SCABSOLUTE: MEASURING SINGLE-CELL PLOIDY AND REPLICATION STATUS

Michael P Schneider[1,2], Geoff Macintyre[3], and Florian Markowetz[1,2,*]

[1]University of Cambridge, Cambridge, United Kingdom
[2]Cancer Research UK Cambridge Institute, Cambridge, United Kingdom
[3]Computational Oncology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[*]Corresponding author: Florian Markowetz, florian.markowetz@cruk.cam.ac.uk

ABSTRACT

Chromosomal instability is a common characteristic of many cancers. Chromosomally instable tumour cells exhibit frequent copy number aberrations (CNAs) and a wide variation in the amount of DNA in cancer cells, referred to as cell ploidy. High levels of ploidy, in particular, are associated with whole genome doubling (WGD), a widespread macro-evolutionary event in tumour history. Individual cells' genomes are also undergoing replication as part of the cell cycle, and this constitutes an important covariate for single-cell genome analysis. Accurate and unbiased measurement of single-cell ploidy and replication status, including WGDs, based on DNA sequencing data is important for many downstream applications, such as detecting genomic variants, quantifying intratumour heterogeneity, and reconstructing tumour evolutionary phylogenies.

Here we present *scAbsolute*, an approach to measure ploidy and replication status in single cells using scalable stochastic variational inference with a constrained Dirichlet Process Gaussian Mixture Model.

We demonstrate its accuracy across three sequencing technologies (10X, DLP, ACT) and different cell lines and tumour samples. We address the problem of identifying cells with double the amount of DNA, but otherwise identical copy number profiles as is the case after WGD, solely based on sequencing information. Finally, we provide a robust and general method for identifying cells undergoing DNA replication.

scAbsolute provides a scalable and unbiased way of ascertaining single-cell ploidy and replication status, paving the way for accurate detection of CNAs and WGDs in single-cell DNA sequencing data.

Introduction

Many common cancers are characterised by chromosomal instability (CIN) and as a result, extensive copy number aberrations (CNAs)[1,2]. CNAs alter the number of copies of genomic regions in a cell, thus creating a background of genomic variation on which evolution can act[3]. CNAs can act as drivers of cancer evolution[4–6] and be used to infer phylogenies[7,8]. Importantly, CNAs have been shown to play a crucial role in cancer treatment and prognosis[9–11], and they correlate with markers of immune evasion and increased activity in proliferation pathways[6,12,13]. As a consequence of CIN, tumour cells, unlike normal cells, vary in ploidy, i.e. they contain varying amounts of DNA.

Large amounts of DNA in a cell or high levels of ploidy, respectively, are often associated with whole genome doubling (WGD)[14]. Previous research has shown how WGD can fuel CIN through abnormal mitosis[15–18]. Cells that have undergone WGD can be genomically unstable and tend to accumulate CNAs more quickly, partly because they appear to be able to better cope with the negative effects of deleterious mutations and ongoing CIN.[14,19–22]. WGD is a common event across cancers[1,23,24] and is associated with poor prognosis[25]. At the same time, fundamental questions are still unanswered, for example mechanisms of WGD, and the interplay between WGD and CIN in the early stages of tumourigenesis[16]. Because of their importance, WGD and CIN are central topics in cancer genomics, but further progress is held back by the lack of accurate methods to identify the ploidy of single cells, which is a crucial prerequisite for many downstream applications, such as quantifying intratumour heterogeneity and phylogenetic reconstruction of tumour evolution, and also heavily impacts single-nucleotide variation (SNV) detection[26–28].

Here, we introduce *scAbsolute* - a computational method specifically targeted at inferring individual cells' ploidy and replication status based on shallow single-cell DNA sequencing data alone. We demonstrate the feasibility of distinguishing cells in different, previously unidentifiable ploidy states, including cells directly after undergoing WGD. Our research improves on existing models for ploidy estimation across different data types.

Ploidy estimation in bulk data The problem of estimating tumour purity and ploidy, as a precursor to further CNA analysis, has been extensively explored in the bulk sequencing setting. The challenge in this setting is estimating the mixing ratio of normal and tumour cells, and the heterogeneity of the tumour cell populations and subclonal populations[29]. Many tools aim to estimate tumour purity and ploidy, and identify subclonal copy-number status[23,27,30–36], but the problem is challenging, because it is underdetermined with multiple mathematically equivalent solutions existing[23]. Some partial improvement can be achieved by using multi-sample bulk sequencing[29]. Importantly, mistakes at the level of ploidy and purity estimation have considerably negative effects on downstream analysis, such as subclonal reconstruction and SNV detection[26].

Single cell technologies Recent advances in single-cell sequencing technologies[37–43] make it possible to measure CNAs in individual cancer cells. An early platform, called Chromium Single Cell CNV[44] and developed by 10X Genomics, relied on whole-genome amplification and a commercial microdroplet platform. Several publicly available data sets were produced by this technology which is no longer commercially available. It has been replaced by protocols developed for shallow WGS single-cell DNA sequencing, including Direct Library Preparation[39,40] (DLP) and Acoustic Cell Tagmentation[43] (ACT). These two technologies rely on amplification-free, single-molecule indexing via direct tagmentation. By using a direct tagmentation step to incorporate index barcodes and sequencing adaptors before subsequent PCR cycles, it is possible to link all original reads to the original single-cell molecules and computationally filter PCR duplicates[39].

Ploidy estimation for single cell data While single cell technologies directly address the purity issue, it is still necessary to estimate an individual cell's ploidy. Existing approaches for single-cell sequencing data either rely on computational steps originally developed for the bulk sequencing setting, or additional experimental information. Most approaches are unable to distinguish between different ploidy solutions in the absence of odd or intermediate copy number states in the data[45]. It is easiest to demonstrate this challenge with a completely normal cell without any CNAs. In the absence of any additional knowledge, there is no possibility to distinguish a normal cell that has just undergone a WGD, or is in G2 phase with a tetraploid genome, from a cell in G1 phase with a diploid genome.This is equally the case in a tumour cell with a more complex genome (Figs. 1 and S1).

Existing computational methods are unable to distinguish these cases. In practice, tools such as HMMCopy[46] are commonly used[39,41,47,48] and serve as the basis for some novel copy number callers[49,50]. Gingko[51] shows improved performance for calling of accurate ploidy on a single cell level[52]. In a further advance, CHISEL[53] estimates haplotype-specific copy numbers based on B-allele frequency (BAF) estimated across 100s and 1000s of cells. Limitations of CHISEL are the requirement to pool a large number of reads together, either by increased sequencing depth or by number of cells, and the need for a matched normal sample or a sufficient number of normal single cells sequenced. Finally, as a BAF-based approach, CHISEL cannot detect recent WGD.

Alternatively, ploidy information can be inferred from experimental information, such as DAPI fluorescence staining and subsequent Fluorescence-activated Cell Sorting (FACS)[43]. However, this approach requires a ploidy control and a sufficient number of cells as input material, and might introduce a bias in the a-priori selection of cells based on their ploidy profile. Laks *et al.* [40] suggest that there is a relationship between ploidy and cell size, as observed via microscopy. However, it is unclear to what extent this can be reliably used to determine absolute cell ploidy.

The challenges of ploidy calling in single-cell data are further aggravated by the fact that different cell cycle phases lead to different overall DNA content and introduce spurious copy number changes. As a result, separating cells undergoing DNA replication in S phase from cells in G1/G2 phase is important to reliably measure copy number status across cells. While this is relatively easy to achieve in a homogeneous sample[40,54], we introduce an approach that generalises to novel cell populations without requiring new training data, and improves on existing experimental evidence based on FACS sorting of DAPI stained cell populations.

Results

The scAbsolute algorithm for calling absolute copy number

The basic idea of *scAbsolute* is to find a transform to convert the values from a scale of read counts per bin to a scale of absolute copy number.

For a cell with its genome split into M fixed-size genomic bins (by default 500 kilobases), we refer to the (unknown) copy number as c_j and the observed per-bin read count as x_j for each bin j. We aim to estimate a scaling factor ρ, so that we can directly measure the ploidy p of a cell:

$$p = \frac{1}{M}\sum_{j=1}^{M} c_j = \frac{1}{M}\sum_{j=1}^{M} \frac{x_j}{\rho} \qquad (1)$$

Equally, we can think of the ρ value as $\rho = \frac{1}{M}\sum_{j=1}^{M} x_j/c_j$. The factor ρ denotes the average reads per copy and per bin and is a measure of per-cell read coverage that we find very useful to compare different cells. The value of ρ is a direct measure of the difference in expected mean read counts between neighbouring copy number states. Note, that our definition of ploidy is directly proportional to the amount of DNA in a cell, and is not referring to the mode of the copy number state distribution in a cell.

scAbsolute directly works on aligned bam files and produces absolute copy number calls across genomic bins. Ploidy and cell-cycle inference takes about 3-10 min per cell (at 500 kilobases bin resolution), and can be naively parallelized across cells. *scAbsolute* proceeds in a series of steps summarised here (Fig. 1); further details can the found in Methods.

Step 1: We use a dynamic programming approach based on the PELT algorithm for change point detection[55] with a negative binomial likelihood to find an initial segmentation of our read counts. It is possible to use other segmentation algorithms at this stage, as long as the quality of the segmentation is reasonably good. Note that our focus here is not on obtaining a highly accurate segmentation, but rather in identifying an accurate ploidy estimate.

Step 2: We then consider the marginal distribution of the segmented read counts. We use a constrained Dirichlet Process Gaussian Mixture Model to fit a mixture of Gaussians to the marginal distributions. The constraint is in forcing the distance between the means of the Gaussians to be constant, analogous to a 1-D grid. The width of the grid then corresponds to a multiple of the scaling factor ρ, since we identify the peaks with the discrete copy number states that are scaled by the per-cell read depth. Because we are working with a single cell, we know that all copy number states occur at integer levels, only.

Step 3: We further constrain the per-cell ploidy solution, i.e. make the solution identifiable, by choosing a multiple of the scaling factor as the correct solution, based on the resulting cell ploidy. Here, we use the solution with minimum L2 error among all solutions within the given ploidy window (by default 1.1-8.0) to select one fit among the limited set of ploidies within the ploidy window.

Step 4: In order to correctly identify cells in which step 3 leads to an incorrect ploidy assignment - such as G2 or WGD cells, or other outliers - we use a reference set of cells for which we compute the genomic read density given an inferred ploidy in a post-processing step. Deviations from the expected read density at the estimated ploidy indicate an incorrect ploidy solution. If any outliers have been detected in this step, they can be refitted by re-applying steps 1-3, with an updated ploidy window.

In the following we present the *scAbsolute* method with three applications: determining per-cell ploidy, identifying previously unidentifiable ploidy cases, and identifying cells undergoing DNA replication.

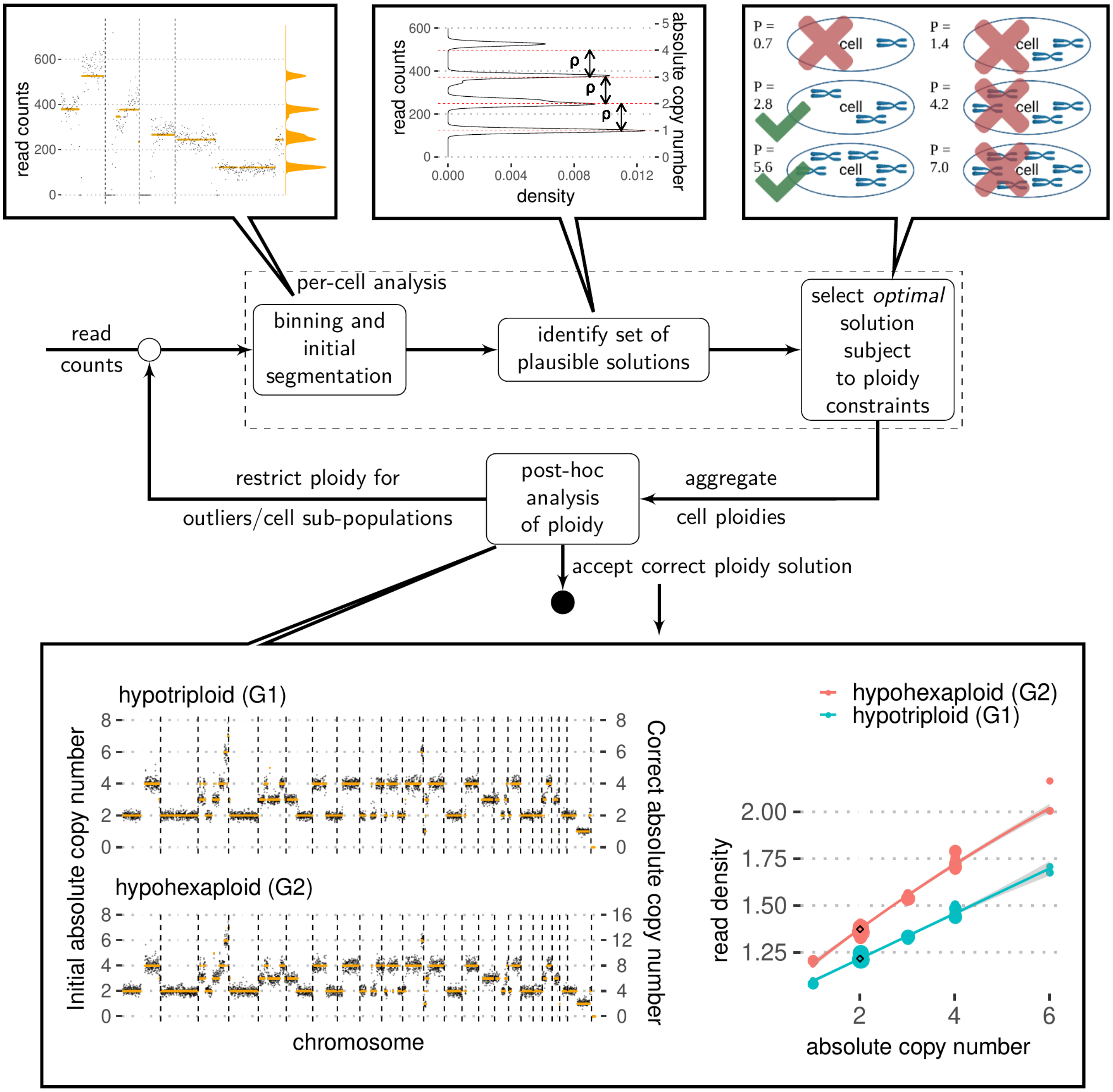


Figure 1: Schematic overview of *scAbsolute* approach. Initially, raw read counts are binned and segmented. The marginal distribution of the segmented read counts is fitted with a constrained Dirichlet Process Gaussian Mixture Model (DPGMM) using stochastic variational inference in order to identify a limited set of plausible solutions. We do this by identifying a constant ρ, that converts the scale from a read count to an absolute copy number scale. There exist a set of values of ρ that lead to equally possible ploidy solutions (in this case triploid or hexaploid solutions). We select the solution with minimum ploidy value subject to per-cell ploidy constraints. Post-hoc analysis of ploidy allows to distinguish previously indistinguishable cell populations based on read density, overcoming the limitation of the above approach. The example shows two (exemplary) copy number profiles for cells in different ploidy states, that have previously been indistinguishable based on copy number profiles alone. We demonstrate that cell ploidy can be determined in the absence of differentiating copy number states and other experimental information, based on per-cell read density alone.

Determining per-cell ploidy

We assessed the performance of *scAbsolute* and three competing computational approaches (HMMCopy, Ginkgo, CHISEL) by comparing computationally predicted ploidy with ploidy estimates derived from experimental sample annotations (Fig. 2). In the case of the 10X cell lines, we have karyotype estimates available, and the mean ploidy estimate based on karyotype analysis provides a good fit with the computational estimate. In the case of the DLP cell lines, we do not have direct evidence (except for the T-47D sample, which we will discuss below), but we find that the estimates are reasonably consistent and in line with the known ploidy of the cell lines. The best experimental evidence is available for the ACT data sets. Here, we have DAPI staining based FACS and selection, and it is reasonable to assume that cells falling outside of the experimentally measured ploidy window are largely ascribed false ploidy. This is not necessarily true for 10X and DLP data, where we only have estimates on the mean of the population, but no information about the ploidy spread of the population and individual cells' ploidies, and so it is difficult to compare methods based on these data.

Consequently, we use the ACT data to compare performance of different computational tools, assuming that ploidy outliers are indeed due to erroneous ploidy prediction. We compare *scAbsolute* to Ginkgo[51], HMMCopy[46] and CHISEL[53] (Fig. 2). Table 1 gives a detailed overview over the prediction results. We compare performance in two ways: First, as the percentage of cells outside an experimental ploidy window of ±0.5 around the peak of the DAPI distribution, which includes uncertainty from segmentation and FACS sorting, but excludes true ploidy changes. Second, as the mean absolute distance across all cells in a sample from the experimental ploidy estimate. For both metrics, we find that *scAbsolute* consistently predicts the correct ploidy solution for the large majority of cells. In only one case (TN5), it performs considerably worse than Ginkgo with an error rate of 17.8% compared to Ginkgo's 10%. This case, and a similar scenario in the TN4 sample, are due to the unidentifiability of the problem, and are discussed in detail below. The three samples with the highest percentage of wrongly assigned ploidies are 40% and 34%, and 31% in the case of HMMCopy (TN5, MB-231, and MB-157), 69%, 66%, and 59% for Ginkgo (TN8, MB-453, and TN6), and 66%, 61% and 44% (TN4, TN8, TN5) for CHISEL, compared with 18%, (TN5) 17% (TN4), and 4% (TN3) for *scAbsolute*.

The use of CHISEL is limited by the requirement to provide phased germline SNPs. Consequently, we can only run the CHISEL analysis on the ACT tumour samples, for which we have bulk exome sequencing of normal tissue as control, but not for the cell line data. CHISEL performs comparably to HMMCopy and Ginkgo, despite using additional information from SNPs. In addition, we run our algorithm on one of the samples originally published with CHISEL that includes raw sequencing data for about 10 000 cells. In general, we find good overlap between the two predictions, in the form of identical predictions (predictions on the diagonal, see Fig. S2). For this sample, we do not have experimental

Table 1: Comparison of ploidy prediction for ACT samples. Ploidy prediction based on *scAbsolute*, Ginkgo, HMMCopy, and CHISEL callers. Ploidy estimate is based on DAPI staining based FACS sorting. We consider an estimate to be an outlier if it falls outside of a ±0.5 window around the experimental ploidy point estimate. We also give the mean of the absolute distance in ploidy between the experimental point estimate and the computational estimate across all cells in parenthesis.

Sample	Ploidy	# of cells	HMMCopy	Ginkgo	CHISEL	*scAbsolute*
			% outlier (mean distance)			
BT-20	2.70	1228	13.4 (0.51)	1.1 (0.28)	-	0.5 (0.19)
MB-157	2.55	1210	31.2 (0.91)	0.5 (0.18)	-	0.9 (0.16)
MB-231	2.41	897	34.4 (0.92)	7.1 (0.36)	-	0.7 (0.03)
MB-453	4.17	1260	8.1 (0.26)	65.6 (1.06)	-	2.8 (0.13)
TN1	3.45	1100	1.7 (0.22)	1.8 (0.18)	27.3 (0.44)	0.5 (0.06)
TN2	3.03	1024	3.8 (0.24)	0.8 (0.15)	24.0 (0.38)	1.0 (0.09)
TN3	3.44	1101	27.2 (0.92)	2.4 (0.20)	23.3 (0.35)	4.1 (0.18)
TN4	3.76	1307	19.8 (0.59)	53.9 (1.18)	65.9 (1.14)	16.8 (0.46)
TN5	2.65	1238	40.1 (1.01)	10.9 (0.34)	43.6 (0.57)	17.8 (0.48)
TN6	3.17	1060	2.9 (0.33)	58.6 (0.84)	11.9 (0.19)	1.2 (0.14)
TN7	3.15	605	4.3 (0.30)	49.3 (1.30)	21.2 (0.32)	3.5 (0.17)
TN8	3.95	1224	2.4 (0.33)	69.4 (0.88)	60.7 (0.73)	2.9 (0.18)

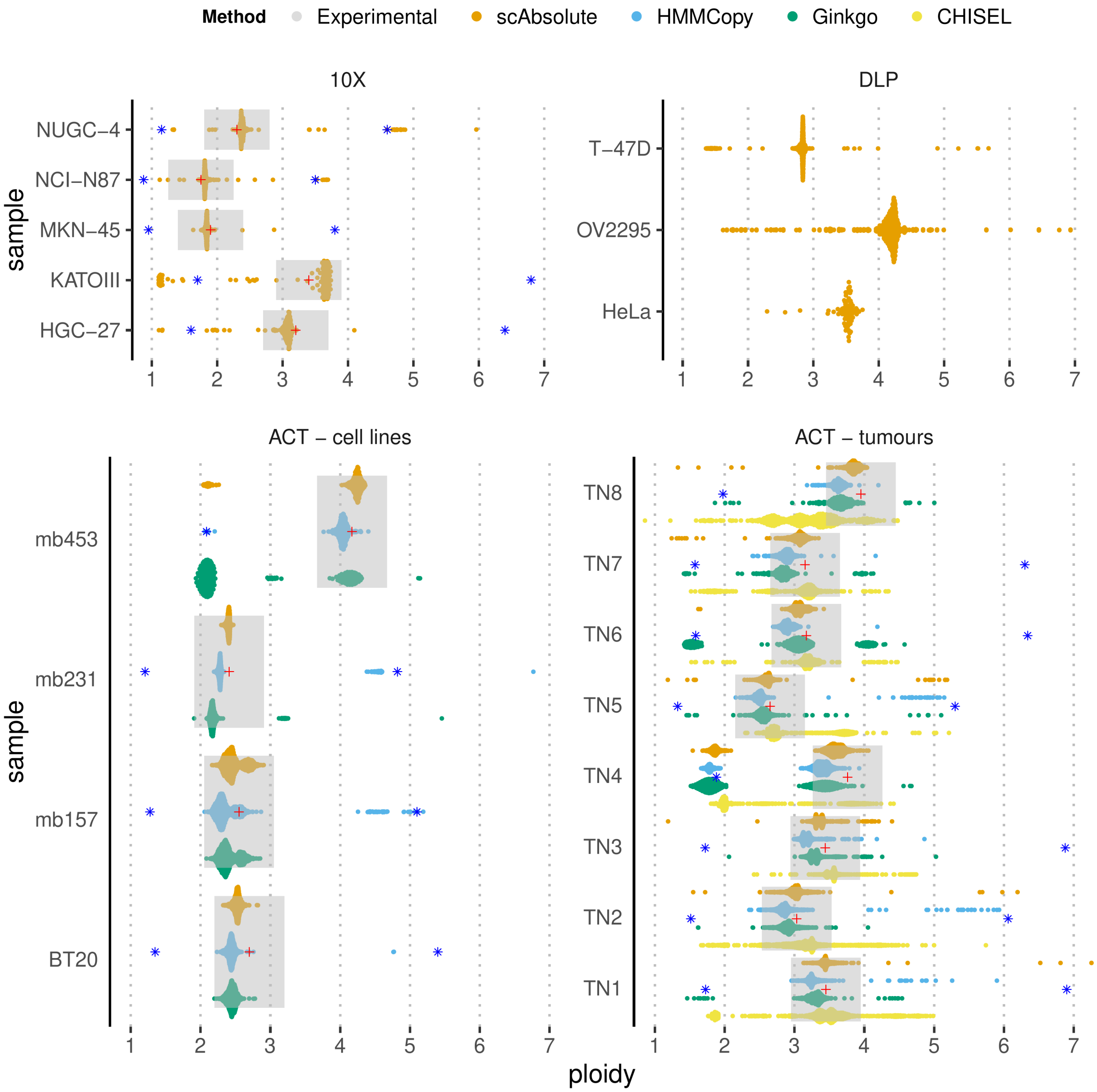


Figure 2: Computational ploidy prediction for cell lines and tumour samples with experimental ploidy annotation. Automatic, per-cell ploidy inference via *scAbsolute* identifies ploidy distribution in accordance with sample annotations across three different sequencing technologies (10X, DLP, ACT). Note that the data has not been quality controlled apart from removal of replicating cells. The annotation for the 10X data is based on karyotype information, the ploidy annotation for ACT data is based on DAPI staining based FACS sorting. We indicate in gray ranges of ±0.5 around the experimental point estimate of the sample ploidy (indicated by the red cross). Blue asterisks indicate ploidy levels of $1/2$ or 2 times the experimental ploidy estimate. In the case of DLP, no experimental annotation is available, but estimates are in accordance with ploidy estimates for the respective cell lines. In the case of T-47D, no method can initially distinguish between different ploidy subpopulations (cells in G1 and G2 phase, respectively), and all cells are matched to the same copy number state (corresponding to G1 phase, see Fig. 3). In the case of ACT, we can compare performance of *scAbsolute* (orange) with HMMCopy (blue) and Ginkgo (green) predictions of per-cell ploidy, and with CHISEL (yellow) in the case of tumour samples, only.

ploidy estimates available, however, we conduct a comparison of outliers to identify differences between the methods. First, we randomly select cells that are predicted to be nearly diploid by *scAbsolute*, and have a ploidy predicted to be greater than 2.5 by CHISEL. We observe CHISEL selects higher ploidy solutions, that are not necessarily supported by copy number levels in some of these instances (Fig. S3). Second, we compare cells that are predicted to be nearly diploid by CHISEL, and have a ploidy predicted to be greater than 2.5 by *scAbsolute* (see Fig. S4). Here, we observe a few cases of highly uneven cell coverage (possibly indicating failed sequencing runs). A second set of cells is classified differently based on additional information from the X chromosome, that is not included in the CHISEL predictions. We also compare the overall copy number profiles for section E of patient S0, as presented in Zaccaria & Raphael [53] and find very similar copy number predictions (Fig. S5).

An additional benefit of *scAbsolute* is its ability to measure ploidy even in some replicating cells. In the case of T-47D cells, about 64% of replicating cells are assigned the same ploidy as the G1 cells (see Fig. S7). This shows that the algorithm is to some extent robust to the noise observed in replicating cells, as can also be seen in example ploidy fits (Fig. S6).

Identifying previously unidentifiable ploidy cases

The main source of outliers, apart from cell quality issues, can be attributed to misclassification of ploidy by a factor of two (Fig. S1). This can be observed by small clusters of cells either at half or twice the ploidy of the cell observed, e.g. in case of the MB-453 cell line or the tumour samples TN4 and TN5 in Fig. 3. Equally, the algorithm (up to step 3) is not capable of distinguishing cells that have undergone WGD or are in G2 phase from their pre-WGD or G1 counterparts, as can be seen from the failure to detect the sizeable G2 subpopulation in the T-47D sample. The T-47D sample has been enriched for cells in different cell cycle states (G1, G2, S phase) using DAPI staining based FACS. Figs. 2 and S8 show consistent and identical, but ultimately wrong, ploidy predictions for the G2 cells in the T-47D sample based on steps 1-3 of *scAbsolute*.

Here, we introduce a novel approach to overcome this limitation, based on the density of reads along the genome (this approach is implemented as step 4 of the algorithm). The approach is applicable to all single-cell DNA sequencing technologies, that do not use pre-amplification and have a sufficient per cell read coverage of about $\rho = 75$ at 500 kilobases resolution with paired-end reads (corresponding to a coverage of about 0.01-0.02 for a normal, human cell). It might potentially be possible to extend this approach to single-end reads, however, we believe that this would require a substantial increase in read depth and there is currently no public data available to test this hypothesis. In particular, we could not include the ACT data discussed previously because it is either single-end read based or does not have sufficient read depth to apply step 4 of the algorithm. Without pre-amplification we can directly reference reads to physical copies of the genome, and this provides the necessary constraint to uniquely identify a ploidy solution. We use the start and end position of a given read, and measure how many other reads overlap with it. Using the fact, that in the absence of pre-amplification, we expect the number of overlapping reads to be limited by the number of physical copies of the molecules, we can use this to build a model of how many overlapping reads there are on average in any genomic bin. To make this approach sufficiently robust, we use a genome-wide measure of read density, i.e. the number of overlapping reads per region of the genome, to create a reference distribution of the expected mean number of overlapping reads. Importantly, this measure of read density depends on two variables, only. An individual cell's ploidy and the sequencing reads per cell as captured by ρ.

First, we obtain a ploidy-normalized per-cell value of read density, by regressing the observed read density across copy number states and chromosomes, and predicting an expected value for a copy number state of 2 (see Fig. 1, bottom panel). This allows us to make a prediction independently of varying copy number states in different cells and even in the absence of a copy number state of 2. Lastly, we need to account for varying per-cell read depths (as measured by the ρ value). We normalize the per-cell value, by fitting a simple quadratic function to the observed read densities (Fig. 3(a)). This makes it possible to directly compare a new cell to the predicted value of read density and observe any strong deviations. Note that the model is currently strongest in the range of $75 - 150$ ρ, as we do have substantially more data points in this range. We provide a model fitted to all publicly available DLP data, that can be freely used to determine if a cell subpopulation is deviating from the expected read density. Here, we make the assumption that the majority of cells are in G1 phase, and we correctly identify the individual per-cell ploidies for the majority of these cells, thus leading to a correct ploidy fit for the majority of cells in the reference set.

To show that this approach generalizes across data sets, we fit a model to a series of high-depth DLP data sets, holding out three single cell libraries: One library based on the T-47D sample with 194 cells in G1, and 151 cells in G2 phase respectively, and two libraries from SA928 (normal cell line) with 522 cells in G1, and 243 cells in G2 phase, respectively. We observe a clear difference in read densities between cells in G1 and G2 phase of the cell cycle, with S phase cells taking a somewhat intermediate position, with some cells more closely aligned to G2 cells possibly corresponding to late-replicating S phase cells, and others to early-replicating cells in S phase. Using the hold-out set,

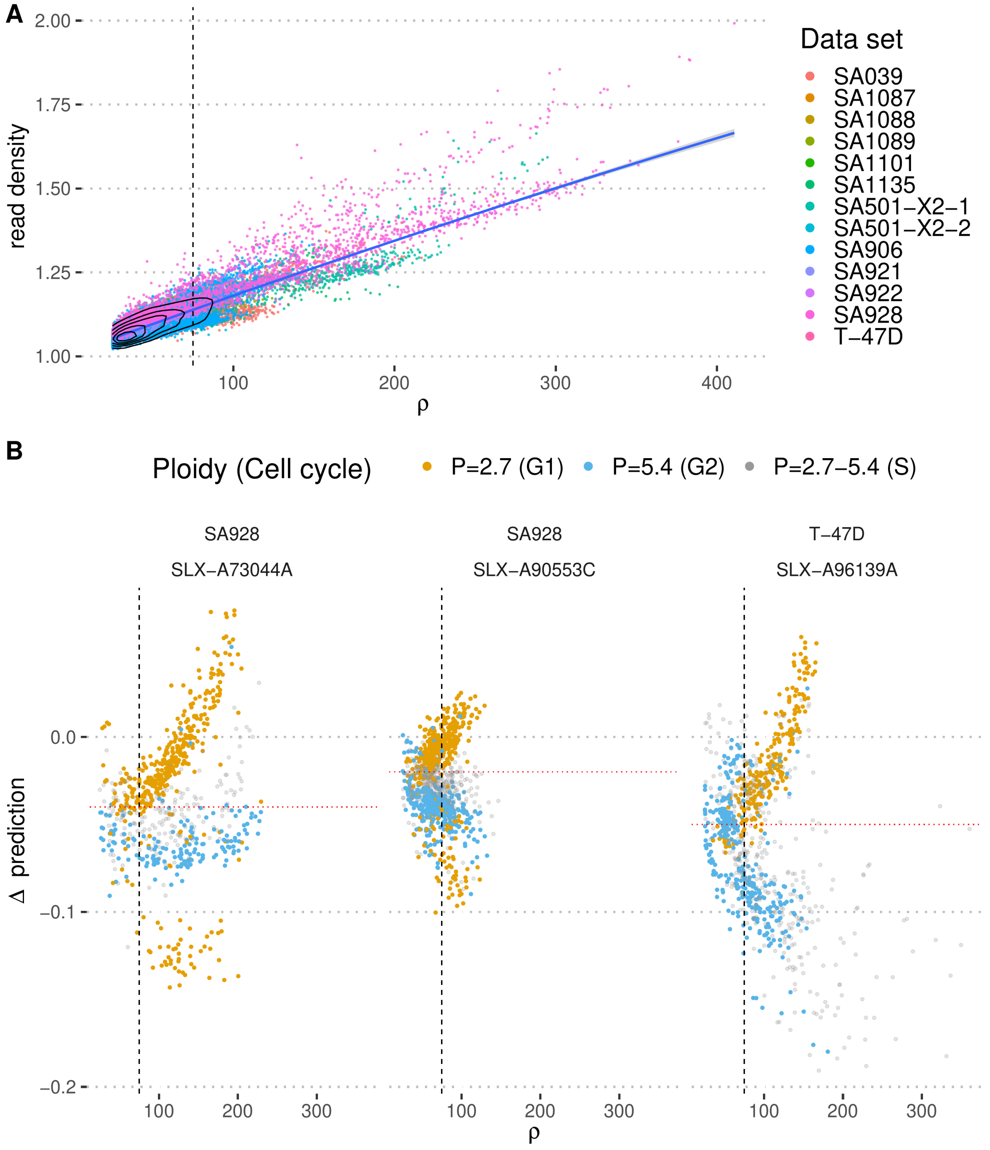


Figure 3: Ploidy estimation for previously unidentifiable ploidy states. A) Model of relationship between read density and read coverage per cell trained on DLP data. We observe a strong relationship that is robust across different sequencing libraries and copy number profiles. We provide a pre-trained model for the DLP technology, trained on a large cohort of cells with sufficient read depth to determine expected read density for a given ρ value. The underlying assumption is that cell fits based on *scAbsolute* are mostly accurate and the majority of cells is in G1-phase. **B)** We evaluate the model (built while holding out the test libraries) by predicting cells in G1 and G2 state of the cell cycle, corresponding to different ploidy states. Cell populations are clearly separated at sufficient read depth (at about 75ρ values). We observe S phase cells among both ploidy populations, possibly indicating early and late replicating cells.

we accurately classify 93% of cells in T-47D, and about 84% of cells in SA928 when looking at cells with a read depth of 75 ρ or higher. Note that in the later case, we can see evidence for potential outlier cells, indicating that this might be an underestimate of actual predictive performance. However, we also note that prediction in the case of normal, diploid cells is somewhat more difficult, as the per-cell regression in these cases is impacted more strongly by outliers.

Identifying cells undergoing DNA replication.

We introduce a basic measure of cycling activity based on fixed-width genomic bin counts and their respective replication time. We measure for each copy number segment in a cell whether we can observe a statistically significant trend in the observed counts as a function of replication time, correcting for the joint effect of GC content and mappability using the partial Mann-Kendall Trend Test. We aggregate over all segments by computing the median of the Mann-Kendall test statistics across all segments weighted by the respective segment length and use this single value as a per-cell measure of cycling activity. The advantages of this measure are: i) applicability across sequencing technologies ii) robustness across a range and mix of copy number profiles (as expected in a complex tumour sample) iii) unsupervised method (no training data required)

Laks *et al.* [40] describe a complex cell-feature based classifier that is able to detect S phase cells with about 90% accuracy. Using only four features (per-cell read depth, measure of overdispersion, cycling activity, and the third quantile of the cycling activity distribution), we achieve a very similar performance (91% accuracy) on the same data using a random forest classifier. Using only the cycling activity measure, in an unsupervised manner, we achieve 88% accuracy. Performance using only this single feature classifier is very different between normal cells (DLP-SA928)

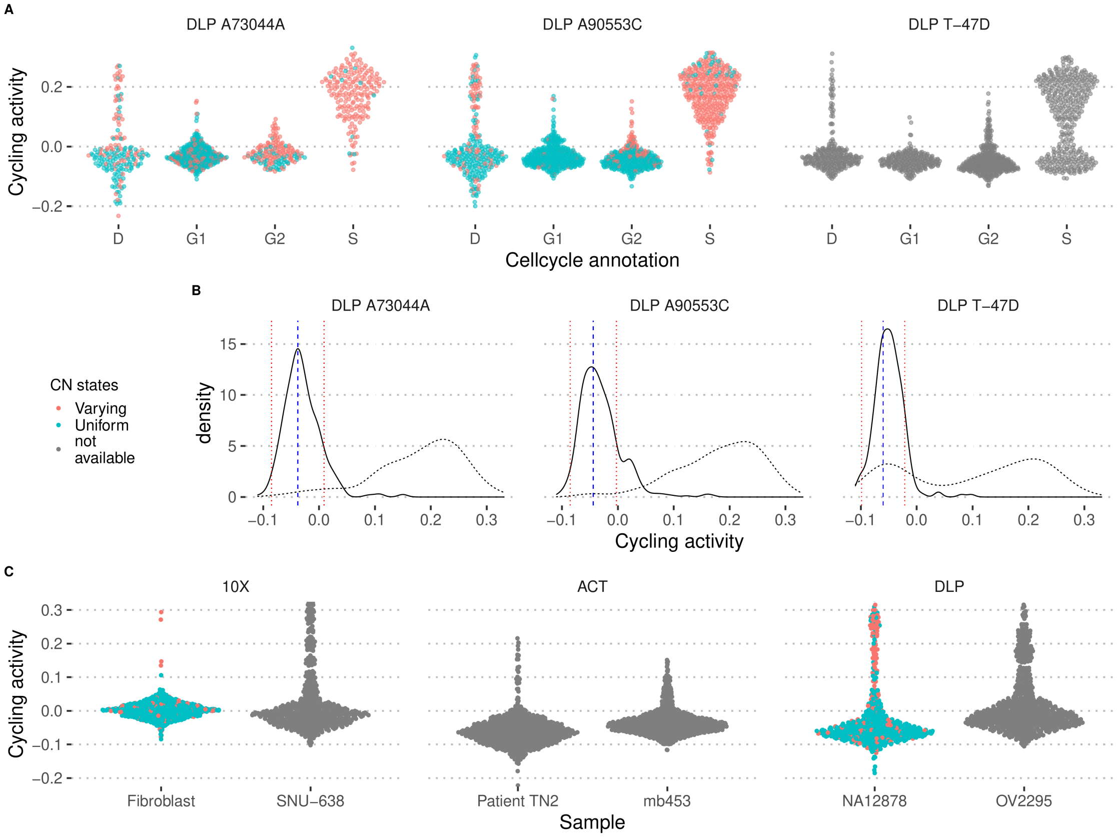


Figure 4: Detection of replicating cells across different sequencing technologies and cell lines. A) Cycling activity for normal, diploid cells (DLP A73044A and DLP A90553C), and a cancer cell line (DLP T-T47D) sequenced with DLP technology. The cellcycle annotation is based on DAPI staining and subsequent FACS sorting of cells. Uniform and varying CN state indicates whether the copy number calls indicate a mostly diploid genome (as in G1/G2), or cycling cells (non-uniform CN state), respectively. This measure can only be reliable estimated for normal, diploid cells, and is not applicable in other samples. **B)** Distribution of cycling activity for G1 (solid line) and S phase cells (dashed line). The mode of the distribution (blue line) is estimated and the left side of the distribution is used to determine a standard deviation that covers the majority of cells in G1 phase (red lines) and to determine an appropriate cutoff value to identify cells in S phase. **C)** Distribution of cycling activity samples covering three different sequencing technologies: 10X, DLP, ACT. While the mode of the distribution differs between samples and sequencing technologies, we observe the same characteristic asymmetric distribution in all cases, except for the 10X Fibroblast sample. In this case, the cells have been cell cycle arrested, and so we do not expect to see any cells in S phase.

at 91% accuracy and T-47D human breast cancer cell line (T-47D) at 82% accuracy. However, we believe this is an underestimate of real performance, due to leakage between cell cycle stages in the FACS step. A closer look at the underlying classification of S phase cells, indicates that there are two distinct populations classified as S phase based on FACS (Fig. S9) Looking at the copy number profiles, we see that the cells classified as S phase according to FACS, but G1 phase according to our classifier appear to be closer to the G1 phase cells based on the UMAP representation of the copy number profiles. Similarly, looking at the raw copy number profiles, there appear to be two different subpopulations among the cells lumped together as S phase cells using FACS. We can replicate the leaking of cells in different phases of the cell cycle when using DAPI staining based FACS. We use Geminin staining to control for cell cycle (Geminin is not expressed in G1 phase, but expressed from the transition from G1 to S phase on), and demonstrate that DAPI staining based selection leads to the inclusion of a relatively large proportion of cells in S phase (see Fig. S10), if the gating is not conducted very carefully.

In order to further validate the approach, we examine two other examples. First, we consider normal diploid cells. In this case, we can assume that the vast majority of CNAs are due to cells undergoing DNA replication. We can observe that the majority of normal cells with CNAs score high on the cycling activity measure (Fig. 4). Similarly, across cell lines and libraries, we observe the typical asymmetric distribution of cycling activity. The only exception (10X Fibroblast cells) have been cell cycle arrested, and so we don't expect to observe any cycling cells in this set.

Secondly, looking at a set of 10 cell lines sequenced with the 10X technology[54], we identify a robust relationship between estimates of number of cycling cells based on scRNAseq and scDNAseq (r = 0.76, p = 0.048) of the same cell lines after a similar number of passages. Similarly, we observe a robust negative relationship between doubling time of a cell line, and our estimate of proportion of cycling cells (r = -0.81, p = 0.008, see Fig. S11).

Discussion

The increasing availability of low-coverage whole-genome sequencing of thousands of individual cells offers an opportunity to study tumour heterogeneity and evolution at an unprecedented resolution. With increasing size of single-cell DNA sequencing data sets and new high-throughput droplet-based protocols under development, we expect scalability to become increasingly relevant. *scAbsolute* is by-design a scalable tool to investigate cell ploidy and replication status at the single-cell level. Single-cell DNA sequencing technologies are still being improved upon, and there exist a number of different sequencing technologies. We demonstrate the general applicability of *scAbsolute* across three recently released protocols, and show an advantage of pre-amplification free approaches in detecting WGD events.

To our knowledge, *scAbsolute* is the first computational approach to overcome the unidentifiability problem associated with WGD events in copy number calling of single cells. This appears to be particularly relevant for the study of early tumourigenesis, by making it possible to study WGD and other ploidy changes in small lesions with very limited numbers of cells. We hope that this will help to further elucidate the role of WGD in cancer, and its contribution to CIN.

The approach is fundamentally different from other approaches originally developed in bulk sequencing settings and recently extended to the single-cell domain[53] that use B-allele frequency (BAF) in order to help identify ploidy solutions. However, by using cell specific haplotype counts, and the high quality total copy number predictions provided by *scAbsolute*, it is possible to estimate allele and haplotype specific copy number states, as recently demonstrated[48,53]. In particular, it is straightforward to estimate the allele specific copy numbers using the BAF as estimator, once the total copy number is known. One limitation of the computational approach is the inability to distinguish cases of cells in G2 phase of the cellcycle from WGD cells. However, it is possible to address this either via recourse to estimates of the relative number of cells in G2 state, or via integration with experimental evidence.

The identification of per-cell ploidy and with it per-cell read depth lies at the basis of many downstream applications, such as copy-number segmentation, estimation of allele- and haplotype-specific copy number states, and the building of tumour phylogenies based on copy-number profiles. By solving the ploidy problem, further progress in developing more accurate and reliable CNAs calling methods seems feasible.

By offering an alternative to the use of FACS with DAPI staining to identify cell populations at different ploidy levels, we open the way for a more unbiased investigation of intratumour heterogeneity. In particular, the choice of cutoffs for FACS might lead to a bias in selecting more homogeneous cell populations and lead to an underestimation of the true level of heterogeneity in tumours. FACS might still be useful in order to reduce the amount of non-informative normal cells sequenced, thus reducing overall costs, and as an experimental validation, however.

Identifying cycling cells is a crucial issue in single cell based approaches. It is necessary in order to get a correct understanding of what constitutes actual tumour heterogeneity at the copy number level, and what are just ephemeral CNAs as a consequence of replication activity. Identifying cycling cells is also an opportunity as a crude measure of clonal fitness and for the identification of cell populations that are highly proliferative. Here, we present a simple, but robust measure of cycling activity and we demonstrate its applicability across a range of cell lines and sequencing technologies. Importantly, we demonstrate some limitations of using DAPI staining based measures of cell cycle status and how this can bias the training of cell cycle classifiers.

Methods

Bin-level quality control Prior to any analysis, we create a set of high quality genomic regions extending existing bin-annotations and specifically targeted at single cell DNA-sequencing (scDNA-seq) data. We use these novel bin annotations across all further experiments. The primary purpose of this step is to only include genomic bins for which the assumption of a linear relationship between copy number status and observed read counts holds.

The set is created by combining data from three different sequencing technologies (10X, DLP, and JBL - an in-house protocol), comprising more than 7000 diploid cells (We do not have access to diploid cells sequenced with the ACT protocol). In order to create a single, unified set of bin annotations, we conduct below analysis separately for the three sequencing technologies, and exclude all bins that fail the quality criterion in any single sequencing technology (see Fig. S12).

To determine bin quality, we only look at diploid cells, so we can disregard the issue of copy number status and segmentation. The set contains no cells in S phase or in G2/M phase (based on FACS sorting, and cell cycle arrest in the case of the 10X cells). In order to have sufficient reads per bin to reliably detect any deviations, and at the same time to have as high as possible a genomic resolution, we conduct this analysis on reads binned at $100kB$ resolution.

First, we normalise the per-cell reads by dividing the number of reads in each bin by the expected number of reads per bin across a cell, creating a normalised read per bin value. Subsequently, we use the median of the normalised reads per bin across all cells sequenced with the same sequencing technology as a per-bin quality metric. Initially, we remove all bins that have a median value of more than 4 or less than 0.10 per bin, and a mappability value smaller than 70 (11% of bins). Note, that the expected value would be 1, so a value of 4 corresponds to a median number of reads falling into a bin four times higher than would be expected on average.

In a second stage, we look at the relationship between GC content and median normalised read counts for each of the sequencing technologies. Separately for each technology, we fit a Generalised Additive Model to characterise the relationship between GC content and normalised read counts, and remove all data points that deviate more than 2 standard deviations (3 standard deviations for the much smaller JBL data set, see Fig. S12). In addition, we use a kernel density estimate to select regions of high density, and remove all cells outside the high-density regions. By using these two criteria, we aim to select only high quality bins that have minimal read count deviation that is not explained by GC content and mappability. Lastly, we use maps of centromere and telomere regions to specifically filter parts of these regions that have not been filtered in the previous steps.

Overall, we remove about 16% of the bins, containing about 9% percent of total reads on the autosomes. For the X and Y chromosomes, we use the existing QDNAseq annotations and run a simplified version of the above pipeline to remove outliers on the X chromosome (only based on density estimates). In the case of the Y chromosome, we remove bins solely based on deviations in the total number of reads observed (Fig. S13).

Initial Segmentation with unknown ploidy We use a dynamic programming approach based on the PELT algorithm[55] to find an initial segmentation of our read counts. We model the read counts with a Negative Binomial distribution[56].

$$Pr(Y = y|m, \alpha) = \frac{\Gamma(y + \alpha^{-1})}{y!\Gamma(\alpha^{-1})} \left(\frac{\alpha m}{1 + \alpha m}\right)^{y} \left(\frac{1}{1 + \alpha m}\right)^{\alpha^{-1}} \tag{2}$$

Here, α denotes a measure of increasing overdispersion; for values of $\alpha \rightarrow 0$, the distribution converges to that for the Poisson[57].

There exists no analytical solution to the Negative Binomial Maximum Likelihood estimate of the overdispersion parameter. We therefore choose a Methods-of-Moments estimator for the parameter α in the cost function[58]. We use this initial segmentation as a stepping-stone in identifying a cell's correct ploidy, and are not interested in an optimal segmentation at this early stage, but rather a robust and reasonably fast approach that achieves reasonable accuracy. We note that optimal segmentation is not the focus of this manuscript, and it is possible to replace the segmentation algorithm with one of the user's choice.

The cost function for a segment $y_j, ..., y_{j+k}$ of length k is defined as

$$C(y_{j,j+k}) = -\log \mathcal{L}(\hat{\mu}, \hat{\alpha}) \tag{3}$$

where $\hat{\mu}$ and $\hat{\alpha}$ are the parameters of the Negative binomial likelihood estimated from the data and $\mathcal{L}$ denotes the negative binomial likelihood function.

The mean is estimated via $\hat{m} = \frac{1}{n}\sum_{i=1}^{n} y_i$, and the overdispersion is estimated as

$$\hat{\alpha}_{\text{MM}} = \frac{s^2 - \bar{m}}{\bar{m}^2} \tag{4}$$

where $\bar{m}$ and s^2 are defined as sample mean and variance, respectively [58].

Ploidy estimation Ploidy estimation describes the identification of copy number segments with underlying copy number states. This is equivalent to finding a scaling factor ρ, so that we can directly measure the ploidy p of a cell:

$$p = \frac{1}{M}\sum_{j=1}^{M} c_j = \frac{1}{M}\sum_{j=1}^{M} \frac{x_j}{\rho} \tag{5}$$

Note that we assume that observed reads scale linearly with copy number state, e.g. we expect to observe on average twice as many reads for copy number state 2 than for copy number state 1, and so on. In our approach, given an initial - not-necessarily exact - segmentation, we consider the marginal distribution of the segmented copy number states, i.e. the spatial correlation of neighbouring bins has been accounted for through the initial segmentation. Since we are dealing with individual cells, in theory all copy number states should appear at integer values (with the exception of bins that contain two or more different copy number states), with some states possibly not observed in a cell (see Fig. 1 for an example for the marginal distribution of segmented counts). In a normal diploid cell which has no CNAs, we would expect to observe a single value scaled by a scalar value $\frac{1}{\rho}$ denoting the read depth of the cell (assuming the cell has two X chromosomes). However, given the inherent measurement noise and imperfect segmentation, in fact one observes a distribution of values around the integer states.

Here, we assume the errors are normally distributed and approximate the marginal distribution of the segmented counts with a (constrained) Gaussian Mixture Model. The constraint on the Gaussians is based on the fact that instead of estimating K means $\mu_1, \ldots, \mu_K$, we restrict the location of the means to $\mu_1 = 1 \cdot \xi, \ldots, \mu_K = K \cdot \xi$. Consequently, we estimate a single parameter ξ and a set of K standard deviations σ_k for the K Gaussians in the model. We might not observe all possible states and we don't know how many states we will observe in any given cell. Consequently, we model the appearance of individual clusters with a Dirichlet Process. The variational distribution of the Dirichlet Process is truncated at T components [59]. In order to further speed up the computation, we use stochastic variational inference [60] and implement the algorithm in TensorFlow [61]. Overall, we estimate the posterior probability $p(\theta|X) = \prod_{k=1}^{K} \pi_k \, \mathcal{N}(\mu_k, \Lambda_k|X)$. Details and the mathematical derivation of the model updates can be found in the appendix.

A major challenge with estimating absolute copy number states is the unidentifiability of a unique solution. There exist many potential solutions for each copy number profile, as it is always possible to shift or scale the solution. For example, consider the case of a perfectly diploid and tetraploid genome. Both are biologically plausible, e.g. in case of a Fibroblast cell in G1/G2 phase, but indistinguishable without any additional mathematical constraints. From a mathematical point, even a biologically implausible triploid solution is equally possible. This problem is less relevant in cancers with a high number of CNAs, as one tends to observe many of the copy number states in a single cell, thus making it easier to identify the correct solution. This makes the problem considerably easier, however, it would still not be possible to detect evolutionary recent whole-genome duplication events or distinguish between cells in G1/G2 phase. As a consequence, the design of our model does not enforce any particular solution. Instead it returns one possible solution, with the constraint, that the means of the Gaussians, i.e. the observed copy number states, occur at a distance that is an integer-multiple of an arbitrary unit distance and the set of solutions lies within a biologically reasonable, user defined ploidy range.

In order to select a biological plausible solution ρ out of the discrete set of possible values $\hat{\xi}$ within the given ploidy range, by default, we select the solution that has a minimal least squared error. In practice, this tends to be the solution with the lowest ploidy within the given ploidy range. For a biologically plausible ploidy range, we chose a minimum ploidy of 1.1. The assumption here is that cells with more copy number loses are probably not viable. As an upper bound, we chose a ploidy of 8 by default. However, this can be flexibly adjusted given other sources of information.

Addressing the unidentifiability problem Up to this point, the approach cannot distinguish between cells in G1 and G2 phase of the cell cycle, or cells directly after WGD. Here, we demonstrate that we can in fact reliably differentiate between these cells, given we have a i) sufficiently high read coverage ii) we use a sequencing technology without pre-amplification step and iii) paired-end read sequencing. Read coverage depends both on the ploidy of the sample, and on the size of the genomic regions that is covered (and mappable to) sequencing reads. It might be possible to avoid the need for paired-end read sequencing, if the coverage is substantially higher. However, we do not have access to any data to test this scenario.

The basic idea behind the approach is to analyse the number of overlapping reads across the genome. For this purpose, we compute a measure of how densely reads are located on the genome, and how many reads physically overlap. Because the genome is not amplified, we then can assume that the number of overlapping, unique reads is directly proportional to and limited by the number of physical copies of the genome at the given location. The only other two parameters we expect to influence the observed read density are the cell ploidy and the per-cell read depth (as measured by the ρ value).

Initially, for each properly mapped read, we count the number of overlapping reads across the genome, using the start and end position of paired-end reads as genomic start and end position, respectively. The size of the regions used is then determined by the fragment size. We compute the average number of overlapping reads across genomic bins, referring to it as read density. In order to account for varying copy number status, to make the estimation process more robust and to enable the comparison across cells with different copy number profiles, we fit a robust linear regression model to approximate the relationship between copy number state and read density (Fig. 1). We also split the data by chromosome, weighting each data point by the number of the bins covered. In order to directly compare cells, we use the predicted value of read density for a copy number state of 2 as per-cell measure of read density. Note that we can even predict this value in the absence of any observed copy number state 2. In the case of normal, diploid cells, we resolve to using the median read density, as the linear regression would otherwise by underdetermined with only a single data point at a copy number state of 2.

In a second step, we use a reference data set of cells to create a model of expected read density for a given read depth. Here, we combine all cells sequenced with the DLP technology in order to create a reference model of expected read density for a given value of ρ. We can then compare the residual read density, measured as the deviation between observed and expected read density at a given value of ρ in order to distinguish cells that have been assigned a wrong ploidy (see Fig. 3).

Detection of replicating cells We devise a simple test statistic to examine if a cell is in the S phase of the cell cycle. The test is based on differences in replication timing for different parts of the genome. In order to quantify replication timing per genomic bin, we use an existing annotation from the Repli-chip from ENCODE/FSU project[62,63], and average the replication times determined in these experiments across multiple cell lines and across genomic regions contained in a bin. This allows us to obtain a single measure of replication time per genomic bin.

Considering the segmented copy number profile S_l, we perform a partial correlation trend test using the *Spearman* rank correlation statistic [64, p. 882]. The test is performed on the raw count data within a given segment, sorted by increasing replication time, while controlling for the per bin GC-content and mappability value. Subsequently, the median of the partial correlation test statistic τ across all segments weighted by segment length is computed. We refer to this single measure as cycling activity. In general, we observe positive values as indicative of cells in S phase, undergoing replication. The distribution is symmetric around its mode with an additional long tail, that we identify with the replicating cells (see Fig. 4).

It is known that GC-content and mappability lead to a bias in the read counts observed across different genomic locations [51, 65]. Here, we fit a Generalised Additive Model (GAM) to estimate and correct for the bias individually per cell. The advantage of this approach is that it gives us a direct estimate for the impact of the covariates on the mean of the read counts observed in each bin. We model the observed read counts x for every bin j as $x_j \sim \text{NegBin}(\mu_j, \alpha)$, with $log(\mu_j) = \beta_0 + \beta_1 \nu + s(gc_j, map_j)$, where ν denotes the segmentation value of a bin, i.e. the median value of read counts across multiple neighbouring bins and gc_j and map_j the GC content and mappability values in bin j, respectively. We model the impact of varying GC and mappability content jointly with thin plate regression splines. We obtain coefficients of the impact of GC content and mappability variation on the mean expression for every cell and every bin, in the form of the coefficients $s(gc_j, map_j)$ that directly relate to the mean-expression and we use these values as covariates in the trend test.

We make use of the characteristic shape of the distribution to classify a cell's replication status across different sequencing technologies and libraries. We chose the threshold dynamically by identifying the mode of the distribution and determining the standard deviation using only the left (negative) part of the cycling activity distribution. By default, we use a cutoff corresponding to two standard deviations from the mode of the distribution as threshold for classifying a cell as being in S phase (Fig. 4(b)). This makes the approach easily applicable to new single cell data without cell cycle annotation, without potentially having to adapt the cell cycle classifier to a new feature distribution.

Code and Data availability

We include three separate datasets, covering three different single-cell sequencing technologies: 10X data[54] and a normal diploid cell line published online as part of the 10X Single Cell DNA sequencing technology demonstration (https://www.10xgenomics.com/), DLP data described in[40] and ACT data[43]. We exclude samples for which the majority of data is below the 25 ρ threshold at 500kb resolution to restrict the influence of segmentation on the ploidy calling.

For HMMCopy, we ran version 0.8.15 of the single cell pipeline to determine ploidy directly from the aligned bam files with default parameters. We ran the latest version of the Ginkgo platform (https://github.com/robertaboukhalil/ginkgo, version:71da01d9b24b1fcd0deb299b416a0fde676b18f7). For CHISEL[53], we directly use the published results in the case of the 10X Breast tumour dataset. For the ACT tumour samples, we ran CHISEL (v1.1.3) in the *nonormal* mode, with germline SNPs imputed and phased using the Sanger Imputation Service based on the normal bulk exome samples available for the ACT tumour samples. We found imputation slightly improved ploidy prediction performance compared with phasing only. Phasing and imputation weere performed using EAGLE2[66] and PBWT[67], respectively.

The source code for *scAbsolute*, and scripts to reproduce all figures and analyses is available at https://github.com/markowetzlab/scAbsolute.git. *scAbsolute* uses the package environment and genome annotations provided by the QDNAseq package[65].

Acknowledgments

We thank Amy Cullen, Dr Anna Piskorz, Dr Thomas Bradley, and Prof James Brenton as well as all members of the Markowetz and Macintyre labs for helpful discussions. We would like to acknowledge the support of The University of Cambridge and Cancer Research UK (C14303/A17197 and A19274). MPS was supported by a Marie-Sklodowska-Curie PhD Fellowship (766030-CONTRA-H2020-MSCA-ITN-2017).

Author contributions

MPS: method development, data analysis, manuscript writing; GM: supervision, manuscript editing; FM: funding, project conception and oversight, manuscript editing.

Supplementary Materials

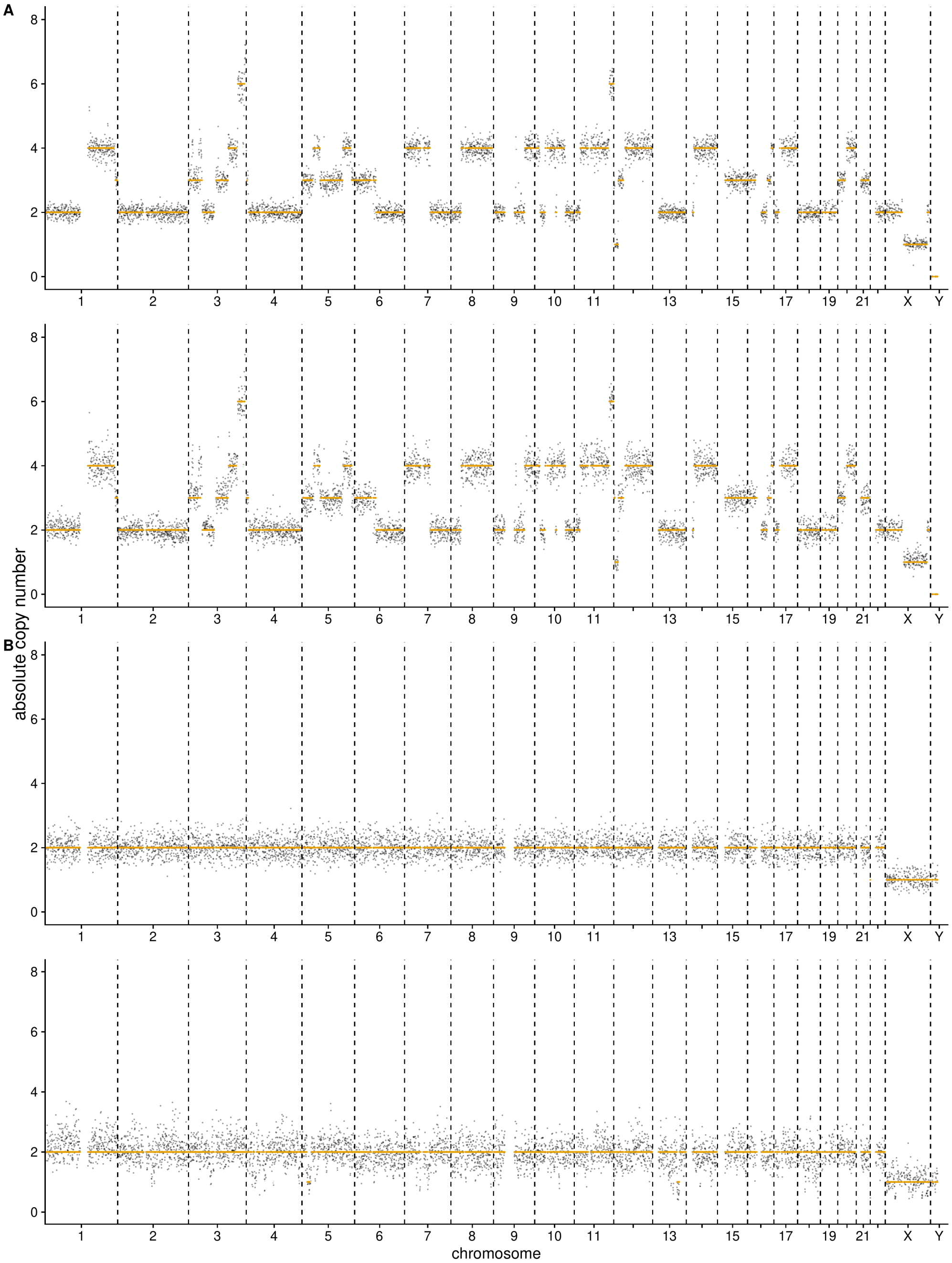


Figure S1: Example cells in G1 and G2 phase of cell cycle (A) Example tumour cell (T-47D) in G1 phase (top panel) and G2 phase (bottom panel) of cell cycle. **(B)** Example normal cell (SA928) in G1 phase (top panel) and G2 phase (bottom panel) of cell cycle. In all cases, cell cycle stage has been verified by DAPI staining based FACS and by subsequent computational analysis. Note, that in case of G2 phase, the initial ploidy solution is off by a multiplicative factor of 2.

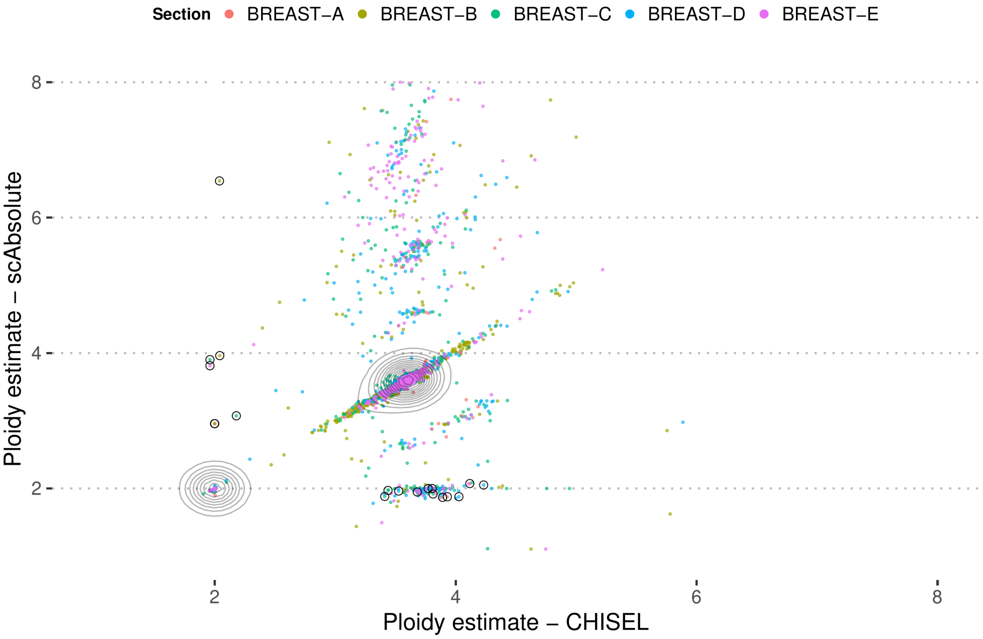


Figure S2: Ploidy predictions for *scAbsolute* and CHISEL on 10X Breast tumour sample. Example cells, shown in Figs. S3 and S4 are marked with black circles. Density areas are indicated on the diagonal, showing a relatively large overlap of predictions in this particular tumour sample.

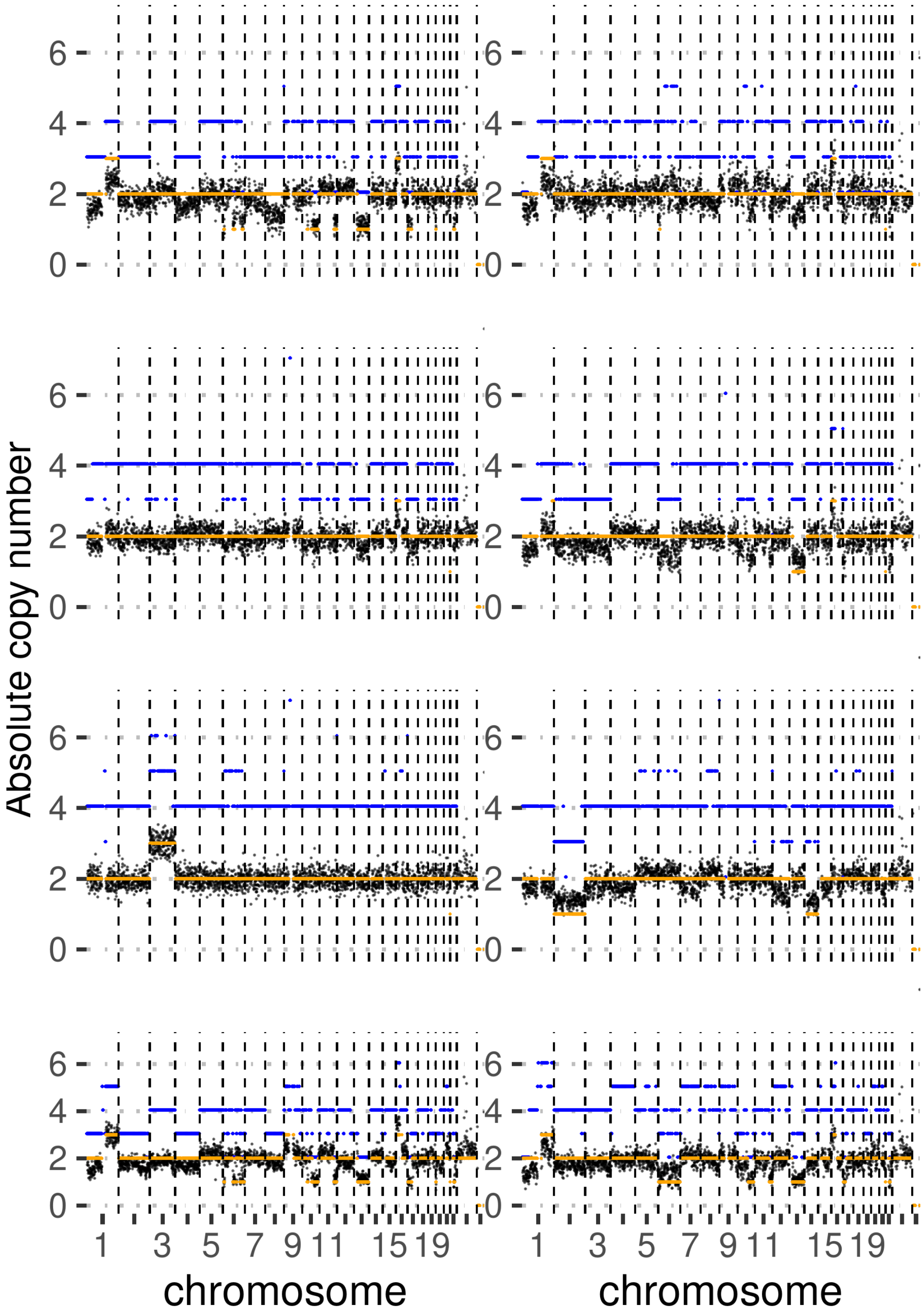


Figure S3: Randomly selected copy number profiles (based on *scAbsolute* segmentation) for diploid CHISEL ploidy predictions that contradict *scAbsolute* predictions. *scAbsolute* predictions are shown in orange, and CHISEL in blue.

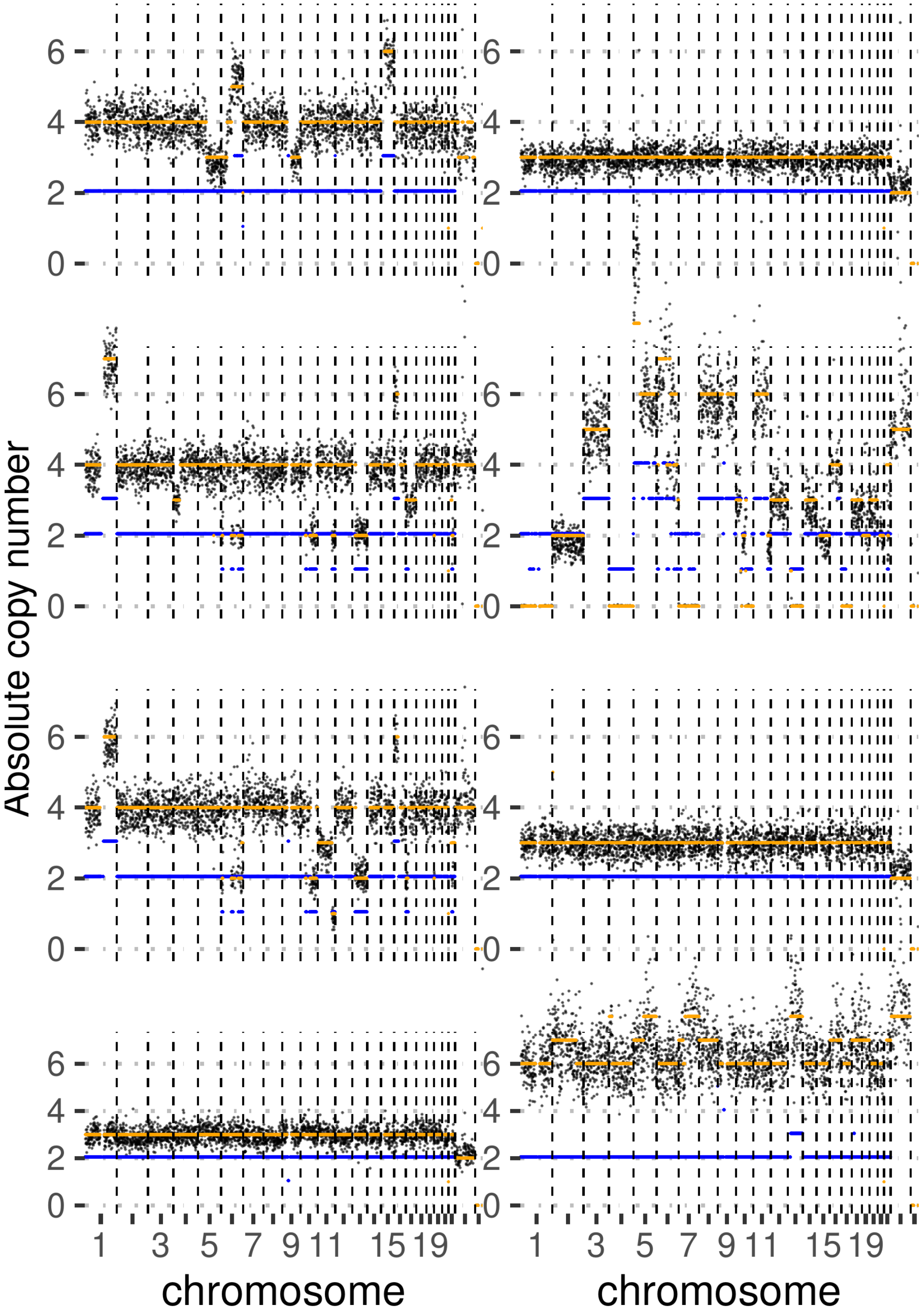


Figure S4: Randomly selected copy number profiles (based on *scAbsolute* segmentation) for diploid *scAbsolute* ploidy predictions that contradict CHISEL predictions. *scAbsolute* predictions are shown in orange, and CHISEL in blue.

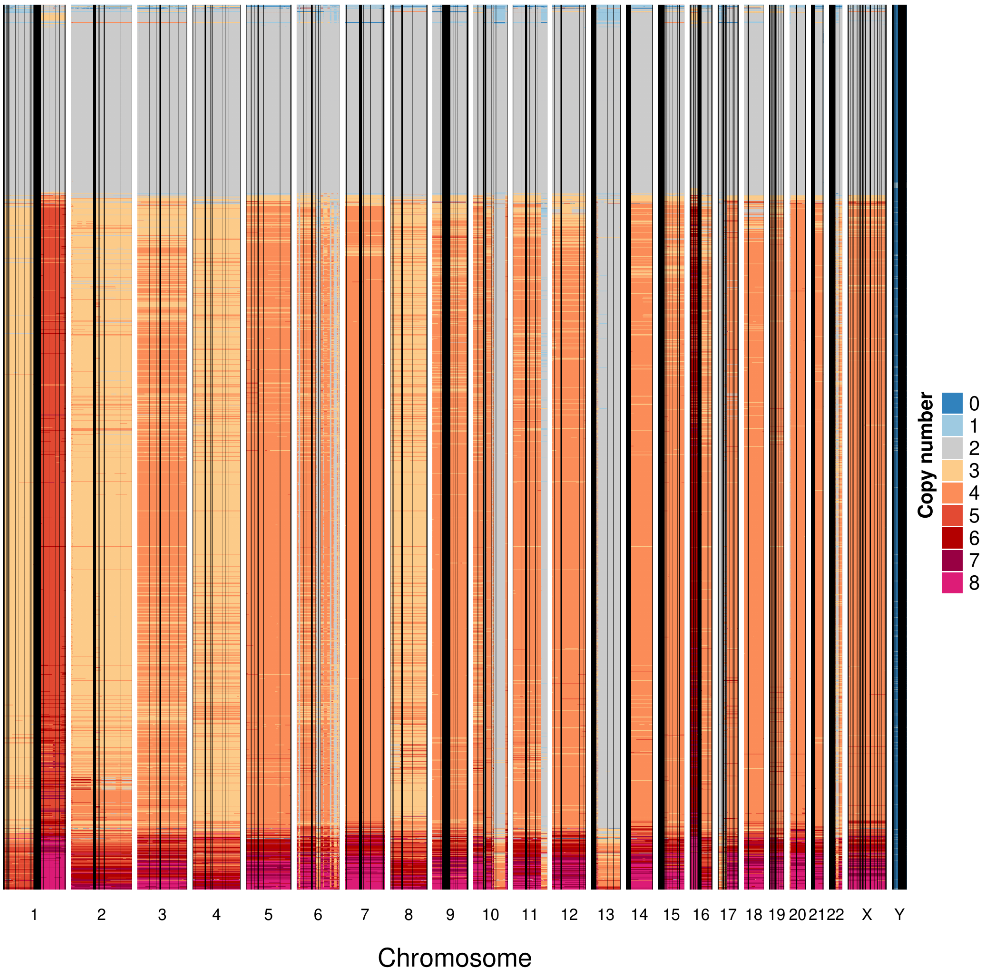


Figure S5: Copy number prediction for Patient S0 (section E) using *scAbsolute*. The prediction reflects the general copy number landscape for the sample as presented in Zaccaria & Raphael [53]. Note that the cells have not been quality controlled, and this explains the small number of ploidy outliers at high and low ploidies. Overall it appears relatively easy to detect higher ploidy states in this dataset, given the number of copy number segments at varying ploidy levels. This might be indicative of a relatively early WGD event leading to the observed copy number profiles.

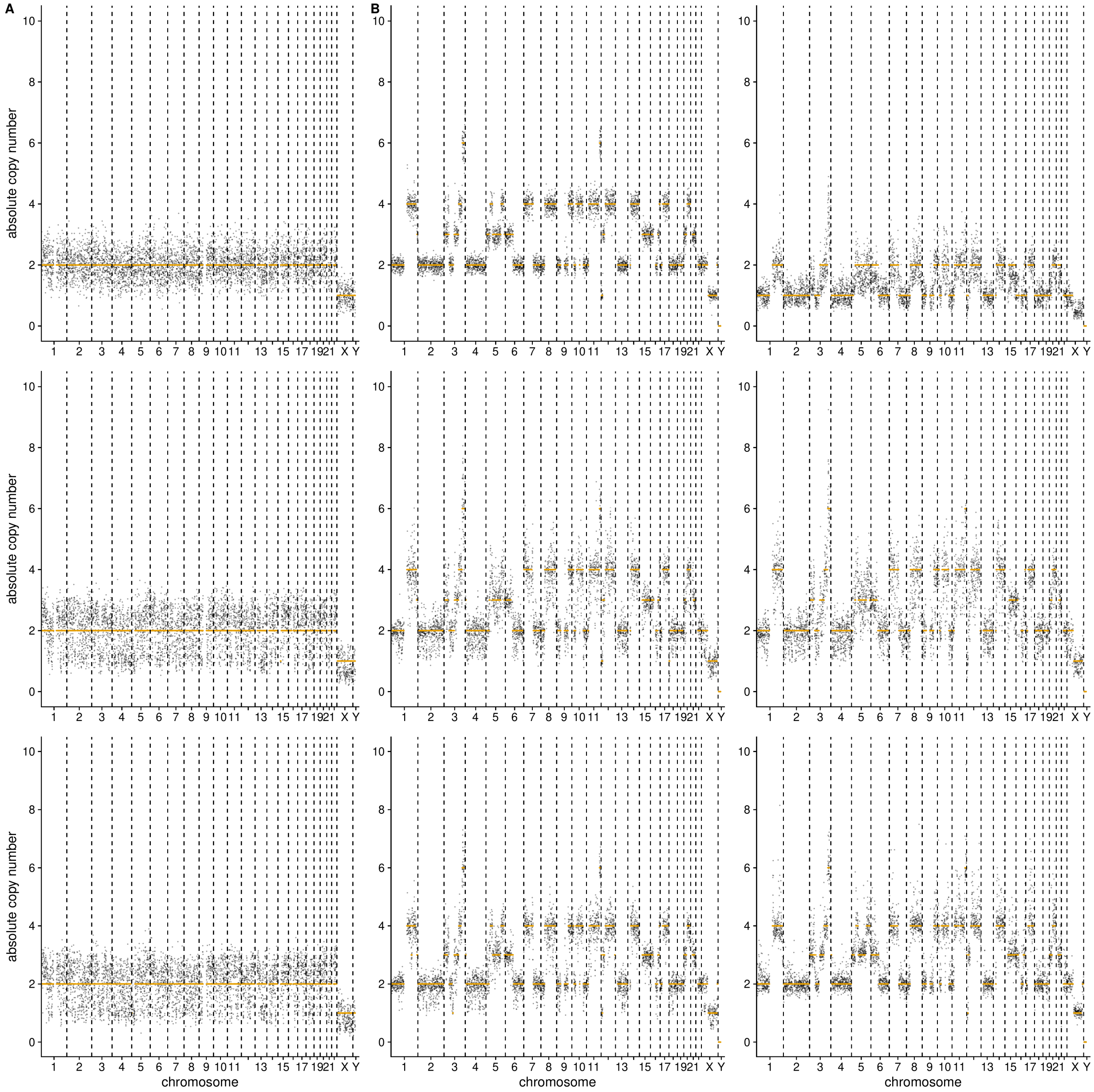


Figure S6: Examples for cells in S phase (A) Normal, diploid cell in G1 phase of cell cycle (top panel) and two examples of cycling cells. **(B)** G1 phase cell (top-left panel) and five examples of cycling cells at various levels of cycling activity measure. In all cases, cell cycle stage has been verified by DAPI staining based FACS and by subsequent computational analysis.

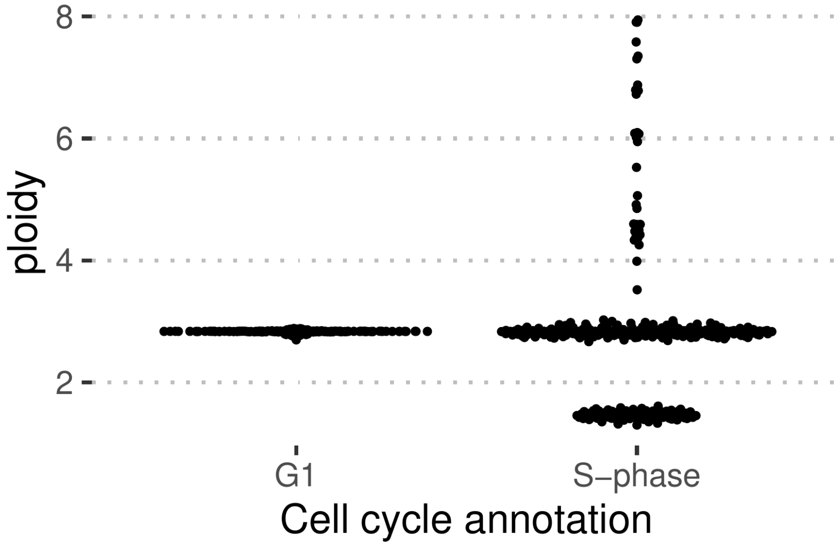


Figure S7: Ploidy prediction for G1 and S phase cells using *scAbsolute*. In this dataset of T-47D cells, *scAbsolute* recovers ploidy for about 63% of S phase annotated cells. Cell cycle was annotated using DAPI staining based FACS sorting.

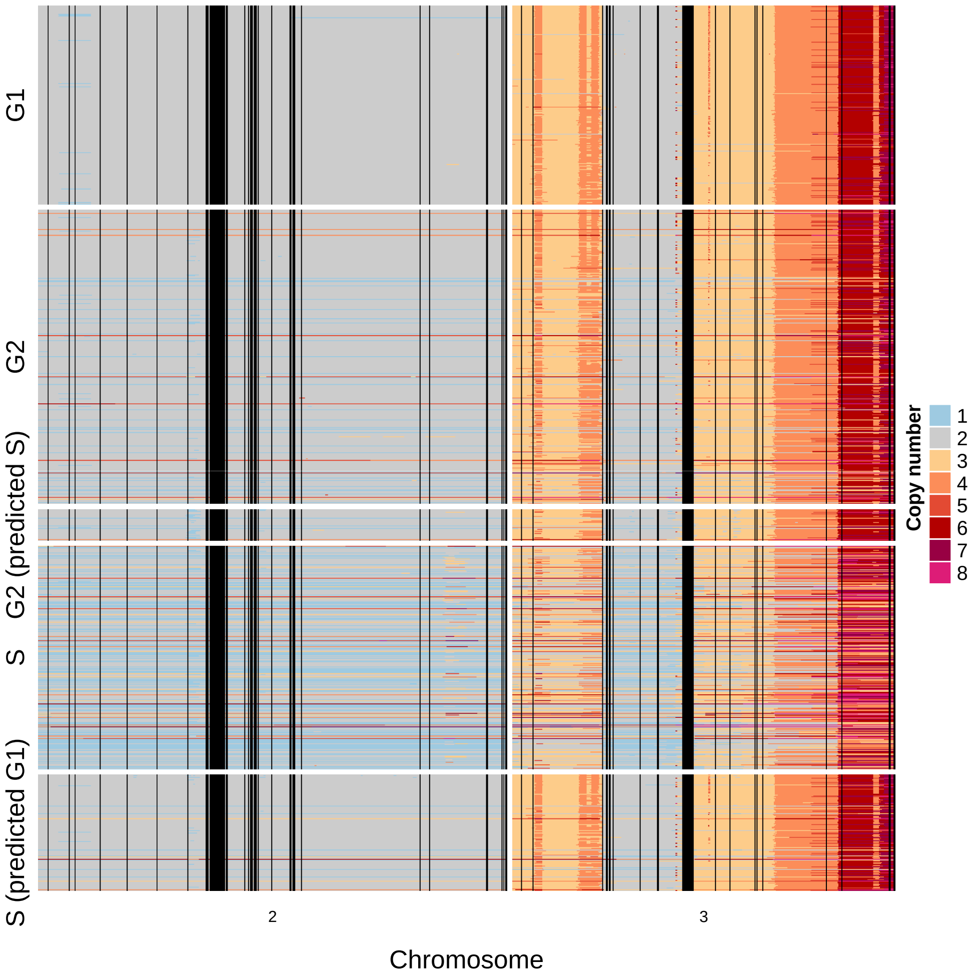


Figure S8: Initial copy number profiles as predicted by scAbsolute for DLP T-47D sample for chromosomes 2 and 3. Overall, the vast majority of cells is independently called with the same ploidy. We see that there is a small subset of cells with the wrong ploidy solution and an even smaller group of cells that is otherwise classified wrongly. The algorithm cannot distinguish between G1/G2 cells. We observe some noise, characteristic of S phase cells that are also mentioned in the original publication among the G2 cell population. One can observe a clear difference in copy number profiles between S phase cells predicted to be in G1 phase of the cell cycle, and S phase cells that are predicted to be in S phase based on cycling activity. This might indicate an issue with the underlying ground truth data.

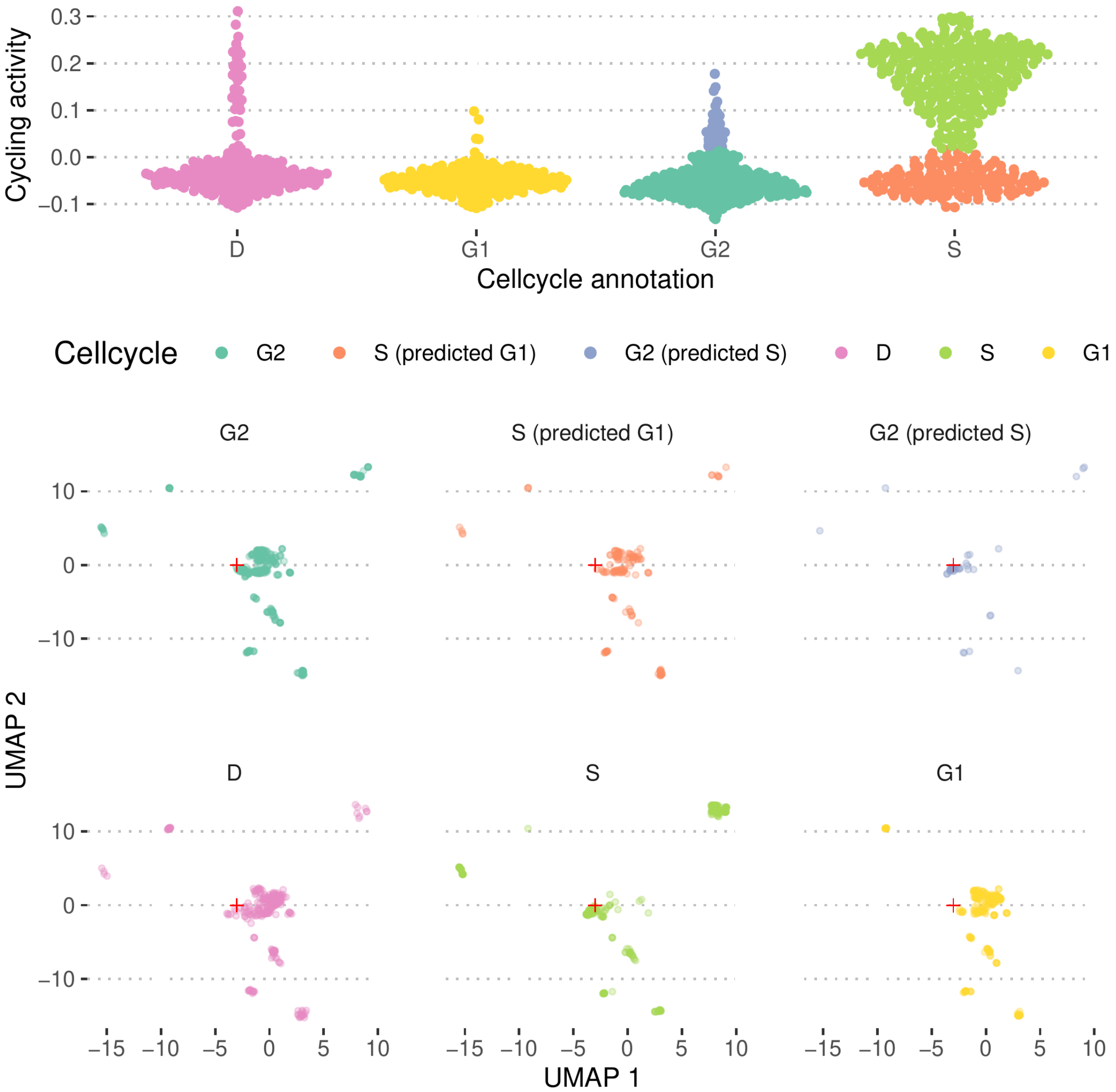


Figure S9: *scAbsolute* predictions for T-47D sample. Top panel shows cycling activity predictions for cells from the T-47D cell line, with DAPI staining based FACS cell cycle annotation on the x-axis. Both for the S phase, and for the G2 phase annotated cells, we observe subgroups that are classified differently by the cycling activity predictor. Looking at copy number profiles in a UMAP representation, we can see that the groups cluster differently based on the cycling activity predictions. Cells that are annotated to be in S phase, but predicted to be in G1 phase (orange), appear to cluster closer with the G1 cells. Similarly, cells in G2 phase that have been predicted to be undergoing replication are clustering more similarly to the cells in S phase. The same pattern can be observed in the raw copy number profiles in Fig. S8.

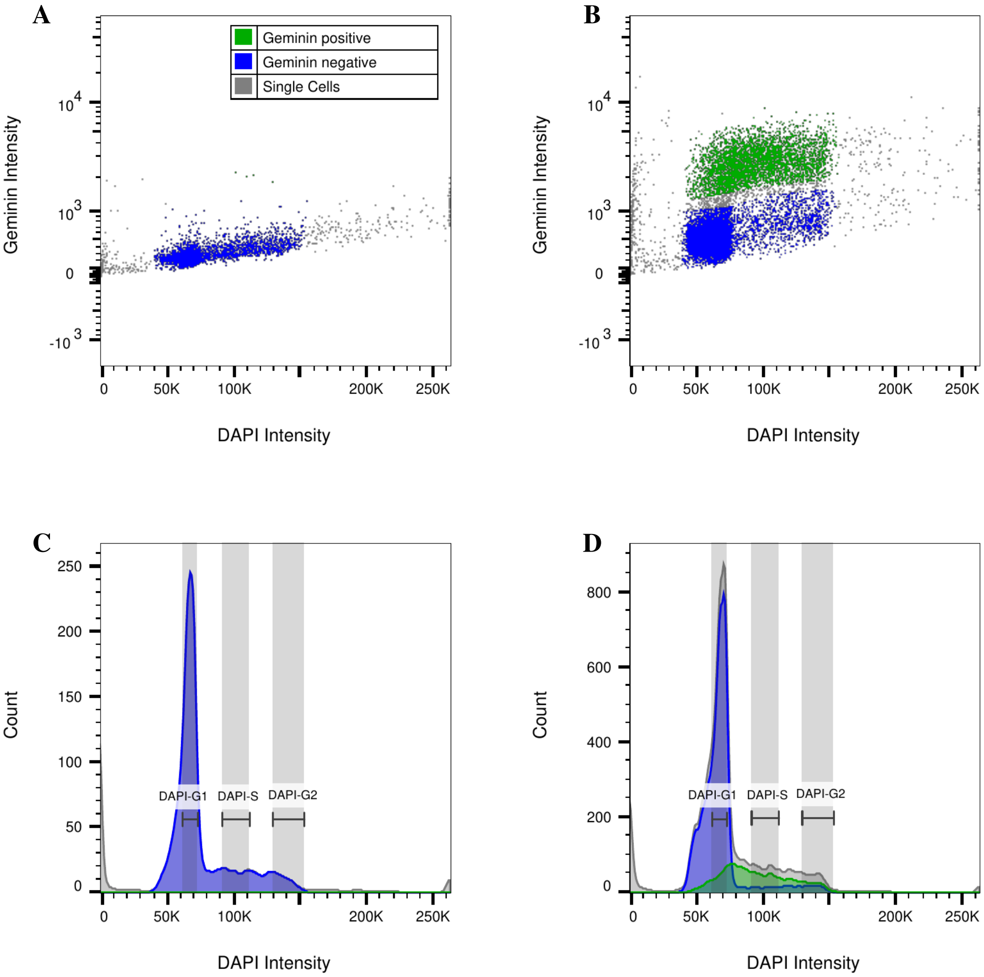


Figure S10: Flow Cytometry Images of normal cells (NA12878) stained with Geminin-AF488 and DAPI for improved G1 cell cycle sorting. Geminin Positive populations – blue; Geminin Negative populations – green. (**A**) Geminin Negative Control - NA12878 stained with DAPI and AF488 secondary antibody. (**B**) NA12878 stained with Geminin/AF488 and DAPI. Manual gating of Geminin positive and Geminin negative populations. (**C**) Cell Cycle curve of Geminin Negative Control using DAPI intensity, with overlay of Geminin gating. DAPI-G1, DAPI-S and DAPI-G2 gating represents original flow cytometry sorting gates using DAPI alone for cell cycle analysis. (**D**) Cell Cycle curve of Geminin stained NA12878 using DAPI intensity. Overlay of Geminin gating reveals Early S phase cells leaking into G1 sorting using DAPI only.

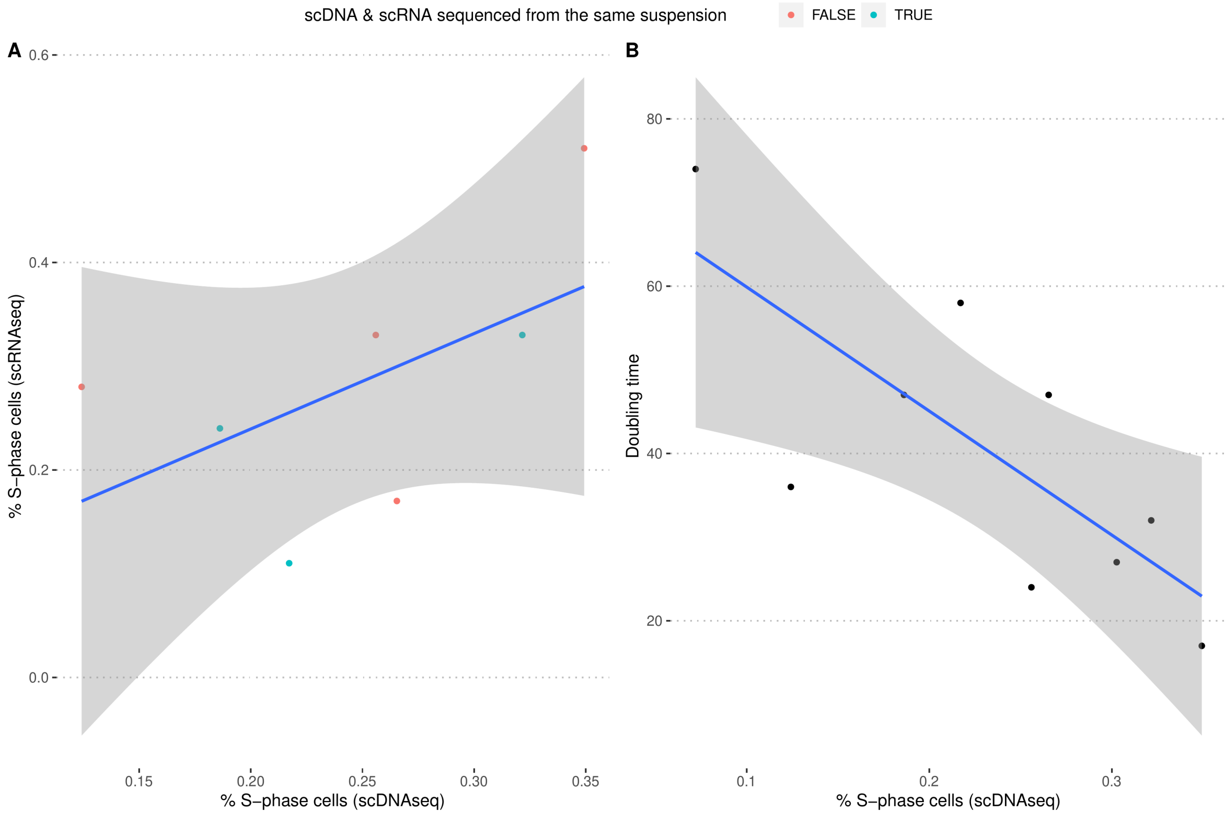


Figure S11: Validation measures for scDNAseq estimates of number of cycling cells in gastric cancer cell lines. **(A)** The number of cycling cells as estimated in scDNAseq data corresponds to estimates of cycling cells based on scRNAseq data. **(B)** Doubling time of cell lines correlates negatively with number of cycling cells as estimated in scDNAseq data.

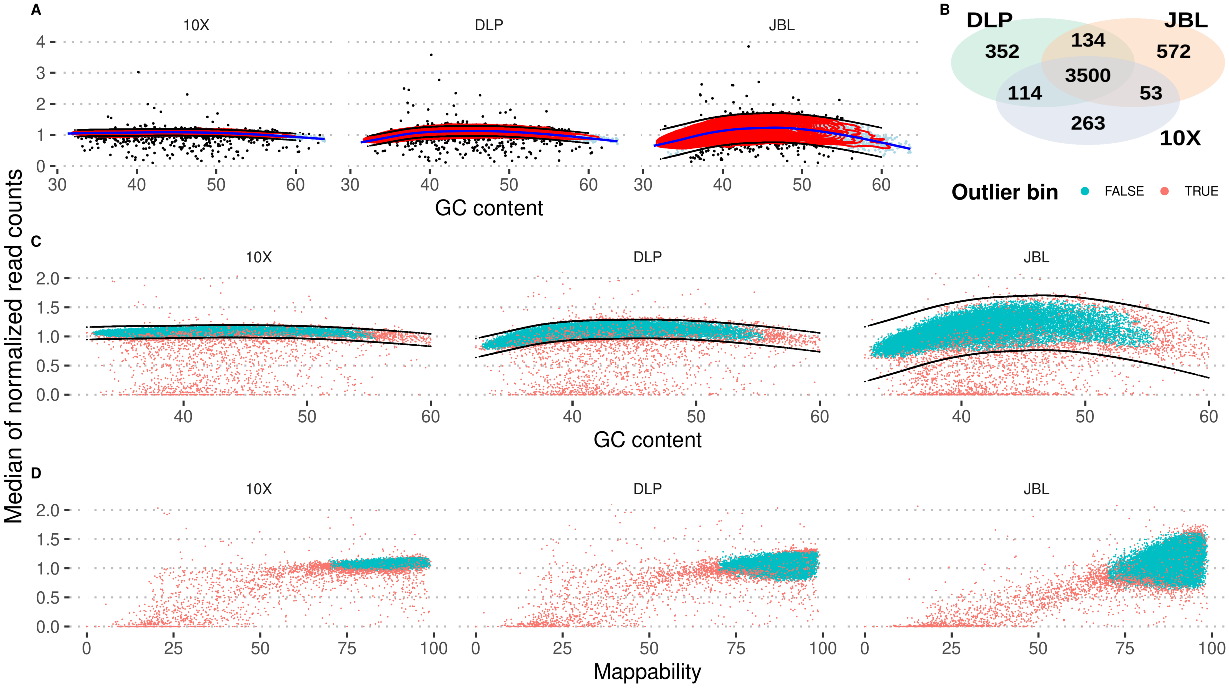


Figure S12: Bin level quality control for autosomes across sequencing technologies. A) Generative additive model smoothing (blue line) and kernel density estimation (red contours) to identify genomic bins that have below or above average median expected read counts. **B)** Number of genomic bins identified as outliers by sequencing technology. We remove the union of all outliers. **C+D)** Genomic bins identified as outliers (in red) across different sequencing technologies as a function of GC content **(C)**, and mappability **(D)**.

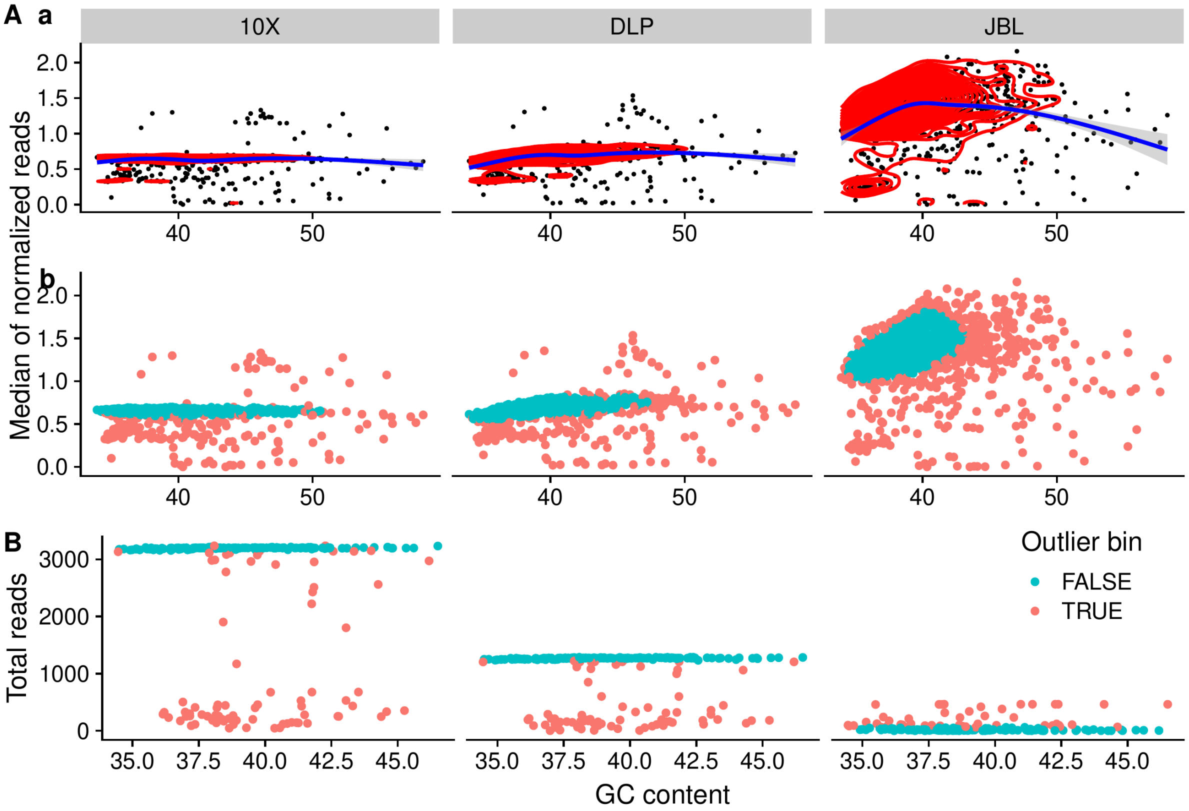
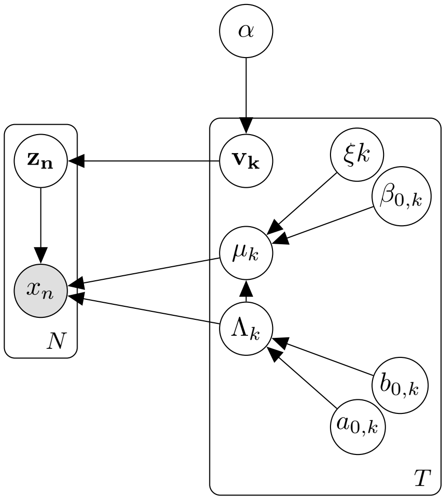


Figure S13: Bin level quality control for sex chromosomes across sequencing technologies. A) a) Generative additive model smoothing (blue line) and kernel density estimation (red contours) to identify genomic bins that have below or above average median expected read counts for the X chromosome. b) Outlier bins are marked in red. **B)** Total (absolute) reads per bin observed on the Y chromosome. Outlier bins (based on deviation from median number of total reads) are marked in red.

1 Appendix

1.1 Derivation of Dirichlet Process Gaussian Mixture Model for scAbsolute algorithm

We implement a DPGMM model, following [59]. We assume throughout the document, that the dimensionality of our problem is 1. We use the mean-field approximation to estimate $p(\theta|X) = \prod_{k=1}^{K} \pi_k \, \mathcal{N}(\mu_k, \Lambda_k|X)$. The model is depicted below. Note that we restrict the means $\mu_k = \xi \cdot k$.



1.1.1 Model and priors

Mean-field approximation

$$Q(V, \mu, \Lambda, Z) = \prod_{k}^{T} q(v_k) q(\mu_k) q(\Lambda_k) \prod_{n=1}^{N} q(z_n) \tag{6}$$

Prior distributions.

$$\begin{aligned} v_k &\sim \text{Beta}(1, \alpha) & (7)\\ \mu_k &\sim \text{Normal}(m_{0,k}, \mathbf{I}) & (8)\\ \Lambda_k &\sim \text{Gamma}(1, 1) & (9)\\ z_n &\sim \text{SBP}(V) & (10)\\ x_n &\sim \text{Normal}(\mu_{z_n}, \Lambda_{z_n}) & (11) \end{aligned}$$

Variational distributions.

$$\begin{aligned} v_k &\sim \text{Beta}(\gamma_{k,1}, \gamma_{k,2}) & (12)\\ \mu_k &\sim \text{Normal}(\xi k, \mathbf{I}) & (13)\\ \Lambda_k &\sim \text{Gamma}(a_k, b_k) & (14)\\ z_n &\sim \text{Discrete}(\rho_n) & (15) \end{aligned}$$

1.1.2 Variational bound

$$\log p(X) \geq \sum_{k=1}^{T} \mathbb{E}_q[\log p(v_k)] - \mathbb{E}_q[\log q(v_k)] \tag{16}$$

$$+ \sum_{k=1}^{T} \mathbb{E}_q[\log p(\mu_k) - \log q(\mu_k)] \tag{17}$$

$$+ \sum_{k=1}^{T} \mathbb{E}_q[\log p(\Lambda_k) - \log q(\Lambda_k)] \tag{18}$$

$$+ \sum_{n=1}^{N} \mathbb{E}_q[\log p(z_n|V) - \log q(z_n)] \tag{19}$$

$$+ \sum_{n=1}^{N} \mathbb{E}_q[\log p(x_n|\mu_{z_i}, \Lambda_{z_i})] \tag{20}$$

In the following, we derive the terms in the variational bound.

v_k terms

$$\mathbb{E}_q[\log p(V|1,\alpha)] = \mathbb{E}_q[\log \prod_{i=1}^{T} V_i] \tag{21}$$

$$= \mathbb{E}_q[\sum_{i=1}^{T} \ln V_i] = \tag{22}$$

$$= \mathbb{E}_q[\sum_{i=1}^{T} \ln \frac{\Gamma(1+\alpha)}{\Gamma(1)\Gamma(\alpha)} V_i^0 (1-V_i)^{(\alpha-1)}] = \tag{23}$$

$$= \mathbb{E}_q[\sum_{i=1}^{T} \ln \Gamma(1+\alpha) - \sum_{i=1}^{T} \ln \Gamma(\alpha) + \sum_{i=1}^{T} (\alpha-1) \ln(1-V_i)] \tag{24}$$

$$= \mathbb{E}_q[T(\ln \Gamma(1+\alpha) - \ln \Gamma(\alpha)) + (\alpha-1) \sum_{i=1}^{T} \ln(1-V_i)] \tag{25}$$

$$= T(\ln \Gamma(1+\alpha) - \ln \Gamma(\alpha)) + (\alpha-1) \sum_{i=1}^{T} \mathbb{E}_q[\ln(1-V_i)] \tag{26}$$

$$= T(\ln \Gamma(1+\alpha) - \ln \Gamma(\alpha)) \tag{27}$$

$$+ (\alpha-1) \sum_{i=1}^{T} [\Psi(\gamma_{i,2}) - \Psi(\gamma_{i,1} + \gamma_{i,2})] \tag{28}$$

$$
\begin{aligned}
\mathbb{E}_q[\log q(V|\gamma_1,\gamma_2)] &= \mathbb{E}_q[\log \prod_{i=1}^{T} V_i] && (29)\\
&= \mathbb{E}_q[\sum_{i=1}^{T} \ln V_i] && (30)\\
&= \mathbb{E}_q[\sum_{i=1}^{T} \ln \frac{\Gamma(\gamma_1+\gamma_2)}{\Gamma(\gamma_1)\Gamma(\gamma_2)} V_i^{\gamma_1-1}(1-V_i)^{(\gamma_2-1)}] && (31)\\
&= \mathbb{E}_q[\sum_{i=1}^{T} \ln\Gamma(\gamma_1+\gamma_2) - \ln\Gamma(\gamma_1) - \ln\Gamma(\gamma_2) && (32)\\
&+ (\gamma_1-1)\ln(V_i) + (\gamma_2-1)\ln(1-V_i)] && (33)\\
&= \sum_{i=1}^{T} \ln\Gamma(\gamma_1+\gamma_2) - \ln\Gamma(\gamma_1) - \ln\Gamma(\gamma_2) && (34)\\
&+ (\gamma_1-1)\mathbb{E}_q[\ln(V_i)] + (\gamma_2-1)\mathbb{E}_q[\ln(1-V_i)] && (35)\\
&= \sum_{i=1}^{T} \ln\Gamma(\gamma_1+\gamma_2) - \ln\Gamma(\gamma_1) - \ln\Gamma(\gamma_2) && (36)\\
&+ (\gamma_1-1)(\Psi(\gamma_{i,1}) - \Psi(\gamma_{i,1}+\gamma_{i,2})) && (37)\\
&+ (\gamma_2-1)(\Psi(\gamma_{i,2}) - \Psi(\gamma_{i,1}+\gamma_{i,2})) && (38)\\
& && (39)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{k=1}^{T}\mathbb{E}_q[\log p(v_k)] - \mathbb{E}_q[\log q(v_k)] &= T(\ln\Gamma(1+\alpha) - \ln\Gamma(\alpha)) && (40)\\
&+ (\alpha-1)\sum_{k=1}^{T}[\Psi(\gamma_{k,2}) - \Psi(\gamma_{k,1}+\gamma_{k,2})] && (41)\\
&- \sum_{k=1}^{T} \ln\Gamma(\gamma_{k,1}+\gamma_{k,2}) + \ln\Gamma(\gamma_{k,1}) + \ln\Gamma(\gamma_{k,2}) && (42)\\
&- \sum_{k=1}^{T}(\gamma_{k,1}-1)(\Psi(\gamma_{k,1}) - \Psi(\gamma_{k,1}+\gamma_{k,2})) && (43)\\
&- \sum_{k=1}^{T}(\gamma_{k,2}-1)(\Psi(\gamma_{k,2}) - \Psi(\gamma_{k,1}+\gamma_{k,2})) && (44)
\end{aligned}
$$

μ_k terms

$$
\begin{aligned}
\mathbb{E}_q[\log p(\mu_k) - \log q(\mu_k)] &= -\mathrm{KL}(Q_{\mu_k}||P_{\mu_k}) && (45)\\
&= -\frac{1}{2}||\xi k - m_{o,k}||^2 = -\frac{1}{2}(\xi k - m_{o,k})^2 && (46)
\end{aligned}
$$

Λ_k terms We use the inverse scale parameter characterization of the Gamma distribution, with $\Lambda_k \sim \mathbb{G}(\Lambda_k|a_{k,0}, b_{k,0})$, for $a_{k,0}, b_{k,0} = 1$, and $\mathbb{E}_q[\Lambda_k] = \frac{a_k}{b_k}$.

$$\mathbb{E}_q[\log p(\Lambda)] = \mathbb{E}_q[\log \prod_{k=1}^{T} \mathbb{G}(\Lambda_k|a_{k,0}, b_{k,0}) \tag{47}$$

$$= \mathbb{E}_q[\sum_{k=1}^{T} \Big(a_{k,0}\log(b_{k,0} + (a_{k,0}-1)\log(\Lambda_k) - b_{k,0}\Lambda_k - \log(\Gamma(a_{k,0}))\Big)] \tag{48}$$

$$= \sum_{k=1}^{T} \Big(a_{k,0}\log(b_{k,0}) + (a_{k,0}-1)\mathbb{E}_q\log(\Lambda_k) - b_{k,0}\mathbb{E}_q\Lambda_k - \log\Gamma(a_{k,0})\Big) \tag{49}$$

$$= \sum_{k=1}^{T} \Big(a_{k,0}\log(b_{k,0}) + (a_{k,0}-1)(\Psi(a_k) - \log(b_k)) - b_{k,0}\frac{a_k}{b_k} - \log(\Gamma(a_{k,0}))\Big) \tag{50}$$

$$\overset{a_{0,k}=b_{0,k}=1}{=} \sum_{k=1}^{T} \frac{a_k}{b_k} \tag{51}$$

$$\mathbb{E}_q[\log q(\Lambda_k)] \overset{\text{neg. Entropy}}{=} -a_k + \log b_k - \log\Gamma a_k - (1-a_k)\Psi(a_k) \tag{52}$$

$$\mathbb{E}_q[\log p(\Lambda_k) - \log q(\Lambda_k)] = a_k - \log b_k + \log\Gamma a_k + (1-a_k)\Psi(a_k) - \frac{a_k}{b_k} \tag{53}$$

z_n **terms** Here, we use the equations as presented by [59, p. 129].

$$\mathbb{E}_q[\log p(z_n|V)] = \sum_{k=1}^{T} f(z_n > k)\ \mathbb{E}_q[\log(1-v_k)] + f(z_n = k)\ \mathbb{E}_q[\log v_k] \tag{54}$$

$$= \sum_{k=1}^{T}\sum_{j=k+1}^{T} \phi_{n,j}\ \Big(\Psi(\gamma_{k,2}) - \Psi(\gamma_{k,1}+\gamma_{k,2})\Big) + \tag{55}$$

$$\phi_{n,k}\ \Big(\Psi(\gamma_{k,1}) - \Psi(\gamma_{k,1}+\gamma_{k,2})\Big) \tag{56}$$

with the following definitions

$$f(z_n = k) = \phi_{n,k} \tag{57}$$

$$f(z_n > k) = \sum_{j=k+1}^{T} \phi_{n,j} \tag{58}$$

$$\mathbb{E}_q[\log v_k] = \Psi(\gamma_{k,1}) - \Psi(\gamma_{k,1}+\gamma_{k,2}) \tag{59}$$

$$\mathbb{E}_q[\log(1-v_k)] = \Psi(\gamma_{k,2}) - \Psi(\gamma_{k,1}+\gamma_{k,2}) \tag{60}$$

$$\mathbb{E}_q[\log q(z_n)] = \sum_{n=1}^{N} \phi_{n,k}\log(\phi_{n,k}) \tag{61}$$

$$\mathbb{E}_q[\log p(z_n) - \log q(z_n)] = \tag{62}$$

$$= \sum_{k=1}^{T} \Big(-\phi_{n,k}\log(\phi_{n,k})\Big) \tag{63}$$

$$+ \sum_{k=1}^{T} \phi_{n,k}\Big(\Psi(\gamma_{k,1}) - \Psi(\gamma_{k,1}+\gamma_{k,2})\Big) \tag{64}$$

$$+ \sum_{j=k+1}^{T} \phi_{n,j}\Big(\Psi(\gamma_{k,2}) - \Psi(\gamma_{k,1}+\gamma_{k,2})\Big) \tag{65}$$

Derivation: (note this leads to a slightly different result than presented in Blei.)

$$P(z_n = k) = v_k \prod_{j=1}^{k-1} 1 - v_j \tag{66}$$

We use the following identities of the Beta distribution: for $B \sim \text{Beta}(b_1, b_2)$

$$\mathbb{E}[\log B] = \Psi(b_1) - \Psi(b_1 + b_2) \tag{67}$$

$$1 - B \sim \text{Beta}(b_2, b_1) \tag{68}$$

$$\mathbb{E}_q[\log(p(z_n|V) - \log(q(z_n))] = \mathbb{E}_z\Big[\mathbb{E}_v[\log V_{z_n} + \sum_{j=1}^{z_n-1} \log(1 - v_j) - \log(\rho_n)]\Big] \tag{69}$$

$$= \mathbb{E}_z\Big[\mathbb{E}_v[\log V_{z_n}] + \sum_{j=1}^{z_n-1} \mathbb{E}_v[\log(1 - v_j)] - \log(\rho_n)\Big] \tag{70}$$

$$= \mathbb{E}_z\Big[\Psi(\gamma_{z_n,1}) - \Psi(\gamma_{z_n,1} + \gamma_{z_n,2}) + \sum_{j=1}^{z_n-1} \mathbb{E}_v[\log(1 - v_j)] - \log(\rho_n)\Big] \tag{71}$$

x_n terms

$$\mathbb{E}_q[\log p(x_n|\mu_{z_n}, \Lambda_{z_n})] \tag{72}$$

$$= \sum_{k=1}^{T} \phi_{n,k} \mathbb{E}_{\Lambda_k}[\mathbb{E}_{\mu_k}[\log P(x_n|\mu_k, \Lambda_k)] \tag{73}$$

$$= \sum_{k=1}^{T} \phi_{n,k} \mathbb{E}_{\Lambda_k}[\mathbb{E}_{\mu_k}[\log \mathbb{N}(x_n; \mu_k, \Lambda_k^{-1})] \tag{74}$$

$$= \sum_{k=1}^{T} \phi_{n,k} \mathbb{E}_{\Lambda_k}\left[\int_{\mu_k} \mathbb{N}(\mu_k; \xi, \mathbf{I}) \log \mathbb{N}(\mathbf{x_n}; \boldsymbol{\mu_k}, \boldsymbol{\Lambda_k^{-1}})\mathbf{d}\boldsymbol{\mu_k}\right] \tag{75}$$

$$= \sum_{k=1}^{T} \phi_{n,k} \mathbb{E}_{\Lambda_k}\left[\int_{\mu_k} (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mu_k - \xi k)^2)\right. \tag{76}$$

$$\left.\Big(-\frac{1}{2}\log(\frac{2\pi}{\Lambda_k}) - \frac{\Lambda_k}{2}(x_n - \mu_k)^2\Big) d\mu_k\right] \tag{77}$$

$$= \sum_{k=1}^{T} \phi_{n,k} \mathbb{E}_{\Lambda_k}\left[-\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\Lambda_k) + \right. \tag{78}$$

$$\left.\left(\frac{1}{2}\int_{\mu_k} (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mu_k - \xi * k)^2)(-\Lambda_k(x_n^2 - 2\mu_k x_n + \mu_k^2) d\mu_k\right)\right] \tag{79}$$

$$= \sum_{k=1}^{T} \phi_{n,k} \mathbb{E}_{\Lambda_k}\left[-\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\Lambda_k) + \right. \tag{80}$$

$$\left.-\frac{1}{2}\Lambda_k((x_n - \xi k)^2 + 1)\right] \tag{81}$$

$$= \sum_{k=1}^{T} \phi_{n,k}\left[-\frac{1}{2}\log(2\pi) + \frac{1}{2}\Psi(a_k) - \frac{1}{2}\log(b_k) + \right. \tag{82}$$

$$\left.-\frac{1}{2}\frac{a_k}{b_k}((x_n - \xi k)^2 + 1)\right] \tag{83}$$

$$\tag{84}$$

We will refer to the following term later on

$$\eta_x = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\Psi(a_k) - \frac{1}{2}\log(b_k) + \tag{85}$$

$$-\frac{1}{2}\frac{a_k}{b_k}((x_n - \xi k)^2 + 1) \tag{86}$$

1.1.3 Variational updates

Updates for v_k These update equations are given in [59, p. 129].

$$\gamma_{k,1} = 1 + \sum_{n=1}^{N} \phi_{n,k} \tag{87}$$

$$\gamma_{k,2} = \alpha + \sum_{n=1}^{N} \sum_{j=k+1}^{T} \phi_{n,j} \tag{88}$$

Updates for ξ

$$\frac{\delta L}{\delta \xi} = \sum_{k=1}^{T} -k(\xi k - m_{0,k}) + \sum_{n=1}^{N}\sum_{k=1}^{T} \phi_{n,k}\frac{k a_k}{b_k}(x_n - \xi k) \tag{89}$$

$$= -\sum_{k=1}^{T} k^2 \xi + \sum_{k=1}^{T} k m_{0,k} + \sum_{n=1}^{N}\sum_{k=1}^{T} \phi_{n,k}\frac{k a_k}{b_k} x_n - \sum_{n=1}^{N}\sum_{k=1}^{T} \phi_{n,k}\frac{k^2 a_k}{b_k}\xi \overset{!}{=} 0 \tag{90}$$

$$\Rightarrow \xi = \frac{\sum_{k=1}^{T} k m_{0,k} + \sum_{n=1}^{N}\sum_{k=1}^{T} \phi_{n,k}\frac{k a_k}{b_k} x_n}{\sum_{k=1}^{T} k^2 + \sum_{n=1}^{N}\sum_{k=1}^{T} \phi_{n,k}\frac{k^2 a_k}{b_k}} \tag{91}$$

Updates for a_k and b_k

$$\frac{\delta L}{\delta a_k} = 1 + \Psi(a_k) + (1 - a_k)\Psi'(a_k) - \Psi(a_k) \tag{92}$$

$$-\frac{1}{b_k} + \frac{1}{2} sum_{n=1}^{N} \phi_{n,k}\Psi'(a_k) - \frac{1}{2}\sum_{n=1}^{N} \phi_{n,k}\frac{1}{b_k}((\xi k - x_n)^2 + 1) \tag{93}$$

$$= 1 + \Psi'(a_k)(1 - a_k + \frac{1}{2}\sum_{n=1}^{N}\phi_{n,k}) - \frac{1}{b_k}(1 + \frac{1}{2}\sum_{n=1}^{N}\phi_{n,k}((\xi k - x_n) + 1) \tag{94}$$

using the fact that Ψ is a monotonous function (95)

$$\Rightarrow a_k = 1 + \frac{1}{2}\sum_{n=1}^{N}\phi_{n,k} \tag{96}$$

$$\Rightarrow b_k = 1 + \frac{1}{2}\sum_{n=1}^{N}\phi_{n,k}((\xi k - x_n)^2 + 1) \tag{97}$$

Computing the partial derivative $\frac{\delta L}{\delta b_k}$ and setting the values of a_k and b_k to the above results satisfies the condition.

$$\frac{\delta L}{\delta b_k} = -\frac{1}{b_k} + \frac{a_k}{b_k^2} - \frac{1}{2}\sum_{n=1}^{N}\phi_{n,k}\frac{1}{b_k} + \frac{a_k}{b_k^2}\sum_{n=1}^{N}((\xi k - x_n)^2 + 1) \overset{!}{=} 0 \tag{98}$$

Updates for ρ_n Here, we need to take into account the constraint $\sum_{k=1}^{T}\phi_{n,k} = 1$ We do this by using Lagrange multipliers.

$$\mathcal{L}(\rho_n, \lambda) = \sum_{k=1}^{T} \big(-\phi_{n,k} \log(\phi_{n,k}) + \sum_{j=k+1}^{T} \phi_{n,j} (\Psi(\gamma_{k,2}) - \Psi(\gamma_{k,1} + \gamma_{k,2})) \tag{99}$$

$$+ \phi_{n,k}(\Psi(\gamma_{k,1}) - \Psi(\gamma_{k,1} + \gamma_{k,2})) + \phi_{n,k}\eta_x \big) - \lambda(\sum_{k=1}^{T} \phi_{n,k} - 1) \tag{100}$$

$$\frac{\delta \mathcal{L}}{\delta \phi_{n,k}} = -1 - \log(\phi_{n,k} + \sum_{j=k+1}^{T} \phi_{n,j}(\Psi(\gamma_{k,2}) - \Psi(\gamma_{k,1} + \gamma_{k,2})) \tag{101}$$

$$+ (\Psi(\gamma_{k,1}) - \Psi(\gamma_{k,1} + \gamma_{k,2})) + \eta_x - \lambda \tag{102}$$

$$\frac{\delta \mathcal{L}}{\delta \lambda} = 1 - \sum_{k=1}^{T} \phi_{n,k} \tag{103}$$

$$\Rightarrow \phi_{n,k} = \frac{\exp(\eta_{z_{n,k}} + \eta_{x_{n,k}} - 1)}{\sum_{k=1}^{T} \exp(\eta_{z_{n,k}} + \eta_{x_{n,k}} - 1)} \tag{104}$$

References

1. Zack, T. I. *et al.* Pan-Cancer Patterns of Somatic Copy Number Alteration. *Nature Genetics* **45,** 1134–1140. doi:10.1038/ng.2760 (10 2013).
2. Lukow, D. A. & Sheltzer, J. M. Chromosomal Instability and Aneuploidy as Causes of Cancer Drug Resistance. *Trends in Cancer.* doi:10.1016/j.trecan.2021.09.002 (2021).
3. McGranahan, N. & Swanton, C. Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell* **27,** 15–26. doi:10.1016/j.ccell.2014.12.001 (2015).
4. Beroukhim, R. *et al.* The Landscape of Somatic Copy-Number Alteration across Human Cancers. *Nature* **463,** 899–905. doi:10.1038/nature08822 (7283 2010).
5. Ciriello, G. *et al.* Emerging Landscape of Oncogenic Signatures across Human Cancers. *Nature Genetics* **45,** 1127–1133. doi:10.1038/ng.2762 (10 2013).
6. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33,** 676–689.e3. doi:10.1016/j.ccell.2018.03.007 (2018).
7. Chowdhury, S. A. *et al.* Algorithms to Model Single Gene, Single Chromosome, and Whole Genome Copy Number Changes Jointly in Tumor Phylogenetics. *PLOS Computational Biology* **10,** e1003740. doi:10.1371/journal.pcbi.1003740 (2014).
8. Schwarz, R. F. *et al.* Phylogenetic Quantification of Intra-tumour Heterogeneity. *PLOS Computational Biology* **10,** e1003535. doi:10.1371/journal.pcbi.1003535 (2014).
9. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The Causes and Consequences of Genetic Heterogeneity in Cancer Evolution. *Nature* **501,** 338–345. doi:10.1038/nature12625 (7467 2013).
10. Sansregret, L., Vanhaesebroeck, B. & Swanton, C. Determinants and Clinical Implications of Chromosomal Instability in Cancer. *Nature Reviews Clinical Oncology* **15,** 139–150. doi:10.1038/nrclinonc.2017.198 (3 2018).
11. Bakhoum, S. F., Danilova, O. V., Kaur, P., Levy, N. B. & Compton, D. A. Chromosomal Instability Substantiates Poor Prognosis in Patients with Diffuse Large B-cell Lymphoma. *Clinical Cancer Research* **17,** 7704–7711. doi:10.1158/1078-0432.CCR-11-2049 (2011).
12. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor Aneuploidy Correlates with Markers of Immune Evasion and with Reduced Response to Immunotherapy. *Science* **355,** eaaf8399. doi:10.1126/science.aaf8399 (2017).
13. Buccitelli, C. *et al.* Pan-Cancer Analysis Distinguishes Transcriptional Changes of Aneuploidy from Proliferation. *Genome Research* **27,** 501–511. doi:10.1101/gr.212225.116 (2017).
14. López, S. *et al.* Interplay between Whole-Genome Doubling and the Accumulation of Deleterious Alterations in Cancer Evolution. *Nature Genetics* **52,** 283–293. doi:10.1038/s41588-020-0584-7 (3 2020).
15. Goupil, A. *et al.* Chromosomes Function as a Barrier to Mitotic Spindle Bipolarity in Polyploid Cells. *Journal of Cell Biology* **219,** e201908006. doi:10.1083/jcb.201908006 (2020).

16. Gemble, S. *et al.* Genetic Instability from a Single S Phase after Whole-Genome Duplication. *Nature* **604,** 146–151. doi:10.1038/s41586-022-04578-4 (7904 2022).
17. Storchova, Z. & Pellman, D. From Polyploidy to Aneuploidy, Genome Instability and Cancer. *Nature Reviews Molecular Cell Biology* **5,** 45–54. doi:10.1038/nrm1276 (1 2004).
18. Storchova, Z. & Kuffer, C. The Consequences of Tetraploidy and Aneuploidy. *Journal of Cell Science* **121,** 3859–3866. doi:10.1242/jcs.039537 (2008).
19. Fujiwara, T. *et al.* Cytokinesis Failure Generating Tetraploids Promotes Tumorigenesis in P53-Null Cells. *Nature* **437,** 1043–1047. doi:10.1038/nature04217 (7061 2005).
20. Dewhurst, S. M. *et al.* Tolerance of Whole-Genome Doubling Propagates Chromosomal Instability and Accelerates Cancer Genome Evolution. *Cancer Discovery* **4,** 175–185. doi:10.1158/2159-8290.CD-13-0285 (2014).
21. Ganem, N. J., Godinho, S. A. & Pellman, D. A Mechanism Linking Extra Centrosomes to Chromosomal Instability. *Nature* **460,** 278–282. doi:10.1038/nature08136 (7252 2009).
22. Quinton, R. J. *et al.* Whole-Genome Doubling Confers Unique Genetic Vulnerabilities on Tumour Cells. *Nature,* 1–6. doi:10.1038/s41586-020-03133-3 (2021).
23. Carter, S. L. *et al.* Absolute Quantification of Somatic DNA Alterations in Human Cancer. *Nature Biotechnology* **30,** 413–421. doi:10.1038/nbt.2203 (5 2012).
24. Dentro, S. C. *et al.* Characterizing Genetic Intra-Tumor Heterogeneity across 2,658 Human Cancer Genomes. *Cell* **184,** 2239–2254.e39. doi:10.1016/j.cell.2021.03.009 (2021).
25. Bielski, C. M. *et al.* Genome Doubling Shapes the Evolution and Prognosis of Advanced Cancers. *Nature Genetics* **50,** 1189–1195. doi:10.1038/s41588-018-0165-1 (2018).
26. Salcedo, A. *et al.* A Community Effort to Create Standards for Evaluating Tumor Subclonal Reconstruction. *Nature Biotechnology* **38,** 97–107. doi:10.1038/s41587-019-0364-z (1 2020).
27. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149,** 994–1007. doi:10.1016/j.cell.2012.04.023 (2012).
28. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine* **376,** 2109–2121. doi:10.1056/NEJMoa1616288 (2017).
29. Zaccaria, S. & Raphael, B. J. Accurate Quantification of Copy-Number Aberrations and Whole-Genome Duplications in Multi-Sample Tumor Sequencing Data. *Nature Communications* **11,** 4301. doi:10.1038/s41467-020-17967-y (1 2020).
30. Van Loo, P. *et al.* Allele-Specific Copy Number Analysis of Tumors. *Proceedings of the National Academy of Sciences* **107,** 16910–16915. doi:10.1073/pnas.1009843107 (2010).
31. Chen, H., Bell, J. M., Zavala, N. A., Ji, H. P. & Zhang, N. R. Allele-Specific Copy Number Profiling by next-Generation DNA Sequencing. *Nucleic Acids Research* **43,** e23. doi:10.1093/nar/gku1252 (2015).
32. Favero, F. *et al.* Sequenza: Allele-Specific Copy Number and Mutation Profiles from Tumor Sequencing Data. *Annals of Oncology* **26,** 64–70. doi:10.1093/annonc/mdu479 (2015).
33. Shen, R. & Seshan, V. E. FACETS: Allele-Specific Copy Number and Clonal Heterogeneity Analysis Tool for High-Throughput DNA Sequencing. *Nucleic Acids Research* **44,** e131. doi:10.1093/nar/gkw520 (2016).
34. Cun, Y., Yang, T.-P., Achter, V., Lang, U. & Peifer, M. Copy-Number Analysis and Inference of Subclonal Populations in Cancer Genomes Using Sclust. *Nature Protocols* **13,** 1488–1501. doi:10.1038/nprot.2018.033 (6 2018).
35. Poell, J. B. *et al.* ACE: Absolute Copy Number Estimation from Low-Coverage Whole-Genome Sequencing Data. *Bioinformatics* **35,** 2847–2849. doi:10.1093/bioinformatics/bty1055 (2019).
36. Sauer, C. M. *et al. Absolute Copy Number Fitting from Shallow Whole Genome Sequencing Data* 2021. doi:10.1101/2021.07.19.452658.
37. Navin, N. *et al.* Tumour Evolution Inferred by Single-Cell Sequencing. *Nature* **472,** 90–94. doi:10.1038/nature09807 (2011).
38. Wang, Y. *et al.* Clonal Evolution in Breast Cancer Revealed by Single Nucleus Genome Sequencing. *Nature* **512,** 155–160. doi:10.1038/nature13600 (7513 2014).
39. Zahn, H. *et al.* Scalable Whole-Genome Single-Cell Library Preparation without Preamplification. *Nature Methods* **14,** 167–173. doi:10.1038/nmeth.4140 (2 2017).
40. Laks, E. *et al.* Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell* **179,** 1207–1221.e22. doi:10.1016/j.cell.2019.10.026 (2019).
41. Vitak, S. A. *et al.* Sequencing Thousands of Single-Cell Genomes with Combinatorial Indexing. *Nature Methods* **14,** 302–308. doi:10.1038/nmeth.4154 (3 2017).

42. Mulqueen, R. M. *et al.* High-Content Single-Cell Combinatorial Indexing. *Nature Biotechnology* **39,** 1574–1580. doi:10.1038/s41587-021-00962-z (12 2021).
43. Minussi, D. C. *et al.* Breast Tumours Maintain a Reservoir of Subclonal Diversity during Expansion. *Nature* **592,** 302–308. doi:10.1038/s41586-021-03357-x (7853 2021).
44. *Single Cell CNV* 10x Genomics.
45. Mallory, X. F., Edrisi, M., Navin, N. & Nakhleh, L. Assessing the Performance of Methods for Copy Number Aberration Detection from Single-Cell DNA Sequencing Data. *PLOS Computational Biology* **16,** e1008012. doi:10.1371/journal.pcbi.1008012 (2020).
46. Lai, D. *et al.* Package 'HMMcopy' (2011).
47. Knouse, K. A., Wu, J. & Amon, A. Assessment of Megabase-Scale Somatic Copy Number Variation Using Single-Cell Sequencing. *Genome Research* **26,** 376–384. doi:10.1101/gr.198937.115 (2016).
48. Funnell, T. *et al.* Single-Cell Genomic Variation Induced by Mutational Processes in Cancer. *Nature,* 1–10. doi:10.1038/s41586-022-05249-0 (2022).
49. Markowska, M. *et al.* CONET: Copy Number Event Tree Model of Evolutionary Tumor History for Single-Cell Data, 2021.04.23.441204. doi:10.1101/2021.04.23.441204 (2021).
50. Salehi, S. *et al.* Cancer Phylogenetic Tree Inference at Scale from 1000s of Single Cell Genomes. doi:10.1101/2020.05.06.058180 (2020).
51. Garvin, T. *et al.* Interactive Analysis and Assessment of Single-Cell Copy-Number Variations. *Nature Methods* **12,** 1058–1060. doi:10.1038/nmeth.3578 (11 2015).
52. Mallory, X. F., Edrisi, M., Navin, N. & Nakhleh, L. Methods for Copy Number Aberration Detection from Single-Cell DNA-sequencing Data. *Genome Biology* **21,** 208. doi:10.1186/s13059-020-02119-8 (2020).
53. Zaccaria, S. & Raphael, B. J. Characterizing Allele- and Haplotype-Specific Copy Numbers in Single Cells with CHISEL. *Nature Biotechnology,* 1–8. doi:10.1038/s41587-020-0661-6 (2020).
54. Andor, N. *et al.* Joint Single Cell DNA-seq and RNA-seq of Gastric Cancer Cell Lines Reveals Rules of in Vitro Evolution. *NAR Genomics and Bioinformatics* **2.** doi:10.1093/nargab/lqaa016 (2020).
55. Killick, R., Fearnhead, P. & Eckley, I. A. Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association* **107,** 1590–1598. doi:10.1080/01621459.2012.737745 (2012).
56. Anscombe, F. J. Sampling Theory Of The Negative Binomial And Logarithmic Series Distributions. *Biometrika* **37,** 358–382. doi:10.1093/biomet/37.3-4.358 (1950).
57. Bliss, C. I. & Fisher, R. A. Fitting the Negative Binomial Distribution to Biological Data. *Biometrics* **9,** 176–200. doi:10.2307/3001850 (1953).
58. Anraku, K. & Yanagimoto, T. Estimation for the Negative Binomial Distribution Based on the Conditional Likelihood. *Communications in Statistics - Simulation and Computation* **19,** 771–786. doi:10.1080/03610919008812887 (1990).
59. Blei, D. M. & Jordan, M. I. Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis* **1,** 121–143. doi:10.1214/06-BA104 (2006).
60. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic Variational Inference. *The Journal of Machine Learning Research* **14,** 1303–1347 (2013).
61. Abadi, M. *et al. TensorFlow: Large-scale Machine Learning on Heterogeneous Systems* 2015.
62. Hansen, R. S. *et al.* Sequencing Newly Replicated DNA Reveals Widespread Plasticity in Human Replication Timing. *Proceedings of the National Academy of Sciences* **107,** 139–144. doi:10.1073/pnas.0912402107 (2010).
63. Thurman, R. E., Day, N., Noble, W. S. & Stamatoyannopoulos, J. A. Identification of Higher-Order Functional Domains in the Human ENCODE Regions. *Genome Research* **17,** 917–927. doi:10.1101/gr.6081407 (2007).
64. Hipel, K. W. & McLeod, A. I. *Time Series Modelling of Water Resources and Environmental Systems* (Elsevier, 1994).
65. Scheinin, I. *et al.* DNA Copy Number Analysis of Fresh and Formalin-Fixed Specimens by Shallow Whole-Genome Sequencing with Identification and Exclusion of Problematic Regions in the Genome Assembly. *Genome Research* **24,** 2022–2032. doi:10.1101/gr.175141.114 (2014).
66. Loh, P.-R. *et al.* Reference-Based Phasing Using the Haplotype Reference Consortium Panel. *Nature Genetics* **48,** 1443–1448. doi:10.1038/ng.3679 (11 2016).
67. Durbin, R. Efficient Haplotype Matching and Storage Using the Positional Burrows–Wheeler Transform (PBWT). *Bioinformatics* **30,** 1266–1272. doi:10.1093/bioinformatics/btu014 (2014).