✉e-mail: PBoutros@mednet.ucla.edu

Box 1 | Lexicon

Branching clones. A nonlinear set of clones descending from a common ancestor (for example, sibling or cousin clones).

Cancer cell fraction (CCF). The fraction of cancer cells from the sequenced sample carrying a set of SNVs, that is, CCF = CP/purity. It can be inferred from the VAF (f) given a sample purity (ρ), the local copy number (N_{T}) and the inferred multiplicity m of the mutations:

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

Cellular prevalence (CP). The fraction of all cells (both tumor and admixed normal cells) from the sequenced tissue carrying a set of SNVs.

Clonal mutation. Mutation present in all the tumor cells of a tumor sample or biopsy.

Clone. A lineage of cells descended from a common ancestor that inherited its genotype. Clones can be characterized by (i) the genotype of the most recent common ancestor of that lineage, which is the set of initial SNVs that will be carried by the descendant cells—that is, detected at the same CP or CCF—and (ii) the fraction of (cancer) cells carrying these SNVs—that is, the CP or CCF of the initial SNVs.

Crossing rule. When performing multisample or multiregion sequencing, when clone A and B are descendant of clone C and the CCF of clone A is higher than the CCF of clone B in one sample but the opposite is true in another sample, then clone A and B must be branching subclones. This rule stems from the more general rule that the shared subclones across samples must have arisen from the same phylogeny, which further constrains the possible phylogenetic relationships between subclones.

Illusion of clonality. A mutation that is clonal in the sequenced tumor sample but is not clonal in the whole tumor.

Infinite sites hypothesis. Hypothesis that the size of the genome can be approximated as infinity. Consequently, mutated positions are only mutated once and never revert to wild type. This approximation results from the observation that, given the large size of the genome, a set of mutations is unlikely to have happened twice during tumor evolution. The infinite sites hypothesis is likely occasionally violated for single nucleotide variants[92], but their frequencies remain very low when considering larger sets of SNVs spread along the genome, such as those making up the genotype of large subclones (see “pigeonhole principle”).

log*R*. Total copy number log ratio, which can be estimated from local normalized tumor to normal read depth $\log_2(R_i) = \log_2\left(\frac{T_i/\bar{T}}{N_i/\bar{N}}\right)$, where the log$R$ at position i is the log-ratio of two normalized depths, the total depth in the tumor or normal at that position (T_i or N_i, respectively) divided by the average depth across positions in the tumor or normal ($\bar{T}$ or $\bar{N}$, respectively).

Linear clones. A set of clones wherein one or more clones is an ancestor of another clone in the set (for example, parent–child clones).

Most recent common ancestor (MRCA). The MRCA is the most recent cell that spawned a set of cells. By extension, the MRCA also refers to the genotype of that ancestor cell. The MRCA of a given tumor is sometimes used to implicitly refer to the MRCA of all cells in a set of sequenced samples. Note that the MRCA of a tumor sample (or set of samples) is not necessarily the MRCA of the whole tumor, owing to the illusion of clonality.

Multiplicity of a mutation. The number of DNA copies bearing a mutation m, which can be estimated from the VAF f, sample purity ρ and total copy number of the region in the tumor cells (N_{T}) as $m = \frac{f}{\rho}(\rho N_{\mathrm{T}} + 2(1-\rho))$. In regions of clonal copy number, the multiplicity of a mutation is a strictly positive integer, so the most likely value can be obtained by rounding to the nearest nonzero integer: $m = \max\left(1, \mathrm{round}\left(\frac{f}{\rho}(\rho N_{\mathrm{T}} + 2(1-\rho))\right)\right)$, where “round” is a function that returns the nearest integer, or by performing probabilistic assignment to integer values. In genomic regions with subclonal copy number alterations, subclonal cell populations may have differing multiplicities. Further, subclonal copy number losses may cause mutations to be lost from some subclones, resulting in multiplicities of zero for these subclones.

Pigeonhole principle. In the context of subclonal reconstruction, the sum of CCFs of branching subclones should be less than the CCF of their parent clone. Indeed, if it was greater, this would mean that mutations have occurred independently in branching lineages. However, according to the infinite sites hypothesis, the same set of random mutations is unlikely to have happened twice independently. Therefore, the smaller subclone must be a descendant of the bigger subclone; that is, they are linear subclones, which is compatible with the infinite sites hypothesis.

Purity, sample purity or tumor purity (*ρ*). The purity is the fraction of cancer cells in the tumor sample. Thus, the cellular prevalence of clonal mutations is the purity. Consequently, the fraction of non-cancer cells in the tissue sample is 1 – ρ.

Subclonal mutation. Mutation that is present in a subset of tumor cells in a tumor sample or biopsy.

Subclone. A clone that is a descendant of the most recent common ancestor of the tumor sample—that is, with associated CCF < 1 in at least one region.

Superclonal cluster. An apparent clone with CCF > 1, usually indicative of germline contamination or purity estimation errors.

Subclonal reconstruction. The exercise of reconstituting the subclonal structure from sequencing data—that is, number of (sub)clones, size of subclones in terms of fraction of cancer cells, and genotype of the subclones, as well as their phylogenetic relationships.

Sufficiency of subclonality. A mutation that is subclonal in the sequenced tumor sample will be subclonal in the whole tumor.

Sum rule. See “pigeonhole principle.”

Variant allele fraction or frequency (VAF). The fraction of mutated reads for a given variant, which is a readout of the proportion of DNA mutated in the sequenced tissue.

Weak parsimony. The vast majority of the SNVs with detectable VAFs are associated with a small number of subclonal lineages.

Table 1 | Checklist of recommended best practices

Recommendation	Rationale
Sequence at high depth (>60×) biopsy samples with the highest pathological purity possible, ideally complemented with deep targeted sequencing of SNVs	Increasing read depth increases the limit of detection for minor subclones and the resolution of CCF estimation[22,23,54]. High purity ensures most of the reads come from the tumor cells, increasing NRPCC.
Ensure the number of SNVs called is sufficient for subclonal reconstruction	A low coding substitution rate can lead to insufficient data for accurate subclonal reconstruction in exome-based studies[23,89].
Sequence multiple regions from a single tumor	Single-region bulk sequencing systematically underestimates the number of subclones, and locally dominant subclones can be mistaken as clonal[13,21,90]. Multiregion sequencing also provides better subclone resolution and allows phylogeny inference.
Minimize germline variant contamination: • Sequence matched normal tissue, ideally from an unrelated tissue source (for example, blood) • Remove known germline variants • Combine multiple SNV detection algorithms • Remove SNVs in genomic regions where read mapping is difficult • Use a panel of normal samples	Germline contamination can lead to false-positive SNVs with high VAF that can be mislabeled as a cluster. Using a consensus call set can improve sensitivity and specificity of variant detection[23].
Call somatic variants with a highly sensitive algorithm	Increased algorithm sensitivity facilitates low-VAF SNV detection, improves clustering accuracy and better captures the degree of tumor heterogeneity. Highly sensitive detection algorithms can also improve the chances of detecting clinically relevant minor subclones[22,23,91]. However, users should be wary of false-positive SNVs, which are often seen at low VAF and may form a low-VAF cluster.
For CNA reconstructions, review solutions for incorrect CP and WGD estimation and adjust accordingly. Optimally, perform experimental ploidy validation.	CNA reconstructions must decide among multiple equally likely ploidy and purity solutions. Ideally, inform CNA calling with experimental ploidy estimates using fluorescence-activated cell sorting, image cytometry or fluorescence in situ hybridization.
Carry out orthogonal copy number estimation	Multiple copy number solutions are usually possible; estimating copy number from WES data can be especially challenging[89].
Perform reconstruction based on CNA + SNV using a method that incorporates a binomial or β-binomial noise model	Binomial and β-binomial noise models better capture the noise in read sampling for a given read depth and CP, improving SNV clustering accuracy.
If possible, use phasing or single-cell sequencing data to support inferred mutation ordering. Ideally, perform multisample sequencing.	An unambiguous phylogeny is not always possible on the basis of the crossing and sum rules only. Phasing or single-cell sequencing information can support or refute a proposed phylogeny[14,16,72]. Our preferred setup is multisample sequencing; intelligent designs combine high and low depth sequencing to minimize cost[18].
Validate subclonal SNVs of interest using high-depth targeted sequencing	SNVs detected in one sample may occur at very low VAFs in another, and high-depth targeted sequencing can detect these rare subclonal populations[17,23,91].

Knowing the proportion of the sampled cells that are cancerous (the sample's purity) allows one to cluster SNVs by the proportion of tumor cells bearing the mutation, called the cancer cell fraction (CCF; Fig. 1d). Some algorithms may then attempt to infer the evolutionary relationships among clones (their phylogeny) on the basis of cluster CCF and mutation co-occurrence, although we believe this is usually advisable only for multisample data.

Inferring CP or CCF from VAFs requires estimating allele-specific copy number, as CNAs drastically affect VAF interpretation (Fig. 1c,d and Box 1). CNAs can be inferred from sequencing data by comparing local read depth in tumor and reference samples (quantified as the $\log_2$ of the tumor and normal depth ratio; $\log R$). The allelic ratio in the tumor relative to the normal sample (quantified as the B-allele frequency; BAF) is also informative as CNAs can alter the allele counts of heterozygous single nucleotide polymorphisms (SNPs) (Fig. 1c). Using these metrics, CNA reconstruction algorithms estimate the purity and ploidy of the sample, identify genomic segments with copy number changes, infer allele-specific copy number, and attempt to distinguish between clonal and subclonal CNAs.

Each of these steps involves uncertainty and can introduce errors into subclonal reconstruction. Indeed, many other sources of error exist upstream of subclonal reconstruction (for example, sequencing, alignment, and variant detection). For example, low tumor purity, errors in sequence alignment and low sequence coverage in specific regions of either the tumor or matched reference normal genomes can all lead to germline variants being misclassified as somatic[11,12] (Fig. 1b). These errors can propagate uncertainty into the subclonal reconstruction results.

Study design and data collection for subclonal reconstruction

The resolution and accuracy of subclonal reconstruction are strongly influenced by how the input data are sampled. Single-sample sequencing, even at high depth, can underestimate the number of subclones, and subclonal populations or mutations can appear clonal (Fig. 2). This is called the illusion of clonality (Box 1)[9,13]. Multiregion sequencing can improve separation of subclones based on CCF differences across samples and thereby facilitate phylogenetic inference[13,14] (Fig. 2a). Increasing sequencing depth improves the precision of CCF estimates and ability to distinguish subclones with similar CCFs. Given these trade-offs, in general, sequencing more samples will improve subclonal reconstruction more than higher-depth sequencing, given a depth sufficient to accurately identify variants and resolve peaks in CCF space.

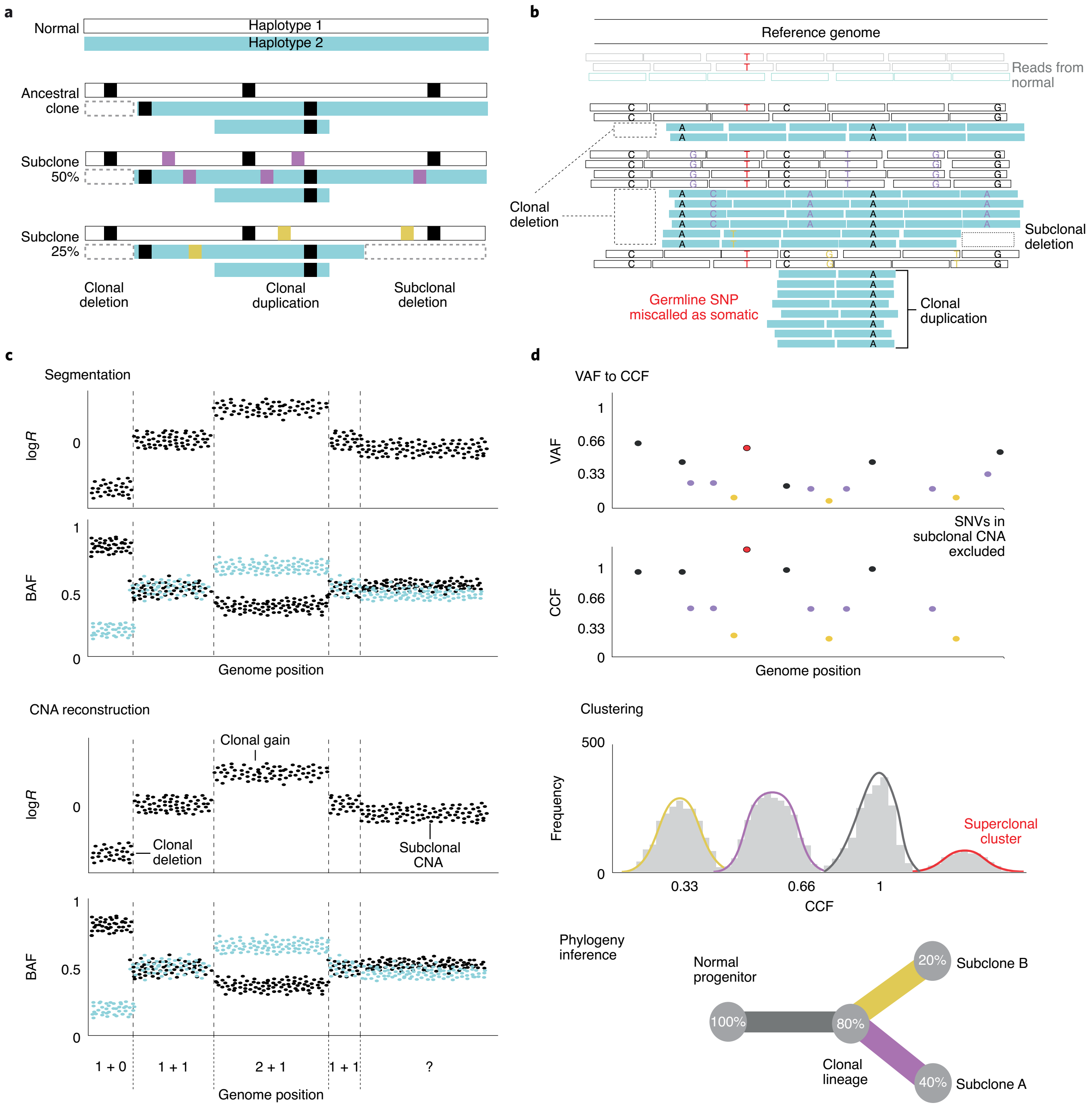
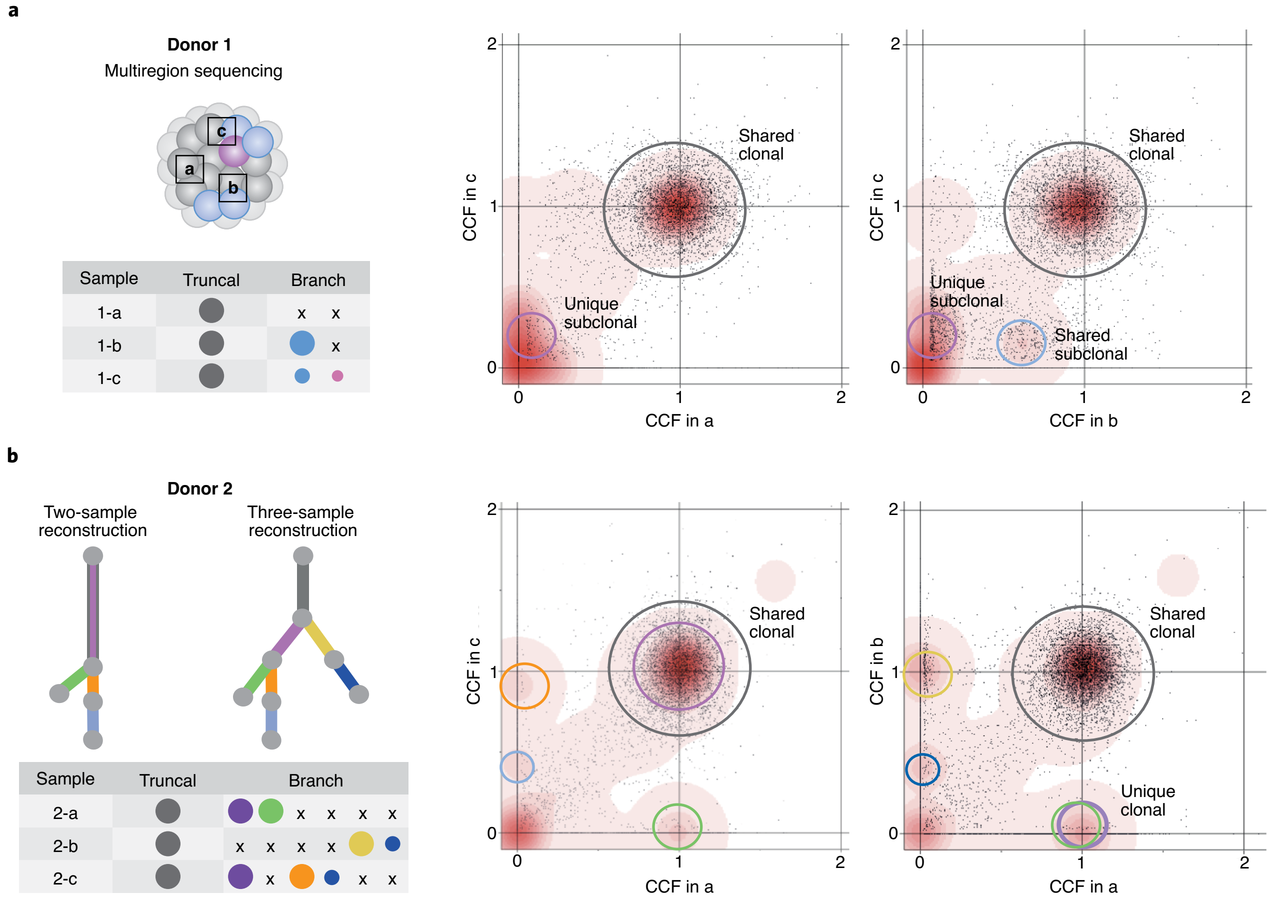


Fig. 1 | Standard workflow and input data for subclonal reconstruction. a, A simplified example of tumor clonal genotypes. We illustrate a tumor containing two subclones at 50% (purple) and 25% (yellow) CCF, both descended from a common ancestral clone (100% CCF, black). The remaining 25% of tumor cells are indistinguishable from the ancestor. **b**, First, somatic mutations are called from aligned reads. Read depth must be much higher (coverage >60×) than illustrated for mutation calling and subclonal reconstruction. Similarly, an elevated local mutation burden is illustrated. A somatic variant caller identifies somatic SNVs by comparing to a matched normal sample, although germline SNP contamination may occur (see main text). **c**, Second, CNA reconstruction is performed. It typically uses read depth and B-allele frequency (BAF) data for heterozygous SNPs. **d**, Third, CNAs are used to translate the measured SNV VAF to a CCF/CP estimate. This procedure relies on an accurate SNV multiplicity estimates (see Box 1); these are typically inaccurate in subclonal CNAs, so we exclude these regions from the analysis. SNV CCFs are then clustered to identify (sub)clonal lineages in the sample. False-positive SNVs or inaccurate CNAs can cause spurious superclonal clusters (that is, those with CCF > 1). Finally, phylogenetic reconstruction infers the ancestral relationships among lineages.

However, optimizing subclonal reconstruction is not typically the only or primary goal of a study. As a result, we recommend that study design match the biological questions being investigated, given technical and financial limitations. For example, if the patient number is large (for example, a clinical trial), single-sample analysis may be appropriate to maximize statistical power for clinical inference. Single-sample studies can put lower bounds on subclonal heterogeneity, as a mutation found subclonally in a single sample is



Sample	Truncal	Branch	
1-a	●	x	x
1-b	●	●	x
1-c	●	●	●

Sample	Truncal	Branch					
2-a	●	●	●	x	x	x	x
2-b	●	x	x	x	x	●	●
2-c	●	●	x	●	●	x	x

Fig. 2 | Subclonal reconstruction using multiple samples. a, Multiple samples can reveal additional subclones. Left: a tumor with three sequenced samples (a, b, c). The table shows clones in each sample with color-coded circles proportional to their CCF in size. Truncal is defined as CCF = 1 in all samples and branch as CCF < 1 in at least one sample. Right: two sample density plots for the tumor. SNV CCFs from each sample are plotted along the axes. Circles indicate clone clusters; red background shows SNV density. SNVs clustered around (1,1) occur in all tumor cells in both samples; subclones on the axes are sample-specific, and clusters off the axes appear subclonal in both samples. For example, a subclonal cluster occurs in ~15% of cells in c but is absent in a. However, region b shows that this cluster was a mixture of two subclones: one unique to c and one shared by b and c. **b**, Sequencing multiple samples clarifies clonal relationships. Left: phylogenetic trees for two- and three-sample subclonal reconstruction from multiregion sequencing. Subclones are represented by color-coded circles, as in **a**. Right: density plots, as in **a**. If we look only at samples a and c, mutations from the purple cluster appear clonal. However, this cluster is absent in sample b and thus subclonal.

sufficient to be deemed subclonal in the whole tumor. Thus, mutations subclonal in single-sample studies are potentially poor clinical targets (the sufficiency of subclonality; Box 1). Single-sample subclonal reconstruction in large cohorts can be useful for showing coarse trends in mutation timing, as clonality errors due to spatial heterogeneity are likely random, but detailed phylogenetic inference will likely be imprecise in these studies[8,15,16].

In other studies, evaluating clonal evolution over time may be critical—for example, in understanding metastatic processes. Samples taken at different times (for example, at diagnosis and relapse) permit inference of temporal features of tumor evolution. Samples taken from different spatial points (for example, different regions of the primary site or metastases) permit inference of lineage frequency changes that can hint at subclone fitness[17]. For example, an intelligent design to look at evolution across many metastatic sites captures variants through high-depth sequencing and follows them spatially through shallow-coverage sequencing[18]. Because of the illusion of clonality, multisample designs are superior when searching for clonal targets—for example, clonal neoantigens[19]. In general, sequencing even one more sample per tumor can help resolve more subclones (Fig. 2a) and clarify phylogenetic relationships (Fig. 2b), and more samples would further improve accuracy[20,21]. However, in practice there are many logistical and physical limitations to multisample sequencing: tumor size, tissue quality, availability of material for research, cost, feasibility of tissue collection in the given clinical setting, and attainable sequencing depth for each sample. As a result, while multisample subclonal reconstruction may be technologically superior, clinical and financial considerations can strongly constrain the possible sample number and space combinations.

Optimizing sequencing depth

When balancing sequencing depth and sample number, an important consideration is that overall sequencing coverage—as well as the specific sequencing technology used, which influences local read depth distribution—determines the sensitivity and specificity of clonal and subclonal SNV detection[22,23]. Both tumor ploidy and purity affect the depth of sequencing coverage needed to detect low-CCF SNVs[22,23].

One useful metric for evaluating sequencing depth for any given tumor is the number of reads per tumor chromosomal copy (NRPCC; Box 2). As NRPCC increases, the signal of true CCF

Box 2 | NRPCC

The number of reads per tumor chromosomal copy (NRPCC) can be defined as

$$\mathrm{NRPCC} = \frac{\rho}{\psi} d$$

where d is the depth of sequencing, ρ is the purity and ψ is the average tumor sample ploidy: that is, $\psi = \rho\psi_T + (1 - \rho)\psi_N$, where ψ_T is tumor ploidy and $\psi_N = 2$ is normal ploidy.

Consider a diploid tumor sample ($\psi_T = 2$) with purity $\rho = 0.5$. In this diploid context, because 50% of the cells are non-tumor cells, 50% of the reads would derive from non-tumor cells. If a mutation occurs on one of the two tumor copies, then ~25% of reads will carry it $\left(\frac{1\times 0.5}{2\times 0.5+2\times 0.5}\right)$. Next consider a $\rho = 0.5$ tumor that undergoes a WGD and becomes tetraploid ($\psi_T = 4$). In this case, two-thirds of the reads derive from tumor cells. Note that mutations that have happened after the WGD will be present on only one of the four tumor copies, and therefore only ~16.7% of reads will carry these clonal mutations $\left(\frac{1\times 0.5}{4\times 0.5+2\times 0.5}\right)$. Thus, the fraction of mutated reads of subclonal mutations in a tetraploid tumor is lower than in a diploid tumor.

One rule of thumb is that most variant detection algorithms will not identify a somatic SNV without at least three variant reads[93].

Let us imagine a tetraploid tumor $\psi_T = 4$, with purity $\rho = 0.7$; that is, $\psi = 0.7 \times 4 + 0.3 \times 2 = 3.4$. An SNV at CCF = 0.33 (present in a third of cancer cells) will be present in a fraction f of the reads quantified by

$$f = \frac{\rho}{\psi}\mathrm{CCF}$$

And the expected number of mutated reads will depend on the depth and is

$$N_{\mathrm{mut}} = f \times d$$

Or, in terms of NRPCC,

$$N_{\mathrm{mut}} = \mathrm{NRPCC} \times \mathrm{CCF}$$

This equation illustrates why the NRPCC is a relevant measure. It defines the expected number of mutated reads at given CCF values, and given that (as a rule of thumb) mutations with $N_{\mathrm{mut}} < 3$ are not being called, it defines the detection threshold, or sensitivity threshold—that is, the power to detect these subclonal mutations.

We model the number of mutated reads as following a binomial distribution:

$$N_{\mathrm{mut}} \sim \mathrm{Bin}(f, d)$$

If we want to select a minimum depth that allows us to detect most of these mutations, the probability of missing or calling them must be low or high, respectively. For example, to not miss more than 5% mutations in that subclone—that is, to have $N_{\mathrm{mut}} < 3$ with probability $P < 0.05$ ($N_{\mathrm{mut}} \geq 3$ with probability $P \geq 0.95$)—we have

$$\begin{aligned} & P(N_{\mathrm{mut}} < 3) < 0.05 \\ & \Rightarrow P\left(\mathrm{Bin}\left(f = \tfrac{0.7}{3.4}0.33, d\right) < 3\right) < 0.05 \\ & \Rightarrow d \geq 91 \end{aligned}$$

The depth of sequencing must thus be greater than 91×, which corresponds to NRPCC = 18.7. For clonal mutations—that is, CCF = 1—the depth must be greater than 29×.

peaks becomes clearer relative to read-sampling noise, and clones are easier to distinguish. In a large single-sample pan-cancer study, most samples with NRPCC > 10 exhibited at least one subclone[15]. An NRPCC of 10 represents a read depth of ~40× in a diploid 50% purity tumor. Some studies have successfully used large numbers of samples sequenced to moderate depth (30–50×), but generally deeper sequencing improves subclonal reconstruction accuracy and resolution[22,24]. Sequencing depth of the reference normal sample is also important to limit false-positive identification of germline variants as somatic ones. This is particularly true in studies sequencing multiple samples for each tumor and using a single reference sample as a control for all, as these false positives can enlarge the apparent number of clonal mutations. Current copy-number detection algorithms appear to be more robust to lower sequencing depths than current SNV detection methods[22,25]. Indeed, some CNA methods are already amenable to single-cell resolution[26], although few methods reconstruct phylogenies from CNAs[27–29].

Sequencing breadth

While in general higher-depth sequencing will improve subclonal reconstruction, it is also important to consider the portion of the genome directly measured—sequencing breadth. Subclonal reconstruction can be applied to whole-genome sequencing (WGS) or to targeted sequencing of genes, either of the whole exome (WES)[30] or subsets of selected genes[31]. The major differences between these approaches are the number of SNVs and indels detected, local depth and variability of coverage, and resolution and accuracy of CNA reconstruction. Moderate depth (30–50×) WGS supports high-quality CNA calls and allows estimation of copy-number CCFs for the characterization of subclonal CNAs. WES detects ~50-fold fewer SNVs and indels and decreases CNA reconstruction resolution and quality, thus hindering subclonal lineage detection[25,32]. Conversely, higher depth of sequencing usually allows better accuracy in CCF estimates, and WES may yield higher-resolution subclonal reconstruction than WGS in samples with many SNVs and indels and few CNAs, particularly if it permits use of multisample data by lowering cost per sample[9,33]. Targeted sequencing using smaller gene panels rarely supports meaningful subclonal reconstruction, unless an initial round of WGS or WES is performed to define SNVs representative of individual subpopulations (cells with nearly identical genotypes)[34]. If there are few genes in the panel, allele-specific CNA reconstruction will be unreliable unless more data (for example, from a SNP chip) are available.

We recommend attempting subclonal reconstruction only with fresh frozen tissue. Standard WGS of formalin-fixed, paraffin embedded (FFPE) samples must contend with variable DNA quality[35], and FFPE-derived artifacts can introduce CNA errors[36]. If FFPE samples must be used, protocols that optimize library preparation and sequencing for accurate SNV and CNA detection from FFPE samples are an active research area[37], but downstream subclonal reconstruction results will need to be interpreted with caution.

The computational workflow for subclonal reconstruction

CNA reconstructions: overview. Most CNA reconstructions use germline SNPs (Fig. 1c) and evaluate their read depth and allele frequencies. For each SNP, copy number state is inferred from the changes in the relative depth—that is, the log$R$—and an imbalance in

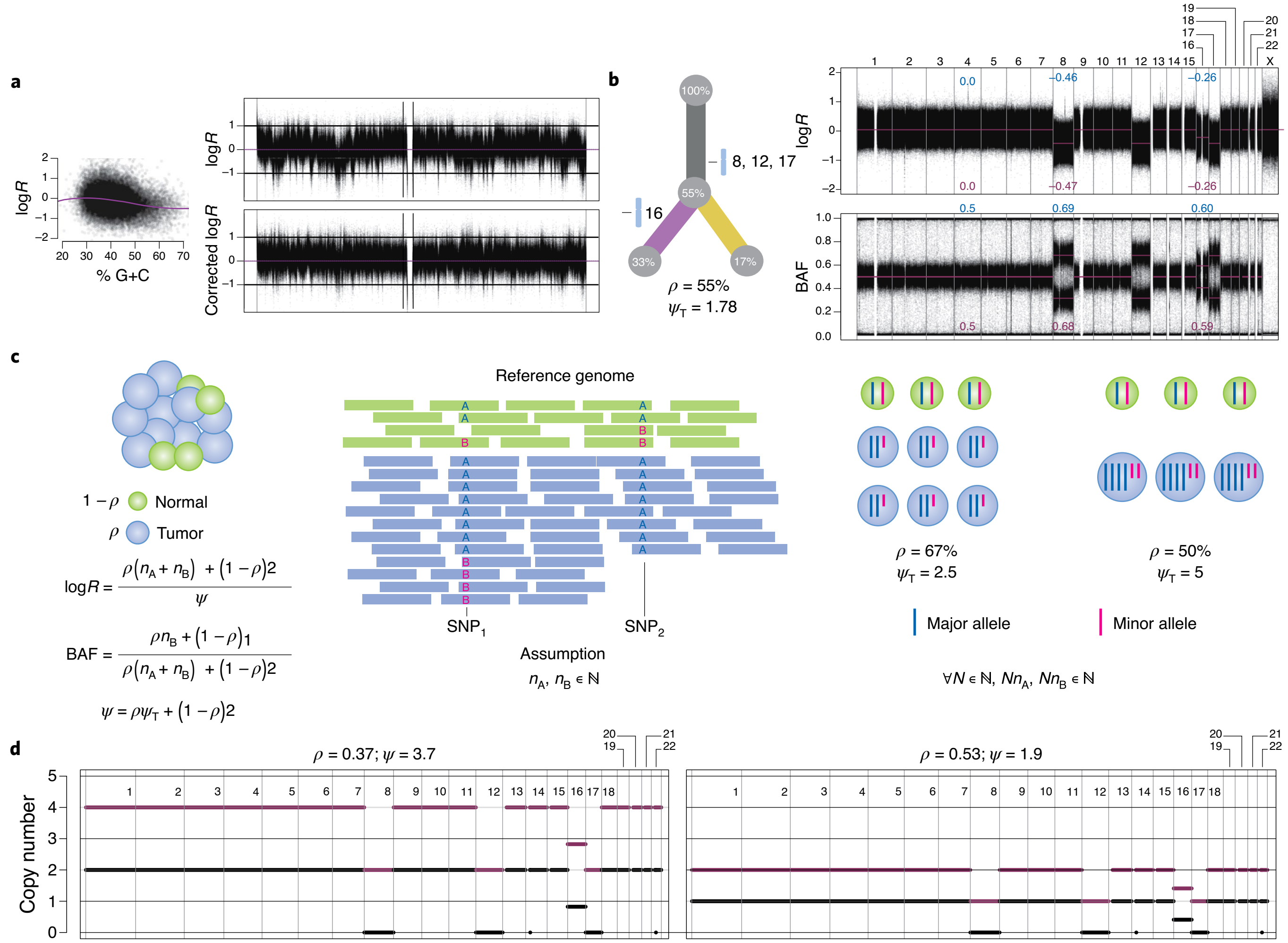


Fig. 3 | CNA reconstructions and uncertainty from whole-genome duplications. a, Effect of G+C content on log*R*. Left: the G+C content (percentage, in 500-kbp bins) around SNPs versus log*R* for a tumor from the Pan-Cancer Analysis of Whole Genomes project[16], with a LOESS (locally estimated scatterplot smoothing) fit (purple). Right: chromosome 22 log*R* before (top) and after G+C and replication timing correction (bottom). **b**, log*R* and BAF reflect relative allele-specific DNA content. Left: the subclonal structure for a tumor with clonal and subclonal chromosomal CNAs. Right: genome-wide log*R* and BAF with expected (blue) and measured (purple) values for CNAs[22]. **c**, Schematic illustration of ploidy ambiguity. The bulk sample contains tumor (blue) and non-tumor (green) cells. The number of reads from each allele from normal and the tumor cells depends on the number of allelic copies. We show a toy example with two heterozygous SNP positions (A and B alleles). log*R* and BAF can be expressed as a function of purity ρ, tumor ploidy ψ_T and the number of major and minor allele copies (n_A and n_B) in the tumor, which clonally should be integers. Combinations of purity and ploidy values that best align n_A and n_B to integers are often used to derive copy number profiles. However, multiple combinations can explain the observed data—namely, multiples of $2\psi_T$ (that is, values a WGD apart). In this example, $\psi_T = 2.5$ and $\psi_T = 2 \times 2.5 = 5$ both explain the data. **d**, Copy number profiles inferred by the Battenberg algorithm. Left: along the genome (*x* axis), copy number of the major (purple) and minor (black) allele (default fit, which favored a WGD solution because it fit the subclonal event on chromosome 16 near integers). Right: same as left but after manual refitting.

the number of maternal and paternal alleles—that is, the BAF—in the corresponding region. Larger allelic imbalances show up as horizontal bands when the SNP BAFs are plotted against genomic position. More subtle differences (typically subclonal CNAs) can be detected by phasing SNPs within a haplotype block using a large phased genome panel[38]. CNA reconstruction algorithms use the presence or absence of BAF and log*R* shifts to segment the genome into regions with constant copy number states. The log*R* is typically noisier than the BAF because it is influenced by local effects such as G+C content and replication timing, whereas BAF is relatively unaffected (Fig. 3a). Many algorithms correct log*R* for these types of covariates[39].

Copy-number calling algorithms make the relatively strong biological assumption that most copy number events should be clonal—that is, they allow one to interpret the log*R* and BAF as being generated by integer copy number values (Fig. 3b). Experimental ploidy validation has suggested that this assumption is satisfied in most cases (for example, breast cancers, ovarian cancers, cancer cell lines[40]) despite most reconstructions bearing at least one subclonal CNA[16]. Concordantly, emerging DNA single-cell sequencing datasets show that CNA-defined subclones present at a single time point differ by only a few genomic segments[26], showing that this assumption is reasonable in many cases.

Segmentation is a critical step in CNA reconstruction because it defines the boundaries of each region of constant copy number state. Segmentation can be done by identifying breakpoints where there is a change in the average read depth and/or BAF. Differences in segmentation lead to many reconstruction differences between methods[41]. Some methods iteratively join segments of fixed size[42]; others use change-point detection algorithms, commonly including circular binary segmentation[43], piecewise constant fitting[44] or hidden Markov models[28,38,45,46]. Structural variant breakpoints can also help inform segmentation[15,47,48].

Once segments are defined, the average major and minor allele copy number for each segment in the cancer's genome can be estimated from its log*R* and BAF, and purity and ploidy can be estimated from clonal CNAs[40] (Fig. 3c). CNA segments with integer or near-integer values are assumed to be clonal: that is, all cancer cells in a sample have the same copy number state in that region. Average copy number values that significantly differ from integers (fractional values) typically indicate subclonal CNAs[38].

CNA reconstruction then requires interpreting fractional average copy numbers as a mixture of whole number states and ascertaining the proportion of cells in each one. Solutions to this problem are intrinsically ambiguous: for any segment, it is always possible to posit a larger copy number and smaller CP that explain the BAF and log*R* equivalently well (Fig. 3c). Purity and ploidy estimates based on CNA reconstructions are ambiguous for the same reason. CNA detection algorithms usually assess reconstructions on the basis of multiple purity and ploidy estimates. Nevertheless, the final estimates may be incorrect and users should carefully evaluate solutions (Fig. 3d).

For clonal CNAs, ambiguity is partially resolved by requiring that the CP of all clonal CNAs must be the same (that is, equal to the purity). However, in experiments of moderate sequencing depth, estimates of exact copy number remain uncertain, as many solutions may be equally likely at any given purity. This can be particularly problematic for estimating copy number in highly amplified regions. There are three main ways to resolve this ambiguity for subclonal CNAs. Genome-wide methods attempt to group subclonal CNAs into subpopulations with the same CP[28,48]. They can correct for errors in individual segments, but are vulnerable to large-scale errors if they group subclonal CNAs into lineages incorrectly. Event-based methods, such as the Battenberg algorithm[38], apply a set of parsimony rules separately to each copy number segment. They make reconstruction errors in segments where their heuristic is wrong, but those errors are restricted to individual segments. Neither of these approaches can robustly infer more than two subclonal copy number states within a single region[41], since they depend on exactly two informative inputs (BAF and log*R*). The third approach assigns subclonal CNAs to subclonal lineages with defined CP—for example, through SNV clustering[46,49]. None of these methods fully resolves the subclonal copy number ambiguity, so we recommend users consider only using SNVs in regions of normal copy number or clonal copy number change for subclonal reconstruction.

CNA reconstructions: troubleshooting. Detecting CNA breakpoints can be challenging, especially in tumors with low effective depth (that is, low NRPCC). Missing a CNA breakpoint can lead to a series of segments with different clonal copy number states being called as a single segment; this segment may be miscalled as normal diploid or subclonal. Noise in sequencing data (for example, from library preparation artifacts) can lead to oversegmentation. Miscalled copy number states can also produce spurious clusters in the final subclonal reconstruction by distorting local CCF estimates. Methods to avoid overfitting include prioritizing BAF over log*R*[38], correcting for G+C content[38,40] and replication timing effects[50], using structural variants as breakpoints[15], and automatically adjusting segmentation parameters[38,40].

An intrinsic ambiguity in CNA reconstruction arises from whole-genome duplication (WGD). Any given CNA reconstruction is equivalent to another with each copy number doubled and purity lowered (Fig. 3c,d). To resolve this, CNA reconstruction methods often return multiple solutions or select tetraploid solutions only when there is positive evidence of them—for example, from odd (1, 3, 5, …) values for major or minor allele copy number states. Because some WGD uncertainty usually persists, we recommend using CNA reconstruction methods that allow the user to set the ploidy and rederive copy numbers and purity to facilitate user-driven assessment of different CNA reconstructions.

Several features in the data can help diagnose purity versus ploidy errors[41]. If the majority of CNAs appear subclonal or known early drivers in the tumor type being studied appear subclonal, the correct purity may be lower than inferred. Missed clonal WGD can sometimes be diagnosed by the presence of subclones with ~50% CCF that contain clonal SNVs acquired after WGD (mutations occurring on one of the four copies), while those occurring at 100% CCF actually represent mutations that occur before WGD and are hence present on two of the four copies. Observing a 50% CCF subclone in multiple samples would further support WGD, as subclones are unlikely to occur at the same CCF across samples otherwise. In the case of multisample reconstruction, purity can be estimated from CNA-adjusted VAFs of SNVs present in all regions[9]. Comparing purity estimates from CNA-only, SNV-only and SNV + CNA subclonal reconstructions can further guide purity inference and is especially useful in tumors with few clonal CNAs (for example, papillary thyroid carcinomas and some leukemias)[15].

Experimental purity and ploidy estimates can also support CNA fitting. CNA detection methods can fit copy number solutions informed by experimental ploidy estimates, as in silico estimates of tumor ploidy have been repeatedly shown to match experimental values[40,51,52]. Experimental purity and ploidy estimates can be obtained through fluorescence in situ hybridization, image cytometry, fluorescence-activated cell sorting or single-cell sequencing[53].

SNV clustering: overview. Before SNVs can be clustered, their measured variant allele frequencies must be transformed into cellular frequencies (CCF or CP) using purity and copy number. It is essential to adjust for CNAs when converting VAFs to cellular frequencies, or to cluster only SNVs in normal diploid regions, as copy number gains and losses will alter the fraction of reads bearing an SNV. Neglecting to adjust for these effects can lead to incorrect clustering (Fig. 1b,d)[54]. Cellular prevalence estimates for SNVs in normal diploid regions can be clustered first to identify the major subclonal lineages. In normal diploid regions, CP is precisely twice VAF, so clustering by VAF and implied CP is equivalent. The cluster with the highest CP can be deemed clonal, and the remaining clusters can be assigned CPs and associated with a subclonal lineage. Errors can arise from incorrect cluster number estimation or SNV-lineage misassignment, potentially shifting the clonal peak, from which purity can be estimated. In general, the clustering principles outlined in this section apply equivalently to indels.

The assumed noise distribution in VAF estimates will influence subclonal reconstruction accuracy. Using an inappropriate noise model can lead to over- or underestimating cluster number. Binomial noise models can capture the influence of the read depth, copy number state and CP on the accuracy of the assessed SNV VAF, while in general fixed-variance Gaussian noise models cannot. Overdispersed binomial models (for example, β-binomial[31] or negative binomial, which are often used for bisulfite, exome or single-cell RNA sequencing data) assume greater variance than standard binomial models, but it is unclear whether they are better suited for subclonal reconstruction from DNA data. Like binomial models, β models are also suited for subclonal reconstruction, as they model VAF directly[24,55].

Care should be taken when translating SNV VAFs to CCF space before clustering, as this requires estimating the multiplicity of each SNV—that is, the number of tumor DNA copies harboring it. SNV multiplicity estimates depend on the accuracy of copy number calls, as before assignment it is necessary to enumerate the space of possible multiplicities. This requires allele-specific copy number estimates, as total copy number is insufficient to set boundaries of possible multiplicities. For example, in the case of copy-neutral loss of heterozygosity, the total copy number is 2, and individual SNVs can have a multiplicity of 1 (for that mutations that occur after the duplication) or 2 (for those mutations that happen before it). By contrast, in balanced diploid regions, the total copy number is again 2, but each individual SNV must have a multiplicity of 1.

If the CCF values of clonal SNVs are computed assuming a multiplicity of 1, then the resulting estimated CCFs will be approximately equal to the (real) multiplicity of the SNV in the clonal lineage (see Supplementary Note). However, subclonal changes in SNV multiplicity equate to subclonal copy number changes, and because subclonal copy number states are ambiguous, the affected SNVs may generate spurious clusters. Nonetheless, CCF clustering can sometimes detect, but not correct, errors in CNA reconstruction. For example, large SNV clusters with CCFs > 1 (superclonal clusters; Box 1) are theoretically impossible for somatic mutations. When they are detected in subclonal reconstruction, this can be diagnostic of failure in detecting the clonal lineage during CNA reconstruction, or of large segments with wrong copy number calls. In the former case, purity will be incorrect (that is, underestimated), and the clonal peak will be shifted upwards in CCF space. In the latter case, SNV multiplicity and thus CCF will be incorrect. These errors often occur in tumors without CNAs or without any clonal CNAs, or as a result of contamination by germline variants (discussed below). In these cases, the CNA-based purity will be underestimated, leading to CCFs > 1 for SNVs in these superclonal lineages.

To address these concerns, a number of methods use generative models of VAFs that incorporate CNA reconstructions. These methods assess the impact of clonal and subclonal CNAs on SNV multiplicity (and their associated changes in VAF) using maximum likelihood[25,31,38,46,56]. Of these, PhyloWGS[25] and LICHEe[56] attempt full phylogenetic reconstructions. CloneHD[46] does not explicitly enforce consistent tree structure, PyClone[31] assumes a single CNA change per segment without reconstructing a clonal tree, and DPClust[38] assigns SNV multiplicities to the most likely value given the CNA reconstruction.

SNV clustering: underlying assumptions. Methods based on SNV clustering rely on several assumptions about cancer evolution. The first assumption is that most SNVs with detectable VAFs are associated with a small number of subclonal lineages (the 'weak parsimony' assumption)[25]. Given the large number of cells in a bulk tumor and the positive mutation rate per cell division, the existence of a very large number of low-prevalence SNVs is uncontroversial. Their low VAF typically precludes detection by typical somatic mutation calling algorithms, as detectable subclonal lineages are primarily established through selective sweeps and early drift[2,57]. This assumption is somewhat controversial, however, as the lowest-VAF cluster may contain a mixture of SNVs coming from numerous parallel lineages growing neutrally[58,59]. Therefore, the lowest VAF cluster might be a mix of subclones, and efforts are ongoing to characterize this to capture non-tail subclones[24].

A second assumption implicitly made by many clustering-based algorithms is that a given genomic position is subject to an SNV only once during the development of an individual tumor and never reverts to the germline state (the 'infinite sites' assumption; Box 1). As a result, each SNV can be uniquely assigned to a specific subclonal lineage[31,56,60]. It is known that the infinite sites assumption can be violated, as exemplified by the existence of parallel acquisition of the same driver SNVs and by triallelic loci. As a result, some methods do not make this assumption[61]. Infinite sites violations occur rarely enough that they are not expected to affect clustering based on hundreds to thousands of SNVs. As a result, this approximation remains widely used for subclonal reconstruction through bulk WGS and WES, and it may even be reasonable for targeted sequencing studies of driver genes.

Subclonal losses of chromosomal segments may also lead to somatic variants disappearing in these lineages, and thereby variants may appear to revert to the germline state, leading to apparent violations of the infinite sites assumption. This situation is relatively frequent, particularly in multisample reconstruction, and may lead to the appearance of spurious clusters. Apart from disregarding mutations in regions showing subclonal copy number events, another elegant way to account for this in subclonal reconstruction is by modeling mutations through a Dollo process, which assumes mutations can occur only once but can subsequently disappear again (only once)[14].

Many methods also make implicit 'hidden' assumptions about the number of clusters or their density. For example, many clustering methods rely on some form of Dirichlet process clustering[14,25,31,38,56,62]. This technique has an important hyperparameter called the concentration parameter, which can be inferred from the data or be constrained with strong priors. We recommend testing different values for these parameters to quantify their impact on the subclonal reconstruction for any given tumor (see Supplementary Note).

SNV clustering: troubleshooting. When two subclonal lineages show similar CPs, their VAF distributions may overlap. As a result, clustering algorithms will merge these lineages in the absence of further information (for example, using more samples). Nonetheless, it is possible to be highly confident about the presence and CP of a lineage while being uncertain about the assignment of many of its SNVs.

Variant calling accuracy will affect clustering accuracy[22,23]. Germline SNPs wrongly called as somatic SNVs have high VAFs (~0.5) (Fig. 1b,d) and may be clustered into their own high-CCF lineage (>100% CCF). These clusters tend to have few SNVs, which have a specific mutational signature (primarily C>T and T>C) and contain few or no CNAs. Filtering against a database of germline variants (for example, dbSNP) and higher-depth normal sample sequencing can reduce the risk of germline contamination, although filtering against a database such as dbSNP can lead to increased numbers of false-negative somatic SNVs, particularly as these databases grow. False-positive SNVs due to sequencing errors have low VAFs and mostly will be assigned to a low-frequency lineage, which can help identify them. They may also have a distinct mutational signature[63]. False negatives in somatic SNV calling can lead to overestimation of the CPs of low frequency subclones because only the higher part of the VAF distribution is observed. CPs can be adjusted for this bias[15], and using a highly sensitive SNV detection algorithm can also help mitigate this effect.

We have observed that algorithms that assume a diploid-copy neutral state can struggle with mutations on the male sex chromosomes. An easy fix is to exclude SNVs on these chromosomes. In theory, these SNVs could be reassigned post hoc with appropriate treatment, although many CNA-reconstruction methods do not report calls on chromosome Y. Tetraploid tumors will likely be annotated as diploid if they have few other CNAs, leading to spurious clusters (see "CNA reconstructions: troubleshooting" above). Artifact clusters also sometimes result from incorrect copy number calls, which can lead to incorrect adjustment of allele frequencies to CPs. Such clusters are easily identified, as all the associated SNVs are located in a single chromosomal region.

Phylogenetic reconstruction. Given the weak parsimony assumption, SNV clusters represent groups of mutations that occurred in one (or a few) subclones at one point in the tumor's evolution, inherited by all of their descendants—that is, members of their subclonal lineages. These clusters represent the ancestors of the subpopulations present in the tumor at the time of sample acquisition. Thus, SNVs clustered in CCF space are assigned to the same lineage. Assigning these mutations to existing subpopulations in the sample requires more information: namely, the ancestor relationships between these lineages or the tumor phylogeny, often represented as a clone tree (Fig. 1d). We believe tree inference should generally be restricted to multisample studies: while subclones can be identified in single samples, they are only weakly informative of the underlying tree (Fig. 2b).

The weak parsimony and infinite sites assumptions restrict the phylogenies consistent with a set of inferred CPs. Assuming SNVs do not revert to their germline state implies that descendent subpopulations inherit all the mutations in their ancestors. The CP of an ancestral lineage must be at least as large as the sum of the CPs of its direct descendants[38] (this concept is known as the pigeonhole principle; Box 1). Consider a common ancestor A with two descendants (lineages B and C), whose relationship is unknown. If CP(B) > CP(C), and CP(B) + CP(C) > CP(A), then B must be an ancestor of C[62]. Tumor phylogenetic methods employ some version of these rules, either explicitly or implicitly[62].

When there are multiple subclonal lineages and few samples, multiple phylogenies are usually consistent with a given set of lineage CPs. Because any set of CPs from a single sample is often consistent with a linear phylogeny, it is typically difficult to unambiguously call a branching phylogeny from single-sample CPs. Branching can be inferred in some cases when the CP of an SNV cluster is incompatible with being either an ancestor or a descendant of a CNA, which can sometimes be detected automatically in a single sample[14,25]. Branching can also be detected in a single sample when nearby or overlapping SNVs on the same chromosome copy are identified as mutually exclusive through read phasing[41] (Fig. 1b). However, with short-read data, it is often difficult to phase enough somatic SNVs for a high-quality reconstruction, though some methods automatically apply this approach[64]. As some phylogenetic ambiguity usually persists, and choice of a single phylogeny is often arbitrary or depends on weakly validated assumptions[60,65], phylogenetic methods should report uncertainty in the reported phylogeny.

In single-sample studies[15], phasing of SNVs can inform branching subclones and the pigeonhole principle helps identify linear subclones, although most phylogenies remain unresolved. Multiple samples can greatly clarify phylogenetic relationships among subclones (Fig. 2b). For example, if CP(B) > CP(C) in one sample but CP(C) > CP(B) in another sample, then these subclones must be cousins or siblings. Thus, though branching is usually impossible to establish using a single sample, it is often possible with just two samples and becomes increasingly easy to identify with more samples. Subclonal reconstruction methods generalize trivially to clustering SNVs in multiple CCF dimensions, and the same principles and assumptions apply. Notably, chromosomal losses leading to variants being lost in subclonal lineages and thereby apparent infinite sites violations are quite common and, if overlooked, can lead to errors in both multi- and single-sample phylogeny reconstruction.

Evolution of the field

DNA sequencing technologies continue to improve, with major ongoing advances in long-read and single-cell sequencing. We anticipate these technologies will be useful for reducing subclonal reconstruction uncertainty and improving accuracy. Long-read sequencing allows more accurate breakpoint detection, copy number state characterization and longer-range phasing relative to short-read sequencing[66–68]. These will all improve CNA reconstruction, reducing downstream errors, while long-range phasing will facilitate phylogeny inference through mutual exclusivity. Single-cell WGS will become increasingly useful for phylogenetic inference as its CNA and SNV detection accuracy improves[69,70]. Both these technologies can be combined with short-read sequencing to reduce costs and leverage their strengths but retain the resolution of high-depth sequencing[71,72]. Subclonal reconstruction algorithms will need to accommodate the data, biases and errors these and other new technologies generate. Carefully characterizing error profiles at each step is critical for interpretation as errors are likely to propagate and affect the final reconstruction.

Both WGS and WES have limited sensitivity for studying intratumor heterogeneity in single samples[73,74]. Bulk sequencing of multiple tumor regions still misses many subclones, particularly when subclones are not evenly distributed across the tumor mass. While cell turnover facilitates subclone mixing, spatial subclone segregation can arise through local differences in the microenvironment and subclone competition. As subclone distributions remain difficult to predict and cannot be exhaustively sampled, emerging liquid biopsy[75] and homogenized tissue sampling technologies[76] can provide a more complete evolutionary profile than even dense tissue sampling by providing a complementary sampling of the tumor with different biases. These technologies are especially useful when detecting minor subclones is a priority (for example, to detect evolving post-treatment resistance). They typically employ targeted gene panels to allow sufficient depth for accurate SNV calling and subclonal reconstruction. However, their sensitivity may depend on tumor features and their efficacy in diverse clinical contexts is not yet clear[17].

Single-cell sequencing can resolve phylogenies in greater detail and with less uncertainty than bulk sequencing[10,14,77,78], and the technology is rapidly developing[79]. As a result of artifacts left during amplification of the genetic material, as well as limited genome coverage, calling SNVs in single-cell data remains an important challenge, with active development[69]. Many alternatives have been explored, such as using the cell's own high-fidelity replicative machinery by expanding single cells into colonies on which to perform bulk whole-genome sequencing[80]; using direct library preparation of thousands of cells[81], which simply skips the error-prone amplification; or using single-cell targeted sequencing approaches to maximize local depth of coverage, also allowing allele-specific copy number inference in single cells[26].

Although inferring the genotypes of each subclone may be more difficult using bulk than single-cell sequencing, bulk-sequencing tissue samples from multiple tumor regions can mitigate some of these limitations[13]. Tor phylogeny reconstruction, we thus recommend combining bulk and single-cell sequencing when possible; this has shown great potential to resolve ambiguities in the phylogenetic tree reconstruction as a result of the single-cell resolution while maintaining high-quality somatic mutation calls from the bulk[14,80]. However, if cells are sampled from a small number of regional biopsies and there is limited clonal mixing, spatial biases may still affect single-cell phylogenetic reconstruction[82].

Subclonal reconstruction is a fast-evolving field. Further work is underway to integrate more data types into subclonal reconstructions (for example, structural variants—which, given the higher noise in their VAF, are assigned to subclones post hoc[83,84]—or epigenetic marks[56]) but these will require careful validation given the uncertainty already present in subclonal reconstruction. Similarly, mechanistic models and approaches have recently been proposed[13,58,59], but remain in their infancy[58,59,85–88]. In the future, combining the phenomenological models discussed with mechanistic models is likely to prove invaluable[24]. A thorough assessment of new and existing methods, particularly for subclonal CNA detection, is acutely needed. Precise metrics of reconstruction accuracy, increasingly realistic synthetic tumor genomes and joint single-cell and bulk tumor sequencing datasets[22,72] are all required to establish community-accepted benchmarks that drive algorithm development and application.

Received: 28 September 2019; Accepted: 9 November 2020;
Published online: 4 January 2021

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
2. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
3. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
4. Hong, M. K. H. et al. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat. Commun.* **6**, 6605 (2015).
5. Mitchell, T. J. et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx Renal. *Cell* **173**, 611–623.e17 (2018).
6. Turajlic, S. et al. Tracking cancer evolution reveals constrained routes to metastases: TRACERx Renal. *Cell* **173**, 581–594.e12 (2018).
7. Andor, N. et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105–113 (2016).
8. Espiritu, S. M. G. et al. The evolutionary landscape of localized prostate cancers drives clinical aggression. *Cell* **173**, 1003–1013.e15 (2018).
9. Jamal-Hanjani, M. et al. Tracking the evolution of non–small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
10. Fittall, M. W. & Van Loo, P. Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome Med.* **11**, 20 (2019).
11. Sendorek, D. H. et al. Germline contamination and leakage in whole genome somatic single nucleotide variant detection. *BMC Bioinformatics* **19**, 28 (2018).
12. Alioto, T. S. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).
13. Sun, R. et al. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.* **49**, 1015–1024 (2017).
14. Salehi, S. et al. ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biol.* **18**, 44 (2017).
15. Dentro, S. C. et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Preprint at *bioRxiv* https://doi.org/10.1101/312041 (2020).
16. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
17. Abbosh, C. et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446–451 (2017).
18. Noorani, A. et al. Genomic evidence supports a clonal diaspora model for metastases of esophageal adenocarcinoma. *Nat. Genet.* **52**, 74–83 (2020).
19. McGranahan, N. et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).
20. Gomez, K. et al. Somatic evolutionary timings of driver mutations. *BMC Cancer* **18**, 85 (2018).
21. Opasic, L., Zhou, D., Werner, B., Dingli, D. & Traulsen, A. How many samples are needed to infer truly clonal mutations from heterogenous tumours? *BMC Cancer* **19**, 403 (2019).
22. Salcedo, A. et al. A community effort to create standards for evaluating tumor subclonal reconstruction. *Nat. Biotechnol.* **38**, 97–107 (2020).
23. Griffith, M. et al. Optimizing cancer genome sequencing and analysis. *Cell Syst.* **1**, 210–223 (2015).
24. Caravagna, G. et al. Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat. Genet.* **52**, 898–907 (2020).
25. Deshwar, A. G. et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).
26. Laks, E. et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179**, 1207–1221.e22 (2019).
27. Schwarz, R. F. et al. Phylogenetic quantification of intra-tumour heterogeneity. *PLOS Comput. Biol.* **10**, e1003535 (2014).
28. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–1893 (2014).
29. El-Kebir, M. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics* **34**, i671–i679 (2018).
30. Zhang, J. et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**, 256–259 (2014).
31. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
32. Shi, W. et al. Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity. *Cell Rep.* **25**, 1446–1457 (2018).
33. Yates, L. R. et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
34. Schuh, A. et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* **120**, 4191–4196 (2012).
35. Boutros, P. C. et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47**, 736–745 (2015).
36. Robbe, P. et al. Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. *Genet. Med.* **20**, 1196–1205 (2018).
37. Chin, S.-F. et al. Shallow whole genome sequencing for robust copy number profiling of formalin-fixed paraffin-embedded breast cancers. *Exp. Mol. Pathol.* **104**, 161–169 (2018).
38. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
39. Deshpande, A., Walradt, T., Hu, Y., Koren, A. & Imielinski, M. Robust foreground detection in somatic copy number data. Preprint at *bioRxiv* https://doi.org/10.1101/847681 (2019).
40. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
41. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harb. Perspect. Med.* **7**, a026625 (2017).
42. Chiang, D. Y. et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6**, 99–103 (2009).
43. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
44. Nilsen, G. et al. Copynumber: efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
45. Lai, D. & Shah, S. HMMcopy: copy number prediction with correction for GC and mappability bias for HTS data. *R Package Version 1* (2012).
46. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7**, 1740–1752 (2014).
47. McPherson, A. W. et al. ReMixT: clone-specific genomic structure estimation in cancer. *Genome Biol.* **18**, 140 (2017).
48. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **14**, R80 (2013).
49. Jiang, Y., Qiu, Y., Minn, A. J. & Zhang, N. R. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **113**, E5528–E5537 (2016).
50. Müller, C. A. et al. The dynamics of genome replication using deep sequencing. *Nucleic Acids Res.* **42**, e3 (2014).
51. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
52. Steele, C. D. et al. Undifferentiated sarcomas develop through distinct evolutionary pathways. *Cancer Cell* **35**, 441–456.e8 (2019).
53. Almendro, V. et al. Genetic and phenotypic diversity in breast tumor metastases. *Cancer Res.* **74**, 1338–1348 (2014).
54. Farahani, H. et al. Engineered in-vitro cell line mixtures and robust evaluation of computational methods for clonal decomposition and longitudinal dynamics in cancer. *Sci. Rep.* **7**, 13467 (2017).
55. Miller, C. A. et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLOS Comput. Biol.* **10**, e1003665 (2014).
56. Popic, V. et al. Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.* **16**, 91 (2015).
57. Sottoriva, A. et al. A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
58. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
59. Williams, M. J. et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* **50**, 895–903 (2018).
60. Strino, F., Parisi, F., Micsinai, M. & Kluger, Y. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res.* **41**, e165 (2013).
61. Marass, F. et al. A phylogenetic latent feature model for clonal deconvolution. *Ann. Appl. Stat.* **10**, 2377–2404 (2016).
62. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* **15**, 35 (2014).
63. Ewing, A. D. et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12**, 623–630 (2015).
64. Zhou, T., Müller, P., Sengupta, S. & Ji, Y. PairClone: a Bayesian subclone caller based on mutation pairs. *J. R. Stat. Soc. Ser. C Appl. Stat.* **68**, 705–725 (2019).
65. El-Kebir, M., Satas, G. & Raphael, B. J. Inferring parsimonious migration histories for metastatic cancers. *Cancer* **2**, 5 (2018).
66. Zamani Esteki, M. et al. Concurrent whole-genome haplotyping and copy-number profiling of single cells. *Am. J. Hum. Genet.* **96**, 894–912 (2015).
67. Mantere, T., Kersten, S. & Hoischen, A. Long-read sequencing emerging in medical genetics. *Front. Genet.* **10**, 426 (2019).
68. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

69. Dong, X. et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* **14**, 491–493 (2017).
70. Martelotto, L. G. et al. Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples. *Nat. Med.* **23**, 376–385 (2017).
71. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
72. Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S. C. & Beerenwinkel, N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat. Commun.* **10**, 2750 (2019).
73. Abécassis, J. et al. Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data. *PLoS One* **14**, e0224143 (2019).
74. Liu, L. Y. et al. Quantifying the influence of mutation detection on tumour subclonal reconstruction. Preprint at *bioRxiv* https://doi.org/10.1101/418780 (2020).
75. Parikh, A. R. et al. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nat. Med.* **25**, 1415–1421 (2019).
76. Litchfield, D. K. et al. Representative sequencing: unbiased sampling of solid tumor tissue. *Cell Rep.* **31**, 107550 (2019).
77. Eirew, P. et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2015).
78. Kim, C. et al. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* **173**, 879–893.e13 (2018).
79. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
80. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
81. Zahn, H. et al. Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods* **14**, 167–173 (2017).
82. Chkhaidze, K. et al. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *PLOS Comput. Biol.* **15**, e1007243 (2019).
83. Eaton, J., Wang, J. & Schwartz, R. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics* **34**, i357–i365 (2018).
84. Cmero, M. et al. Inferring structural variant cancer cell fraction. *Nat. Commun.* **11**, 730 (2020).
85. Noorbakhsh, J. & Chuang, J. H. Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nat. Genet.* **49**, 1288–1289 (2017).
86. Tarabichi, M. et al. Neutral tumor evolution? *Nat. Genet.* **50**, 1630–1633 (2018).
87. Heide, T. et al. Reply to 'Neutral tumor evolution?'. *Nat. Genet.* **50**, 1633–1637 (2018).
88. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Reply: Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nat. Genet.* **49**, 1289–1291 (2017).
89. Zare, F., Dow, M., Monteleone, N., Hosny, A. & Nabavi, S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics* **18**, 286 (2017).
90. Gerlinger, M. et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).
91. Vinci, M. et al. Functional diversity and cooperativity between subclonal populations of pediatric glioblastoma and diffuse intrinsic pontine glioma cells. *Nat. Med.* **24**, 1204–1215 (2018).
92. Kuipers, J., Jahn, K., Raphael, B. J. & Beerenwinkel, N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.* **27**, 1885–1894 (2017).
93. Rieber, N. et al. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS One* **8**, e66621 (2013).

Acknowledgements

A.S. was supported by an NSERC CGS. M.T. and P.V.L. are supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202) and the Wellcome Trust (FC001202). M.T. is a postdoctoral fellow supported by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie grant agreement no. 747852-SIOMICS). P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of the Francis Crick Institute. M.N.L. was supported by a Junior Research Fellowship (Trinity College, University of Oxford). P.C.B. was supported by a Terry Fox Research Institute New Investigator Award and a CIHR New Investigator Award. Q.M. is supported by an Associate Investigator Award from the Ontario Institute of Cancer Research and holds a Canada CIFAR AI chair. This work was supported by the NIH/NCI under award numbers P30CA016042 (P.C.B.) and P30-CA008748 (Q.M.), and through support from the ITCR (1U24CA248265-01) to P.C.B.

Author contributions

M.T., A.S., A.G.D., M.N.L., J.W., D.C.W., Q.D.M., P.V.L. and P.C.B. wrote the text. D.C.W., Q.D.M., P.V.L. and P.C.B. oversaw the completion of this work.

Competing interests

P.C.B is a member of the Scientific Advisory Boards of BioSymetrics Inc. and Intersect Diagnostics Inc. M.T., A.S., A.G.D., M.N.L., J.W., D.C.W., Q.D.M., and P.V.L. declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-01013-2.

Correspondence should be addressed to P.C.B.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Lin Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.