

# Principles of reconstructing the subclonal architecture of cancers

Stefan C. Dentro<sup>1,2</sup>, David C. Wedge<sup>3</sup>, and Peter Van Loo<sup>2,4\*</sup>

<sup>1</sup>Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, United Kingdom

<sup>2</sup>The Francis Crick Institute, 1 Midland Road, London NW1 1AT, United Kingdom

<sup>3</sup>Big Data Institute, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, United Kingdom

<sup>4</sup>Department of Human Genetics, University of Leuven, B-3000 Leuven, Belgium

\*Corresponding author

Stefan C. Dentro : [sd11@sanger.ac.uk](mailto:sd11@sanger.ac.uk), +44 (0)1223 834244

David C. Wedge : [david@well.ox.ac.uk](mailto:david@well.ox.ac.uk), +44 (0)1865 287512

Peter Van Loo : [peter.vanloo@crick.ac.uk](mailto:peter.vanloo@crick.ac.uk), +44 (0)20 3796 1719

## **Abstract**

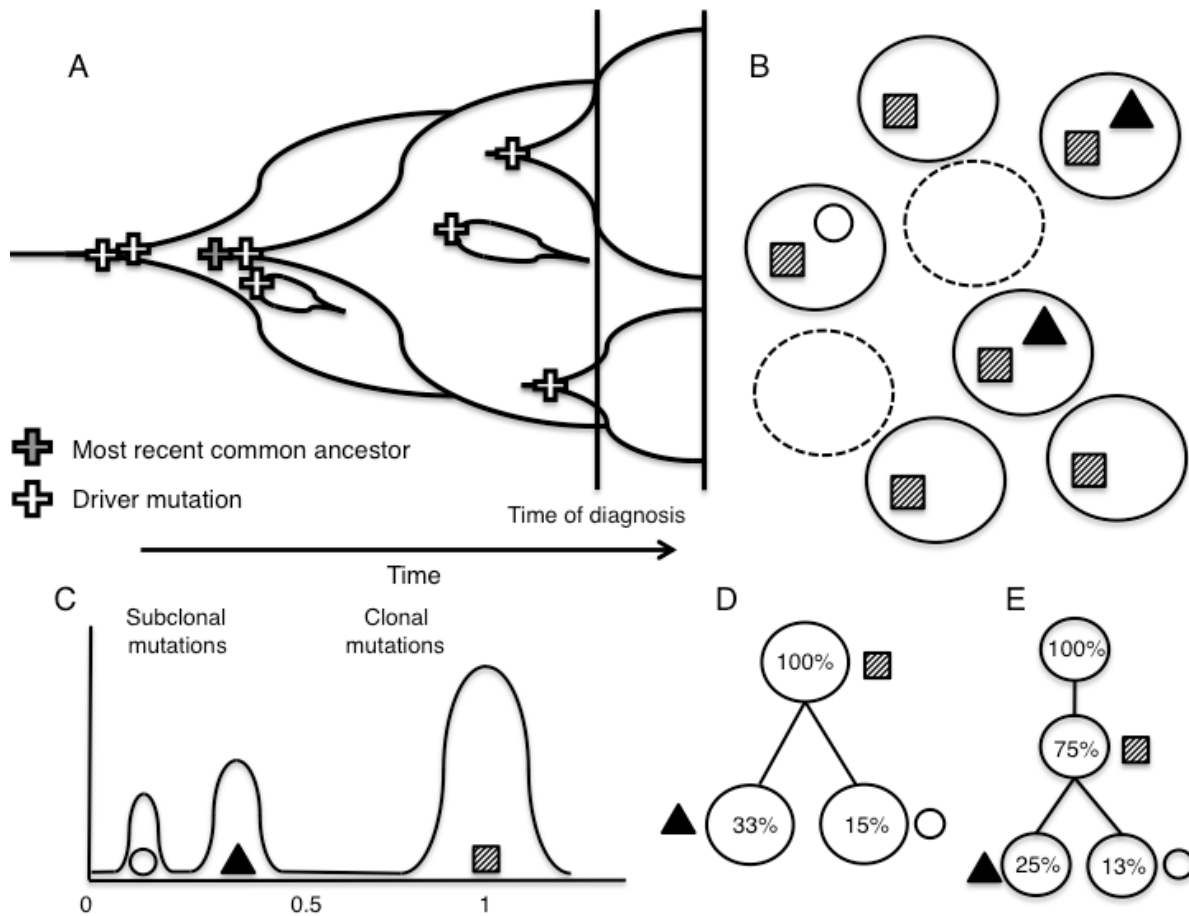
Most cancers evolve from a single founder cell through a series of clonal expansions that are driven by somatic mutations. These clonal expansions can lead to several co-existing subclones sharing subsets of mutations. Analysis of massively parallel sequencing data can infer a tumour's subclonal composition through the identification of populations of cells with shared mutations. We describe the principles that underlie subclonal reconstruction through single nucleotide variants (SNVs) or copy number alterations (CNAs) from bulk or single cell sequencing. These principles include estimating the fraction of tumour cells for SNVs and CNAs, performing clustering of SNVs from single- and multi-sample cases, and single cell sequencing. The application of subclonal reconstruction methods is providing key insights into tumour evolution, identifying subclonal driver mutations, patterns of parallel evolution and differences in mutational signatures between cellular populations, and characterizing the mechanisms of therapy resistance, spread and metastasis.

## Introduction

Cancers evolve through the acquisition of changes in the genome and epigenome of their cells (Nowell 1976; Tabin *et al.* 1982). Some of these mutations provide the cell in which they occurred with an evolutionary advantage over other cells and are known as *driver* mutations, while other mutations (*passenger* mutations) are assumed to have a neutral effect (Stratton *et al.* 2009; Garraway and Lander 2013). A tumour cell with a selective advantage is better suited to its local micro-environment and can therefore proliferate quicker than other cells and generate more daughter cells. This process is called *clonal expansion* (Greaves and Maley 2012; Vogelstein *et al.* 2013). This interplay between mutation and selection allows a tumour to evolve and adapt to a changing environment.

Part of a tumour's evolutionary story can be inferred through massively parallel sequencing of tumour samples (Fig. 1). Mutations that have occurred before the *most recent common ancestor* (MRCA) are carried by all tumour cells in a sample and can be used as markers of the clonal population (Campbell *et al.* 2008). As the tumour develops further, it continues to acquire more driver and passenger mutations. A tumour cell that acquires an additional driver mutation and embarks on a clonal expansion will generate a subpopulation of cells bearing mutations that are not shared by all cells in the tumour (Figure 1A). Such a subclonal cell population can therefore be identified through a set of shared mutations.

Figure 1B shows a schematic example, where the 'square' mutations are carried by all tumour cells and are therefore clonal, and the 'triangle' and 'circle' mutations are present only in a subpopulation of tumour cells. A tumour sample usually also contains non-tumour cells, such as stromal cells, immune cells and fibroblasts, that do not share any genomic mutations with the tumour clones. The fraction of tumour cells in a sequencing sample is known as the *purity* or *cellularity*.



**Figure 1.** During cancer evolution, a tumour acquires driver mutations (marked with a plus sign) that can initiate clonal expansions (A). Over time, a number of these clonal expansions can occur, resulting in the rise of subpopulations of cells harboring distinct sets of mutations. Tumour samples typically consist of a mixture of tumour cells with mutations (solid lines) and normal cells without mutations (dashed lines, B). Some mutations are carried by all tumour cells (marked with a square), while others are present in a subset of cells (triangle and circle). Using allele frequencies of mutations obtained from sequencing data and accounting for copy number aberrations, an estimate of the fraction of tumour cells carrying each mutation can be obtained (C). A set of mutations can then be used as a marker for a population of cells, allowing estimation of the fraction of tumour cells of the corresponding subclone. Clustering algorithms can be applied to obtain the cancer cell fractions of each subclone. The relationship between subclones can be visualised as a tree (D and E). Some methods perform this clustering in fraction of tumour cells space (D), others in the space of fraction of all cells (E).

For each of the somatic mutations, an associated *variant allele frequency* (VAF) can be calculated. Besides the tumour purity and the fraction of tumour cells carrying a mutation, the VAF also depends on copy number changes. For example, a mutation that has occurred before a gain is carried by two out of three chromosome copies, while a mutation that occurred after the gain is carried by one out of three chromosome copies. It is important to account for all these factors when using VAF values to infer a tumour's subclonal architecture, as mutations from the same subclonal population may show different VAFs due to copy number changes.

Accounting for the factors above, it is useful to represent the propensity of mutations or mutation clusters through their *cellular prevalence* (CP, the fraction of cells carrying the mutation(s) in the sample), or their *cancer cell fraction* (CCF, the fraction of tumour cells carrying the mutation(s)). In Figure 1C, the CCF space is shown, where the clonal mutations (denoted by squares) now appear in a cluster around 1.0, as they are found in 100% of tumour cells, and the subpopulation denoted by triangles that consists of 33% of tumour cells now appears at 0.33. Subclonal reconstruction can be performed by clustering these mutations, here resulting in a 'square' and a 'triangle' cluster. A key underlying assumption is that each mutation has occurred only once during the lifetime of a tumour, which is referred to as the *infinite sites assumption* (Nik-Zainal *et al.* 2012; Beerenwinkel *et al.* 2014). A tumour's subclonal architecture can be represented by a phylogenetic tree (Fig 1D and 1E).

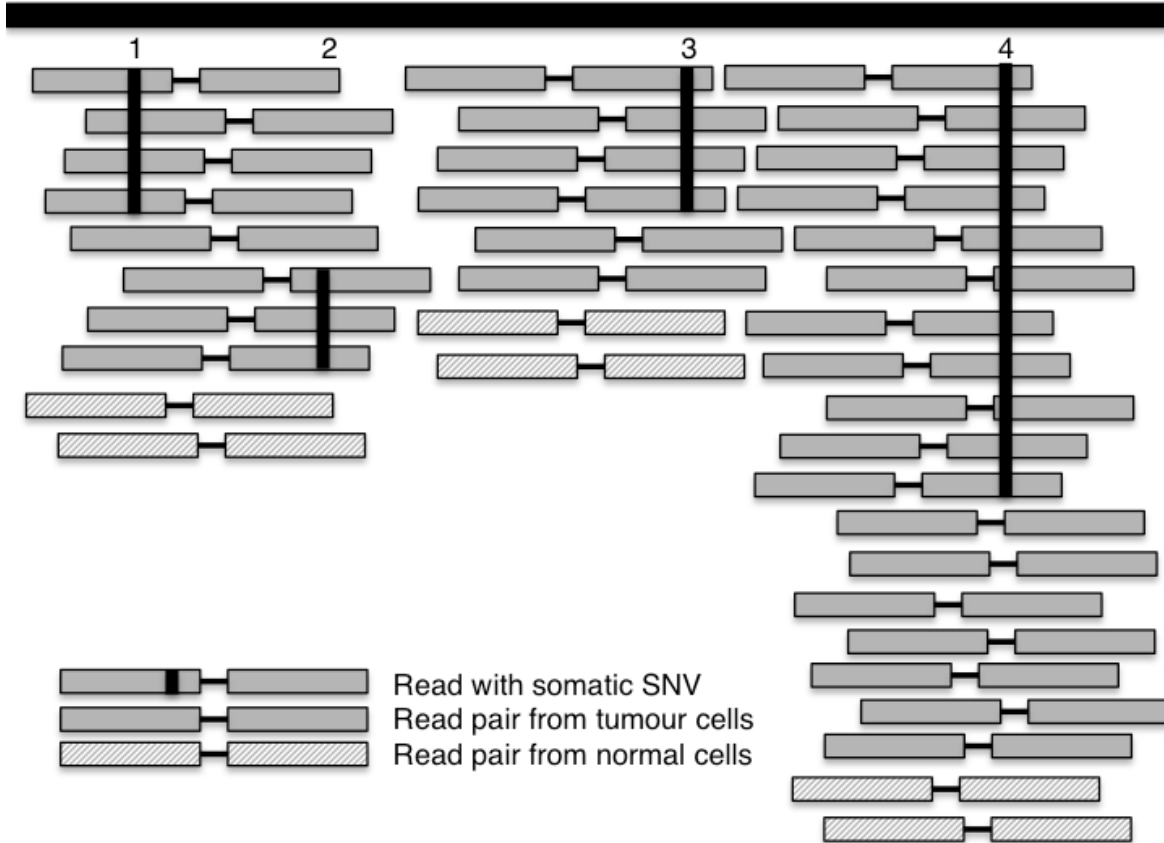
Here, we describe the principles of reconstructing the subclonal architecture of cancers from massively parallel sequencing data. Subclonal reconstruction methods build upon the principles described above to reconstruct the subclonal architecture of tumours, starting from either single nucleotide variants (SNVs) or copy number alterations (CNAs), or both. We describe how cancer cell fractions can be calculated and outline the principles behind SNV- and CNA-based subclonal reconstruction methods, using data from single-biopsies, multi-sampling and single cell sequencing. Finally, we outline which biological insights have been obtained through these methods and outline future directions.

## Estimating Cancer Cell Fractions

Cancer cell fractions can be estimated from VAFs of SNVs. Massively parallel sequencing results in short reads, which can then be aligned to a reference genome, followed by SNV

calling. Both the variant and reference alleles of an SNV are supported by a number of reads,  $r_{mut}$  and  $r_{ref}$  respectively. The VAF of SNV  $i$ ,  $f_i$ , can straightforwardly be calculated as:

$$f_i = \frac{r_{mut,i}}{r_{mut,i} + r_{ref,i}}, \text{ Eq. 1}$$



**Figure 2.** Allele frequencies of SNVs must be transformed to Cancer Cell Fractions, accounting for copy number changes, before they can be clustered to identify subclonal populations. This illustration shows 4 SNVs in different (sub)clonal populations and in regions with different copy number states, to illustrate this principle. SNVs 1 and 2 are clonal and subclonal respectively and appear in a non-aberrated copy number state. SNV 3 coincides with a subclonal deletion, with the SNV falling on the retained allele (i.e. the other allele is subclonally deleted). SNV 4 has occurred before a gain and is therefore carried by two chromosome copies. Even though SNV 1, 3 and 4 are clonal, their allele frequencies differ due to copy number alterations.

However, mutation clustering to identify (sub)clonal populations cannot be performed directly using VAFs, as copy number changes impact allele frequencies. Figure 2 shows four SNVs in a sequencing sample that consists of 80% tumour cells and 20% normal cells. SNV 1 is clonal and occurs in a region with a normal diploid copy number state. This mutation is therefore carried by approximately half the reads that represent tumour DNA. SNV 2 is subclonal and also occurs in a region of normal diploid copy number. As both copy number and normal cell contamination are equal for both SNV 1 and 2, their allele frequencies are directly comparable and proportional to the fraction of tumour cells by which they are carried. SNV 3 falls into an area that was subclonally lost. As the subclonal loss has occurred on the other allele, this SNV's VAF is increased compared to SNV 1. SNV 4 is clonal, falls into an area that is clonally gained and is on the gained allele. Its VAF is therefore higher than that of SNV 1. If these SNVs were clustered in VAF space, SNVs 3 and 4 would be mistaken for evidence of additional mutation clusters, while they in fact belong to the clonal cluster.

This example illustrates that the copy number state of an SNV, also called its *multiplicity*, is key to understanding VAF distributions of mutations. Estimating the multiplicity of an SNV is challenging, as it requires establishing the copy number state of a single base. Copy number callers often estimate copy number states for large stretches of DNA, which might not accurately represent the copy number state exactly at the base of the SNV. To assist with resolving this issue, it is helpful to consider the product of mutation multiplicity  $m_i$  of a mutation  $i$  and its cancer cell fraction  $CCF_i$ :

$$u_i = CCF_i m_i, \text{ Eq. 2}$$

Let us consider the properties of  $u_i$ : a clonal SNV will have a CCF of 1.0 (i.e. 100% of tumour cells) and in each cell the number of chromosome copies,  $m_i$ , is an integer. It follows from the above equation that for clonal mutations  $u_i \geq 1$ . A subclonal mutation has a CCF less than 1.0 (for example 0.4, or 40% of tumour cells) and can only be carried by a single chromosome copy (unless also affected by a subclonal CNA), therefore  $m_i = 1$ . It follows that  $u_i < 1$  for subclonal mutations. We can use these observations to obtain  $m_i$  from  $u_i$ :

$$m_i = \begin{cases} |u_i|, & u_i \geq 1 \\ 1, & u_i < 1 \end{cases}, \text{ Eq. 3}$$

Furthermore,  $u_i$  can be written as a function of the fraction of tumour cells  $\rho$  with a total number of chromosome copies in tumour cells at locus  $i$ ,  $n_{tot,t,i}$ , and a fraction of normal cells  $1 - \rho$  with a total number of chromosome copies in normal cells at locus  $i$ ,  $n_{tot,n,i}$ :

$$u_i = f_i \frac{1}{\rho} [\rho n_{tot,t,i} + (1 - \rho) n_{tot,n,i}] , \text{ Eq. 4}$$

In the formula above,  $\rho$  and  $n_{tot,t,i}$  can be obtained through copy number analysis,  $f_i$  can be calculated from  $r_{mut}$  and  $r_{ref}$  using Eq. 1, and the  $n_{tot,n,i}$  values are considered known (typically 2). This equation therefore provides us with a way to calculate  $u_i$  and by extension to obtain the multiplicity of the SNV.

SNV 1 in figure 2 for example is clonal and has 4 reads reporting the variant and 6 reporting the reference allele. The purity is 0.8 (80% of total cells are tumour cells) and the total copy number of both the tumour and normal cells is 2. Its  $u_i$  therefore becomes:

$$\frac{4}{4 + 6} * \frac{1}{0.8} * [0.8 * 2 + 0.2 * 2] = 1.000$$

Which translates into a CCF of 1.0 via eq. 3. While for SNV 4 it yields:

$$\frac{11}{11 + 9} * \frac{1}{0.8} * [0.8 * 3 + 0.2 * 2] = 1.925$$

Which also translates into a CCF of 1. SNV 4 illustrates that  $u_i$  must be rounded to obtain the multiplicity of a clonal SNV. It differs slightly from the expected value 2 because of variability in the number of reads due to limited sequencing depth. A similar mutation with 12 variant reads out of 20 would lead to an estimate of 2.100.

The accuracy of the multiplicity estimate in practice depends on the accuracy of the VAF and local copy number. Slight deviation in the VAF due to read sampling can result in minor deviation of the multiplicity estimates, as illustrated in the example above. Incorrect copy number profiles may also result in large errors if, for example, the CNA profile has been called as diploid instead of tetraploid. Ambiguity in estimating whole genome duplications is a difficult problem in copy number analysis. If a copy number profile is erroneously called as diploid then SNVs carried by two chromosome copies will be estimated to have a multiplicity of 1, while SNVs on 1 chromosome copy will become subclonal as they appear to be on 0.5 copies (e.g. exactly half of tumour cells). The CCF space will therefore show an SNV cluster at exactly 0.5,



while the copy number profile may also contain subclonal CNAs at exactly 50% of tumour cells. The uncertainty may be mitigated through the application of a key assumption: a CNA profile is thought to be in its normal state (diploid) unless substantial evidence of a whole genome duplication is available (i.e. the most parsimonious diploid state is assumed unless there is evidence otherwise). However in rare cases, when whole genome duplications occur late and are not followed by other copy number alterations, they leave no traces in the data and it is mathematically impossible to infer from the data available that they occurred.

We now have obtained a series of formulas to calculate CCF from a VAF and copy number profile. First, we obtain  $u_i$  through eq. 4 and then calculate the multiplicity and CCF using eqs. 3 and 2 respectively. Some methods cluster SNVs as a fraction of all cells in the tumour (CP space) and therefore need to make one more step:

$$CP_i = CCF_i \rho, \text{ Eq. 5}$$

Finally, some methods adjust the multiplicity to address SNVs that may appear subclonal due to a subclonal deletion. In these cases it is unknown whether the SNV occurred first and was then deleted in a fraction of cells, or the SNV occurred after the deletion. It is important to account for such subclonal deletions (e.g. by appropriately adjusting multiplicity estimates), and ensure that these subclonal deletions do not result in the inference of spurious subclonal populations.

### **SNV-based subclonal reconstruction**

SNV-based reconstruction methods cluster SNVs with a similar CCF or CP, derived from VAF values as described in the last section. However, the VAF of a SNV - and therefore also its CCF - can be a relatively coarse measure and is a function of local sequencing depth, which should be taken into account when clustering SNVs. For example, if the SNV falls in a region of diploid copy number with a depth of 20 reads in a sample with 50% tumour cells, its CCF changes by 0.2 when a variant read is added or removed (e.g. 3 mutant reads correspond to a CCF of 0.6, while 4 mutant reads correspond to a CCF of 0.8). If the same SNV is sequenced to 80X depth, one additional variant read would change the CCF by only 0.05. Tumours are often sequenced at 30X average coverage or higher, but this coverage is not constant across the genome. Due to this discrete sampling of mutant and non-mutant reads, and the variability of the sequencing depth,

CCF estimates of mutations from specific (sub)clones will show a distribution of values. For example, clonal mutations will display a range of CCF values around 1.0 (Figure 1C).

A suitable error model can account for this variability. The number of variant reads can be seen as the number of successes of  $N$  independent coin tosses, where  $N$  is the total read depth. The number of successes (variant reads) can therefore be modelled through a binomial distribution with  $r_i$  the number of reads reporting the variant at location  $i$ ,  $r_{tot,i}$  the total depth at location  $i$  and  $p_i$  the probability of observing a mutant read:

$$r_i \sim \text{Bin}(r_{tot,i}, p_i), \text{ Eq. 6}$$

Both  $r_i$  and  $r_{tot}$  are observed in the data.  $p_i$  can be considered the product of two factors: the proportion of reads one expects to see if the mutation is fully clonal,  $\zeta_i$ , and the true fraction of tumour cells carrying the mutation  $\pi_i$ :

$$p_i = \zeta_i \pi_i, \text{ Eq. 7}$$

$\zeta_i$  can be calculated from the tumour purity and the copy number state of the locus, as detailed above. Take for example a clonal SNV in a balanced diploid copy number region in a sequencing sample consisting of 80% tumour cells. The SNV is heterozygous and therefore expected to be carried by half the reads that represent tumour DNA. The expected proportion of reads is therefore  $0.5 * 0.8$ , i.e. 0.4. If the region has three copies and the SNV is carried by two copies, one expects two thirds of the reads representing tumour DNA to be carrying the variant allele, making the expected fraction  $2 * 0.8 / (3 * 0.8 + 2 * 0.2)$ , i.e. 0.57.

The key estimate in subclonal reconstruction is the true fraction of tumour cells that are carrying mutation  $i$ ,  $\pi_i$ . Many methods (Nik-Zainal *et al.* 2012; Landau *et al.* 2013; Roth *et al.* 2014; Jiao *et al.* 2014; Deshwar *et al.* 2015) use a Dirichlet Process, which models subclonal fractions as:

$$\pi_i \sim \text{DP}(\alpha P_0), \text{ Eq. 8}$$

where  $\text{DP}(\alpha P_0)$  is a Dirichlet Process with a given probability distribution  $P_0$  and a dispersion parameter  $\alpha$ . A realisation of a Dirichlet Process (DP) can be seen as a distribution over a (possibly) infinite sample space, or alternatively as a sampling from an unknown number of unknown distributions (Dunson 2010). This approach allows co-estimating both the number of contributing distributions  $n$  (the number of cellular populations) and their properties (fraction of

tumour cells and number of mutations they contain). The observed sampling  $P_0$  represents  $n$  of the (possibly) infinite number of distributions and can be used to estimate  $n$  (i.e. cellular populations) through the stick-breaking representation (Sethuraman 1994). Stick-breaking implies that the real probability distribution  $P$  can be expressed as follows:

$$P = \sum_{h=1}^{\infty} \omega_h \pi_{\theta_h}, \theta_h \sim P_0, \text{ Eq. 9}$$

where  $\pi_{\theta_h}$  is a location in CCF space and  $\omega_h$  represents the probability weight of cluster  $h$ :

$$\omega_h = V_h \prod_{l < h} (1 - V_l), \text{ Eq. 10}$$

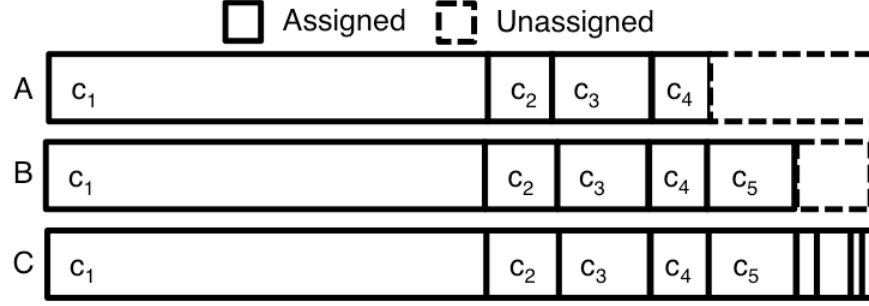
with

$$V_h \sim \text{Beta}(1, \alpha), \text{ Eq. 11}$$

The  $V_h$  represent parts of a unit length stick that are iteratively broken off from the remaining stick. The  $V_h$  get increasingly smaller as more parts are broken off, providing a discrete representation of an infinite space.

Figure 3 symbolizes the stick at various iterations of the stick-breaking procedure. Figure 3A and 3B show the stick after 4 and 5 breaks respectively, while Figure 3C shows it after completion. Each substick represents a fraction of the total weight (number of SNVs) of a cluster and can be assigned a CCF through resampling using the assigned SNVs. Then for each SNV and for each substick, a likelihood can be calculated representing the probability that that SNV is generated by that substick, taking the characteristics of the SNV, the stick location and its associated weight into account. After assigning all SNVs, the weights are updated such that they reflect the overall likelihood across SNVs.

The DP models an appropriate number of clusters because the assigned SNVs (influenced by the cluster weight) are used to resample the cluster CCF and the weight represents the fraction of total SNVs assigned to the cluster. By repeating this process over many iterations, the weight and SNV assignments will accumulate in certain locations that correspond to the estimated clusters. Therefore, the DP has the advantage that the number of clusters does not have to be specified *a priori*, making it ideally suited to this problem.



**Figure 3.** The stick-breaking property of the Dirichlet Process is used to estimate the number of mutation clusters in the data. For each mutation, a stick of arbitrary length is broken into randomly sized bits that represent a cluster. At point A, breaks have been introduced, corresponding to clusters  $c_1$ - $c_4$ . B shows the stick after introducing break 5, while C shows the completed stick-breaking procedure. The size of each broken part represents the weight associated with a cluster and influences the mutation assignments, where a high weight makes it more likely that a mutation is assigned to that cluster. These weights are updated after probabilities for each cluster have been obtained for each mutation, eventually converging on a solution.

Subclonal reconstruction also depends on the ability to call subclonal SNVs in a sequenced tumour. The number of reads required to call a SNV depends on the properties of the SNV caller (outside the scope of this text), and on the sequencing error rate distribution. As a rough rule of thumb, 3 mutant reads are typically required to detect an SNV, and mutations present in small fractions of tumour cells may be missed. The coverage at which the tumour was sequenced, the admixture of tumour and normal cells in the sequencing sample and the total amount of DNA from each tumour cell all contribute to the ability to detect clonal and subclonal mutations. The following formula combines these three factors into a power metric:

$$p_s = c_s \frac{\rho}{\rho\psi_t + (1-\rho)\psi_n}, \quad \text{Eq. 12}$$

Here,  $c_s$  is the sequencing coverage of the tumour sample,  $\rho$  is the tumour purity and  $\psi_t$  and  $\psi_n$  are the ploidy of the tumour and normal cells respectively (the amount of genomic material per cell, expressed in number of haploid genome copies).  $p_s$  is equivalent to the number of reads per chromosome copy and represents the expected number of reads reporting a clonal SNV. If for example  $p_s$  equals 10 and an SNV can be detected when there are 3 mutant reads, then (as an approximation) mutations present in >30% subclones can be detected.

The Dirichlet Process provides a flexible framework that has a built-in mechanism that restricts it from creating a large number of clusters, can incorporate a suitable error model to address variability due to read sampling and does not require specification of the number of clusters. Many methods are based on the above principles, including PyClone (Roth *et al.* 2014), PhyloSub (Jiao *et al.* 2014) and PhyloWGS, (Deshwar *et al.* 2015). Alternatively, SciClone uses a Variational Bayesian Mixture model that does not require a Markov Chain Monte Carlo approach, but does require specification of the number of clusters (Miller *et al.* 2014). CloneHD is based on a Hidden Markov Model and couples SNV and CNA data to perform subclonal reconstruction (Fischer *et al.* 2014).

### Copy number-based subclonal reconstruction

Subclonal reconstruction can also be performed using copy number changes. Somatic copy number callers often use read depth and/or the imbalance in the number of the maternal and paternal alleles to estimate copy number aberrations (Van Loo *et al.* 2010; Carter *et al.* 2012; Fischer *et al.* 2012; Nik-Zainal *et al.* 2012; Ha *et al.* 2014). To observe allelic imbalances, it is helpful to look at the B-allele frequency (BAF) of a germline heterozygous SNP. For sequencing data the BAF can be calculated as:

$$BAF_i = \frac{r_{B,i}}{r_{A,i} + r_{B,i}}, \quad \text{Eq. 13}$$

where  $r_{A,i}$  and  $r_{B,i}$  represent the total reads reporting allele A and B respectively. Alternatively, the BAF can be expressed as a function of the number of chromosome copies of allele A and B ( $n_A$  and  $n_B$  respectively):

$$BAF_i = \frac{n_{B,i}}{n_{A,i} + n_{B,i}}, \quad \text{Eq. 14}$$

A germline heterozygous SNP will have a BAF of approximately 0.5 in the absence of any copy number changes. Deviations from 0.5 therefore can be used to detect somatic aberrations.

As tumours are often admixed with normal cells, establishing the copy number state of an aberration based on the deviation of BAF requires estimating the fraction of tumour cells in the sample (the tumour purity). The number of chromosome copies in the formula above should therefore be split into a contribution of  $\rho$  tumour cells and  $(1 - \rho)$  normal cells:

$$BAF_i = \frac{\rho n_{B,t} + (1-\rho)n_{B,n}}{\rho(n_{A,t} + n_{B,t}) + (1-\rho)(n_{A,n} + n_{B,n})}, \text{ Eq. 15}$$

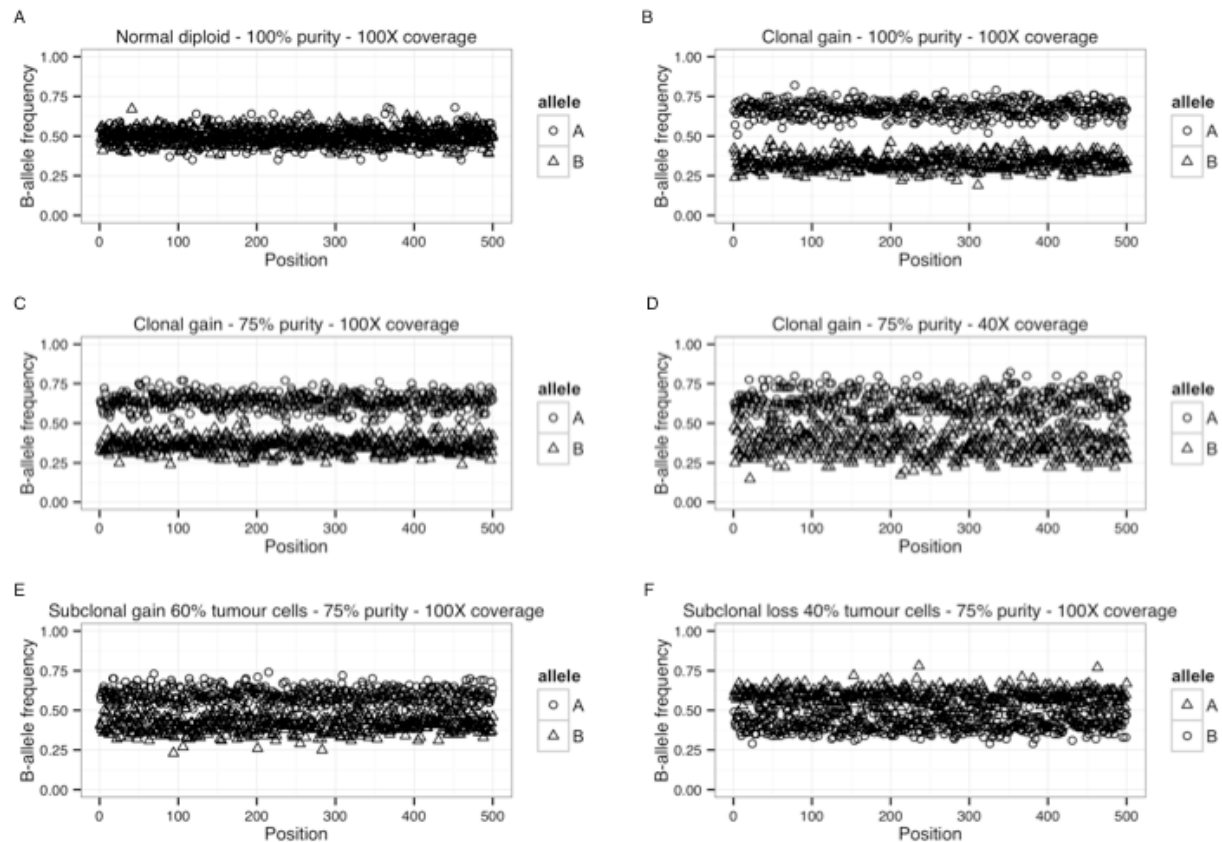
where  $\rho$  represents the tumour purity,  $n_{A,t}$  and  $n_{B,t}$  the number of chromosome copies in tumour cells and  $n_{A,n}$  and  $n_{B,n}$  the number of chromosome copies in normal cells. Several methods have been developed to co-estimate clonal copy number states and tumour purity based on these allele-specific signals (Van Loo *et al.* 2010; Carter *et al.* 2012; Ha *et al.* 2014).

Tumours that exhibit much clonal genomic instability will show deviation of the BAF for large proportions of the genome. In such tumours, the BAF values show clear levels corresponding to different clonal states, which translates into more usable signal for methods that co-estimate copy number states and tumour purity. However, genomes that show large amounts of subclonal genomic instability will show a range of different BAF values and will be more difficult to fit.

Figure 4 shows allele frequency values for a number of example cases that are affected by copy number changes and different normal cell admixtures. Panel A shows a region with no copy number alterations in a tumour that has no normal cell infiltration. One expects both alleles to be present in equal proportions, resulting in allele frequencies of 0.5. Panel B shows a region with a clonal gain. The bands representing allele A and B are clearly separated, with allele A representing two thirds of the total chromosome copies and allele B one third. Panel C contains a similar gain, but in a sample with 75% tumour purity, resulting in a smaller difference between the bands. Panel D shows the gain, again with 75% tumour cells, but now the coverage is reduced from 100X (as in panels A, B and C) to 40X. The bands appear to be overlapping as lowering the depth increases the noise and widens the bands. Panel E shows an example where the gain is subclonal in 60% of tumour cells resulting in further overlap of both bands. And finally panel F shows a subclonal loss in 40% of tumour cells.

Figure 4 illustrates that the allele frequencies of individual SNPs are subject to statistical variation and this noise increases with lower coverage. Combining SNPs into haplotype blocks through phasing can mitigate this effect (Carter *et al.* 2012; Nik-Zainal *et al.* 2012). Through haplotype phasing, information can be combined across multiple SNPs within a region of copy number change, by matching alleles across SNPs. For example, for SNP  $i$ , allele A may correspond to the maternal allele, while for SNP  $i + 1$ , allele B may correspond to the maternal allele. If these are combined appropriately, smaller deviations of the BAF from the normal state

can be detected, and higher precision copy number changes, including subclonal copy number changes, can be inferred.



**Figure 4. B-allele frequencies (BAF) of germline heterozygous SNPs can be used to identify copy number aberrations. A-F show that the BAF is noisy, and that it gets increasingly more difficult to separate the bands as the purity or coverage goes down and when the aberration is subclonal. To reduce the noise, SNPs can be phased to determine which allele is the B-allele. By combining the SNPs over longer stretches of DNA it becomes possible to detect subclonal aberrations.**

Once exact allele frequencies of segments have been calculated after haplotype phasing, subclonal copy number changes can be detected. As a first step, for each segment, one can determine whether the BAF value of this segment can be explained by a clonal copy number change. Deviation of the observed exact allele frequency from the theoretical allele frequency can be used to identify a segment having a subclonal copy number state, i.e. a combination of

two or more populations of tumour cells with different copy number states, in addition to a population of normal cells.

When such a segment is fit with a clonal copy number state, the multiple subclonal states are combined into a single (integer) representation. For example, if the real copy number state of the segment is 2+1 (2 copies of one parental allele and 1 copy of the other allele) in 80% and 1+1 in 20% of tumour cells (i.e. on average 1.8+1), its clonal fit will likely be 2+1 in 100% of tumour cells (1.8+1 rounded up). The observed allele frequency will therefore deviate from the frequency expected under the clonal copy number fit, allowing us to infer that the segment cannot be explained with a clonal copy number state.

Formally, given allele-specific copy number values  $n_A$  and  $n_B$  (integer if clonal, non-integer if subclonal), there are four options for the theoretical clonal allele frequency  $\hat{h}_f$  (assuming diploid copy number in the normal cell population):

Allele A and B are both rounded down:

$$\hat{h}_f = \frac{\rho[n_B] + 1 - \rho}{\rho([n_A] + [n_B]) + (1 - \rho)2}, \quad \text{Eq. 16}$$

Allele A is rounded up and B is rounded up:

$$\hat{h}_f = \frac{\rho[n_B] + 1 - \rho}{\rho([n_A] + [n_B]) + (1 - \rho)2}, \quad \text{Eq. 17}$$

Allele A is rounded down and B is rounded up:

$$\hat{h}_f = \frac{\rho[n_B] + 1 - \rho}{\rho([n_A] + [n_B]) + (1 - \rho)2}, \quad \text{Eq. 18}$$

Allele A and B are both rounded up:

$$\hat{h}_f = \frac{\rho[n_B] + 1 - \rho}{\rho([n_A] + [n_B]) + 2(1 - \rho)}, \quad \text{Eq. 19}$$

Subclonal segments can be identified by testing the observed allele frequency  $h_f$  against the theoretical  $\hat{h}_f$  values and accepting a segment as subclonal if the observed  $h_f$  is significantly different from  $\hat{h}_f$  in all four scenarios.

After inferring that the data for a given segment cannot be explained by any realistic clonal copy number state and, therefore, this segment must be a combination of two or more subclonal



populations with different copy number states, one can estimate the combination of subclonal copy number states for the segment. This depends on the different copy number states at the locus and their respective fractions of tumour cells. This problem has multiple solutions, as there can be any number of subclones with distinct subclonal copy number states. However, for any given segment, the most parsimonious assumption is that there are only two distinct copy number states, and that those copy number states differ at most by one chromosome copy (i.e. are separated by only one copy number event). Formally, if a fraction of tumour cells  $\tau$  shows copy number state  $n_{A,1} + n_{B,1}$  and a fraction of tumour cells  $1-\tau$  shows copy number state  $n_{A,2} + n_{B,2}$ ,  $\tau$  can be calculated as:

$$\tau = \frac{1-\rho+\rho n_{B,2}+2h_f(1-\rho)-h_f\rho(n_{A,2}+n_{B,2})}{h_f\rho(n_{A,1}+n_{B,1})-h_f\rho(n_{A,2}+n_{B,2})-\rho n_{B,1}+\rho n_{B,2}}, \quad \text{Eq. 20}$$

The principles outlined above are implemented in the Battenberg algorithm (Nik-Zainal *et al.* 2012). Other BAF-based methods apply similar metrics to detect deviation from clonal copy number. There are two different approaches to establish these values: event-based or population-based. Event-based callers, such as the Battenberg algorithm, aim to establish these values for each segment individually (Carter *et al.* 2012; Nik-Zainal *et al.* 2012), while population-based callers aim to explain as many segments as possible with a single subclonal fraction (Fischer *et al.* 2012; Ha *et al.* 2014).

It is also possible to estimate total copy number from read depth alone by binning reads across the genome and comparing the relative differences between bins with a matched normal sample. The advantage of methods such as Battenberg that rely heavily on BAF values is that allele frequencies are less affected by various biases that affect read depth (such as wave bias related to GC content and/or replication timing (Diskin *et al.* 2008; Koren *et al.* 2012)), as these biases affect both alleles equally and will therefore be cancelled out in the BAF calculation.

## Principles of phylogenetic tree reconstruction

Evolutionary relationships between subclonal populations can be inferred as well. Phylogenetic trees are often constructed building upon the *pigeonhole principle*, which states that if there are  $m$  containers (pigeonholes) and  $n$  items (pigeons) to be stored then there must be a container with more than one item if  $n > m$ . In subclonal reconstruction terms, the pigeonhole principle

states that no sum of subpopulations can exceed the CCF or CP of their ancestor (Beerenwinkel *et al.* 2014). For example, consider a subclonal reconstruction with mutation clusters at 100%, 80% and 40% of tumour cells. The pigeonhole principle determines that as  $100\% + 80\% > 100\%$ , the 80% cluster must represent a cellular population that is a descendant of the 100% population. Furthermore, as  $80\% + 40\% > 100\%$ , the population at 40% must be a descendant of the population at 80%. Therefore, the pigeonhole principle dictates a linear phylogeny for this example. In contrast, if the second cluster was found to represent 50% of tumour cells instead of 80%, the population at 40% could either be a descendant of the 50% population, or a parallel population (directly descending from the 100% population), as  $50\% + 40\% < 100\%$ .

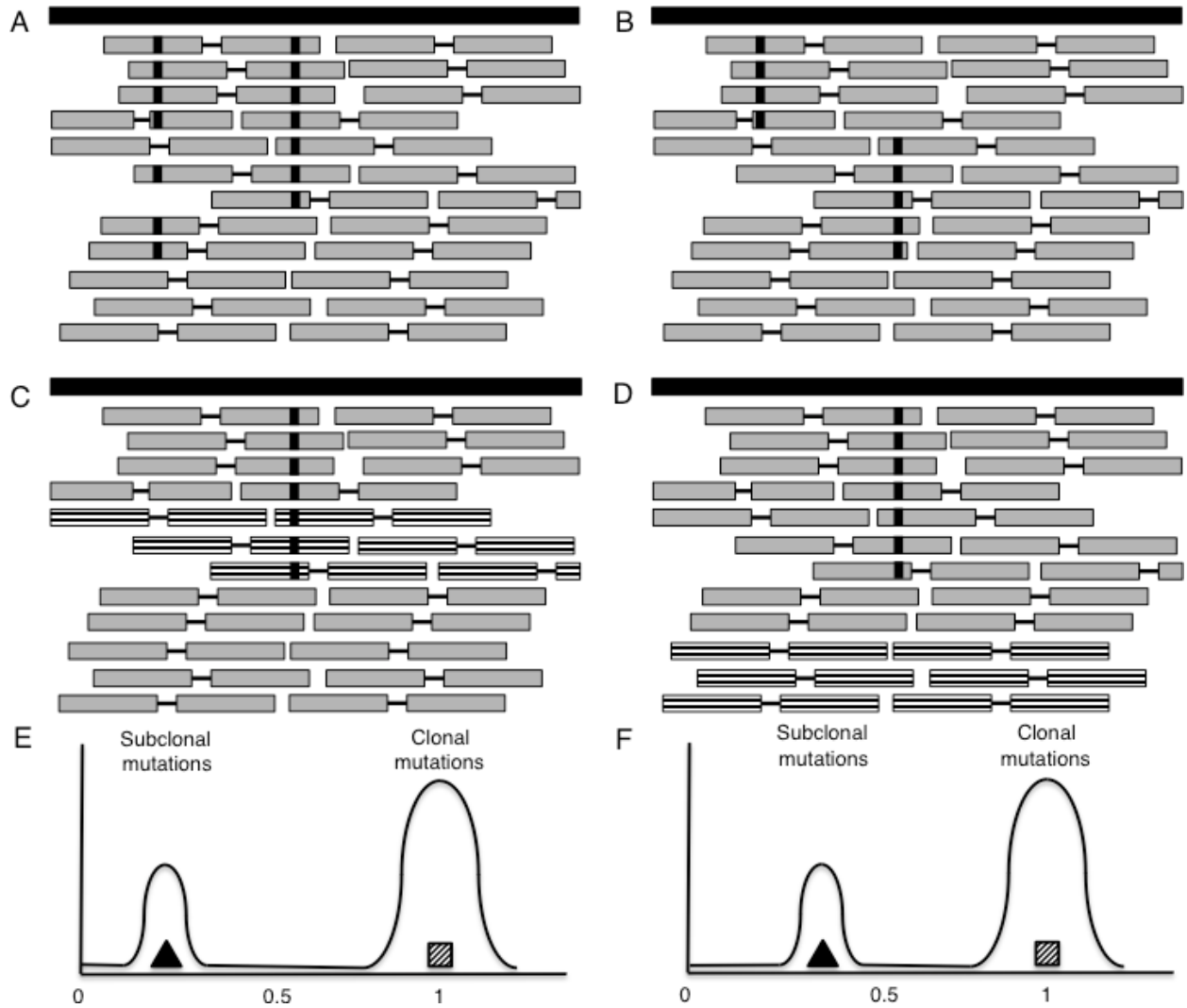
A corollary of the pigeonhole principle is that mutations in clusters with low CCF values are not necessarily part of the same subclone. For example, a tumour with two separate subclonal cell populations at 40% will appear to have only a single 40% subclone. Indeed, mutations that each of the founder cells of the subpopulations acquired at the start of the two separate clonal expansions will both appear at a CCF of 0.40. In such a scenario, all these mutations are clustered together and as there is no further information about the separate subclones are therefore (parsimoniously) assumed to represent a single cellular population. However, in such case, the pigeonhole principle does infer that there can be no more than two such subpopulations.

Both populations can however be separated through mutation phasing, or when a second sample is available. For example, when one of the two subclones has expanded into a metastasis, the metastasis sample will show these mutations at a CCF of 1.0, while the mutations in the other subclone are completely absent.

Mutation phasing can provide evidence for the existence of two separate cellular populations when a pair of mutations cannot have occurred in the same cell. For example, when a subclonal mutation is found on an allele that is lost in a specific subclone, one can infer that that mutation cannot be present in that subclonal lineage.

In general, to help resolve the tree topology, one could use SNVs that cannot have occurred in the same cell and therefore must be markers of distinct cellular lineages that correspond to branching evolution. One can phase SNVs against each other or against heterozygous SNPs and use fraction of tumour cells estimates to assess whether the aberrations can have occurred in the same lineage. Consider the phasing of two SNVs that are close enough to be spanned by a single

read pair. If there are read pairs showing both variants, then the SNVs must belong to the same lineage (be part of the same node in the tree or show an ancestor-descendant relationship), as there must then be at least one cell that has both variant alleles (Fig. 5A).



**Figure 5.** A pair of SNVs that are close enough to be covered by a single read pair and that have occurred in the same lineage appear as read pairs that contain both variant alleles (A). If the SNVs have originated in different lineages they appear in read pairs that contain the variant allele of one SNV and the wild type allele of the other (B). The VAF of a subclonal SNV that has been subclonally deleted (as illustrated by the striped reads that represent the deleted copies in C) is ‘shifted’ (E). An SNV that has occurred on the retained allele (i.e. the other allele is subclonally deleted, D) will not be shifted (F).

Using the same principle, one can also sometimes exclude that two mutations are on the same lineage, thereby showing that the tumour must have a branching phylogeny. Consider two SNVs (SNV 1 and 2) that are close enough to be spanned by a single read pair, in a region with only one chromosome copy in the tumour cells. If there are no read pairs reporting both variant alleles, but there are read pairs showing (i) SNV 1 and a wild type allele of SNV 2 and (ii) SNV 2 and a wild type allele of SNV 1, then the SNVs must be part of different branches of the phylogenetic tree (Fig. 5B).

It is also possible to infer the phylogenetic relationship between subclones by phasing SNVs and copy number changes. This principle is based on the observation that, for linear phylogenies, clusters of subclonal mutations that occur on a deleted allele will show different ('shifted') VAF values than clusters of subclonal mutations that occur on the other (retained) allele (illustrated in Fig. 5C and 5E). If the subclonal copy number change occurs in a different branch of the phylogenetic tree, no such shift should appear (Fig. 5D and 5F).

Nik-Zainal *et al.* (2012) first used phasing to create branched evolutionary stories, but to date there is no published method that can automatically leverage the potential of mutually exclusive mutations. In addition, it must be emphasised that with a haploid genome size of 3 Gb and typically a few thousand mutations per tumours, with current read lengths, mutation pairs that can be phased using the principles above are rare.

### **Single-cell-based approaches**

The advent of single cell sequencing theoretically gives access to a more fine-scaled level of tumour heterogeneity than bulk sequencing data. By sequencing a large number of single cells, one can gain in-depth understanding of the diversity within a tumour sample. However, in contrast to bulk sequencing methods, single cells are often sequenced at a much lower coverage. Further, the data suffers from uneven coverage and allele dropouts, errors introduced during whole genome amplification resulting in false positive SNV calls, and in some cases two cells are inadvertently sequenced together (called *doublets*) (Hou *et al.* 2012; Zong *et al.* 2012; Voet *et al.* 2013; Van Loo and Voet 2014). Subclonal reconstruction methods therefore must account for missing values due to mutant alleles being missed due to a lack of coverage or allelic dropout. A doublet effectively fuses two cells together and therefore contains mutations from

both cells. A subclonal reconstruction method might take this as misleading evidence where some cells carry one of the sets of mutations, while a doublet contains both.

Various approaches have recently been published to estimate the tumour phylogeny from single cell data (Kim *et al.* 2014; Yuan *et al.* 2015; Jahn *et al.* 2016; Ross and Markowitz 2016; Roth *et al.* 2016). Some methods infer phylogenies of cellular populations, in a similar manner to methods for bulk data. Others build mutation trees that instead show the order in which mutations have occurred without establishing cellular populations. These methods all start from a genotype matrix in which all mutations in all cells are represented. They (1) co-cluster evidence from bulk sequencing and single cell data using a Dirichlet Process (Yuan *et al.* 2015), (2) construct mutation lineages by a pair-wise mutation ordering (Kim and Simon 2014), (3) use heuristics to find a basic tree topology followed by clustering of single cells and further refinement (Ross and Markowitz 2016), or (4) build mutation trees using a Markov chain Monte Carlo based approach (Jahn *et al.* 2016). The recently developed Single Cell Genotypes (Roth *et al.* 2016) implements a robust feature allocation model to identify subclones with shared genotypes and to infer the genotype of each subclone. This method elegantly accounts for missing data due to single cell sequencing limitations such as allelic dropout, as well as for the occurrence of doublets.

### **Multi-sample-based approaches**

Obtaining multiple samples from the same donor allows for extraction of more detailed subclonal reconstructions. These datasets can consist of multiple tumours taken from different sites (e.g. multiple primary sites, primary and metastasis), multiple samples from the same tumour or multiple samples from the same cancer that represent different time points (e.g. primary and relapse).

Multiple sampling strategies provide a series of advantages. Consider a tumour that has two subclones that each comprise 20% of tumour cells. A single sample analysis will not be able to separate the two groups of mutations as both occur in 20% of tumour cells. But if in another sample the cellular prevalence of the two subclones does vary, one can separate the two groups of mutations. In addition, having multiple samples may help resolve tree topologies. In single sample cases it is often not possible to resolve phylogeny, as more rare subclones may be placed

in multiple positions in the tree. By applying the pigeonhole principle across the samples for each subclone, one can often rule out various configurations where a subclone may fit in multiple places in one sample, but not the other. Finally, with multiple sampling strategies, mutations with low allele fractions in one sample can be confirmed (or detected) in another sample where they have higher allele fractions due to higher tumour purity or higher CCF.

Approaches based on a DP can be extended into multiple dimensions (Bolli *et al.* 2014). The read counts across  $n$  samples can be modelled as independent draws from  $n$  Binomial distributions.

$$r_{i,1} \sim \text{Bin}(r_{tot,1}, p_{i,1}) , \quad \text{Eq. 21}$$

$$r_{i,n} \sim \text{Bin}(r_{tot,n}, p_{i,n}) , \quad \text{Eq. 22}$$

The stick-breaking procedure is performed across the samples where a cluster has a single weight (representing the number of mutations), but a separate location in each of the samples. Posteriors are obtained across samples by calculating the total probability for each mutation for each cluster under consideration. Finally, the DP can be used to jointly perform clustering and infer phylogenetic relationships between the clusters by interleaving two stick-breaking procedures (Ghahramani *et al.* 2010).

Several methods for single sample analysis, including PyClone (Roth *et al.* 2014), SciClone (Miller *et al.* 2014) and CloneHD (Fischer *et al.* 2014), can be used to analyse multiple samples. Furthermore, automated tree inference has been implemented in PhyloSub (Jiao *et al.* 2012) and extended to include SNVs in copy number aberrant regions in PhyloWGS (Deshwar *et al.* 2015).

### **Biological insights obtained through subclonal architecture reconstruction**

Subclonal reconstruction analysis has recently been used to reveal insights into the complexities of tumour evolution in a number of cancer types. Various papers focussed on a single cancer type report vast differences in heterogeneity between patients, where known genes are mutated early in one case, but late in another (Yates *et al.* 2015) and that some tumours can show evidence of rapid evolution, while other tumours in the same cohort show a stable balance between subclones (Schuh *et al.* 2012).

The application of treatment can introduce a phase of rapid tumour evolution (Landau *et al.* 2013; Landau *et al.* 2015), where mutations in known drivers are observed to be subclonal (Landau *et al.* 2013; Bolli *et al.* 2014; Gerlinger *et al.* 2014). Mechanisms of resistance can be acquired in parallel in different lesions (Gerlinger *et al.* 2014; Gudem *et al.* 2015), subclones can persist through treatment (Schuh *et al.* 2012) and the existence of a subclonal driver mutation can be an independent risk factor for disease progression (Landau *et al.* 2013).

A primary tumour can contain observable signs of metastatic and treatment resistance potential before onset (Yates *et al.* 2015) and in some cases can contain patterns that predict the evolutionary progression (Landau *et al.* 2015). Mutational processes can differ between clones and subclones through spatially (De Bruin *et al.* 2015) and temporally (Bolli *et al.* 2014) separated samples from the same cancer. Gudem *et al.* (2015) reported metastasis-to-metastasis seeding in a number of lethal metastatic prostate cancers and Cooper *et al.* (2015) observed clonal expansions in morphologically normal cells in multi-focal prostate tumours.

These separate studies hint that intra-tumour heterogeneity is widespread and that tumours of the same cancer type can differ greatly. McGranahan *et al.* (2015) reported that subclonal mutations in known drivers are common across 5000 exome-sequenced tumours representing 10 cancer types from The Cancer Genome Atlas (TCGA). Andor *et al.* (2016) performed subclonal reconstruction on 1165 exome-sequenced tumours from TCGA and report that 86% of tumours across 12 cancer types contain at least one subclone. Larger studies as well as systematic pan-cancer studies are required to further our insight into whether there are distinguishable patterns of tumour evolution.

Through large-scale international efforts such as TCGA and the International Cancer Genome Consortium (ICGC), a wealth of whole exome and whole genome sequencing data has been generated, most of which presently has not been mined from the perspective of evolution. We expect that efforts such as the ICGC Pan-Cancer Analysis of Whole Genomes have the potential to significantly broaden our understanding of tumour's subclonal architecture and evolutionary history. For subclonal inference, whole-genome sequences show clear advantages over exome sequences, as they allow detection of nearly two orders of magnitude more mutations, and as more detailed (and subclonal) copy number changes can be inferred. We expect that large-scale

multi-sampling whole-genome sequencing approaches across cancer types will lead to key evolutionary insights. However, the cost of sequencing is at present still a limiting factor.

There is also a need for smaller, more focussed studies with the right data to further deepen our understanding of the factors that play a role in tumour progression, how they interact together and what that means for patient care. Subclonal reconstruction methods have already enabled important discoveries that are of direct clinical interest. We expect that these types of studies will play a key role in further advancing our understanding of tumour evolution.



## References

- Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, Ji HP, Maley CC. 2016. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* **22**:105–113.
- Beerenwinkel N, Schwarz RF, Gerstung M, Markowitz F. 2015. Cancer Evolution: Mathematical Models and Computational Inference. *Systematic Biology*, **64**(1):e1–e25.
- Bolli N, Avet-Loiseau H, Wedge DC, Van Loo P, Alexandrov LB, Martincorena I, Dawson KJ, Iorio F, Nik-Zainal S, Bignell GR, et al. 2014. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature Communications*, **5**. doi: 10.1038/ncomms3997.
- Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, Follows GA, Green AR, Futreal PA, and Stratton MR. 2008. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci* **105**(35):13081–13086.
- Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, et al. 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, **30**(5):413–421.
- Cooper CS, Eeles R, Wedge DC, Van Loo P, Gundem G, Alexandrov LB, Kremeyer B, Butler A, Lynch AG, Camacho N, et al. 2015. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nature Genetics*, **47**(4):367–372.
- De Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, Jamal-Hanjani M, Shafi S, Murugaesu N, Rowan AJ, et al. 2014. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, **346**(6206):251–256.

Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. 2015. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, **16**:35.

Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K. 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research*, **36**(19):e126–e126.

Dunson DB. 2010. Nonparametric Bayes Applications to Biostatistics. *Bayesian Nonparametrics*. Hjort NL, Holmes C, Müller P, Walker SG. Cambridge University Press, Cambridge.

Fischer A, Vzquez-Garca I, Illingworth CR, Mustonen V. 2014. High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Reports*, **7**(5):1740–1752.

Garraway L, Lander E. 2013. Lessons from the Cancer Genome. *Cell*, **153**(1):17–37.

Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, Fisher R, McGranahan N, Matthews N, Santos CR, et al. 2014. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, **46**(3):225–233.

Ghahramani Z, Jordan MI, Adams RP. 2010. Tree-Structured Stick Breaking for Hierarchical Data. *Adv. Neural Inf. Process. Syst. (NIPS)*, **23**:19–27.

Greaves M, Maley CC. 2012. Clonal evolution in cancer. *Nature*, **481**(7381):306–313.

Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, Brewer DS, Kallio HML, Högnäs G, Annala M, et al. 2015. The evolutionary history of lethal metastatic prostate cancer. *Nature*, **520**(7547):353–357.

Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, Melnyk N, McPherson A, Bashashati A, Laks E, et al. 2014. TITAN: Inference of copy number architectures in clonal cell populations from tumor whole genome sequence data. *Genome Research*, **24**(11):1881-1893.

Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, Li F, Wu K, Liang J, Shao D, et al. 2012. Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. *Cell*, **148**(5):873–885.

Jahn K, Kuipers J, Beerenwinkel N. 2016. Tree inference for single-cell data. *Genome Biol* **17**: 86.

Jiao W, Vembu S, Deshwar AG, Stein L, Morris Q. 2014. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, **15**(1):35.

Kim KI, Simon R. 2014. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics*, **15**:27.

Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. 2012. Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. *The American Journal of Human Genetics*, **91**(6):1033–1040.

Landau D, Carter S, Stojanov P, McKenna A, Stevenson K, Lawrence M, Sougnez C, Stewart C, Sivachenko A, Wang L, et al. 2013. Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell*, **152**(4):714–726.

Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, Kluth S, Bozic I, Lawrence M, Bttcher S, et al. 2015. Mutations driving CLL and their evolution in progression and relapse. *Nature*, **526**(7574):525–530.

McGranahan N, Favero F, de Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. 2015. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*, **7**(283):283ra54–283ra54.

Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, Vij R, Tomasson MH, Graubert TA, Walter MJ, et al. 2014. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Comput Biol*, **10**(8):e1003665.

Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, et al. 2012. The life history of 21 breast cancers. *Cell*, **149**(5):994–1007.

Nowell PC. 1976. The clonal evolution of tumor cell populations. *Science* **194**:23–28.

Ross EM, Markowetz F. 2016. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, **17**:69.

Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Ct A, Shah SP. 2014. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, **11**(4):396–398.

Roth A, McPherson A, Laks E, Biele J, Yap D, Wan A, Smith MA, Nielsen CB, McAlpine JN, Aparicio S, Bouchard-Ct A, Shah SP. 2016. Clonal genotype and population structure inference from single-cell tumor sequencing. *Nature Methods*. doi: 10.1038/nmeth.3867.

Schuh A, Becq J, Humphray S, Alexa A, Burns A, Clifford R, Feller SM, Grocock R, Henderson S, Khrebtukova I, et al. 2012. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* **120**: 4191–4196.

Sethuraman J. 1994. A Constructive Definition of Dirichlet Priors. *Statistica Sinica* **4**:639–650.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. *Nature*, **458**(7239):719–724.

Tabin CJ, Bradley SM, Bargmann CI, Weinberg RA, Papageorge AG, Scolnick EM, Dhar R, Lowy DR, Chang EH. 1982. Mechanism of activation of a human oncogene. *Nature*, **300**(5888):143–149.

Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, et al. 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci* **107**:16910–16915.

Van Loo P, Voet T. 2014. Single cell analysis of cancer genomes. *Curr Opin Genet Dev* **24**:82–91.

Voet T, Kumar P, Van Loo P, Cooke SL, Marshall J, Lin M-L, Esteki MZ, Van der Aa N, Mateiu L, McBride DJ, et al. 2013. Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic Acids Research*, **41**(12):6119–6138.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. 2013. Cancer Genome Landscapes. *Science*, **339**(6127):1546–1558.

Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, Aas T, Alexandrov LB, Larsimont D, Davies H, et al. 2015. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine*, **21**(7):751–759.

Yuan K, Sakoparnig T, Markowetz F, Beerenwinkel N. 2015. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biology*, **16**:36.

Zong C, Lu S, Chapman AR, Xie XS. 2012. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science*, **338**(6114):1622–1626.