McPherson *et al. Genome Biology* (2017) 18:140
DOI 10.1186/s13059-017-1267-2

Genome Biology

METHOD Open Access

ReMixT: clone-specific genomic structure estimation in cancer

Andrew W. McPherson[1,2], Andrew Roth[3,4], Gavin Ha[5,6], Cedric Chauve[7], Adi Steif[1], Camila P. E. de Souza[1,2], Peter Eirew[1], Alexandre Bouchard-Côté[8], Sam Aparicio[1,2], S. Cenk Sahinalp[9,10] and Sohrab P. Shah[1,2*]

Abstract

Somatic evolution of malignant cells produces tumors composed of multiple clonal populations, distinguished in part by rearrangements and copy number changes affecting chromosomal segments. Whole genome sequencing mixes the signals of sampled populations, diluting the signals of clone-specific aberrations, and complicating estimation of clone-specific genotypes. We introduce ReMixT, a method to unmix tumor and contaminating normal signals and jointly predict mixture proportions, clone-specific segment copy number, and clone specificity of breakpoints. ReMixT is free, open-source software and is available at http://bitbucket.org/dranew/remixt.

Keywords: Cancer genomics, DNA sequencing, Tumour heterogeneity, Genomic rearrangement, Copy number variation

Background

Chromosomal rearrangements pattern the genomes of cancer cells. Owing to various forms of DNA repair deficiency, such structural variations accumulate on cell division, leading to genome instability in the life histories of cancer cells. Coupled with evolutionary selection and clonal expansion, genomic instability and consequent segmental aneuploidies mark expanded cell populations within a tumour, forming important components of their genotypes. Within each tumour, branched evolution produces mixed populations of tumour cells with ancestrally related, but divergent chromosomal structures.

Accurate detection and quantification of genomic structural changes in a population of cancer cells as measured by bulk, whole genome sequencing (WGS) remains a significant computational challenge. The process of DNA extraction from a tumour sample pools and admixes molecules from the input material without labelling the assignment of DNA to its parent cell. The resulting sequencing data represent a randomly sampled subset of DNA fragments from the admixed pool, leaving the problem of unmixing the structural rearrangements which mark the constituent clones in the input material. The key difficulty of the problem is that the admixed pool dilutes the signal of genomic rearrangements and copy number alterations in the data, often to a level approaching that of the experimental noise.

Rearrangements and copy number changes are intrinsically linked, with unbalanced rearrangements producing changes in copy number, and loss or gain of rearranged chromosomes resulting in segment-specific copy changes. Rearrangement breakpoints representing tumour-specific adjacencies can be predicted with reasonable accuracy from WGS data using a variety of tools [1–4]. However, existing methods for copy number analysis do not consider tumour-specific adjacencies, and instead model segments as adjacent only if they are adjacent in the reference genome [5–9]. This results in only partial ability to leverage the spatially correlated nature of the data to borrow statistical strength.

We propose that breakpoints provide the potential for a more comprehensive model of genome structure. Knowledge of long-range connectivity between segments of a cancer genome provides the opportunity to simultaneously analyse breakpoints and copy number in a unified

*Correspondence: sshah@bccrc.ca
[1] Department of Molecular Oncology, BC Cancer Agency, 675 West 10th Avenue, Vancouver, BC, Canada
[2] Department of Pathology and Laboratory Medicine, University of British Columbia, 2329 West Mall, Vancouver, BC, Canada
Full list of author information is available at the end of the article

model and to reconstruct the true genomic topology. Integrating both copy number and breakpoints also provides additional information about each breakpoint: whether the breakpoint is real or a false positive, the prevalence of the breakpoint in the clone mixture, and the number of chromosomes harbouring the breakpoint per clone. A natural hypothesis then emerges: a comprehensive model of genome structure will improve both copy number inference and biological interpretation through reconstructed tumour genomes.

Some progress has been made on more comprehensive modelling of genome structure in tumour clones. Mahmoody et al. [10] propose an algorithm to infer missing adjacencies in a mixture of rearranged tumour genomes; however, they do not model copy number. Zerbino et al. [11] propose a framework for sampling from the rearrangement history of tumour genomes. Oesper et al. [12] propose PREGO, a method for inferring the copy number of segments and breakpoints using a genome graph-based approach, though they do not model normal contamination or tumour heterogeneity, limiting applicability of their method to real tumour data. More recently, Li et al. [13] formulate a Markov random field model of allele-specific copy number change and apply their method, Weaver, to samples harbouring a single tumour clone and contaminating normal cells.

We propose ReMixT, a method for jointly inferring clone mixture proportions, clone- and allele-specific segment copy numbers, and clone-specific breakpoint copy number from WGS data. We formulate the problem as a posterior inference problem on a probabilistic graphical model. Our model captures the spatial correlation both between segments that are adjacent in the reference genome in addition to correlations between segments adjacent in the tumour genome as nominated by predicted breakpoints. We describe an algorithmic solution using structured variational inference. Importantly, our algorithm is similar in complexity to a breakpoint-naive hidden Markov model (HMM) of segment copy number. We leverage haplotype blocks to more accurately measure allele-specific read counts and infer allele-specific copy number for each clone.

We assert that joint inference of all three features of genome sequencing described above will result in more accurate prediction compared to independent inference. Knowledge of rearrangement breakpoints will prevent the smoothing over of copy number changes produced by true rearrangements. Incorrect smoothing of highly rearranged chromosomes may have detrimental effects on the estimation of mixing proportions and variance parameters, as the model would be forced to compensate for an unexpected increase or decrease in read depth across the smoothed chromosomes. Finally, post hoc prediction of rearrangement breakpoint copy number based on segment copy number may fail if the exact locations of associated copy number transitions are not identified, particularly for rearrangements present in a minor fraction of clones.

We show using simulations that a more complete model of genome structure that includes breakpoint information results in improved inference of mixture proportion and segment copy number over an otherwise equivalent HMM combined with post hoc annotation. Performance improvements are most dramatic when the proportion of one clone is small. We benchmark ReMixT against TITAN [5], THetA2 [14], Battenberg [8], and CloneHD [7] using a novel framework for generating realistic partially simulated WGS datasets from an existing WGS dataset. As further validation, we applied ReMixT to four primary tumour samples from a patient with high-grade serous ovarian cancer (HGSOvCa) and performed single cell breakpoint sequencing on a subset of the clone-specific breakpoints. Next we applied ReMixT to a primary breast cancer sample and its derived mouse xenograft samples, recapitulating previously described [15] clonal dynamics identified using deep sequencing of single nucleotide variants (SNVs). Finally, we analysed two HGSOvCa cell lines, providing examples of how ReMixT-predicted clone-specific breakpoints can phase disparate subclonal genomic regions into partial tumour chromosomes towards fully reconstructing clone-specific cancer genomes.

Results

The ReMixT model of genome structure

We consider the problem of predicting segment and breakpoint copy number given WGS data from tumour and matched normal samples. Assume as input a set of alignments of uniquely mapped concordant reads and a set of putative breakpoints predicted from discordant reads. Given N segments indexed by n, $n \in \{1 \ldots N\}$; K breakpoints indexed by k, $k \in \{1 \ldots K\}$; and assuming M clones indexed by m, $m \in \{1 \ldots M\}$, we aim to predict the following:

1. Mixture proportions of tumour clones and normal cells ρ_m
2. Clone- and allele-specific copy numbers of genomic segments c_{nm}
3. Clone-specific copy number of rearrangement breakpoints b_{km}

Data preprocessing

Preprocessing of tumour WGS data produces measured total and allele-specific read counts for a set of genomic segments in addition to tumour-specific adjacencies between those segments. First, the genome is partitioned into regular length segments, with segments containing the breakends of input breakpoints further partitioned such that each breakend coincides with a segment boundary. Total read counts are obtained by counting the number of uniquely aligned paired-end reads fully contained within each segment. Next, haplotype blocks are predicted from single nucleotide polymorphisms (SNPs) using shapeit2 [16] and a 1000 Genomes reference panel. Reads containing heterozygous SNPs are assigned to haplotype blocks, and haplotype block counts are aggregated within segments, resulting in per-segment allele-specific read counts. GC and mappability biases contribute significant variance to segment read counts. We use a position-specific model [17] to calculate a bias-adjusted *effective length* for each segment, where segments with shorter effective lengths are statistically less well represented by read counts. For visualization purposes, we calculate *raw* major and minor copy numbers for each segment from observed depths and allele ratios and inferred normal and tumour depth. Additional details are provided in Additional file 1: Sections 1.1 and 1.2.

Probabilistic model

We propose a probabilistic model of genome structure and a structured variational inference algorithm for calculating the optimal clone mixture and segment and breakpoint copy number (Fig. 1). Below we focus on a model of total copy number and defer the details of the allele-specific model and modelling of outliers to Additional file 1: Section 1.3. Let $p(x|c, h, l, \theta)$ be the likelihood of observed total read count x given per clone segment copy number c, segment length l, global likelihood parameters θ, and per clone haploid read depths h. The haploid read depths encode both the mixture and depth of sequencing and are specified as reads per nucleotide for a single copy of a segment. The expected read count μ_n of segment n is a linear combination of the segment length, clone-specific copy number, and clone-specific haploid read depth, summed over clones (Eq. 1):

$$\mu_n = l_n \sum_m h_m c_{nm} \tag{1}$$

A reasonable starting point is to assume read counts are Poisson distributed [18] ($x_n \sim \text{Pois}(\mu_n)$); however, we show in Additional file 1: Section 1.2.3, that a two-component negative binomial mixture provides a significantly better fit to real data.

Let $p(C, B|O, \lambda)$ be the joint probability of segment and breakpoint copy number (C and B respectively) given breakend orientations O. We assume the copy numbers of a sequence of segments have the Markov property-given breakpoint copy number, and represent the resulting chain structure as a product of un-normalized transition factors[1]. A breakpoint with breakend interposed between

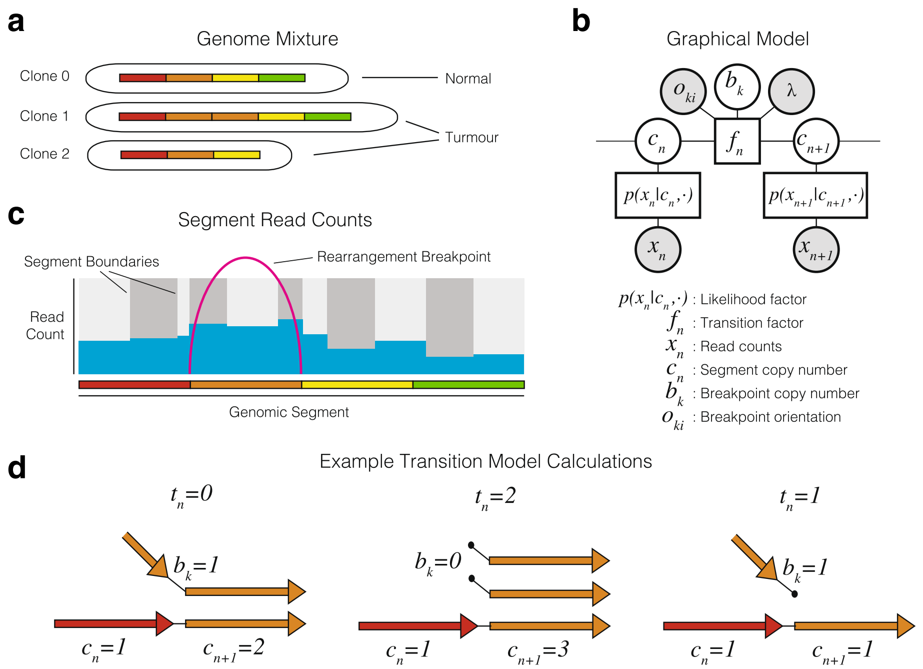


Fig. 1 An overview of the ReMixT Method. **a**) Bulk sequencing is applied to a mixture of cells modeled as a set of clones of unknown proportion each with distinct sets of chromosomes with unknown structure. **b**) Observed data include binned read counts per segment, and rearrangement breakpoints connecting segment ends. **c**) The ReMixT graphical model as a factor graph. **d**) Calculation of the transition factor involves calculating the number of *telomeres* *t*, the number of segment ends left unconnected to another segment end in the model

two segments will result in a copy number transition between those segments. For instance, a transition in copy number is expected between two segments to either side of the start of a deletion, with the difference in segment copy number equal to the number of chromosomes harbouring the deletion event, or equivalently, the number of copies of the deletion breakpoint. A mismatch in segment and breakpoint copy number implies that at least one segment end is left disconnected (Fig. 2d). We call these free ends *telomeres*, and define the transition factors of our probability model in terms of the number of telomeres t implied by the segment and breakpoint copy number. Without a breakpoint, the number of telomeres is simply the absolute difference in copy number between adjacent segments $t(c, c') = |c - c'|$. Depending on its orientation, a positive copy number for a breakpoint may explain some or all of the difference in copy number between adjacent segments. The number of telomeres at a transition coincident with a breakpoint can thus be calculated as $t(c, c', b', o) = |c - c' - o \cdot b|$, with orientation $o \in \{-1, +1\}$. For multiple clones, t may be a more complex function of the copy number differences for each clone (see Additional file 1: Section 1.4).

Define transition factors $f(c, c', b|o, \lambda) = e^{-\lambda t(c,c',b|o)}$, and let k_n be the index of the breakpoint interposed between segment n and $n + 1$. Write the joint probability over the observed read counts and segment and breakpoint copy number as given by Eq. 2:

$$\begin{aligned} p(X, C, B|h, L, O, \theta, \lambda) &= p(X|C, L, h, \theta)p(C, B|O, \lambda) \\ &\propto \prod_{n=1}^{N} p(x_n|c_n, h, l_n, \theta) \\ &\quad \times \prod_{n=1}^{N-1} f(c_n, c_{n+1}, b_{k_n}|o_n, \lambda) \end{aligned} \tag{2}$$

Exact inference in the ReMixT model is intractable due to additional dependencies introduced by modelling the long-range connectivity of breakpoints.

Structured variational inference

We are seeking to infer the posterior probability $p(z|x)$ of the unobserved model variables z given observed data x. The variational inference approach seeks to approximate an intractable posterior $p(z|x)$ with a more tractable family of distributions $q(z)$, typically characterized by an increased number of parameters and fewer dependencies [19]. An optimal $q(z)$ is computed by minimizing the Kullback-Leibler (KL) divergence between $p(z|x)$ and $q(z)$ as given by Eq. 3:

$$\begin{aligned} D_{\mathrm{KL}}\left(q(z)|p(z|x)\right) &= \int q(z) \log\left(\frac{q(z)}{p(z|x)}\right) dz \\ &= \log p(x) - \int q(z)p(x, z)dz \\ &\quad + \int q(z) \log q(z)dz \\ &= \log p(x) - \mathbb{E}_q\left[p(x, z) - \log q(z)\right] \end{aligned} \tag{3}$$

The expectation given in the final form of Eq. 3 forms a lower bound on the model evidence $p(x)$, since $D_{\mathrm{KL}}\left(q(z)|p(z|x)\right)$ is positive and approaches zero for a perfect approximation. Importantly, the difficult problem of directly minimizing the KL divergence is equivalent to the easier problem of maximizing this evidence lower bound (ELBO). The mean field approximation assumes a distribution $q(z) = \prod_i q_i(z_i)$ that factorizes over single model variables. In structured variational inference, each z_i is a disjoint set of model variables, allowing q to have a more complex dependency structure that better approximates the posterior [20, 21]. Independence between factors of q allows for application of a coordinate descent algorithm that iteratively maximizes the ELBO with respect to each q_i using general updates given by Eq. 4:

$$\log q^*(z_j) = \mathbb{E}_{\prod_{j \neq i} q_j(z_j)}[\log p(x, z)] + \text{const} \tag{4}$$

We approximate the posterior $p(C, B, h, \theta|X, L, O, \lambda)$ using a distribution q with factorization given by Eq. 5:

$$q(C, B, h, \theta) = q(h)q(\theta)q(C)\prod_k q_k(b_k) \tag{5}$$

Taking a variational expectation maximization (EM) approach, we specify the distributional form of $q(h)$ and $q(\theta)$ to be the Dirac delta function, and compute point estimates for those parameters. Applying Eq. 4 to $q(C)$ results in Eq. 6[2]:

$$\begin{aligned} \log q^*(C) &= \sum_B \left(\prod_k q(b_k)\right) \log p(X, C, B, h, \theta|L, O, \lambda) \\ &\quad + \text{const} \\ &= \sum_n \zeta_n(c_n) + \sum_{n=1}^{N-1} \zeta_n(c_n, c_{n+1}) + \text{const} \end{aligned} \tag{6}$$

$$\zeta_n(c_n) = \log p(x_n|c_n, h, l_n, \theta) \tag{7}$$

$$\zeta_n(c_n, c_{n+1}) = \sum_b q_{k_n}(b) \log f(c_n, c_{n+1}, b|o_n, \lambda) \tag{8}$$

By inspection, the probability distribution $q^*(C)$ given by Eq. 6 has a chain topology equivalent to an HMM, with an emission calculated as a function of the read count likelihood and transition matrices calculated by modifying f according to $q_{k_n}(b)$ (Eqs. 7 and 8). The emission and transition terms $\zeta_n(c_n)$ and $\zeta_n(c_n, c_{n+1})$ define the variational parameters of $q(C)$. The sum product algorithm can be used to calculate the single and pairwise posterior

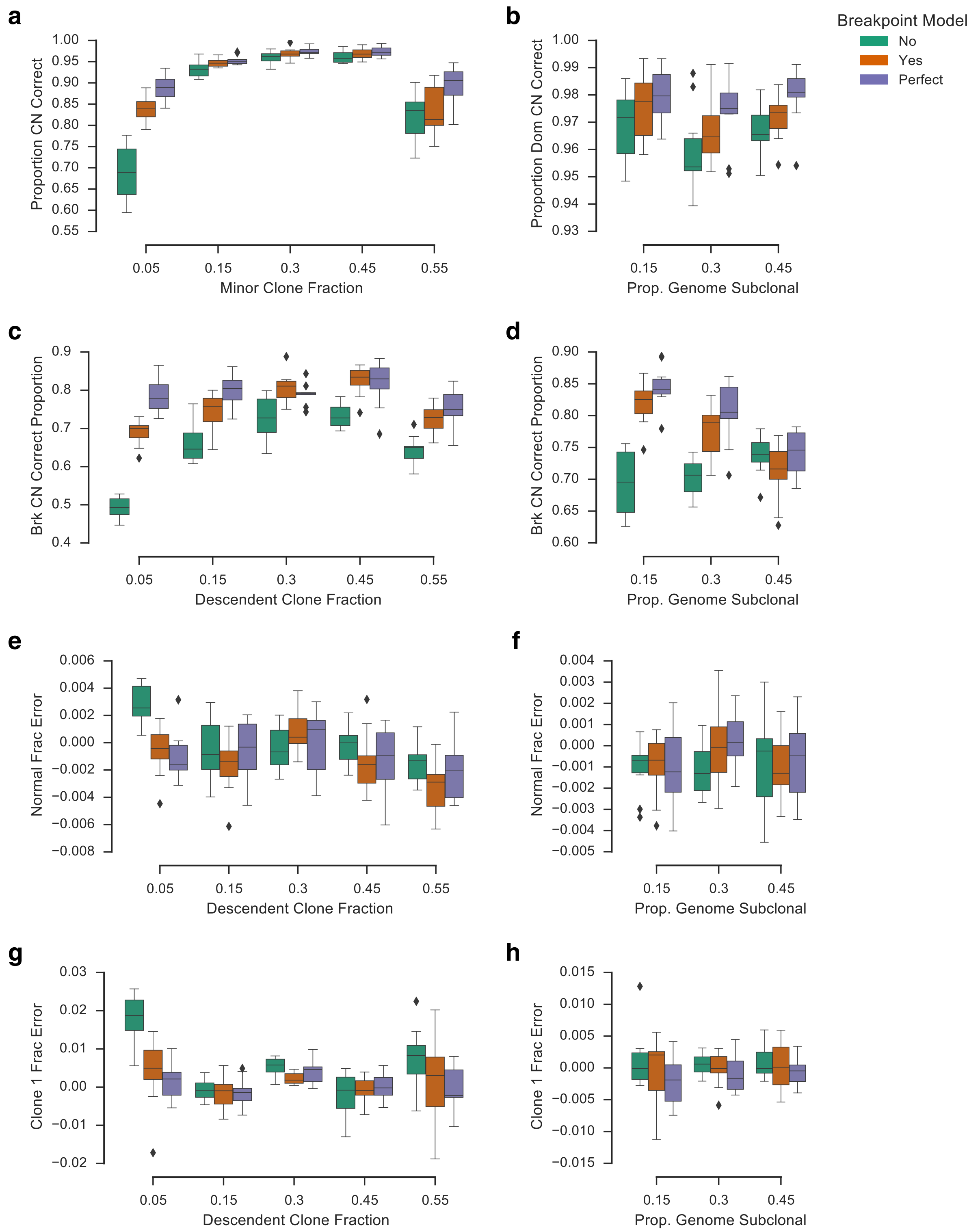


Fig. 2 Simulation results for the integrated breakpoint model and an equivalent hidden Markov model (*HMM*) with postprocessing to infer breakpoint copy number. Also shown are results for the breakpoint model with perfect initialization. Two sets of simulations were performed, varying fraction of the descendant tumour clone (*left column*) and proportion of the genome with divergent copy number (*right column*). Boxplots show proportion of the genome (**a**, **b**) and proportion of breakpoints (**c**, **d**) for which the tool correctly called clone-specific copy number, in addition to relative normal fraction error (**e**, **f**) and relative minor clone fraction error (**g**, **h**). *Boxes* show the interquartile (*IQR*) range with a *line* depicting the median. Whiskers extend 1.5 × IQR above quartile 3 and below quartile 1. *Diamonds* show positions of outlier data points

marginal probabilities of $q(C)$, denoted $\gamma_n(c)$ and $\gamma_n(c,c')$ respectively. The posterior marginals of $q(C)$ will appear in the updates of the other factors of q, as shown below.

Applying Eq. 4 to optimize $q_k(b_k)$ results in Eq. 9:

$$\begin{aligned}\log q_k^*(b_k) &= \sum_C q(C)\log p(X,C,B,h,\theta|L,O,\lambda) + \text{const}\\ &= \sum_{n:k_n=k}\sum_c\sum_{c'}\gamma_n(c,c')\log f(c,c',b_k|o,\lambda)\\ &\quad + \text{const} \qquad (9)\end{aligned}$$

Intuitively, the variational updates for $q(C)$ and $q_k(b_k)$ described above involve first updating the transition matrices of an HMM, weighting specific transitions that correspond to copy number changes induced by high-probability breakpoint copy number states, and then updating breakpoint copy number states according to the probabilities over adjacent segments in the HMM.

Since the entropy of a delta function is constant, optimal estimates of h and θ involve minimizing only the $\mathbb{E}_q\left[\log p(x,z)\right]$ term of the ELBO. Read counts are independent of breakpoints given segment copy number; thus, the expectation is calculated over $q(C)$ only (Eq. 10). Minimization is accomplished by computing derivatives with respect to the parameters and using quasi-Newton methods to find a local minimum.

$$\begin{aligned}\mathbb{E}_q\left[\log p(x,z)\right] &= \sum_C q(C)\log p(X,C,B,h,\theta|L,O,\lambda)\\ &= \sum_n\sum_c \gamma_n(c)\log p(x_n|c,h,l_n,\theta) \qquad (10)\end{aligned}$$

Realistic simulations of bulk genome sequencing

We developed a principled method of simulating rearranged genomes that fulfilled three important criteria. First, the simulated tumour genomes were required to have been produced by a known evolutionary history composed of duplication, deletion, and balanced rearrangement events applied successively to an initially non-rearranged normal genome. Second, the copy number profile of the simulated tumour genome should be reasonably similar to those of previously observed tumours. Third, the simulated data should be subject to the same biases seen in real genome sequence data.

To satisfy the first two criteria, we developed a sampling framework for generating realistic evolutionary histories based on a scoring and re-sampling strategy (see Additional file 1: Section 2.1). This first step produces a set of rearrangements, in addition to per-clone per-segment copy numbers. WGS read-level data are generated from segment copy numbers in one of two possible ways. For *segment count simulations*, read counts are simulated directly from a likelihood model given simulated segment copy number. For *aligned read re-sampling*, individual reads are re-sampled from a very high depth *source* normal genome dataset based on simulated segment copy number. By using an appropriate likelihood model, segment count simulations can be used to generate read counts with a distribution that reflects the over-dispersion and outliers in real data. Aligned read re-sampling datasets are computationally more intensive to generate, but are able to produce read count data with GC and mappability bias similar to that of the source dataset. See Additional file 1: Section 2.2 for additional details.

Breakpoint model improves inference for segment count simulations

We first sought to understand the benefit of an integrated breakpoint model using segment count simulations. We compared the ReMixT model with an equivalent breakpoint-naive HMM followed by post hoc breakpoint copy number calculation. For the breakpoint-naive model, we first infer segment copy number using the ReMixT model with breakpoint copy number at zero. We then use a simple greedy algorithm (see Additional file 1: Section 2.5) to perform a post hoc computation of the breakpoint copy number based on the segment copy number inferred using the HMM. As variational inference is sensitive to initialization, we also included results using the ReMixT breakpoint model with perfect initialization. We performed our evaluation on two sets of simulations, one in which we varied the proportion of the genome simulated to be subclonal, and one in which we varied the descendant clone fraction (see Additional file 1: Section 2.3 for details)[3].

We evaluated the breakpoint model and the HMM on the model's ability to recover the true clonal mixture, segment copy number, and breakpoint copy number (Fig. 2). Mixture prediction was assessed by calculating the relative deviation of the predicted normal fraction and descendant clone fraction from the simulated values. Segment and breakpoint copy number prediction was assessed by calculating the proportion of segments/breakpoints for which the true clone-specific copy number was recovered by the method.

For both segment and breakpoint copy number prediction, the breakpoint model outperformed the baseline HMM. The proportion of segment copy number called correctly was significantly higher for the breakpoint model for all simulations with the exception of those simulations with a descendant clone fraction of 55% (paired t test, p value $<$ 0.05, Fig. 3a and b). Additionally, the proportion of breakpoints with correctly predicted copy number was significantly higher for the breakpoint model for all simulations with the exception of those with the proportion of the genome subclonal set at 45% (paired t test, p value $<$ 0.05, Fig. 3c and d). Improvement with respect to prediction of minor clone fraction was observed for descendant clone fractions 0.05 and 0.3 (paired t test,

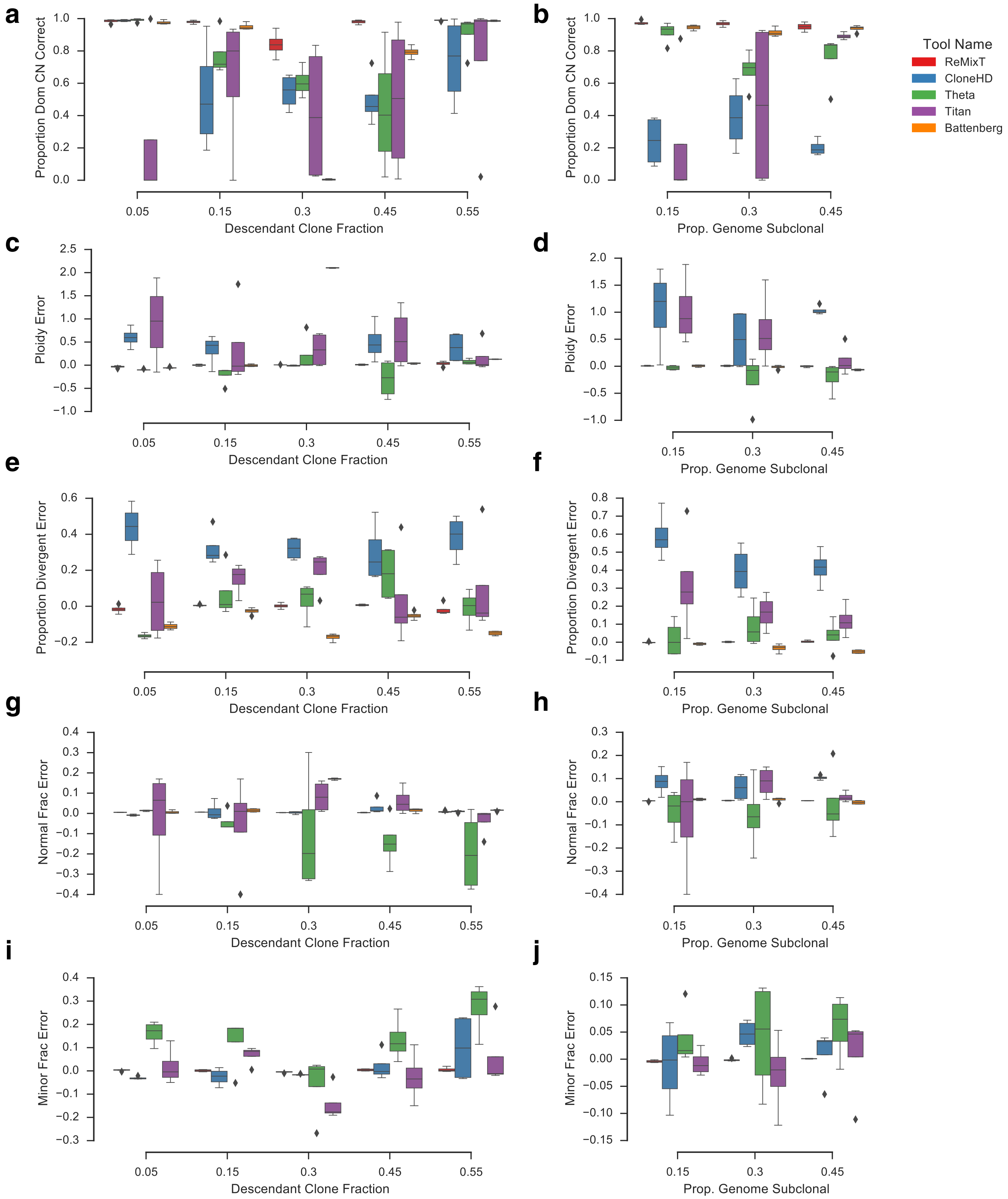


Fig. 3 Performance comparison of ReMixT with CloneHD, TITAN, Battenberg, and THetA using read re-sampling simulations. Two sets of simulations were performed, varying fraction of the descendant tumour clone (*left column*) and proportion of the genome with divergent copy number (*right column*). *Boxplots* show proportion of the genome for which the tool correctly called the copy number of the dominant clone (**a**, **b**), relative mean ploidy error compared to simulated (**c**, **d**), relative proportion divergent error compared to simulated (**e**, **f**), relative normal fraction estimation error compared to simulated (**g**, **h**), and relative minor clone fraction estimation error compared to simulated (**i**, **j**). Battenberg was excluded from the minor clone fraction benchmark, as it does not produce a global estimate of this parameter. *Boxes* show the interquartile (*IQR*) range with a *line* depicting the median. Whiskers extend 1.5 × IQR above quartile 3 and below quartile 1. *Diamonds* show positions of outlier data points

p value $<$ 0.05, Fig. 3g). No improvement was observed with respect to normal fraction prediction, though we did observe a decrease in accuracy for descendant clone fraction 0.55 (paired t test, p value $=$ 0.03, Fig. 3e). Perfect initialization showed improved results over our current initialization method, indicating additional room for improvement with respect to this aspect of the algorithm.

Comparison with existing copy number inference methods

We used our aligned read re-sampling framework to compare the performance of ReMixT to four existing methods for subclonal copy number inference: TITAN [5], CloneHD [7], Battenberg [8], and THetA2 [12, 14]. We performed our comparison on two sets of genome mixtures, one in which we varied the proportion of the genome simulated to be subclonal, and one in which we varied the descendant clone fraction. We used aligned read re-sampling to produce realistic simulated datasets using 200X sequencing of the NA12878 hapmap individual provided by Illumina [22]. Each tool was run with default parameters according to available instructions (see Additional file 1: Section 4 for details).

Performance of the four tools varied significantly across each measure (Fig. 3). CloneHD was unable to recover the copy number of the dominant clone with reasonable accuracy for a majority of the simulations ($<$ 43% accurate for 50% of simulations). In general, CloneHD copy number results showed a higher mean ploidy and higher divergent proportion (proportion of the genome predicted to have clonally divergent copy number) than simulated results (average 37% higher and 44% higher respectively). However, in many instances, CloneHD was able to estimate normal fraction with reasonable accuracy (within 6.6% of simulated for 50% of the simulations). Minor clone fraction estimation was less accurate (within 28% of simulated for 50% of the simulations). Our results imply that CloneHD is prone to over-fitting, producing unrealistic copy number profiles.

THetA, by contrast, produced solutions accurate with respect to mean ploidy (within 6.5% of simulated for 75% of simulations) and, to a lesser extent, divergent proportion (within 20% of simulated for only 25% of simulations). Additionally, THetA copy number predictions were more consistent in their accuracy, with the dominant copy number predicted with greater than 81% accuracy for 50% of the simulations. The normal fraction estimation error was in general higher than for the other tools (within 17% of simulated for 50% of simulations). THetA's estimated descendant clone fractions were also less accurate than those of the other tools (within 21% of simulated for only 25% of simulations).

TITAN's results were the most variable, with dominant copy predicted accurately for a large number of simulations ($>$ 88% for 25% of simulations) but poorly for many other simulations ($<$ 21% for 25% of simulations). As with CloneHD, TITAN appeared to over-fit for a subset of the simulations, producing solutions for which mean ploidy and divergent proportion were higher than simulated ($>$ 28% higher than simulated ploidy for 25% of simulations and $>$ 66% higher than simulated divergent proportion for 50% of simulations). TITAN estimated normal fractions with low error for a majority of simulations (within 5% of simulated for 50% of simulations), though prediction of minor clone fractions was more variable (error greater than 19% of simulated for 75% of simulations).

Battenberg's results were the most consistent of the competing tools. For the simulations with 50/50 tumour mixtures, Battenberg produced a solution at double the simulated ploidy, highlighting the unidentifiability of this particular scenario. Excluding the 50/50 tumour mixture simulations, Battenberg predicted dominant copy number within 3% for 75% of the simulations and ploidy within 4% for 75% of the simulations. Battenberg in general under-estimated the divergent proportion, 13% lower than simulated for 75% of simulations. Normal fractions were also accurate, within 6% of simulated for 100% of simulations, excluding 50/50 mixtures. Battenberg does not estimate minor clone fraction and was thus excluded from such analyses.

ReMixT consistently outperformed the four competing tools on all measures. For 75% of the simulations, ReMixT was able to infer integer copy number for both clones with greater than 91% accuracy. Lower accuracy results were obtained for 50/50 tumour mixtures, primarily due to the inherent ambiguity of assigning copy numbers to specific clones for such mixtures. Normal fraction estimation was slightly biased, and was over-estimated by 1.4% of simulated on average, though never by more than 2.6%. As expected, minor clone fraction estimation was less accurate for mixtures with the smallest simulated minor clone fractions, up to 50% of simulated, averaging 5%. For the remaining simulations minor clone fraction estimation error averaged 0.6% with a maximum of 8%.

Targeted single cell validation of clone-specific breakpoints

Next we sought to establish the accuracy of breakpoint copy number inference in a realistic setting using targeted single cell sequencing in a set of specially separated high-grade serous ovarian tumour samples [23]. The set of samples included two obtained from the patient's right ovary, one from the left ovary, and one from the omentum (Fig. 5b). Each sample was whole genome sequenced to an approximate depth of 30X.

We hand-selected 12 breakpoints associated with putative copy number changes for validation by targeted single cell sequencing (Fig. 4). Specifically, for each of the 12 candidate breakpoints, at least one breakend coincided

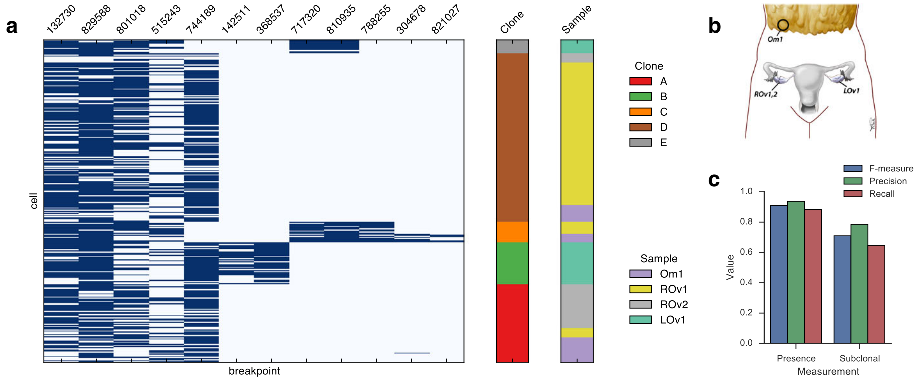


Fig. 4 Single cell validation of ReMixT results for 12 breakpoints in 294 cells from 4 HGS Ovarian tumour samples: Omentum 1 (Om1), Right Ovary 1 and 2 (ROv1 and ROv2), and Left Ovary 1 (LOv1). (**a**) Breakpoint (*x*-axis) by cell (*y*-axis) presence (*dark blue*) / absence (*light blue*) with cells annotated by sample of origin and clone as inferred by the Single Cell Genotyper. (**b**) Approximate anatomic location of the 4 tumour samples. (**c**) F-measure, precision and recall for ReMixT calls of breakpoint presence and subclonality

with a transition in copy number in at least one sample, where copy number was inferred using an earlier version of ReMixT [23]. In addition, we selected 60 somatic and 24 germline single nucleotide changes based on their utility as clonal markers [23]. Targeted single cell sequencing was performed as previously described [23], cells were clustered into clones using the Single Cell Genotyper [24], and breakpoints were assigned to clones if they were present in at least three cells of that clone. Joint analysis of the breakpoint and single nucleotide data produced a robust estimate of the clonal genotypes with respect to the targeted breakpoints (Fig. 4a).

Next we evaluated the ability of ReMixT to accurately determine which breakpoints were present/absent and clonal/subclonal in each sample. We calculated the *F* measure for present/absent and clonal/subclonal calls (Fig. 4c). *F* measure values were similar to results obtained from running ReMixT on aligned read re-sampling simulations.

Tracking clonal expansions using clone-specific breakpoints

Several previous studies have used clone-specific SNVs to identify patterns of clonal evolution [25], infer patterns of cancer cell dissemination to metastatic sites [23, 26], and track expansion and contraction of tumour clones over time and in response to therapy [27] and in response to xenograft passaging [15]. We sought to evaluate the utility of clone-specific breakpoints predicted by ReMixT for investigating clonal evolution in successive xenograft passages. To this end, we analysed primary and xenograft tumour samples derived from a patient with breast cancer (SA501 from [15]). Our analysis focused on four samples, the primary tumour sample and three xenograft samples labelled X1A, X3A, and X3F. The relationship between these four samples and the additional two un-sequenced xenograft samples X2A and X2F is shown in Fig. 5b.

For validation of X3F clone-specific copy number changes, we used recently published single cell WGS data [28]. We inferred total integer copy number and performed phylogenetic analysis using previously described techniques [15, 28]. Three major clones were identified. Proportions of cells assigned to each clone were 0.82, 0.11, and 0.07 for clones A, B, and C respectively. Clones B and C were highly similar and formed a distinct clade; thus, for this analysis we merged clones B and C. For clone A and merged clone BC, we reconstructed clone copy number profiles by selecting the most prevalent copy number within each clone for each segment. Segments with copy number 6 or higher were removed, as specific copy number states above 5 could not be inferred using available techniques.

ReMixT analysis using default parameters estimated a clonal mixture of 0.85 for the dominant clone and 0.15 for the minor clone. Clone-specific copy numbers matched single cell copy number for 91% of the genome. Accuracy was highest for segments in lower copy number states ($\leq$ 3 total copies). Segments with higher copy number ($\geq$ 4 total copies) and no clonal divergence were frequently predicted as subclonal by ReMixT, evidence that ReMixT over-fits some segments with higher copy number (Fig. 5c). Additional disparity appeared to be the result of noisy segments in lower copy states predicted as subclonal.

Next we identified a set of high confidence subclonal breakpoints for analysis of clonal dynamics in the xenograft passages. We smoothed segments smaller than 100 kb and aggregated adjacent segments with the

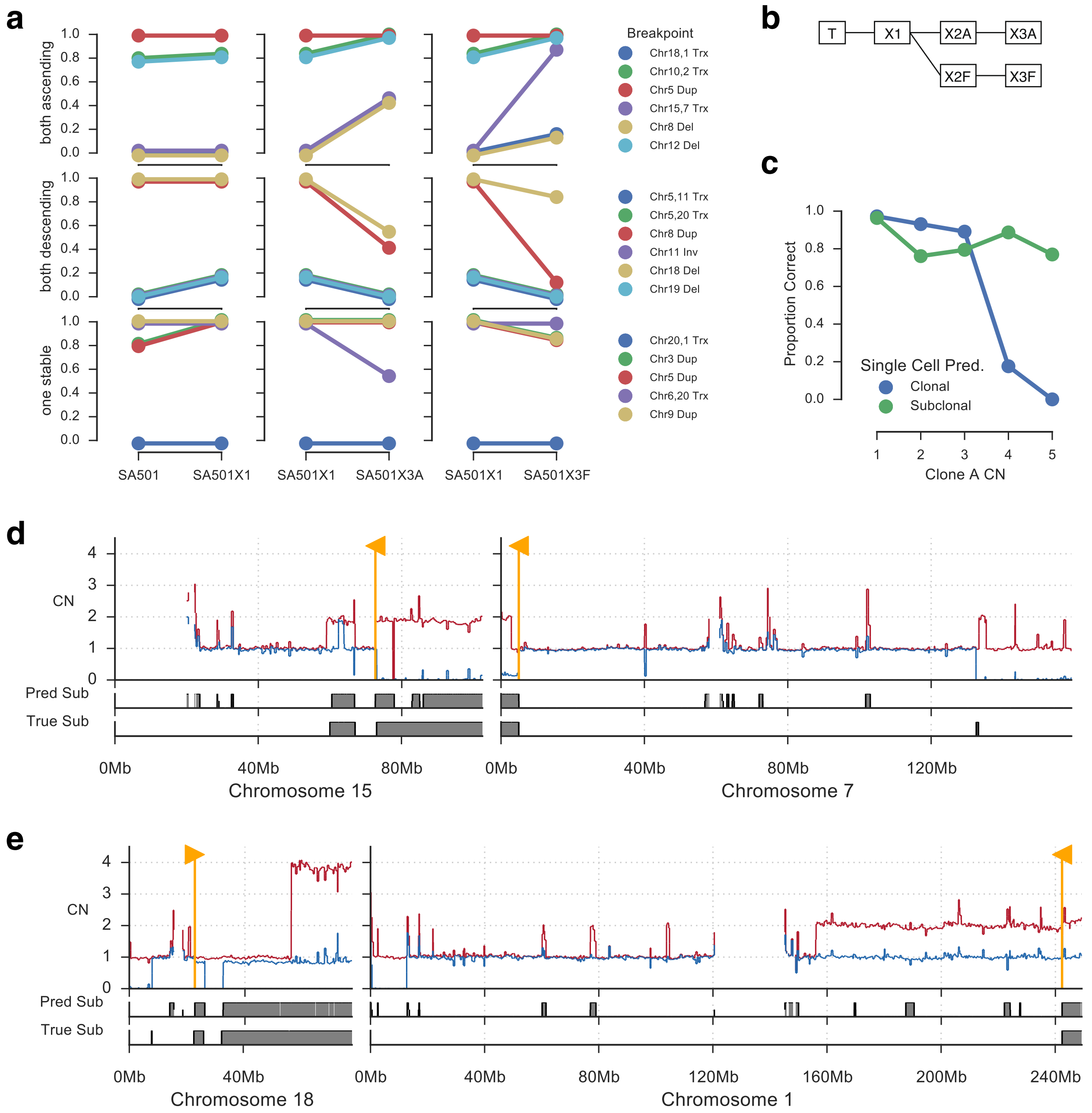


Fig. 5 Tracking clonal expansions in xenograft passages. **a** Breakpoints identified by ReMixT as clone-specific were classified according to their clonal prevalence change between SA501X1A and replicate xenograft passages SA501X3A and SA501X3F. All breakpoints could be classified as ascending in both SA501X3A and SA501X3F, descending in both, or stable in at least one. Shown are the clonal prevalence changes between pairs of samples for which WGS was available. **b** Relationship between primary tumour sample T and xenograft passages X*. **c** Accuracy of copy number inference for X3F based on single cell whole genome sequencing. Shown is the proportion of regions with correctly predicted copy number (*y*-axis) for each clone A copy number (*x*-axis), split between clonal and subclonal (*blue/green*) as determined from single cell data. **d** Copy number profile (*top*) for chromosomes 7 and 15 showing corroboration between single cell (*bottom*) and ReMixT (*middle*) subclonal copy number prediction. *Yellow flags* show the location of translocation breakpoints predicted to be subclonal by ReMixT. **e** Similarly, chromosomes 1/18 translocation breakpoints predicted to be subclonal by ReMixT. Copy number plots show raw major (*red*) and minor (*blue*) copy numbers

same allele-specific difference between clone copy numbers. We then removed segments with length less than 1 Mb or copy number greater than 4. Breakpoints were selected if they were predicted to be subclonal, and were immediately adjacent at each breakend to a segment with subclonal copy number from the above set of filtered high confidence segments. This technique was used to identify 17 subclonal breakpoints in one of X1, X3A, X3F, and X5 or the primary tumour sample. In X3F, the ReMixT copy number matched the single cell copy number for 84% of the 1-Mb regions to either side of each breakend. For 11 of the predictions, corroboration was >92%, and for the remaining predictions, corroboration was closer to 50%, indicating a lack of corroboration on one side of each breakend. Included in the set of breakpoints were inter-chromosomal translocations linking subclonal segments on disparate chromosomes, indicative of clone-specific loss or gain of rearranged tumour chromosomes (Fig. 5d and e).

Patient SA501 was previously shown to have exhibited reproducible patterns of clonal expansions across multiple replicate xenografts using a combination of targeted bulk and single cell sequencing of SNVs [15]. In particular, X3A and X3B showed similar patterns of clonal expansions for clusters of SNVs used as clonal markers. We sought to establish whether the same clonal dynamics were evident in X3F, and whether those clonal dynamics could be understood using clonal-specific breakpoints. To that end, we classified each of the high confidence subclonal breakpoints according to whether they exhibited the same expansion patterns from X1 to X3A and X1 to X3F. Of the 17 high confidence breakpoints, 6 could be classified as ascending in both X3A and X3F, 6 as descending in both X3A and X3F, with the remaining stable from X1 to either X3A or X3F (Fig. 5a). Strikingly, we did not identify any conflicting breakpoints, those ascending in X3A and descending in X3F or vice versa.

Assembling tumour chromosomes using subclonal breakpoints

We applied ReMixT to WGS data from two tumour-derived cell line samples and a matched normal sample obtained from a patient with HGSOvCa [29]. The two cell lines are derived from an ascites sample (DAH354) and a primary tumour sample (DAH355) obtained during debulking surgery. Cell line samples and matched normals were sequenced to approximately 30X and analysed with ReMixT using default parameters. Tetraploid solutions were selected based on ploidy evidence from preliminary single cell sequencing experiments for DAH355 (data not shown).

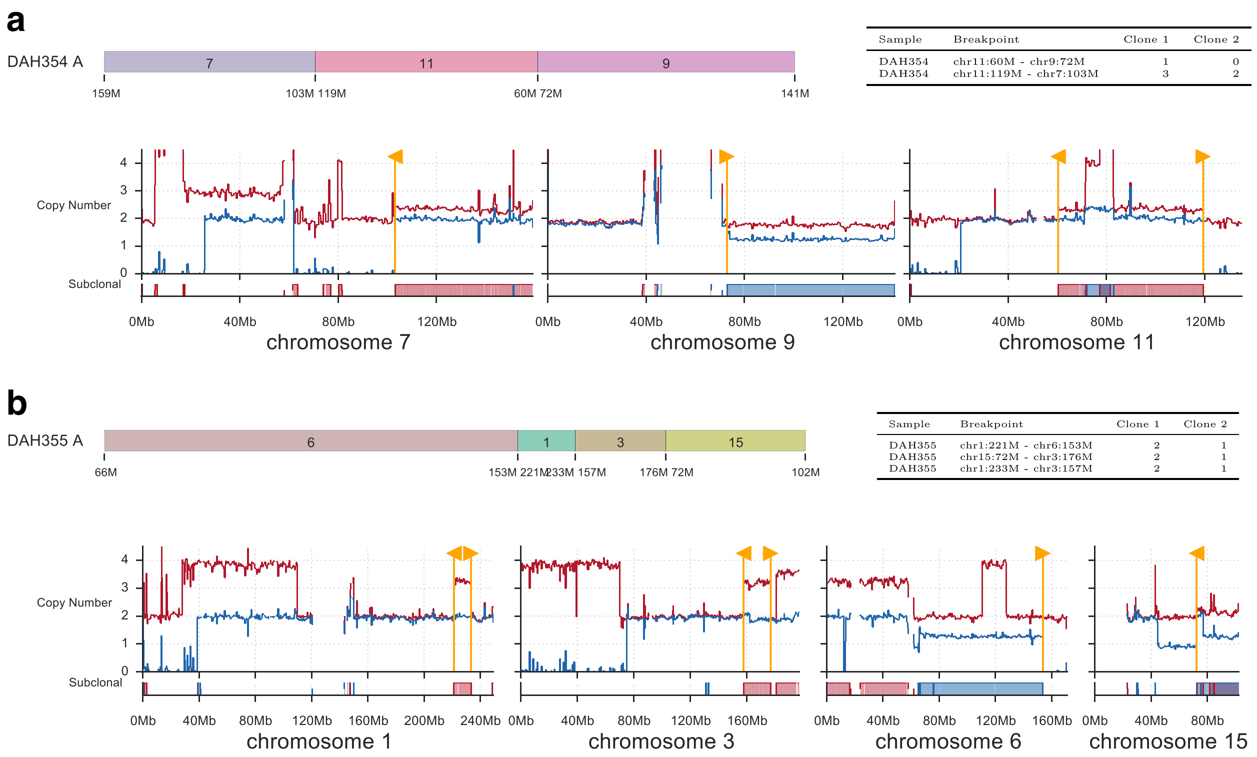
As expected of HGSOvCa, the copy number profiles of the cell line samples showed substantial evidence of genome instability. For both samples, the fraction of the genome predicted to be diploid heterozygous was insignificant, and the fraction of the genome with loss of heterozygosity was 40% and 35% for DAH354 and DAH355 respectively. Both DAH354 and DAH355 showed evidence of multiple genomically distinct clonal populations, with dominant clone fractions of 0.7 and 0.61 respectively, and fraction of the diploid genome predicted as subclonal as 14% and 32% respectively. A total of 348 somatic breakpoints were identified by deStruct [4], of which 278 were determined to be present (positive copy number) by ReMixT in one or both samples. A total of 97 breakpoints were predicted to have clone-specific copy number in one or both samples, with 17 having clone-specific copy number in both samples.

In both DAH354 and DAH355, we observed several clone-specific translocations adjacent to large segments with clonally divergent copy numbers. As with SA501, we suspected that the loss or duplication of a single tumour chromosome would result in multiple clonally divergent segments across the reference genome. We thus searched for clonally divergent segments connected by subclonal breakpoints as a method for understanding the structure of tumour chromosomes with divergent copy number across the clonal population (Fig. 6). In DAH354, we identified a tumour chromosome composed of three segments from reference chromosomes 7, 11, and 9 (Fig. 6a), and in DAH355, we identified a tumour chromosome composed of four segments from reference chromosomes 6, 1, 3, and 15 (Fig. 6b).

Discussion

We have demonstrated that ReMixT improves both inference and interpretation of copy number changes and genomic rearrangements. Improved accuracy was observed for prediction of clone fraction, clone specific copy number, and clone specificity of breakpoints. We show how breakpoint copy number changes can be used a markers of clonal populations, and used to track clonal population dynamics in the same way as SNVs. By linking clone specific copy number changes to breakpoints we show how targeted single cell sequencing can be used to jointly profile clonal genotypes in SNV and copy number space. Furthermore, we are able to reconstruct partial tumour chromosomes lost or gained in sub-populations of cells.

Although our method shows performance gains over other methods, further improvements are possible. The performance of our variational inference algorithm is highly dependent on the quality of the initialization. Improvement may be gained using more sophisticated or informed initialization methods, or extensions to variational inference using annealing or MCMC. Our current implementation is limited to two tumour clones, largely due to the increased computational complexity of modelling additional clones. An approximating distribution factorized per clone would solve the complexity issue within the context of structured variational inference, however based on our own experimentation, such a factorization exacerbates the initialization problem and was found to be infeasible. Thus improvements to the



Sample	Breakpoint	Clone 1	Clone 2
DAH354	chr11:60M - chr9:72M	1	0
DAH354	chr11:119M - chr7:103M	3	2

Sample	Breakpoint	Clone 1	Clone 2
DAH355	chr1:221M - chr6:153M	2	1
DAH355	chr15:72M - chr3:176M	2	1
DAH355	chr1:233M - chr3:157M	2	1

Fig. 6 Inference of partial tumour chromosome assemblies based on linking subclonal segments and breakpoints. Two assembled chromosomes are shown for cell lines DAH354 (**a**) and DAH355 (**b**). Shown for each assembled chromosome is a schematic of the segments involved (*top left*), a table of breakpoint copy number predicted by ReMixT (*top right*), and a chromosome copy number plot (*bottom*). Each copy number plot shows raw major (*red*) and minor (*blue*) copy numbers (*top axis*), in addition to prediction of subclonality (*bottom axis*)

variational inference method may also allow for the use of a more factorized approximation, removing the limitation on the number of clones.

Conclusions

Traditionally, classes of genomic aberration have been predicted and characterized independently, with post-hoc analysis to determine correlation between events in each class. However, there are clear dependencies between classes of aberrations with respect to their generation via mutational processes and their observation using genome sequencing. A number of existing methods partially leverage class dependencies[7, 30, 31], and the development of ReMixT represents a further step towards a comprehensive model of genomic aberrations in tumour populations. We anticipate further benefit may be gained from jointly modelling copy number changes, rearrangements, SNPs and SNVs, all within the context of an appropriate phylogenetic model. Future research leveraging the patterns of genome damage and the totality of somatic alterations in a cancer's evolutionary history to elucidate its biologic and mutagenic properties will derive benefit from ReMiXT's improved accuracy in structural alteration detection and interpretation.

Endnotes

[1] A product of normalized conditional probabilities and a prior probability for the first segment would also be possible, though we believe integration of breakpoints into the model would be less intuitive.

[2] Assuming uniform improper priors over h and θ, we have $\log p(X, C, B|h, \theta, L, O, \lambda) = \log p(X, C, B, h, \theta|L, O, \lambda) + \text{const}$.

[3] We maintained a distinction between ancestral/descendant clone mixtures of x / $1 - x$ and the reversed $1 - x$ / x clone mixture, as results for these mixtures differ.

Additional file

Additional file 1: Supplementary methods, supplementary analyses, and additional experimental results. (PDF 35000 kb)

Acknowledgements
We thank Dr. Anne-Marie Mes-Masson, CHUM Research Centre (CRCHUM), Montreal, Canada for providing the high-grade serous ovarian cancer cell lines.

Funding

This work was supported by a Discovery Frontiers project grant, The Cancer Genome Collaboratory, jointly sponsored by the Natural Sciences and Engineering Research Council (NSERC), Genome Canada (GC), the Canadian Institutes of Health Research (CIHR), and the Canada Foundation for Innovation (CFI) to SPS and SCS. In addition, we acknowledge generous long-term funding support from the BC Cancer Foundation. The SPS and SA groups receive operating funds from the Canadian Breast Cancer Foundation, the Canadian Cancer Society Research Institute (impact grant 701584 to SA and SPS), the Terry Fox Research Institute (Program Project Grants program on forme fruste tumours), the CIHR (grant MOP-115170 to SA and SPS), and the CIHR Foundation (grant FDN-143246 to SPS). SPS and SA are supported by Canada Research Chairs. SPS is a Michael Smith Foundation for Health Research scholar. Additional funding is provided by NIH GM108348 and by the Indiana University Precision Health Grand Challenge Initiative to SCS.

Availability of data and materials

ReMixT is written in Python and C++. Source code is available at http://bitbucket.org/dranew/remixt, released under the MIT license. The version of ReMixT used for this manuscript has been published on http://zenodo.org with doi: 10.5281/zenodo.819479.

High-grade serous ovarian cancer

We used previously published data [23] available from the European Genome-phenome Archive (http://www.ebi.ac.uk/ega/) under accession number [EGA:S00001000547].

Breast cancer

We used previously published data [28] available from the European Genome-phenome Archive (http://www.ebi.ac.uk/ega/) under accession number [EGA:S00001002170].

Authors' contributions

AM developed and implemented the algorithm, performed the analysis, and wrote the manuscript. AS analyzed the SA501 single cell data. PE was responsible for development of the xenograft models. SPS contributed to the manuscript text. AR, GH, CC, CD, AB, SCS, and SA contributed ideas during development of the model and analysis of the results. All authors read and approved the final manuscript.

Ethics approval and consent to participate

High-grade serous ovarian cancer.

Ethical approval was obtained from the University of British Columbia (UBC) Research Ethics Board (H08-01411 NGS Huntsman). Women undergoing debulking surgery (primary or recurrent) for carcinoma of ovarian, peritoneal, and/or fallopian tube origin were approached for informed consent for the banking of tumour tissue. All experimental methods comply with the Helsinki Declaration.

Breast cancer

Anonymized tumour tissue from women aged 26–82 undergoing surgery or diagnostic core biopsy was collected with informed consent, according to procedures approved by the UBC Research Ethics Board (H06-00289 Breast Tumour Tissue Repository and H13-01125 Breast Xenograft Aparicio). All experimental methods comply with the Helsinki Declaration.

Patient-derived xenografts

Female NOD/SCID interleukin-2 receptor gamma null (NSG) and NOD Rag-1 null interleukin-2 receptor gamma null (NRG) mice were bred and housed at the Animal Resource Centre at the British Columbia Cancer Research Centre and the Biological Resource Unit at the Cancer Research UK Cambridge Research Institute. Surgery was carried out on mice between the ages of 5–10 weeks. All experimental procedures were approved by the University of British Columbia Animal Care Committee and the University of Cambridge Animal Welfare and Ethical Review Committee.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

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

Author details

[1]Department of Molecular Oncology, BC Cancer Agency, 675 West 10th Avenue, Vancouver, BC, Canada. [2]Department of Pathology and Laboratory Medicine, University of British Columbia, 2329 West Mall, Vancouver, BC, Canada. [3]Department of Statistics, Oxford University, 24-29 St Giles, Oxford, United Kingdom. [4]Ludwig Institute for Cancer Research, Oxford University, Old Road Campus Research Building, Headington, Oxford, United Kingdom. [5]Dana-Farber Cancer Institute, 450 Brookline Ave, Oxford, Boston, USA. [6]Eli and Edythe L. Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA, USA. [7]Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada. [8]Department of Statistics, University of British Columbia, 2329 West Mall, Vancouver, BC, Canada. [9]Vancouver Prostate Centre, 2660 Oak Street, Vancouver, Canada. [10]Department of Computer Science, Indiana University Bloomington, 107 S. Indiana Avenue, Bloomington, IN, USA.

Received: 16 March 2017 Accepted: 3 July 2017
Published online: 27 July 2017

References

1. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. Delly: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):333–9.
2. Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res. 2009;19(7):1270–8.
3. Layer RM, Chiang C, Quinlan AR, Hall IM. Lumpy: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):84.
4. McPherson A, Shah SP, Sahinalp SC. deStruct: accurate rearrangement detection using breakpoint specific realignment. bioRxiv. 2017. https://doi.org/10.1101/117523.
5. Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, Melnyk N, McPherson A, Bashashati A, Laks E, Biele J, Ding J, Le A, Rosner J, Shumansky K, Marra MA, Gilks CB, Huntsman DG, McAlpine JN, Aparicio S, Shah SP. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. 2014;24(11):1881–93.
6. Oesper L, Mahmoody A, Raphael BJ. THetA: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. Genome Biol. 2013;14(7):80.
7. Fischer A, Vázquez-García I, Illingworth CJR, Mustonen V. High-definition reconstruction of clonal composition in cancer. Cell Rep. 2014;7(5): 1740–52.
8. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, Shlien A, Cooke SL, Hinton J, Menzies A, Stebbings LA, Leroy C, Jia M, Rance R, Mudie LJ, Gamble SJ, Stephens PJ, McLaren S, Tarpey PS, Papaemmanuil E, Davies HR, Varela I, McBride DJ, Bignell GR, Leung K, Butler AP, Teague JW, Martin S, Jönsson G, Mariani O, Boyault S, Miron P, Fatima A, Langerød A, Aparicio SAJR, Tutt A, Sieuwerts AM, Borg Å, Thomas G, Salomon AV, Richardson AL, Børresen-Dale AL, Futreal PA, Stratton MR, Campbell PJ, Breast Cancer Working Group of the International Cancer Genome Consortium. The life history of 21 breast cancers. Cell. 2012;149(5): 994–1007.
9. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. Control-freec: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28(3):423–5.
10. Mahmoody A, Kahn CL, Raphael BJ. Reconstructing genome mixtures from partial adjacencies. BMC Bioinforma. 2012;13 Suppl 19:9.
11. Zerbino DR, Paten B, Hickey G, Haussler D. An algebraic framework to sample the rearrangement histories of a cancer metagenome with double cut and join, duplication and deletion events. arXiv. 2013. arXiv:1303.5569v1.
12. Oesper L, Ritz A, Aerni SJ, Drebin R, Raphael BJ. Reconstructing cancer genomes from paired-end sequencing data. BMC Bioinformatics. 2012;13 Suppl 6:10.
13. Li Y, Zhou S, Schwartz DC, Ma J. Allele-specific quantification of structural variations in cancer genomes. Cell Syst. 2016;3(1):21–34.
14. Oesper L, Satas G, Raphael BJ. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. Bioinformatics. 2014;30(24):3532–40. doi:10.1093/bioinformatics/btu651.

15. Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, Gelmon K, Chia S, Mar C, Wan A, Laks E, Biele J, Shumansky K, Rosner J, McPherson A, Nielsen C, Roth AJL, Lefebvre C, Bashashati A, de Souza C, Siu C, Aniba R, Brimhall J, Oloumi A, Osako T, Bruna A, Sandoval JL, Algara T, Greenwood W, Leung K, Cheng H, Xue H, Wang Y, Lin D, Mungall AJ, Moore R, Zhao Y, Lorette J, Nguyen L, Huntsman D, Eaves CJ, Hansen C, Marra MA, Caldas C, Shah SP, Aparicio S. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature. 2015;518(7539):422–6.
16. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012;9(2):179–81.
17. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):72.
18. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. Detecting copy number variation with mated short reads. Genome Res. 2010;20(11): 1613–22.
19. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. 2016. arXiv preprint arXiv:1601.00670.
20. Saul LK, Jordan MI. Exploiting tractable substructures in intractable networks. Adv Neural Inform Process Syst. 1996;486–92.
21. Ghahramani Z, Jordan MI, Smyth P. Factorial hidden Markov models. Mach Learn. 1997;29(2-3):245–73.
22. Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, Iqbal Z, Chuang HY, Humphray SJ, Halpert AL, et al. A reference dataset of 5.4 million human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. bioRxiv. 2016;055541.
23. McPherson A, Roth A, Laks E, Masud T, Bashashati A, Zhang AW, Ha G, Biele J, Yap D, Wan A, Prentice LM, Khattra J, Smith MA, Nielsen CB, Mullaly SC, Kalloger S, Karnezis A, Shumansky K, Siu C, Rosner J, Chan HL, Ho J, Melnyk N, Senz J, Yang W, Moore R, Mungall AJ, Marra MA, Bouchard-Côté A, Gilks CB, Huntsman DG, McAlpine JN, Aparicio S, Shah SP. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. Nat Genet. 2016;48(7):758–67.
24. Roth A, McPherson A, Laks E, Biele J, Yap D, Wan A, Smith MA, Nielsen CB, McAlpine JN, Aparicio S, Bouchard-Côté A, Shah SP. Clonal genotype and population structure inference from single-cell tumor sequencing. Nat Methods. 2016;13(7):573–6.
25. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G, Bashashati A, Prentice LM, Khattra J, Burleigh A, Yap D, Bernard V, McPherson A, Shumansky K, Crisan A, Giuliany R, Heravi-Moussavi A, Rosner J, Lai D, Birol I, Varhol R, Tam A, Dhalla N, Zeng T, Ma K, Chan SK, Griffith M, Moradian A, Cheng S.-W. G, Morin GB, Watson P, Gelmon K, Chia S, Chin SF, Curtis C, Rueda OM, Pharoah PD, Damaraju S, Mackey J, Hoon K, Harkins T, Tadigotla V, Sigaroudinia M, Gascard P, Tlsty T, Costello JF, Meyer IM, Eaves CJ, Wasserman WW, Jones S, Huntsman D, Hirst M, Caldas C, Marra MA, Aparicio S. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486(7403):395–9.
26. Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, Brewer DS, Kallio HML, Högnäs G, Annala M, Kivinummi K, Goody V, Latimer C, O'Meara S, Dawson KJ, Isaacs W, Emmert-Buck MR, Nykter M, Foster C, Kote-Jarai Z, Easton D, Whitaker HC, ICGC Prostate UK Group, Neal DE, Cooper CS, Eeles RA, Visakorpi T, Campbell PJ, McDermott U, Wedge DC, Bova GS. The evolutionary history of lethal metastatic prostate cancer. Nature. 2015;520(7547):353–7.
27. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, DiPersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012;481(7382):506–10.
28. Zahn H, Steif A, Laks E, Eirew P, VanInsberghe M, Shah SP, Aparicio S, Hansen CL. Scalable whole-genome single-cell library preparation without preamplification. Nat Methods. 2017;14(2):167–73.
29. Létourneau IJ, Quinn MCJ, Wang LL, Portelance L, Caceres KY, Cyr L, Delvoye N, Meunier L, de Ladurantaye M, Shen Z, Arcand SL, Tonin PN, Provencher DM, Mes-Masson AM. Derivation and characterization of matched cell lines from primary and recurrent serous ovarian cancer. BMC Cancer. 2012;12:379.
30. Greenman CD, Pleasance ED, Newman S, Yang F, Fu B, Nik-Zainal S, Jones D, Lau KW, Carter N, Edwards PAW, Futreal PA, Stratton MR, Campbell PJ. Estimation of rearrangement phylogeny for cancer genomes. Genome Res. 2012;22(2):346–61.
31. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumours. Genome Biol. 2015;16:35.