OPEN ACCESS
CellPress

Cell Reports
Resource

# High-Definition Reconstruction of Clonal Composition in Cancer

Andrej Fischer,[1,*] Ignacio Vázquez-García,[1,2] Christopher J.R. Illingworth,[3] and Ville Mustonen[1,*]
[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
[2]Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK
[3]Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK
*Correspondence: af7@sanger.ac.uk (A.F.), vm5@sanger.ac.uk (V.M.)
http://dx.doi.org/10.1016/j.celrep.2014.04.055


## SUMMARY

**The extensive genetic heterogeneity of cancers can greatly affect therapy success due to the existence of subclonal mutations conferring resistance. However, the characterization of subclones in mixed-cell populations is computationally challenging due to the short length of sequence reads that are generated by current sequencing technologies. Here, we report cloneHD, a probabilistic algorithm for the performance of subclone reconstruction from data generated by high-throughput DNA sequencing: read depth, B-allele counts at germline heterozygous loci, and somatic mutation counts. The algorithm can exploit the added information present in correlated longitudinal or multiregion samples and takes into account correlations along genomes caused by events such as copy-number changes. We apply cloneHD to two case studies: a breast cancer sample and time-resolved samples of chronic lymphocytic leukemia, where we demonstrate that monitoring the response of a patient to therapy regimens is feasible. Our work provides new opportunities for tracking cancer development.**

## INTRODUCTION

Cancer develops via the accumulation of genetic alterations during an evolutionary process (Stratton et al., 2009). Recent years have seen a torrent of genetic data from cancer genomes generated at different levels of resolution ranging from low-density genotyping array data for gauging copy-number profiles to whole-genome sequencing to capture all genetic aberrations. These data have been hugely informative in discovering driver mutations that are causally responsible for the development and progression of cancer (Garraway and Lander, 2013; Vogelstein et al., 2013; Wheeler and Wang, 2013). However, the ascent of cancer genomics has not been without formidable challenges. For instance, a breast cancer can harbor thousands of point mutations together with some smaller number of large-scale copy-number alterations (Nik-Zainal et al., 2012a; Stratton et al., 2009). Out of these variants, most are likely to be of no great relevance for the cancer phenotype of the cell and are considered passengers. Even focusing solely on nonsynonymous coding variants would leave tens to hundreds of mutations for further analysis depending on the cancer type (Garraway and Lander, 2013; Vogelstein et al., 2013; Wheeler and Wang, 2013). But such a drastic filtering would also carry the risk of missing some important mutations, as was underlined by a recent discovery of *TERT* promoter mutations driving melanoma (Horn et al., 2013; Huang et al., 2013). An obvious computational challenge is to prioritize candidate causal variants for follow-up functional validation (Gonzalez-Perez et al., 2013). The sheer volume of data is of help in achieving this aim in a virtuous cycle; for instance, it is now possible to determine region-specific mutation rates by pooling gene-activity data to construct a baseline model for subsequent driver-gene detection (Lawrence et al., 2013). These high-resolution statistical models are pushing forward the field of cancer genomics as a whole.

While progress has been made in understanding the vast number of mutations in sequenced tumor samples and the processes generating them (Alexandrov et al., 2013a, 2013b; Fischer et al., 2013; Lawrence et al., 2013; Nik-Zainal et al., 2012a), another layer of variability has been discovered in the form of subclonal population structure. It is often the case that a sample of cells from a single tumor can not be considered as an isogenic lineage of cancer cells with stromal contamination, not even to a first approximation (Burrell et al., 2013). The fraction of cancerous cells rather consists of a collection of subclones, with private and shared mutations, related by their joint evolutionary history going back to the most recent common ancestor (Nik-Zainal et al., 2012b) (see Figure 1). Clonal heterogeneity can be detected using next-generation DNA sequencing (Shah et al., 2009) and has important biological and medical implications (Aparicio and Caldas, 2013; Bedard et al., 2013). First, naive sample extraction strategies will lead to an underestimation of real tumor heterogeneity. Second, subclones can be resistant to a particular therapy and are then amplified in a process called competitive release, whereby the drug eradicates any susceptible competitors (Greaves and Maley, 2012; Wargo et al., 2007). Rather than waiting for de novo resistance mutations to emerge, cancer likely escapes using existing subclonal variation (Bozic et al., 2012). As a result, clonal dynamics and changes in clonal composition can inform therapy, highlighting the importance of

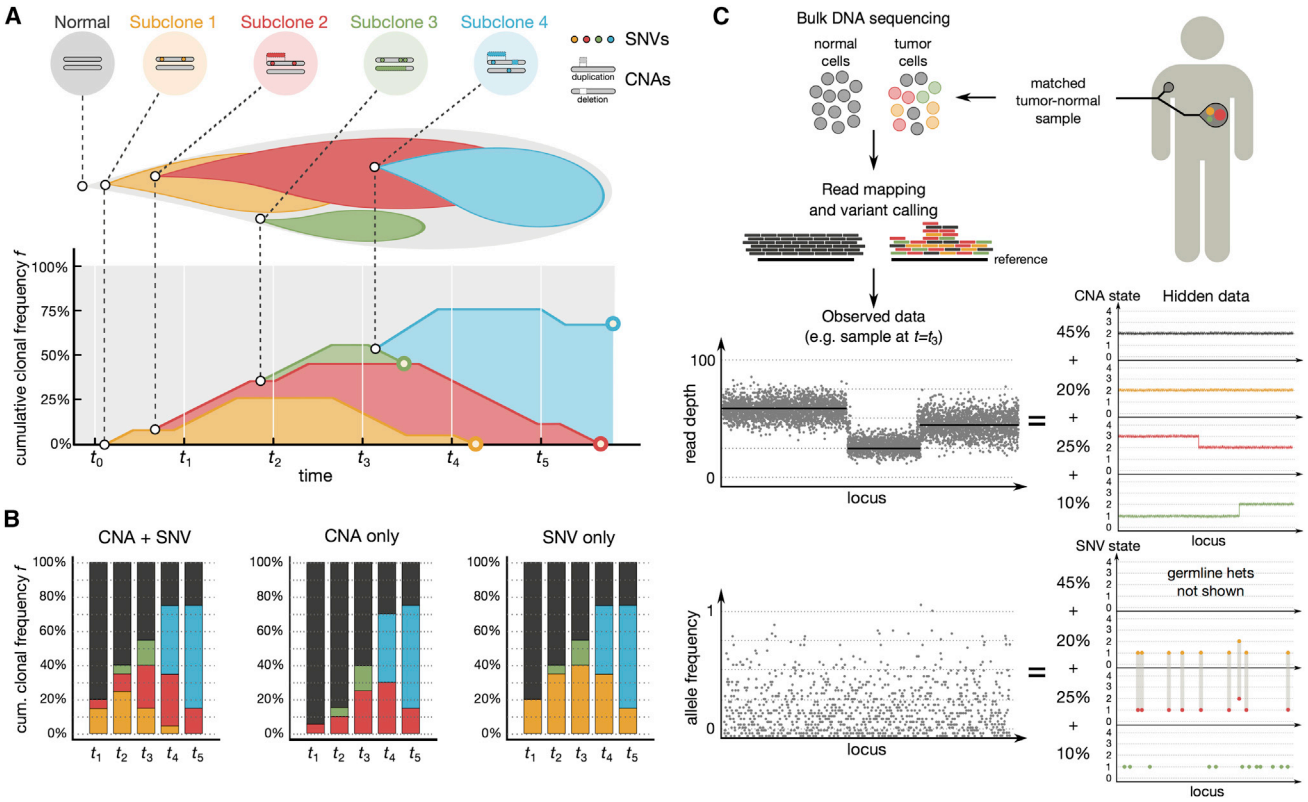


**Figure 1. Reconstruction of Clonal Heterogeneity**

(A) Schematic view of subclonal diversification. In this example, mutations in daughter cells of a single founder cell (left) diverge into subclones (reflected by different colors). A point mutation occurs early on with a subsequent gain of a chromosome arm and a short deletion at a later stage, each followed by clonal expansion (subclones 1, 2, and 4). A short-lived lineage arises independently (subclone 3).

(B) In the left column, the demarcation of the clonal lineages using CNA and SNV information is shown. Middle and right columns show different decompositions using only one of these data types.

(C) DNA sequencing of a cell population in a tumor sample and a matched normal. The different data layers (left) can be used to infer the underlying population structure (right); vertical lines highlight shared SNVs.

monitoring cancer progression. While emerging single-cell technologies are showing great promise, it is still not possible to sequence individual cells routinely to capture the full information about their genotype and copy-number profiles (e.g., Navin et al., 2011; Potter et al., 2013; Shapiro et al., 2013). This leaves the field reliant on short-read sequencing as the main experimental assay for cancer genomics in the near future. Therefore, computational inference of subclonal population composition from short-read data is an important challenge.

Existing computational methods have so far mostly focused on the decomposition of the sample into tumor and normal cells, estimating its purity while trying to account for an aberrant ploidy of the tumor cells. Early attempts of purity and mean ploidy estimation were designed for SNP array data and used relative read depth and/or B-allele fractions (such as Rasmussen et al., 2011; Song et al., 2012; Van Loo et al., 2010; Yau et al., 2010). The value of using correlations along the genome was realized in some methods employing hidden Markov models (HMMs) (Greenman et al., 2010; Li et al., 2011; Liu et al., 2010; Sun et al., 2009). Several methods for purity estimation have been reviewed and compared (Mosén-Ansorena et al., 2012). More recent computational methods try to leverage the large amount of information gathered in next-generation sequencing (NGS) to estimate tumor purity and characterize tumor ploidy (Carter et al., 2012; Chen et al., 2013; Larson and Fridley, 2013; Su et al., 2012). While these methods increasingly use probabilistic modeling, including HMMs (Ha et al., 2012), to account for noisy data, they do not infer individual subclonal fractions and copy-number profiles and often include only one or two of the available data types. If methods for purity estimation assume a fully clonal tumor population, they can give unreliable results if there is considerable subclonality. Models that try to account for possible subclonality can produce more robust estimates (Carter et al., 2012; Chen et al., 2013; Larson and Fridley, 2013).

More recently, a few studies have started to infer the subclonal structure from NGS data. In the analysis of breast cancer genomes (Nik-Zainal et al., 2012b), the histogram of observed single-nucleotide variants (SNVs) has been explained with a small number of mutation clusters using a Dirichlet process. These clusters are then used to manually derive a consistent phylogenetic tree. This ansatz has recently been extended to the case of multiple samples (Bolli et al., 2013). The THetA

algorithm (Oesper et al., 2013) uses genome-wide segmented read-depth information to find mixtures of subclonal copy-number profiles. For the limit case of a fully clonal tumor, THetA outperforms previous purity estimators and runs efficiently. Since the inference in THetA is based on read depth alone, there can be several equivalent copy-number profiles explaining the data. The PyClone algorithm (Roth et al., 2014), on the other hand, tries to deconvolve the tumor into subclones based on somatic SNVs from deep sequencing using a hierarchical Bayesian clustering model that can incorporate local copy-number information. Within the same category, the recent PhyloSub algorithm uses deeply sequenced SNVs and phylogenetic tree constraints to infer subclonal frequencies (Jiao et al., 2014).

We here describe a probabilistic algorithm, cloneHD, to perform subclone reconstruction from short-read data. Our algorithm offers three qualitatively new additions that differentiate it from existing methods.

First, our method addresses the clonal inference problem using data of multiple types, both at the level of copy-number aberrations (CNAs), using read depths and B-allele fractions (BAFs) (denoted cna-mode, cna-baf-mode) and at the level of somatic SNVs (denoted snv-mode). Inferences are performed with a set of coupled hidden Markov models *jointly* across all data types, which can greatly improve the evidence for one of several competing solutions. However, the resulting clonal decompositions need not be the same at the two levels. For instance, two clones with identical copy-number profiles can still have different somatic mutations. In snv-mode, cloneHD tries to find this alternative partitioning while respecting the overall copy-number status of the population. We can also perform an integrative analysis to seek a clonal decomposition jointly at the level of CNAs, BAFs, and SNVs (i.e., cna-snv-mode and cna-baf-snv-mode).

As a result of this, the method generates inferences of the number of clones detected in the sample, their population frequencies across time and/or space, and subclone-specific posterior probabilities of copy-number profiles and somatic variant genotypes. To our knowledge, the calculation of *consistent* locus- and subclone-specific posterior probabilities across all three levels, namely total copy number, B-allele status, and SNV genotype, has not been done before. This reconstruction facilitates subclone-specific computational analyses at high definition and so opens new ground for exploration.

Finally, our algorithm is designed to take account of data where multiple samples have been sequenced from a single patient. Our approach exploits correlations across time (longitudinal data) or across space (multiregion and/or metastatic samples).

To achieve computational efficiency, the algorithm employs a fuzzy data segmentation scheme, which coarse-grains the data while retaining most of the correlation information (see Experimental Procedures, Supplemental Experimental Procedures, and Figure S1 for details). The inference of two subclones in a single whole-genome cancer sample at full data resolution (1 kb) with hundreds of segments can thus be performed within minutes running cloneHD on a standard personal computer.

In the following, we demonstrate the performance of the approach using simulated data and two case studies: a breast cancer sample (Nik-Zainal et al., 2012b) and time-resolved samples of chronic lymphocytic leukemia (Schuh et al., 2012).

## RESULTS

### Benchmarking against Simulated Data

We first measured the performance of cloneHD by running it against simulated data to demonstrate that the algorithm can successfully infer the number of subclones and their frequencies and reconstruct their copy-number and somatic SNV states. The simulation ensemble consisted of 100 data sets, where each sample had a purity of 0.7 (i.e., 30% of the DNA was derived from noncancerous cells) and further contained two cancer subclones of variable sizes. In our simulations, we constructed evolutionary trajectories where the initially smaller subclone gained in frequency over time while the other subclone decreased. Their copy-number profiles and SNV genotypes, and therefore their subclonal identity, remained the same at all times. Each simulated genome was composed of 20,000 loci, out of which a mean of 2,500 loci contained somatic mutations and a mean of 2,342 loci were germline heterozygous (generating B-allele counts). These simulated cancers were "sequenced" at a depth of 15X (fold coverage, the average number of reads representing a nucleotide) per haploid chromosome at up to three time points in their evolution (for detailed description of the simulations, see Supplemental Experimental Procedures). With real data, the effective sequencing depth is not exactly known. In cloneHD, this parameter is learned as a sample-specific mass, defined as the mean sequencing depth per haploid chromosome. Figure 2 shows one representative simulated data set and its particular explanation using cloneHD.

We considered two main performance measures for the inferences: reconstruction fidelity and mean error in clone frequencies per sample point. We defined the fidelity of an inference as the amount of posterior probability per locus assigned to its true state. For instance, fidelities close to one mean that the algorithm has correctly reconstructed the copy-number profile or somatic SNV genotypes for a subclone. Although this metric is a useful indicator for the overall performance of subclonal reconstruction it has some limitations. For example, even with perfect clonal frequencies, at low sequencing depths and/or a small number of samples, a substantial uncertainty about the hidden state remains and cannot be removed without more data. This is especially the case for the somatic SNV genotype state, which in general has no persistence along the genome. Such uncertainty reflects the inherent limits of inference rather than any shortcomings of the reconstruction algorithm. In Figure 3A, we show that cloneHD successfully reconstructed the clonal copy-number states and somatic SNV genotypes from the simulated data, obtaining fidelities close to the maximum achievable given the noise level. As expected, the performance increases when more samples (time points in the simulations) and/or data types (cna-mode, cna-baf-mode, cna-baf-snv-mode) are added.

We note that our model selection criterion (Bayesian information criterion [BIC]) undercalled the number of clones in up to 13 of 100 runs (cna-mode). With additional time points, this underestimate disappeared and there was some overcalling in up to 6 of 100 runs (cna-baf-mode). However, the clones that were identified in the miscalled runs have meaningful fidelities, while some smaller clones are missed. In analyzing real data, we suggest

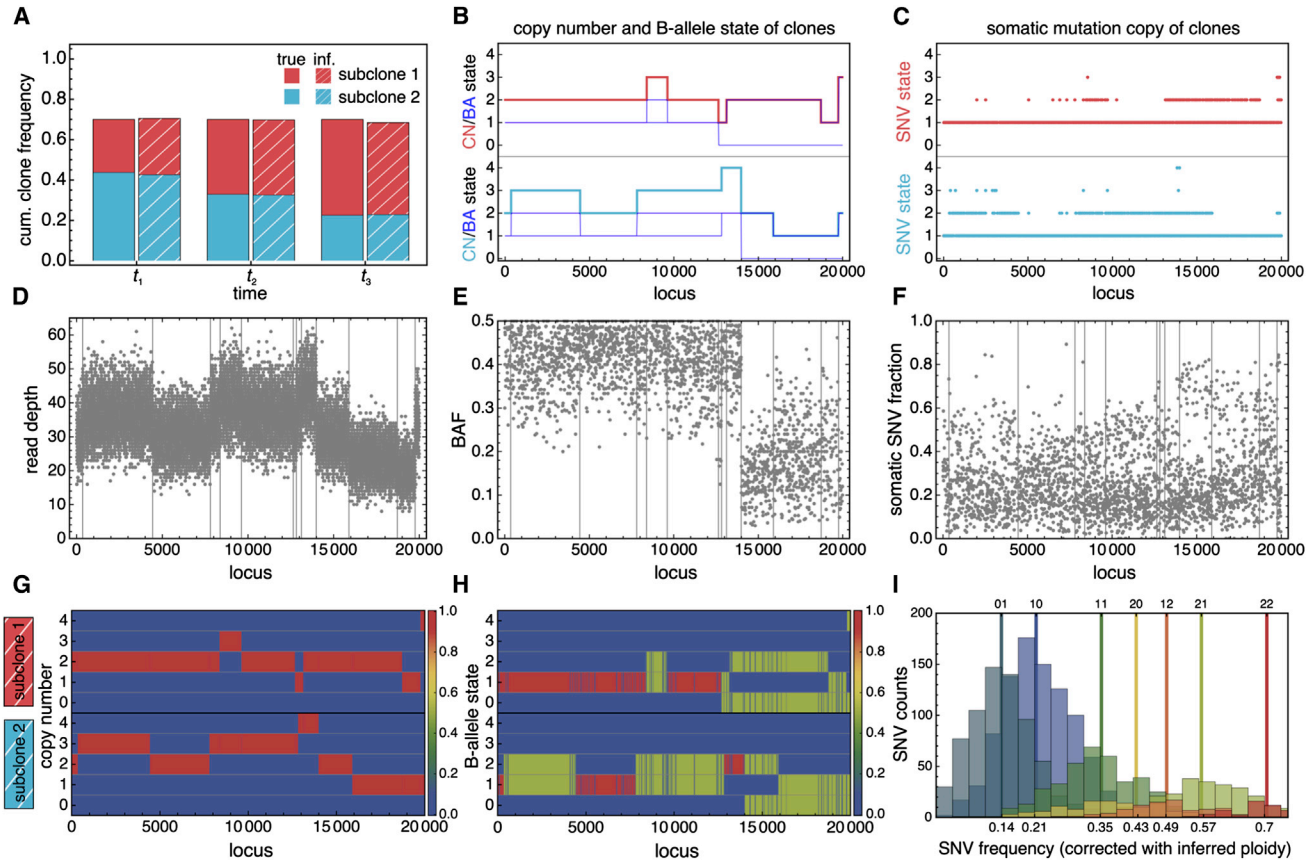


**Figure 2. Example Reconstruction of a Simulated Data Set**

(A)–(C) show the hidden data that are inferred from the sequencing experiment.

(A) An example simulated data set containing two subclones (red and blue) with fractions summing up to 0.7, a contamination from normal cells of 0.3, and with three time points.

(B) True underlying copy-number states for the cancer subclones, where the B-allele copy number is shown in dark blue (two values reflect that we do not know the parental chromosome of origin for these variants).

(C) Somatic variants for each subclone.

(D–F) Read-depth track, here 15X per haploid chromosome, from the mixture of normal plus subclones that can be used to infer copy-number profiles of the subclones (we show true copy-number changes as vertical lines to guide the eye) (D), B-allele counts (fractions plotted) for the mixture help to decide between balanced and unbalanced copy-number changes (E), and somatic mutation counts (fractions plotted) (F). (D)–(F) show only the data for time point $t_1$, while there are two more sets of data guiding the inference (not shown). The bottom row shows the cloneHD inference output using cna-baf-snv-mode for this data set. As already shown in (A), inferred clone frequencies closely match the input.

(G) The posterior probability of subclone-specific copy-number states closely matches the true profiles shown in (B) (there is only a short segment that is wrongly assigned).

(H) The posterior probability of subclone-specific B-allele states closely matches the true profiles shown in (B).

(I) For each SNV, the observed allele fraction was scaled by one half of its local mean total copy number and a genotype state was randomly assigned based on the cloneHD posterior. The histograms for the most prevalent states are shown. The vertical lines denote for each genotype state the predicted frequency in diploid DNA (even for genotypes higher than 2, where this number could go beyond one).

that BIC be regarded as an informed heuristic, with emphasis being placed on the stability (or lack thereof) of the solution when changes are made to the total number of subclones or the copy-number range. Figure 3B shows that the mean error per sample between the true frequencies and the inferred ones is small and decreased as a function of sample points and with the addition of data types. This result is clearly not independent of the fidelity and shows how closely the underlying subclonal dynamics can be learned. We also note that inferences where the mass was not accurately captured often show poor fidelities because the solution found differs from the correct copy-number profile by an overall shift (typically by one copy).

In summary, cloneHD can successfully reconstruct subclonal frequencies and the underlying copy-number profiles and SNV genotypes from complex simulated mixtures. Although these simulated data sets provide a demanding test for our algorithm, they are not ideal; it is not clear how comparable they are to real cancer cell populations. Biological data sets cannot be expected to follow specific emission models verbatim. However, our choices for the simulations were set with the motivation

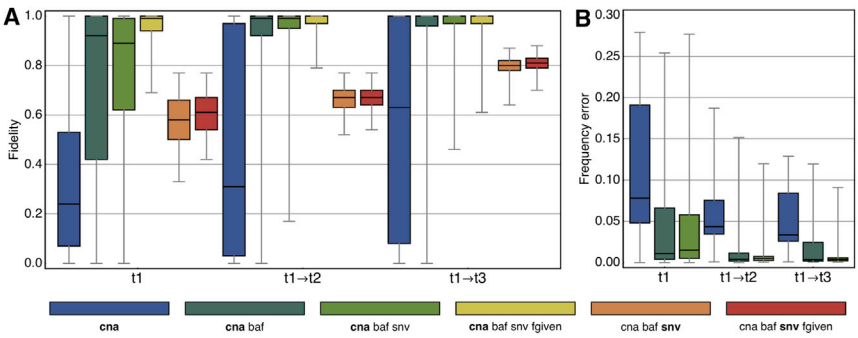


**Figure 3. Benchmarking of cloneHD against a Simulated Data Set**

Inferences of 100 evolutions with two cancer subclones and a purity of 0.7 demonstrate strong performance in the reconstruction of subclonal copy-number profiles, genotypes, and frequencies. In the box plots, bars denote minimum and maximum values, while areas show upper and lower quartiles. Horizontal black lines denote median values.

(A) Fidelities of copy-number state and SNV genotype (bold text in the legend) as a function of the number of samples ($t_1$, $t_1$-$t_2$, $t_1$-$t_3$) and data types considered, including the case where the correct frequencies are given (denoted fgiven). Using more data types (e.g., cna-baf-mode instead of cna-mode) and using more samples each help achieve a higher performance.

(B) The mean errors of inferred frequencies per subclone, averaged over time points, show an increasingly accurate inference of subclonal trajectories.

of reproducing the complexities observed in real data. Therefore, we believe that our simulations are valuable in providing an assessment of the performance of our method. We next applied cloneHD to two real tumor data sets that have been thoroughly studied earlier and that can be considered as cases where the real solution is already known, at least to a first approximation.

## Inference from a Normal-Tumor Pair: Subclones in a Single Breast Cancer

The 188X breast cancer sample PD4120a has been used as a showcase data set, demonstrating that the cellular composition and evolutionary history of a tumor can be retraced from whole-genome sequencing (WGS) to considerable detail (Nik-Zainal et al., 2012a, 2012b). In this first extensive analysis, as many pieces of evidence from different data types as possible were collected to draw a comprehensive picture of its subclonal structure and the life history most compatible with it. The ambition of cloneHD is to automate some of the steps of this analysis, while being routinely applicable to whole-cancer-genome data at moderate sequencing depth. As compared to Nik-Zainal et al. (2012b), we note that we are currently not factoring in the phasing of somatic mutations to individual chromosomes using germline SNVs in conjunction with a large number of haplotypes from an external database, such as the 1000 Genomes Project (1000 Genomes Project Consortium, 2012). The special status of PD4120a suggests its use as a real data benchmark, as has been done in a previous attempt to resolve the subclonal structures of cancer samples (Oesper et al., 2013).

The data used for the inference consisted of (1) the integer mean read depth in each of the 2,727,971 windows of 1 kb genome-wide for both PD4120a and its matched normal sample, PD4120b; (2) read counts of both alleles at the 1,116,088 originally heterozygous loci; and (3) read counts of both alleles at the 70,690 somatic SNV loci. The matched normal sample was used to prefilter the data and to derive a read-depth bias field, reflecting technical rather than biological sources of variation in the read depth. This modulation was observed with high agreement in the tumor read-depth data and could be included in the cloneHD analysis (see Figure S1, where we also describe the heuristic prefiltering steps we used to mask out centromeric and telomeric regions as well as very short-scale variation). Coarse-graining the data in a way that neighboring segments are in different copy-number states with a probability of 1% or greater resulted in 580 segments genome-wide. We then performed the inference of subclonal structure incrementally adding data types (cna-mode, cna-baf-mode, cna-baf-snv-mode, snv-mode). We here report the findings for two and three subclones with up to four chromosome copies. Explanations with a single subclone could all be decisively ruled out.

In cna-mode, cloneHD consistently found two subclones at fractions of 0.65 and 0.096 with a mass of 107.7. The two subclones are mostly diploid with several single-copy gains and losses along the genome. Finding this solution, however, depends on penalizing states with zero total copies, for which the prior expectation is smaller on biological grounds. Without this penalty, the best solution (0.33, 0.096, and mass 107.7) finds the bigger subclone at about half the size and visiting states 0, 2, and 4 instead, suggesting a spurious solution.

In cna-baf-mode, cloneHD found two subclones at fractions of 0.69 and 0.11 (mass 102.1). The copy-number profile of the larger clone (shown in Figure 4) is the same as the one found in cna-mode, whereas the smaller subclone is shifted up by one copy and could plausibly be explained as tetraploid with losses of one and up to two copies in some of its chromosomes. This is in close agreement with the solution in (Nik-Zainal et al., 2012b) and further discussed in (Oesper et al., 2013). This solution also confirms the balanced and unbalanced loss of two copies of chromosomes 2 and 7 in the minor clone, respectively.

In cna-baf-snv-mode, trying to decompose the population across all data levels, cloneHD returned the cna-baf solution as the best explanation with almost identical fractions and mass. In the course of the inference, another solution was transiently visited with fractions of 0.62 and 0.092 (mass 103.4), which explained the SNV data slightly better at the expense of the first two data layers.

Finally, in snv-mode, using the local copy-number information from the best cna-baf-mode solution, cloneHD found support for three subclones at fractions 0.47, 0.23, and 0.084, summing to about the same purity as the cna-baf solution. This fact is not predetermined by using the copy-number constraint. In Figure 4E, the goodness of fit of this solution is shown, assigning each SNV to a genotype according to the cloneHD posterior probability and comparing the observed SNV allele fraction, corrected for local ploidy, to the one predicted by the model. An

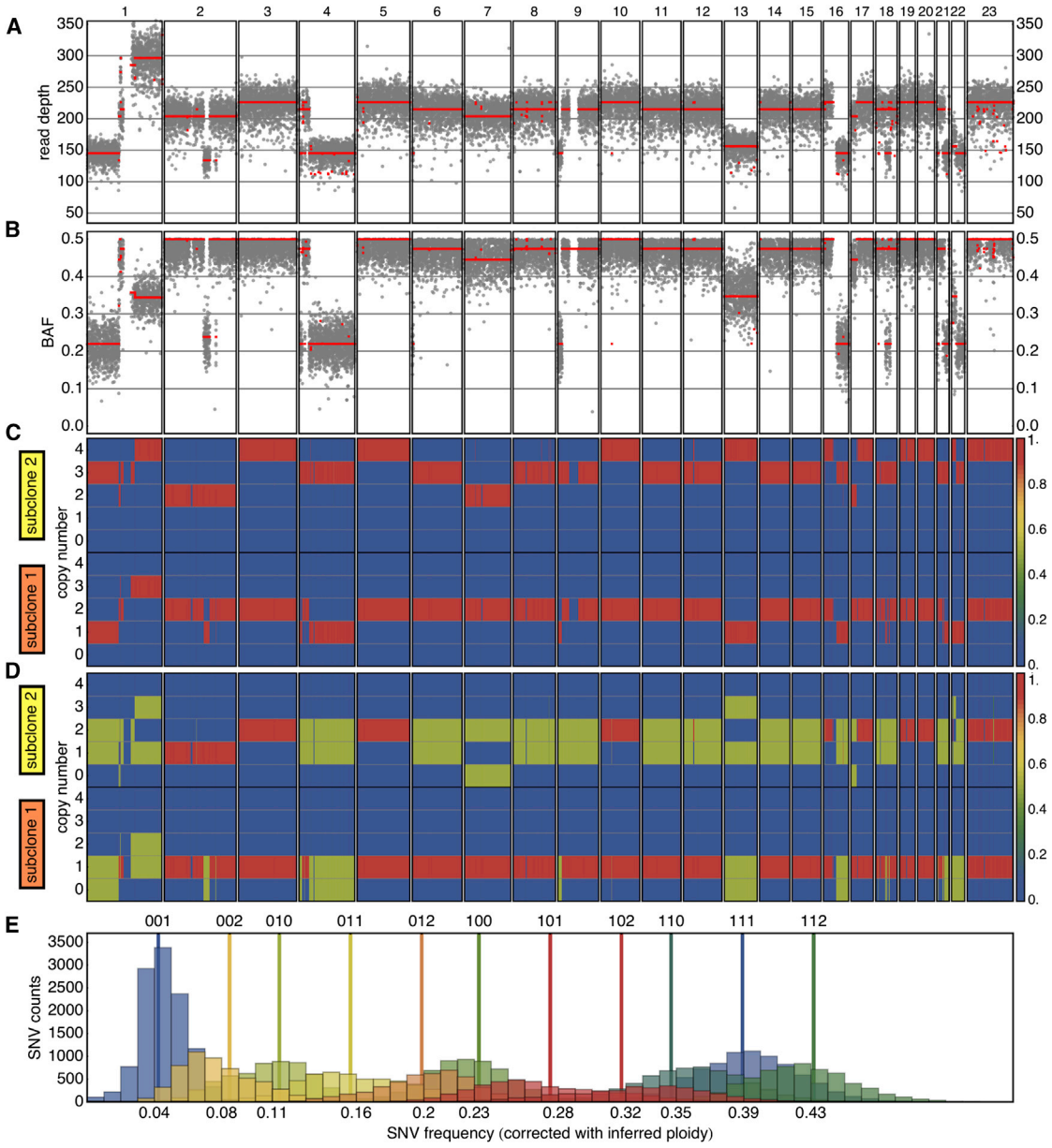


**Figure 4. The 188X Breast Cancer PD4120a and its Interpretations**

(A and B) The goodness of fit of the best explanation found with cloneHD in cna-baf-mode (subclonal fractions of 0.11 and 0.68 and a mass of 102.3, red line) for the read-depth data (A, gray dots, corrected for bias field) and the B-allele data (B, frequencies reflected at 0.5).

(C and D) The posterior distribution of the total copy-number states (C) and the B-allele copy number (D) for the cloneHD solution show the larger subclone with large scale deletions in chromosomes 1p, 4q, 13, 16q, 21q, and 22 as well as duplication of 1q. The smaller subclone 2 has several chromosomes in three and four copies and a copy-neutral loss of heterozygosity in chromosomes 7 and 17p.

(E) In snv-mode, there is support for three subclones at fractions 0.47, 0.23, and 0.08. The SNV goodness of fit is shown in terms of genotype-specific histograms (see Figure 2I).

interpretation of this result is that the larger cna-baf subclone is split in two smaller sets when considering also SNVs.

To compare our findings with the ThetA result, we fixed the subclonal fractions found therein (0.619 and 0.101) in cloneHD. In cna-mode, the optimal mass was learned to be 113.2 (not penalizing zero-copy states), leading to a good explanation of the CNA data. The copy-number profile found in this case is in almost perfect agreement with Oesper et al. (2013). However, for both BAF and SNV data, this candidate is clearly a poor explanation. At closer inspection, we note that it differs from the minor-tetraploid solution by shifting the smaller subclone two copies downward. This shift-by-two operation could, in principle, leave both the CNA and BAF level unaffected if all chromosomes were balanced in the minor allele. The decisive

**Table 1. Statistical Evidence for Different Subclones in the 188X Breast Cancer PD4120a**

| Mode | $f_1$ | $f_2$ | $M$ | $-\mathcal{L}(CNA)$ | $-\mathcal{L}(BAF)$ | $-\mathcal{L}(SNV)$ |
|---|---|---|---|---|---|---|
| cna | 0.651 | 0.096 | 107.7 | 11,443,900 | 3,134,300 | 360,100 |
| cna[a] | 0.329 | 0.096 | 107.7 | 11,476,400 | 4,190,100 | 379,500 |
| cna-baf | 0.687 | 0.109 | 102.1 | 11,449,300 | 3,018,500 | 351,900 |
| cna-baf-snv[b] | 0.687 | 0.111 | 102.0 | 11,450,300 | 3,017,700 | 351,200 |
| cna-baf-snv[c] | 0.617 | 0.092 | 103.4 | 11,463,500 | 3,041,800 | 347,400 |
| THetA (cna)[a] | 0.619 | 0.101 | 113.4 | 11,447,000 | 3,046,900 | 376,441 |

Overview of several candidate explanations of PD4120a in terms of two subclones. The first columns show the way in which a particular solution was found, the subclonal fractions, and mass parameter. The next three columns show the log-likelihood values (rounded to 100 units) for the different data tracks. Note that the best cna-mode solution fails to explain the BAF data.

[a]Not penalizing zero-total-copy states.

[b]This is the solution shown in Figures 4A–4D.

[c]This solution represents a recurring minor solution.

chromosomes are 7 and 17p, which are much better explained with a copy-number-neutral loss of heterozygosity. It is a common feature that solutions that were strong competitors at one level are ruled out completely when trying to explain data at the next (see Table 1).

This analysis sheds some further light on this fascinating and highly complex cancer genome and the different explanations put forward in Nik-Zainal et al. (2012b) and Oesper et al. (2013). It also clearly demonstrates the added value of using all available data sets in a comprehensive and integrated inference framework.

## Temporally Correlated Samples: Clonal Dynamics in Chronic Lymphocytic Leukemia

We next analyzed a chronic lymphocytic leukemia (CLL) whole-genome-sequence data set. CLL exhibits extensive clinical and biological heterogeneity, and none of the conventional treatments are curative (Alsolami et al., 2013). Furthermore, subclonality adversely affects clinical outcomes for CLL patients (Landau et al., 2013). Our case study data set consists of samples of a matched normal and five separate longitudinal tumor samples reported in Schuh et al. (2012) (patient ID CLL003). The time points correspond to changes in therapeutic regimen: (a) before chlorambucil; (b) before fludarabine, cyclophosphamide, rituximab; (c) immediately after six cycles of fludarabine, cyclophosphamide, rituximab; (d) before ofatumumab; and (e) after ofatumumab.

The patient CLL003 was studied by the authors in detail via targeted deep sequencing of the coding variants observed with WGS to a mean depth of 100,000X in order to reconstruct the clonal evolution of the tumor. Here, we used only the WGS data consisting of 4,406 SNVs and genome-wide read-depth data (in 20 kb windows with 10 kb overlap) to infer the clonal evolution of CLL003. Figures 5A and 5B show these SNV frequencies and read-depth profiles across time.

We first ran cloneHD on the data in cna-mode and identified three subclones (and a normal) as shown in Figure 5C. The inferred temporal evolution closely matches the one obtained from targeted deep sequencing and presented by Schuh et al. (who identified a fourth subclone that is at a very small frequency only at time point (a) before disappearing; Schuh et al., 2012).

Interestingly, running cloneHD with snv-mode, using the copy-number information from the cna-mode analysis, gave a very closely matching temporal evolution (see Figure 5C). The main difference to the cna-mode solution is that the small green subclone is here at slightly higher frequencies. Indeed, this snv-mode solution is even closer to the one reported by Schuh et al. with the green subclone present in a finite fraction at all time points. Such consistency across levels need not be the case, as the cell population can be split in different ways at the level of somatic CNAs and SNVs as shown in Figure 1. The importance of this consistency in terms of the respective roles of CNAs versus SNVs driving these subclones is not presently clear and suggests a direction for future investigation.

Finally, the result of the inference in cna-snv-mode is consistent with the other two being almost identical to the cna-mode-solution. Inspection of the copy-number posterior for the green subclone, which is visible only in time points (a), (c), and (e), suggests that it represents a fitting of noise: these three time points possibly have a slightly different technical bias so that the bias field, derived from the matched normal sample, cannot fully correct for it. In Figure 5D, we show genome-wide copy-number profiles for the dominant red and blue subclones.

In Figure 5E, we also report the posterior genotype probabilities for each coding mutation predicted to have a functional effect (Ensembl Variant Effect Predictor was used to select these mutations; McLaren et al., 2010). We note that using snv-mode does not force the clone decomposition to be the same at the CNA and SNV levels, so the discrepancy of the role of the green subclone could be biological. However, owing to the greater weight of the CNA data in the combined inference, we suspect that the SNV data supporting the green subclone are overwhelmed and that subclone is used instead to fit the bias field discrepancy. This is further supported by the SNV part of the log likelihood, which is ~5,000 units better when the green subclone is at frequencies given by the snv-mode compared to those from the cna-snv-mode.

## Assessing the Potential of Near-Real-Time Monitoring of Clonal Dynamics

The CLL003 results presented here demonstrate the power of cloneHD to quantify the subclonal evolution and the subclone

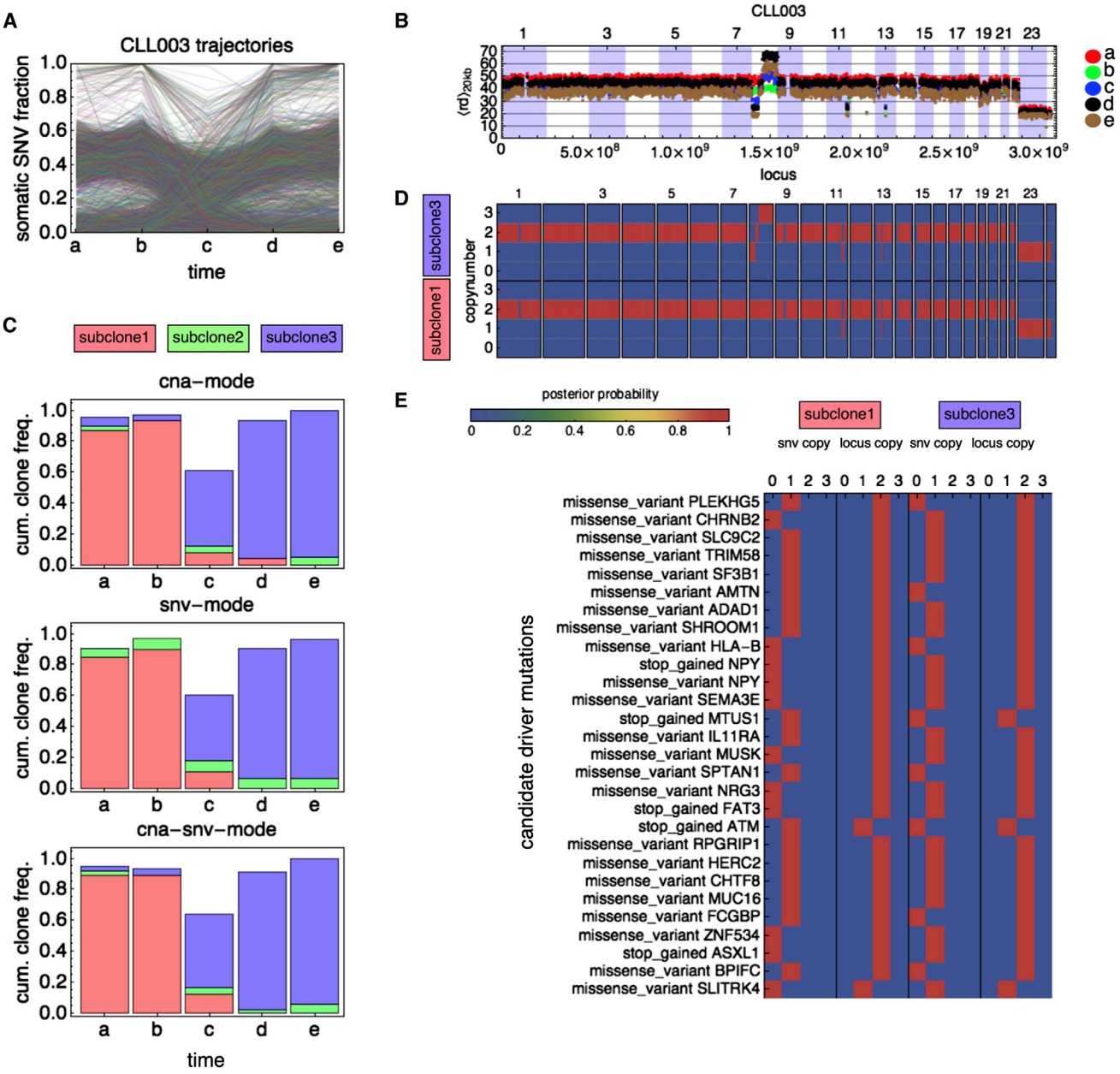


**Figure 5. Clonal Dynamics in Chronic Lymphocytic Leukemia**

(A and B) SNV trajectories for patient CLL003 (A) and genome-wide read-depth tracks (every 100th point shown) (B) across five time points together with a matched normal sample (not shown) form the input data.

(C) cloneHD identified three cancer subclones and a normal (white area) for this cancer. The evolutions inferred from SNV and CNA data are in close agreement.

(D) Genome-wide subclone-specific copy-number states for the major subclones red and blue have different aberration at chromosome 8 and shared ones at chromosomes 11 and 13.

(E) Subclone-specific SNV and locus copy-number state for variants that were identified as having a possible functional effect using the Ensembl Variant Effect Predictor.

genotypes from a WGS data set, facilitating the interpretation of tumor progression under a variable drug regimen. Such subclone trajectories can potentially be used to quantitatively study the underlying fitness landscape of CLL evolution under drugs once more similar data sets are analyzed. However, the analysis presented so far was purely retrospective in nature and thus would not have been of direct clinical utility in providing decision-making support for the clinician. To have an idea of the potential clinical utility, we formed partial data sets consisting of all data up to a given time point to mimic a real-time monitoring scenario. Figure 6 shows the results of these inferences running cloneHD in snv-mode.

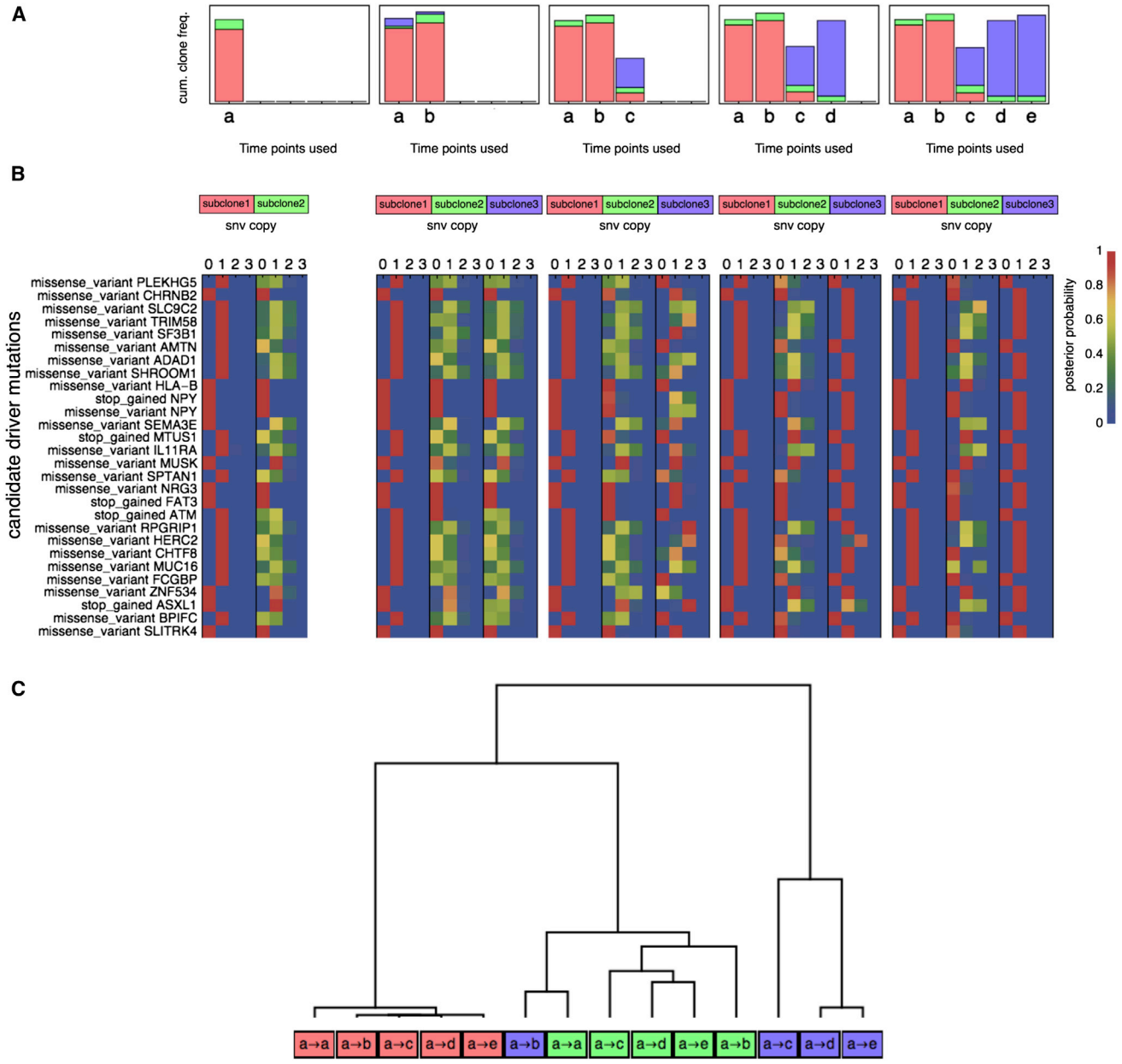


**Figure 6. Mimicking a Near-Real-Time Monitoring Scenario by Performing Inference on Partial Data**

(A) The time-development of subclonal evolution using only partial data recapitulates faithfully what could be inferred using all the data.

(B) Posterior probabilities for candidate driver SNVs are fully fixed for the red subclone after the first observation. For the other two subclones, more observations are needed: using the time points (a)–(c) is enough to see the emerging blue subclone close to its complete data inference counterpart.

(C) Clustering genome-wide SNV posterior probabilities across all partial inferences quantifies the relationships between the subclones. The red subclone is identified from the beginning as a separate one, whereas the blue subclone is clustered with the green subclone after two time points. From time point (c) onward, the blue subclones form their own clade.

Having only the first time point identifies the red subclone decisively with no further improvement with adding more data. For the green subclone, some point mutations (e.g., *SEMA3E* and *ASXL1*) are seen already at this early point, but most mutations have substantial uncertainty associated with them (Figure 6). Using data up to time point (c), it is apparent that the red subclone has substantially declined whereas the proportion of healthy cells in the third sample is larger (white area in the figure). In addition, there is a large fraction of the blue subclone present, and at this time point many of its coding mutations would already be correctly assigned. The last two time points further consolidate the genotype of the blue subclone and improve the green subclone, which is the smallest of all three.

In summary, cloneHD was applied to a time-resolved genome-wide data set and recovered an evolutionary history as was inferred using targeted deep sequencing (Schuh et al., 2012). In this case, close agreement was found between the patterns of evolution inferred independently using copy-number alterations or somatic SNVs. However, this will not necessarily be the case in general. Interestingly, the blue subclone is not seen at time points (a) and (b) when using only SNV data but is clearly manifest (albeit at small frequency) once CNA data are included. This detection in an early time point seems to be driven by chromosome 8 loss and gain events (see Figures 5B and 5D). These loss/gain events are just about visible to the human eye from the read-depth track at time point (b). Finally, analyzing partial data sets to emulate a real-time monitoring scenario, cloneHD could reveal information of potential clinical relevance.

## DISCUSSION

The difficult path from collecting mutational events using DNA-sequencing to elucidating subclonal cancer progression can be traversed. In contrast to the problem of identifying driver mutations, here the numerous passenger mutations are an asset. They faithfully report the evolution of a cancer genome, although they can sometimes be compatible with more than a single history. We have shown here that such degeneracy is greatly reduced when the tumor is observed at varying stages of its evolution, when subclonal frequencies are different. We have also shown the great benefit of performing a simultaneous analysis using several available data types (i.e., read depths, B-allele counts, and somatic SNV counts). Our reanalysis of the breast cancer sample PD4120a demonstrated the value of such an integrated analysis. Our analysis of a longitudinal data set of CLL demonstrates that its clonal progression could be deciphered using the whole-genome sequencing data without needing extra targeted deep sequencing as done by Schuh et al. (2012). For this patient, we also performed a mimic of a real-time monitoring scenario that could reveal clinically important information. Both of these results—whole-genome sequencing data suffices and real-time monitoring is informative—are proofs of concept and should be used as an encouragement to design prospective studies where patients' responses to therapies are monitored in real time via WGS.

We developed cloneHD in a way that user-specified constraints can be easily included, such that competing explanations can be ruled out using several distinct sources of information. For instance, external estimates for a lower bound on the sample purity could be used. In the case that histopathological image analysis has revealed cell fractions of different molecular phenotypes, cloneHD can assign somatic SNVs and copy-number variants to specific subclones according to these given fractions. Comparing the population fractions derived from image analysis, or any other phenotyping, to those obtained from the genetic data alone presents an interesting avenue for future research.

As a statistical inference program, cloneHD has some important limitations. The role of model complexity is central to most of them. While the real underlying complexity of a system (here, a tumor cell population) can be very large, noise in the observed data (due to finite sequencing depth) allows one to reconstruct only some major features of that complexity. One must find a balance between the need to explain all the structure visible in the data and the danger of overfitting it with a model that is too flexible. The BIC model selection criterion that we use in cloneHD tries to find this compromise and is validated with extensive simulations, where we know the true system complexity. For real data sets, however, this criterion should be regarded as an informed heuristic and should be supplemented with considerations of reconstruction quality and stability as well as biological consistency. For example, the algorithm might find a spurious solution with compensatory copy-number state changes across subclones, which is very unlikely on biological grounds but might serve to opportunistically maximize the total log likelihood.

To build intuition on the solution space, we have also included a systematic scan mode to cloneHD in order to visualize the log-likelihood landscape. This mode is practical for single samples with only few global parameters to be scanned over. For multiple samples and with increasing knowledge about the evolutionary dynamics of cancers, one could further constrain the subclonal fractions to follow trajectories that depend on much fewer parameters.

Another limitation comes with the use of explicit emission models, such as Poisson and Binomial distributions, to connect noisy data to the underlying genomic states (see Supplemental Experimental Procedures). Data for which these models are not valid approximations should not be included in cloneHD.

The computational efficiency of cloneHD is achieved with a fuzzy data segmentation scheme: HMMs are allowed to change their state only at loci where the data itself support a certain minimum jump probability. If this threshold is set too high, some true transitions might be missed, leading to incorrect reconstructions. If it is set too low, too many segments are introduced, slowing the algorithm down. We found a jump probability of 1% or greater to be a good compromise. Once all the parameters are learned, however, one can recalculate posterior distributions with cloneHD where every locus is allowed to be in every state. This is the highest definition achievable.

Lastly, cloneHD does not explicitly enforce a consistent tree structure for the subclones along the genome. Especially for SNVs, such a constraint might be very useful. In the present setup, this would, however, require integration over all possible trees, a calculation outside the current scope of cloneHD.

In the future, studies with both temporally and spatially resolved sequence data of tumor cell populations are likely to become ubiquitous. Computational methods able to exploit the information in such correlated samples are clearly needed. Until single-cell sequencing methods mature or disruptive technologies for bulk sequencing with very long reads emerge, inferences as performed here will be necessary. For this period, we hope that cloneHD will help to generate useful insights into subclonal cancer evolution.

Beyond cancer progression, subclonality is common to asexual evolution, potentially giving cloneHD a much broader scope for application. Many features in the evolution of cancer are shared with bacterial, viral, or parasitic populations, including asexual reproduction as well as clonal expansion and

competition. Clonal heterogeneity has been observed both in laboratory populations (e.g., Lang et al., 2013) and wild populations within the host (e.g., Bryant et al., 2013; Lieberman et al., 2014). cloneHD can also be useful in deciphering genotypes in fully clonal isolates with added complexity due to copy-number variation (e.g., see Figure S2).

We here presented cloneHD, an algorithm for the probabilistic inference of subclonal copy-number profiles, genotypes, and population frequencies. cloneHD can be used to perform an integrative analysis of somatic CNAs, B-allele variants, and somatic SNVs across multiple correlated samples. Using simulations, a single breast cancer, and time-resolved CLL data we have demonstrated the ability of cloneHD to quantify and track subclonal progression in cancers.

## EXPERIMENTAL PROCEDURES

A full exposé of the mathematical details and the implementation of the algorithm is given in the Supplemental Experimental Procedures, so we focus here on some key conceptual points only. cloneHD is a probabilistic framework to resolve the subclonal structure of a cell population from NGS data. This data usually comes at three levels: the read-depth data (number of reads mapping to different loci in the genome) contain information about the (aberrant) copy-number profiles that are present in the cancer cell population, the B-allele count data (number of reads reporting a minor allele at an originally heterozygous locus) contain additional information about the copy-number states by differentiating between balanced and unbalanced copy-number changes, and the somatic mutation data (number of reads reporting a somatic nucleotide variant not seen in normal cells) contain further information about the size of subclonal fractions in the sequenced sample and their somatic mutation genotypes.

The cloneHD setup is capable of performing a joint inference on several samples of the same tumor, e.g., from longitudinal (Schuh et al., 2012) or multifocal (Gerlinger et al., 2012) sequencing studies. It assumes that the same $n$ subclones are present in all of these $N_s$ samples but at possibly different population fractions $f_j^s$ ($s = 1 \ldots N_s$, $j = 1 \ldots n$). Having the same set of subclonal copy-number profiles and genotypes realized at different relative proportions can greatly help in resolving tumor structure.

Because most haplotype information is lost in the sequencing process, cloneHD tries to leverage the correlations along the genome that remain in the read depth and minor allele count data by modeling these with hidden Markov models, where their emission properties couple the hidden, locus- and subclone-specific copy-number profiles $c_{ij}$, minor allele genotypes $b_{ij}$, and somatic SNV genotypes $g_{ij}$ to all the observed data ($i = 1 \ldots L$, where $L$ is the number of observations in a data set). Global parameters that are jointly learned across all data types are the subclonal fractions $f_j^s$ and, for CNA data, the sequencing yield $M^s$ per haploid DNA (which we call *mass*). These cellular fractions and masses are sample specific. The global parameters are determined by maximizing the total log-likelihood of all the observed data. A given estimate of $f$ and $M$ determines a posterior distribution for the hidden states $c$, $b$, and $g$ for every single observation. This high-resolution information can then be used to perform further subclone-specific mutation data analysis.

The greatest improvement of cloneHD over existing methods is that it couples the different data layers and enforces a consistent explanation of all the data in a hierarchical fashion. Proposed estimates of $f$ and $M$ lead to a posterior distribution over total copy-number states per subclone along the genome, e.g., showing strong evidence for a deletion of one particular chromosome copy in subclone 1: $c_{i1} = 1$. At originally heterozygous loci in that region, the minor allele genotypes must be consistent with this fact, e.g., $b_{i1} \leq 1$. In general, these consistency constraints are probabilistic, with the copy-number profile posterior distribution $\gamma_i(c_i)$ informing the BAF and SNV genotype prior distributions at each locus (see Supplemental Experimental Procedures).

Increasing the proposed number of subclones $n$ and the maximum copy number $c^{max}$ that their respective copy-number profiles can visit greatly increases the model complexity of cloneHD, with the hidden state space dimensionality growing exponentially. We use the Bayesian information criterion (BIC) as a heuristic model selection scheme. The BIC penalty term below aims to capture model complexity not only by the number of free parameters but also by the number of states that are available to explain the data.

$$BIC = 2(\mathcal{L}_{CNA} + \mathcal{L}_{BAF} + \mathcal{L}_{SNV}) - k \log(L_{CNA} + L_{BAF} + L_{SNV})$$

$$k \equiv (c^{max} + 1)^n + N_s(n + 1)$$

Additionally, the goodness of fit (the average geometrical distance of data points to the model prediction) can also be used as a model comparison criterion and is included in the output.

The sizeable model complexity of cloneHD requires the data to be efficiently organized to avoid wasting computational effort. Previous algorithms have chosen to segment the read-depth data in some meaningful form on usually large length scales (Oesper et al., 2013; Van Loo et al., 2010). With the aim to retain as much of the correlation information as possible, we have implemented a fuzzy data segmentation scheme that is scale-free. This is done by a stand-alone program, filterHD (described in Supplemental Experimental Procedures), which is a continuous state-space HMM in the spirit of the well-known Kalman filter (Kalman, 1960) but adapted for integer observations and employing a jump-diffusion propagator. filterHD not only is a powerful probabilistic smoothing algorithm but also produces a posterior jump probability track, highlighting regions of the data where real jumps in the emission rate could have occurred. Allowing the HMM in cloneHD to make state transitions only at sites with nonnegligible posterior jump probability effectively segments the read-depth data into blocks that are still probabilistically connected.

The diffusive part of the filterHD dynamical model is used to learn a potential read-depth bias. This technological bias results in modulations of the read-depth profile, which are not caused by real discrete copy-number changes in some parts of the cell population. If sequencing data of a matched normal sample are available and if the read-depth bias in both normal and tumor samples is the same, then filterHD can produce a high-quality estimate of this bias field that can then be included into the cloneHD inference. Since filterHD is a probabilistic framework, one can assert this assumption quantitatively in terms of likelihoods.

### Code Availability

The latest version of the cloneHD software, including filterHD, as well as extensive documentation, can be found at: https://github.com/andrej-fischer/cloneHD.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and two figures and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.04.055.

## ACKNOWLEDGMENTS

We would like to acknowledge the Wellcome Trust for support under grant numbers 098051 and 097678. A.F. is in part supported by the German Research Foundation (DFG) under grant number FI 1882/1-1. C.I. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society under grant number 101239/Z/13/Z. We would like to thank A. Schuh, J. Becq, and J.-B. Cazier for help with the CLL data; P. Van Loo and D. Wedge for help with the breast cancer data, discussions, and comments on an earlier version of the manuscript; P. Campbell for discussions; and C. Greenman, I. Tomlinson, O. Krijgsman, and S. Schiffels for comments on an earlier version of the manuscript.

Received: December 19, 2013
Revised: March 26, 2014
Accepted: April 24, 2014
Published: May 29, 2014

## REFERENCES

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, Å., Børresen-Dale, A.-L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013a). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013b). Deciphering signatures of mutational processes operative in human cancer. Cell Rep. *3*, 246–259.

Alsolami, R., Knight, S.J., and Schuh, A. (2013). Clinical application of targeted and genome-wide technologies: can we predict treatment responses in chronic lymphocytic leukemia? Per. Med. *10*, 361–376.

Aparicio, S., and Caldas, C. (2013). The implications of clonal genome evolution for cancer medicine. N. Engl. J. Med. *368*, 842–851.

Bedard, P.L., Hansen, A.R., Ratain, M.J., and Siu, L.L. (2013). Tumour heterogeneity in the clinic. Nature *501*, 355–364.

Bolli, N., Avet-Loiseau, H., Wedge, D.C., Van Loo, P., Alexandrov, L.B., Martincorena, I., Dawson, K.J., Iorio, F., Nik-Zainal, S., Bignell, G.R., et al. (2013). Heterogeneity of genomic architecture and evolution in multiple myeloma. Nat. Commun. *5*, 1–13.

Bozic, I., Allen, B., and Nowak, M.A. (2012). Dynamics of targeted cancer therapy. Trends Mol. Med. *18*, 311–316.

Bryant, J.M., Harris, S.R., Parkhill, J., Dawson, R., Diacon, A.H., van Helden, P., Pym, A., Mahayiddin, A.A., Chuchottaworn, C., Sanne, I.M., et al. (2013). Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. Lancet Respir. Med. *1*, 786–792.

Burrell, R.A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. Nature *501*, 338–345.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. *30*, 413–421.

Chen, M., Gunel, M., and Zhao, H. (2013). SomatiCA: identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data. PLoS ONE *8*, e78143.

Fischer, A., Illingworth, C.J., Campbell, P.J., and Mustonen, V. (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biol. *14*, R39.

Garraway, L.A., and Lander, E.S. (2013). Lessons from the cancer genome. Cell *153*, 17–37.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. *366*, 883–892.

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G.R.S., Creixell, P., Karchin, R., Vazquez, M., Fink, J.L., Kassahn, K.S., Pearson, J.V., et al.; International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group (2013). Computational approaches to identify functional genetic variants in cancer genomes. Nat. Methods *10*, 723–729.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306–313.

Greenman, C.D., Bignell, G., Butler, A., Edkins, S., Hinton, J., Beare, D., Swamy, S., Santarius, T., Chen, L., Widaa, S., et al. (2010). PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics *11*, 164–175.

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

Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. Science *339*, 959–961.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science *339*, 957–959.

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

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. J. Basic Eng. *82*, 35–45.

Landau, D.A., Carter, S.L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M.S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell *152*, 714–726.

Lang, G.I., Rice, D.P., Hickman, M.J., Sodergren, E., Weinstock, G.M., Botstein, D., and Desai, M.M. (2013). Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature *500*, 571–574.

Larson, N.B., and Fridley, B.L. (2013). PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. Bioinformatics *29*, 1888–1889.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Li, A., Liu, Z., Lezon-Geyda, K., Sarkar, S., Lannin, D., Schulz, V., Krop, I., Winer, E., Harris, L., and Tuck, D. (2011). GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays. Nucleic Acids Res. *39*, 4928–4941.

Lieberman, T.D., Flett, K.B., Yelin, I., Martin, T.R., McAdam, A.J., Priebe, G.P., and Kishony, R. (2014). Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. Nat. Genet. *46*, 82–87.

Liu, Z., Li, A., Schulz, V., Chen, M., and Tuck, D. (2010). MixHMM: inferring copy number variation and allelic imbalance using SNP arrays and tumor samples mixed with stromal cells. PLoS ONE *5*, e10909.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069–2070.

Mosén-Ansorena, D., Aransay, A.M., and Rodríguez-Ezpeleta, N. (2012). Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. BMC Bioinformatics *13*, 192.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al.; Breast Cancer Working Group of the International Cancer Genome Consortium (2012a). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979–993.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al.; Breast Cancer Working Group of the International Cancer Genome Consortium (2012b). The life history of 21 breast cancers. Cell *149*, 994–1007.

Oesper, L., Mahmoody, A., and Raphael, B.J. (2013). THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. *14*, R80.

Potter, N.E., Ermini, L., Papaemmanuil, E., Cazzaniga, G., Vijayaraghavan, G., Titley, I., Ford, A., Campbell, P., Kearney, L., and Greaves, M. (2013). Single-cell mutational profiling and clonal phylogeny in cancer. Genome Res. *23*, 2115–2125.

Rasmussen, M., Sundström, M., Göransson Kultima, H., Botling, J., Micke, P., Birgisson, H., Glimelius, B., and Isaksson, A. (2011). Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. Genome Biol. *12*, R108.

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

Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S.M., 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.

Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., et al. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature *461*, 809–813.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet. *14*, 618–630.

Song, S., Nones, K., Miller, D., Harliwong, I., Kassahn, K.S., Pinese, M., Pajic, M., Gill, A.J., Johns, A.L., Anderson, M., et al. (2012). qpure: A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. PLoS ONE *7*, e45835.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Su, X., Zhang, L., Zhang, J., Meric-Bernstam, F., and Weinstein, J.N. (2012). PurityEst: estimating purity of human tumor samples using next-generation sequencing data. Bioinformatics *28*, 2265–2266.

Sun, W., Wright, F.A., Tang, Z., Nordgard, S.H., Van Loo, P., Yu, T., Kristensen, V.N., and Perou, C.M. (2009). Integrated study of copy number states and genotype calls using high-density SNP arrays. Nucleic Acids Res. *37*, 5365–5377.

Van Loo, P., Nordgard, S.H., Lingjærde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., et al. (2010). Allele-specific copy number analysis of tumors. Proc. Natl. Acad. Sci. USA *107*, 16910–16915.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Wargo, A.R., Huijben, S., de Roode, J.C., Shepherd, J., and Read, A.F. (2007). Competitive release and facilitation of drug-resistant parasites after therapeutic chemotherapy in a rodent malaria model. Proc. Natl. Acad. Sci. USA *104*, 19914–19919.

Wheeler, D.A., and Wang, L. (2013). From human genome to cancer genome: the first decade. Genome Res. *23*, 1054–1062.

Yau, C., Mouradov, D., Jorissen, R.N., Colella, S., Mirza, G., Steers, G., Harris, A., Ragoussis, J., Sieber, O., and Holmes, C.C. (2010). A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome Biol. *11*, R92.