

# SCONCE: A method for profiling Copy Number Alterations in Cancer Evolution using Single Cell Whole Genome Sequencing

Sandra Hui<sup>1,✉</sup> and Rasmus Nielsen<sup>1, 2, 3,✉</sup>

<sup>1</sup>Center for Computational Biology, University of California, Berkeley, Berkeley, 94720, USA

<sup>2</sup>Department of Integrative Biology, University of California, Berkeley, Berkeley, 94720, USA

<sup>3</sup>Department of Statistics, University of California, Berkeley, Berkeley, 94720, USA

Copy number alterations are a significant driver in cancer growth and development, but remain poorly characterized on the single cell level. Although genome evolution in cancer cells is Markovian through evolutionary time, copy number alterations are not Markovian along the genome. However, existing methods call copy number profiles with Hidden Markov Models or change point detection algorithms based on changes in observed read depth, corrected by genome content, and do not account for the stochastic evolutionary process. We present a theoretical framework to use tumor evolutionary history to accurately call copy number alterations in a principled manner. In order to model the tumor evolutionary process and account for technical noise from low coverage single cell whole genome sequencing data, we developed SCONCE, a method based on a Hidden Markov Model to analyze read depth data from tumor cells using matched normal cells as negative controls. Using a combination of public datasets and simulations, we show SCONCE accurately decodes copy number profiles, with broader implications for understanding tumor evolution. SCONCE is implemented in C++11 and is freely available from <https://github.com/NielsenBerkeleyLab/sconce>.

copy number alterations | CNAs | cancer genomics | aneuploidy

Correspondence: [sandra\\_hui@berkeley.edu](mailto:sandra_hui@berkeley.edu)  
[rasmus\\_nielsen@berkeley.edu](mailto:rasmus_nielsen@berkeley.edu)

## 1. Introduction

In cancerous cells, somatic driver and passenger single nucleotide polymorphisms (SNPs) and copy number alterations (CNAs) accumulate over time. CNAs are extremely common across cancer types (1, 2).

Many large scale cancer studies are done with bulk samples, and many methods and evaluation techniques (3, 4) have been developed to identify copy number alterations in bulk sequencing, especially for low coverage data (5) and tumor heterogeneity deconvolution (6). However, bulk sequencing averages mutations across many cells and loses the granularity and detail single cell sequencing (SCS) can provide. Using single cell sequencing, we can evaluate these mutations on a cell by cell level and treat each cell as an individual in a population. However, the SCS process is technically challenging and produces noisy low coverage data, due to challenges like cell dissociation, small amounts of starting DNA, and non uniform whole genome amplification (7). Although the rapidly increasing availability of single cell RNA

sequencing (scRNA-seq) of tumors can yield insights into tumor subpopulations (8) and relevant biological pathways and processes (9, 10), using scRNA-seq for calling CNAs is limited to areas of the genome that are expressed at the time of sequencing and does not directly measure genomic copy number. However, single cell whole genome DNA sequencing data promises to circumvent these problems, despite the inherent noisiness of the data.

The main components of CNA calling are detecting contiguous regions of the genome with the same ploidy, called segments, and determining the absolute copy number, or ploidy, of each segment. Previous approaches to calling CNAs using single cells have been based on Hidden Markov Models (HMMs) and change point detection (11). For example, HMMcopy use a Hidden Markov Model to segment tumor genomes using GC and mappability corrected tumor reads, normalized by matched normal cells. Although HMMcopy was originally designed for array comparative genomic hybridization data (12, 13), it's been widely used for single cell sequencing data (11, 13).

CopyNumber was also designed for microarray use, and uses normalized and log transformed copy number measurements rather than raw read counts to detect breakpoints from changes in genome coverage. However, although this method outputs segments, it does not output absolute copy number calls. One strength of CopyNumber, however, is that it can be run in individual and multi sample modes. In the multi sample mode, breakpoints are forced to be shared across all samples (14).

AneuFinder, which was designed for calling CNAs in single cell whole genome sequencing data, uses a trained HMM to model copy number state using a negative binomial distribution (15). In newer versions, Aneufinder uses change point detection analysis to find changes in read coverage (16). To determine absolute copy number, each segment is normalized and scaled such that the mean bin count matches a known ploidy, which is determined from a DNA quantification technique, such as flow cytometry (17). If overall ploidy is not known, a scalar is fit such that all segments get an integer copy number (15).

Ginkgo uses variably sized bins for GC correction and removes outlier "bad" bins based on a fixed set of diploid cells (18), then employs circular binary segmentation (19) to de-

tect breakpoints in normalized read counts and scales ploidy estimates to call absolute copy number. Ginkgo can also cluster cells and build phylogenetic trees (18).

The method SCNV automatically identifies and uses diploid cells as a null error model, and adapts SeqCBS (20) for use in single cells by pooling diploid cells, calibrating model cutoffs using the pooled diploid cells, and discretizing copy number calls (21). SCNV then uses a bin free method based on change point detection on two nonhomogeneous Poisson processes (20) to segment the genome and identify CNAs. This allows for greater resolution of CNAs which might be obscured by choice of bin size boundaries (21).

SCOPE uses a Poisson latent factor model, based on CODEX (22), to normalize read counts, and then uses a log-likelihood ratio test across multiple samples (23) to detect shared breakpoints, with the segmentation stopping rule defined by a cross-sample modified Bayes Information Criterion (24). This allows SCOPE to use cell specific and shared sample information to better estimate technical noise (25).

CHISEL phases SNP haplotypes (26) of fixed size, and uses cell specific read depth ratios and allele specific frequencies to cluster bins across cells in order to call allele specific copy numbers. This allows CHISEL to call CNAs that are aligned with the observed allelic balance, but also makes it prone to errors caused by allelic drop out from low sequencing coverage (27).

SCICoNE corrects read counts for GC and mappability across bins and cells, then uses a likelihood based model to detect breakpoints shared across cells by combining adjacent bins with similar copy number states. SCICoNE then builds a CNA based tree without the infinite sites assumption, allowing for an arbitrary number of CNAs at a site (28). However, the CNA calling procedure precedes and is independent of the tree reconstruction (29).

All of these methods, except for SCNV, require dividing the reference genome into adjacent bins of variable or uniform size. All methods use bin or cell specific GC and mappability corrections to adjust read counts and mask out "bad" bins that exhibit extremely high or low coverage due to centromeres, telomeres, or highly repetitive regions. However, only SCNV and SCOPE utilize detailed bin specific coverage information from diploid cells, and none are based on explicit stochastic models of tumor evolution. An objective of this paper is to develop models for CNA calling based on explicit models of tumor evolution. The rationale is that the use of such explicit models of evolution might improve inferences similarly to what has been observed in models of molecular evolution used in phylogenetics (30–32).

Because tumor cells evolve forward in time from an ancestral diploid state through mutations that only depend on the current state of the cell, copy number alterations are inherently governed by a (possibly time-inhomogeneous) temporal Markov process. However, the read distribution observed along the length of the genome (the spatial process) is not Markovian. To realize this, consider a mutation within a segment of DNA with ploidy 4 that reduces the ploidy from 4 to 3. When moving from the left to the right along the

length of the genome, the ploidy would then go from 4→3→4. There are two transitions (breakpoints) caused by the same single CNA. In many other situations, the rate of mutation from 3→4 (as in the second breakpoint) might be low, however, because the chromosome previously was in state 4, the rate of transition back from 3 to 4 is in fact high in our example. The process along the length of the genome is not Markovian because copy number alterations may have finite length and each mutation may induce two breakpoints.

Even though the spatial process is not Markovian, the HMM framework is computationally convenient. An aim of this paper is, therefore, to develop Markovian approximations of the spatial process that can be used for inference. We present SCONE (Single Cell cOpy Numbers in Cancer), a method based on modeling the temporal Markovian evolutionary process and deriving a best approximating spatial HMM from this process. SCONE also uses diploid data as a null to model the technical noise in single cell sequencing data and can robustly learn model parameters and detect copy number alterations. We show on simulated data that the method more accurately estimates the ploidy states of a cell than previous state-of-the-art methods, and we analyze real data to show that the observations from simulated data are mirrored by similar differences among methods in analyses of real data.

## 2. Theory and Methods

**2.1. Simulations.** In order to robustly evaluate SCONE, we provide two simulation models, one based on line segments and one based on bins. In particular, the line segment model simulates the evolutionary process behind CNAs without assuming any bins, but treating the genome as a line segment. The binned model divides the genome into discrete bins when simulating the evolutionary process. We consider the line segment process to be the more realistic evolutionary model. Of note, the provided simulation models are derived differently from the Hidden Markov Model, described in Section 2.3. We simulate data and estimate parameters and copy number calls under different models, to avoid biasing method comparisons towards our method. See Supplement S1 for full simulation details.

**2.2. Simulation Datasets.** We simulated 4 datasets under the line segment model and 6 datasets under the binned model in order to generate a variety of types and quantity of copy number events. Specifically, each dataset had 100 tumor cells and 100 diploid cells, where read counts from diploid cells were averaged together to form the null model. See Supplement S1.4 for full simulation parameter values.

**2.3. Hidden Markov Model.** In order to simultaneously segment the tumor genome and call absolute copy numbers, we use a Hidden Markov Model along the length of the genome. We define the state space of the HMM as the integer tumor ploidy in a given genomic bin, from 0 up to a user specified  $k$  (suggested  $k = 10$ ), and the alphabet as the integer observed tumor read depth in that bin.

We model emission probabilities for tumor read counts per bin with a negative binomial distribution (interpreted here as an overdispersed Poisson). We incorporate the mean diploid read count for each bin into the emission probabilities, in order to normalize for technical noise and sequencing bias. Let the tumor read depth in window  $i$  for tumor cell  $A$  be represented by random variable  $X_{iA}$ , such that

$$\mathbb{E}(X_{iA}) = \lambda_{iA} = \left( \rho_{iA} \times \frac{\mu_i}{2} \right) \times s_A + \varepsilon$$

$$X_{iA} \sim \text{NegBinom}(\lambda_{iA}, \sigma_{iA}^2 = a\lambda_{iA}^2 + b\lambda_{iA} + c)$$

where  $\rho_{iA}$  is the ploidy in window  $i$  for cell  $A$ ,  $\mu_i$  is the mean diploid read depth in window  $i$ ,  $\varepsilon$  is a constant sequencing error term,  $s_A$  is a cell specific library size scaling factor, and  $\{a, b, c\}$  are constants learned from diploid data, such that the emission probability for an observed read depth,  $x_{iA}$ , is given by the specified negative binomial distribution. See [Library Size Scaling Factors](#) for  $s_A$  calculations and [Negative Binomial Mean and Variance Calculations](#) for  $\{a, b, c\}$  calculations.

For the HHM, the initial probability vector is defined as the steady state distribution of the Markov chain. The log-likelihood of the observed tumor data is calculated using the forward algorithm and summed across all chromosomes for a given cell. The HMM is reset to the initial probability vector at the beginning of each chromosome to maintain chromosomal independence.

**2.4. Joint evolutionary process process of two bins forward in time.** In Supplement S1, we described two principled models of CNA evolution. However, neither of these models have the property that they are Markovian along the length of the genome. To construct an approximating process that is Markovian, we will first construct a process affecting two bins. This process will effectively be similar to the described binned process, but it is parameterized slightly differently out of convenience. From this description of the joint evolution of two bins, we will then derive the approximating Markov process used for HMM inference of copy number state.

Consider two adjacent bins in the genome on one lineage,  $(U, V) \in \{(0, 0), (0, 1), \dots, (k, k)\}$ , where  $U$  is the ploidy in bin  $i$ , and  $V$  is the ploidy in bin  $i + 1$ . The ploidies in these bins change through evolutionary history according to rate parameters  $\{\alpha, \beta, \gamma\}$ :

$$\begin{aligned} \alpha &= \text{rate of } \pm 1 \text{ CNA} \\ \beta &= \text{rate of any CNA} \\ \gamma &= \text{rate ratio of CNAs affecting both } U \text{ and } V \end{aligned}$$

These rates are encoded in a transition rate matrix  $\mathbb{Q}$ :

$$q_{(U,V),(U',V')} = \begin{cases} \gamma(\alpha + \beta) & \text{if } (U', V') = \begin{cases} (U + n, V + n) \\ (U - n, V - n) \end{cases}, n = 1 \\ \gamma\beta & \text{if } (U', V') = \begin{cases} (U + n, V + n) \\ (U - n, V - n) \end{cases}, n > 1 \\ \alpha + \beta & \text{if } (U', V') = \begin{cases} (U + n, V) \\ (U - n, V) \\ (U, V + n) \\ (U, V - n) \end{cases}, n = 1 \\ \beta & \text{if } (U', V') = \begin{cases} (U + n, V) \\ (U - n, V) \\ (U, V + n) \\ (U, V - n) \end{cases}, n > 1 \\ 0 & \text{otherwise} \end{cases}$$

From this rate matrix  $\mathbb{Q}$ , the time dependent transition probabilities  $\mathbb{P}$  are calculated via the matrix exponential as

$$P_{(U,V),(U',V')}(t) = e^{\mathbb{Q}t}$$

This gives the probability of observing a transition from  $(U, V)$  to  $(U', V')$  in time  $t$ .

**2.5. Discrete process (Markovian approximation) along the genome.** We convert the forward-in-time process for two bins into a Markov model along the length of the genome with transition probability matrix  $\mathbb{M}_t = \{m_{i,i',t}\}$ ,  $i, i' \in \mathbb{S}$ , i.e. we identify the probability of moving from state  $i$  to  $i'$  along the genome, after a given evolutionary time  $t$ . Under the assumption that the cell has an ancestral diploid state at time  $t = 0$ , we set  $(U, V) = (2, 2)$  and  $(U', V') = (i, i')$ . By normalizing over all states  $W$  in  $\mathbb{S}$ , the one-step transition probabilities of the discrete approximating Markov process along the length of the genome are given by

$$m_{i,i',t} = \frac{P_{(2,2),(i,i')}(t)}{\sum_{W \in \mathbb{S}} P_{(2,2),(i,W)}(t)}$$

This time dependent transition matrix approximates a non-Markovian process using an evolutionary time-informed HMM. The advantage of using this model over more generic HMMs is that information about the ancestral diploid state is included in the model specification allowing more accurate inference of ploidy state. While the model is only an approximation, as it ignores the non-Markovian nature of any realistic model of CNA changes along the genome, we will evaluate it using the aforementioned non-Markovian simulation models.

**2.6. Model Training.** The model training has four steps. We first estimate the constants,  $\{a, b, c\}$ , used to model the emission probabilities, from the diploid data. Second, for each tumor cell,  $A$ , we quickly estimate an unconstrained transition

matrix, initial probability vector, and library size scaling factor,  $s_A$ , using a modification of the Baum Welch algorithm. Third, the model rate parameters,  $\{\alpha_A, \beta_A, \gamma_A, t_A\}$ , are then fit to the estimated transition matrix using least squares. Fourth, the initial estimates for  $\{s_A, \alpha_A, \beta_A, \gamma_A, t_A\}$  are refined using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithm to maximize the forward loglikelihood of the observed tumor read depths. See Supplement S2 for full model training details.

**2.7. Real Data preprocessing.** We applied SCONE to a published dataset, consisting of 34 diploid cells (as determined by cell sorting), and 4 tumor subpopulations (24, 24, 4, and 8 cells, respectively) from one triple negative breast cancer patient (33), a cancer type with prevalent CNAs (34).

We first applied standard preprocessing and quality control steps to the sequencing data: trimming adapters and low quality read ends (35, 36), removing low complexity and short reads (37), and removing PCR duplicates (38). After cleaning up the sequencing data, reads were aligned to the reference genome (hg19) using bowtie2 (39), and reads with q scores less than 20 were removed (40).

The reference genome was binned into uniformly sized bins, and cell specific read depth was counted in each dataset using bedtools (41). Per window read depth was averaged across diploid cells, and the  $\{a, b, c\}$  constants for the Negative Binomial distribution were calculated (see Supplement S2.1 for full details).

**2.8. Other methods.** In order to evaluate the accuracy of the inference procedure, we compared to HMMcopy (12, 13), CopyNumber (14), and AneuFinder (15, 16). We limited our comparison to methods that do not require bam files or SNPs, as our simulation model does not create bam files or model SNPs for simplicity. For both real and simulated datasets, we used the averaged diploid cells as the matched normal sample to determine the somatic copy number for each tumor cell. To run HMMcopy, the `HMMsegment` function (default parameters) was used to segment each cell, and copy numbers were extracted from the resulting `state` element – 1.

The normalized and log-transformed copy number estimates from HMMcopy were used as input for CopyNumber. Then, missing data were imputed, using the `constant` method, and the `winsorize` function was used to remove outliers. To run in single sample mode, the `pcf` function was used with parameters `return.est=T`, `normalize=T`, `digits=6`, and the exponentiated estimates element was extracted for the copy number estimates. To run in multi sample mode, the `multipcf` function was run with parameters `return.est=T`, `digits=6`, and copy numbers were similarly extracted from the exponentiated estimates result. Additionally, because CopyNumber does not output absolute copy number calls, we scaled CopyNumber results to minimize the sum of squared differences from the true ploidy in simulated datasets to create a ploidy estimate for comparison purposes.

The procedure for running AneuFinder differed slightly between simulations and real data. See [Scripts to run other](#)

[methods](#) for full scripts.

**2.8.1. Simulations.** Because our simulation model does not incorporate GC or mappability into simulated read depth, we did not use the GC and mappability corrections in HMMcopy (and subsequently CopyNumber) in order to avoid over-correcting. We used the averaged diploid read counts for the matched normal sample to detect somatic CNAs only. Similarly, in AneuFinder, we skipped the GC and mappability corrections by running the `findCNVs` function with default parameters (`method="edivisive"`, `R=10`, `sig.lvl=0.1`), and extracting the `copy.number` element from the model segments.

**2.8.2. Real Data.** With real data, we ran HMMcopy by first doing read correction (`correctReadcount`, default parameters) on both the tumor data and averaged diploid cells, then ran as described above.

To run AneuFinder, we ran the `Aneufinder` function with 250,000 binsize, all chromosomes, GC correction, and hg19 assembly. As in simulations, copy number calls were extracted from the `copy.number` element from the `edivisive` model segments.

### 3. Results

**3.1. GC content and mappability.** Because GC content and sequence mappability can bias read distributions, many methods explicitly incorporate corrections for GC content and sequence mappability. However, any technical noise that would affect the tumor sequencing would also affect the diploid sequencing, so in SCONE, these corrections are already directly accounted for in our emission probabilities via the diploid mean.

To verify this, we examined prediction accuracy of expected tumor read counts per window with different amounts of information. For window  $i$ , let  $\mu_i$  be the mean diploid read count,  $\zeta_i$  be the GC content, and  $\eta_i$  be the mappability from the Duke Uniqueness of 35bp Windows from ENCODE/OpenChrom (UCSC accession wgEncodeEH000325) (42, 43). For each tumor cell,  $A$ , from the previously published data in (33), we predicted the  $i$ th window tumor read depth,  $x_{iA}$ , using various linear regressions on  $\{\mu_i, \zeta_i, \eta_i\}$ , then calculated the sum of squared differences between predicted and actual tumor read depths. Boxplots of the summed squared differences per cell are shown in Figure 1 and empirical cumulative distribution function (ECDF) plots are shown in Supplemental Figure S1 for A:  $x_{iA} \sim \mu_i$ , B:  $x_{iA} \sim \mu_i + \zeta_i$ , C:  $x_{iA} \sim \mu_i + \eta_i$ , D:  $x_{iA} \sim \mu_i + \zeta_i + \eta_i$ , E:  $x_{iA} \sim \zeta_i$ , F:  $x_{iA} \sim \eta_i$ , G:  $x_{iA} \sim \zeta_i + \eta_i$ .

The sum of squared differences remains consistent across models that incorporate the diploid mean (models A, B, C, and D), and have overlapping ECDF plots, while the sum of squared differences increases for models that depend solely on GC content and mappability (models E, F, and G). Because adding the GC content and mappability did not perform significantly differently from the diploid mean alone (two sample KS-test on the cumulative distribution of summed

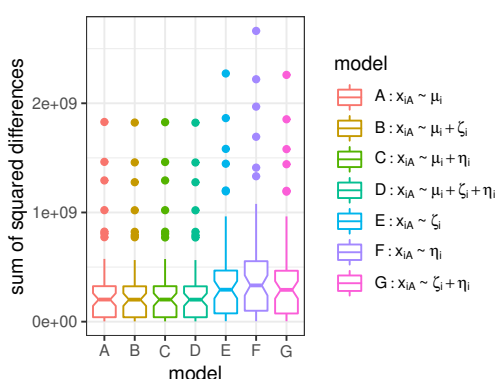


Figure 1: For each linear regression, a boxplot of the sum of squared differences between the predicted read count and observed read count for each tumor cell in (33) (uniformly sized 250kb bins) is shown. No statistically significant difference in error is observed by adding GC or mappability information to the diploid null model.

squared differences,  $D = 0.033333$ ,  $p\text{-value} = 1$ ), we conclude using the diploid mean is sufficient, and do not add GC or mappability corrections. This conclusion is robust to changes in window size and binning method (ie. uniformly sized bins vs variably sized bins with equal numbers of uniquely mappable bases).

**3.2. Error rates.** To compare the accuracy of each copy number calling method, we compared the sum of squared differences (SSD) between true copy number and estimated copy number across ten simulation datasets. Recall that these datasets were simulated under a more general non-Markovian model (see Supplement S1 for simulation details and Supplement S1.4 for parameter values).

For each cell, the SSD was calculated across all 12,397 windows (number of uniform non-overlapping 250kb windows in hg19). Overall, SCONE has similar or lower error rates than AneuFinder, and consistently significantly lower error rate than HMMcopy and CopyNumber (see Figure 2).

For example, in Simulation Set A (consisting of many small overlapping CNAs per cell, under the line segment model; Figure 2A), the median SSD for SCONE is 3493.41 and 67.59, for  $k = 5$  and  $k = 10, 15$ , respectively, which is lower than the median SSD for AneuFinder, at 103.32. Meanwhile, the median SSD for CopyNumber (in multisample and individual modes, respectively) was 5312.37 and 5286.02, and the median SSD for HMMcopy (which does not output absolute copy number calls, and so was optimally scaled) was 26779.47. Of note, because SCONE cannot call ploidies above the user specified  $k$ , its error rate is significantly higher for  $k = 5$  when the true simulated ploidy is greater than 5.

Scaling problems can also arise if  $k$  is set too low. For Simulation Set I (consisting of very short spiky CNAs, under the binned model; Figure 2I), the median SSD for SCONE for  $k = 5$  is 1917.50, while the median SSD for  $k = 10, 15$  drops to 101.50. The median SSD for AneuFinder is over three times worse at 352.00, while the median SSDs for HMMcopy and CopyNumber (multisample and individual modes) are orders of magnitude worse, at 4727.50, 5715.67, and 5595.00, respectively. In both of these simulation sets, despite the higher median SSD for SCONE at  $k = 5$ , the

median SSD consistently drops for  $k = 10, 15$ . Because higher values of  $k$  result in a higher run times without significant gain in accuracy, we recommend setting  $k = 10$ .

In other simulations, AneuFinder has scaling problems that SCONE does not. In Simulation Set C (consisting of mainly deletions, under the line segment model; Figure 2C), the median SSD for SCONE is 3.84 and 3.9 for  $k = 5$  and  $k = 10, 15$ . The median SSD was 6384.32, 1845.24, and 1822.80 for HMMcopy and CopyNumber (multisample and individual modes), respectively. However, the median SSD for AneuFinder is orders of magnitude higher, at 21726.87. Upon closer inspection, AneuFinder incorrectly doubles the ploidy for the majority of the cells in this simulation set (see Supplemental Figure S3C for an example decoding and Supplemental Figure S2C).

To check if the differences in median SSD between methods were due to scaling issues, we also rescaled all copy number calls to minimize the SSD between simulated ploidy and estimated ploidy for all methods. With this optimal rescaling, SCONE consistently outperforms or is on par with other methods (see Supplemental Figure S2).

Although the median SSD for SCONE with  $k = 5$  in Simulation Set A decreases from 3493.81 to 2900.76, rescaling does not address the underlying limitation of  $k$  being too small. The median SSDs for the other methods for Simulation Set A also decrease, but not significantly (see Supplemental Figure S2A). Similarly, under Simulation Set I, fixing the incorrect scaling for SCONE with  $k = 5$  causes the median SSD to drop from 1917.50 to 1649.83, but it doesn't address the root problem of  $k$  limiting the ploidies SCONE can call.

In contrast, the median SSD for AneuFinder for Simulation Set C drops significantly from 21726.87 to 4.33, while the median SSD for SCONE remained constant. This shows AneuFinder's high median SSD for Simulation Set C was due to incorrect scaling, rather than incorrect breakpoint detection and segmentation.

However, although HMMcopy median SSD values decreased with optimal scaling, they never dropped into the same range as SCONE and AneuFinder, implying there are non-scaling related reasons behind the high median SSD values. The median SSD values for CopyNumber did not change, as its output was already scaled because it does not report absolute ploidies.

**3.3. Genome wide decodings.** By plotting the genome wide copy number profile for a representative cell from each simulation set, we can learn more about the specific differences between methods that lead to differing error rates. Of note, the value of  $k$  must be set high enough to allow a wide enough ploidy range in SCONE (suggested  $k = 10$ ). For brevity, only genome decodings for SCONE (with  $k = 10$ ) and AneuFinder are shown in the main text (see Supplemental Figure S3 for decodings with other programs and other values of  $k$  for SCONE).

In some cases where the maximum  $k$  is set too low, the error rate from SCONE is high because it can't estimate high

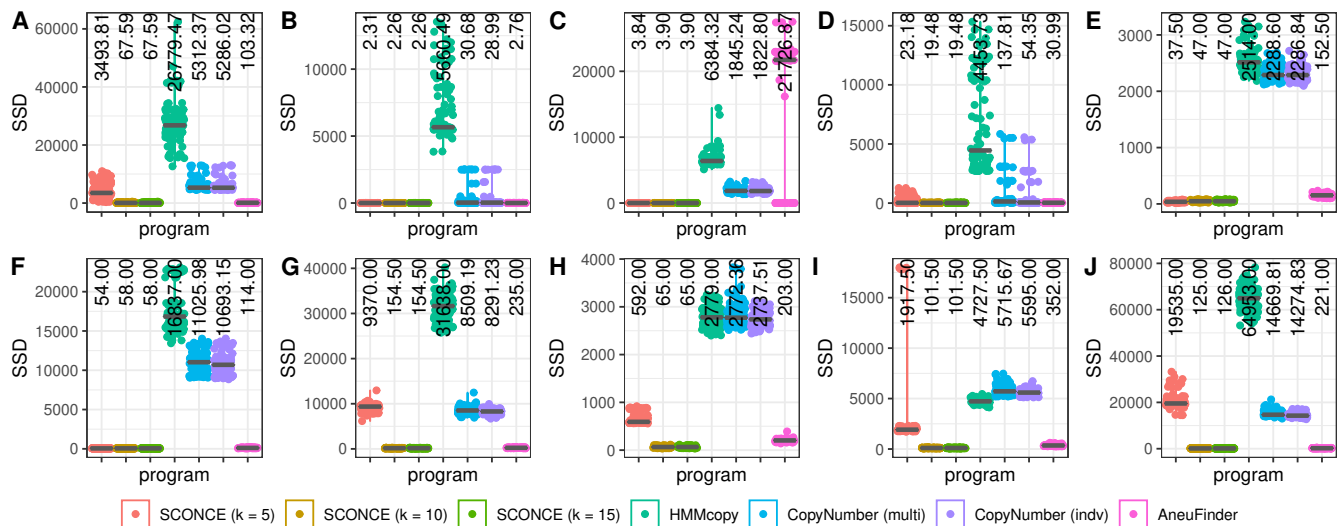


Figure 2: For each method, the sum of squared differences (SSD) between simulated ploidy and estimated ploidy is shown across different parameter sets. Each dot represents the error for one cell and the median SSD is shown with a gray line and printed at the top of each column. SCONCE consistently has SSD values that are lower or on par with other methods.

enough ploidies (Simulation Set A, with many small overlapping CNAs; Figure 2A, Supplemental Figure S3A). This can be seen in chromosome 3, where the true ploidy reaches a maximum of 8, but SCONCE's ploidy estimates are limited to  $k$ .

In other cases, setting  $k$  too low causes the library size scaling factor to be estimated incorrectly (Simulation Set I, with very short spiky CNAs, Figure 2I, 3A, Supplemental Figure S3B). Specifically, for  $k = 5$ , SCONCE incorrectly reports ploidy of 1 instead of 2 for most of the genome. However, once the value of  $k$  is high enough ( $k = 10, 15$ ), SCONCE consistently recovers the simulated ploidy with a lower error rate than AneuFinder. In particular, AneuFinder misses small CNAs (ranging from 1 to 5 bins in width), that SCONCE does not miss, such as in chromosomes 9 and 17 (shown with arrows in Figure 3A). These results are consistent across simulations with approximately equal rates of insertions and deletions (Supplemental Figure S3A, S3B) and in simulations with mostly insertions (Supplemental Figure S3D).

Additionally, in simulations with mostly deletions (Simulation Set C, under the line segment model), AneuFinder consistently and incorrectly doubles the estimated ploidy, leading to a high error rate, while SCONCE does not (Figure 2C, Supplemental Figure S3C). Specifically, AneuFinder mainly calls ploidies of  $\{0, 2, 4\}$ , instead of  $\{0, 1, 2\}$ . When we optimally scaled all copy number estimates to minimize the SSD, AneuFinder's error rates dropped, thereby verifying the existence of a scaling problem. Even with this optimal scaling, SCONCE continued to have lower error rates than other methods (Supplemental Figure S2).

Furthermore, SCONCE considerably outperforms methods like HMMcopy and CopyNumber in regions of 0 read coverage. By using the diploid null model, we are able to separate between true deletions and areas that have missing data due to sequencing noise (Simulation Set A, with many small overlapping CNAs, Supplemental Figure S3A;

Simulation Set C, with mostly deletions, Supplemental Figure S3C). We note that this problem observed in the real data was not contributing to the performance of HMMcopy and CopyNumber in the simulated data, as no regions with missing data were simulated.

For real data (33), copy number estimates from a representative cell (SRR054570) from SCONCE (with  $k = 10$ ) and AneuFinder are shown in Figure 3B (see Supplemental Figure S4 for copy number estimates from each method for cell SRR054570 and another representative cell, SRR053675). Of note, because we specifically incorporate diploid data as our null model, SCONCE makes the most parsimonious calls, rather than assuming copy number 0, in regions that are hard to sequence or map and have no diploid data. For example, in regions around centromeres and telomeres, AneuFinder often calls 0 ploidy when there's no observed diploid or tumor reads. However, SCONCE uses the lack of diploid and tumor reads to predict no change in copy number. This can be clearly seen in Figure 3B in the centromeres of chromosomes 1, 9, and 16, and in the telomeres of chromosomes 13, 14, 15, 21, and 22. Additionally, by examining Supplemental Figure S4B, small CNAs (between 5 and 22 250kb windows in length, on chromosomes 9, 10, 12, 13, and 18) are consistently missed by AneuFinder, while SCONCE calls these. CNAs larger than 99 windows in length, however, are consistently called well by both SCONCE and AneuFinder. These results recapitulate the results seen from simulations.

## 4. Discussion

CNAs are an important driver in cancer evolution, and accurately detecting them on a single cell level can deepen our understanding of tumorigenesis. In this paper, we derive several models of copy number alterations for inference and simulation. We show that using HMMs derived from models of the evolutionary process that generate CNAs,

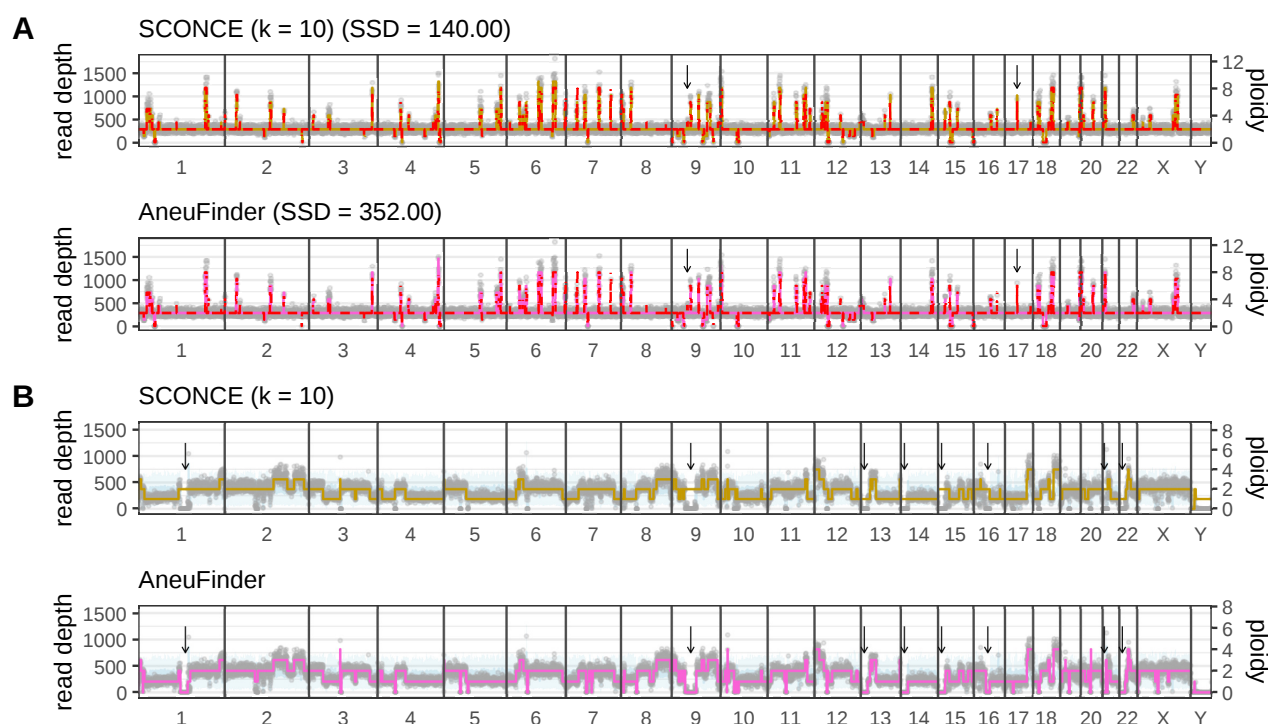


Figure 3: Genome wide copy number decodings are shown for representative cells from simulations and real data. Simulation Set I (very short spiky CNAs under the binned simulation model) is shown in panel A, and cell SRR054570 from (33) is shown in panel B. Genomic window is plotted along the x-axis, per window read depth is shown along the left y-axis, and ploidy is plotted along the right y-axis. Black vertical lines denote chromosome boundaries, gray dots represent observed tumor read depth in each window, the red dotted line denotes the true ploidy from simulation (where applicable), the light blue line shows the mean diploid read count, the light blue band shows  $\pm 1$  standard deviation in the diploid read count, and the colored lines denote the copy number decoding from each method. Black arrows highlight regions with differences in CNA calls between SCONE and AneuFinder. In panel A, small CNAs in chromosomes 9 and 17 are called by SCONE, but not by AneuFinder. In panel B, centromeres in chromosomes 1, 9, and 16 and telomeres in chromosomes 13-15, 21, and 22 are called with ploidy 0 by AneuFinder, but SCONE makes the more parsimonious call of ploidy 2.

more accurate inferences of CNA could be obtained. The method for inference based on these models, SCONE, is available as an open source computer package at <https://github.com/NielsenBerkeleyLab/scone>.

One limitation of SCONE is that it requires data from diploid cells sequenced on the same platform as the tumor cells. While this increases accuracy by accounting for platform specific biases and single cell sequencing errors, it also increases sequencing costs to sequence diploid cells, which may not be directly of interest to investigators. Alternatively, as in the (33) dataset and in other methods (27) utilizing other datasets, sequenced tumor cells that are determined to be diploid by other means (such as via cell sorting) can be relabeled as diploid cells.

One of the key strengths of SCONE over competing methods is its principled Markovian approximation of a non-Markovian process. This allows for future interpretations and applications of model parameters to understand tumor evolution. Specifically, SCONE learns transition rate parameters  $\{\alpha, \beta, \gamma\}$ , tree branch length  $t$ , and library size scaling factors, but these values are not used directly outside of the copy number profile decoding. Understanding these transition rates in the context of using these tree branch lengths to build phylogenies is the subject of future work.

Compared to other methods, SCONE has increased sensitivity in calling very small CNAs, particularly those smaller than 5500kb. Additionally, in cells with substantial copy

number losses, SCONE can accurately create copy number profiles without erroneous ploidy doublings. This is due to SCONE's method of estimating library sizes using the Viterbi decoding to account for how changes in the copy number profile necessarily impact the library scaling factor.

Furthermore, because SCONE uses the averaged diploid data as a null model, in regions with zero tumor read coverage, it can differentiate between genomic loss and sequencing noise, which other methods can not do. In particular, in regions with diploid coverage but no tumor reads, SCONE calls 0 ploidy, and in regions without coverage in either the diploid cells or the tumor cell, SCONE makes the most parsimonious call. This increases CNA calling accuracy of hard to sequence regions, such as telomeres, centromeres, and repetitive regions.

In conclusion, we present an accurate and principled evolutionary model for calling copy number alterations in single cell whole genome sequencing of tumors, with implications for broader applications.

## 5. Acknowledgements

**5.1. Funding.** This work was supported by the National Institutes of Health [R01GM138634-01 to R.N.].

## 6. Code Availability

SCONCE is implemented in C++11 and is freely available from <https://github.com/NielsenBerkeleyLab/sconce>. See Supplement S6 for full details.

## 7. Bibliography

- Rameen Beroukhi, Craig H. Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S. Boehm, Jennifer Dobson, Mitsuyoshi Urashima, Kevin T. Mc Henry, Reid M. Pinchback, Azra H. Ligon, Yoon-Jae Cho, Leila Haery, Heidi Greulich, Michael Reich, Wendy Winckler, Michael S. Lawrence, Barbara A. Weir, Kumiko E. Tanaka, Derek Y. Chiang, Adam J. Bass, Alice Loo, Carter Hoffman, John Prensner, Ted Liefeld, Qing Gao, Derek Yecies, Sabina Signoretti, Elizabeth Maher, Frederic J. Kaye, Hidefumi Sasaki, Joel E. Tepper, Jonathan A. Fletcher, Josep Taberner, José Baselga, Ming-Sound Tsao, Francesca Demicheli, Mark A. Rubin, Pasi A. Janne, Mark J. Daly, Carmelo Nucera, Ross L. Levine, Benjamin L. Ebert, Stacey Gabriel, Anil K. Rustgi, Cristina R. Antonescu, Marc Ladanyi, Anthony Letai, Levi A. Garraway, Massimo Loda, David G. Beer, Lawrence D. True, Aikou Okamoto, Scott L. Pomeroy, Samuel Singer, Todd R. Golub, Eric S. Lander, Gad Getz, William R. Sellers, and Matthew Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature* 2010 463:7283, 463(7283):899–905, 2 2010. ISSN 1476-4687. doi: 10.1038/nature08822.
- Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C. Drento, Santiago Gonzalez, Daniel Rosebrock, Thomas J. Mitchell, Yulia Rubanova, Pavana Anur, Kaixian Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kertine Kleinheinz, Ignacio Vázquez-García, Kerstin Haase, Lara Jerman, Subhaji Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G. Livitz, Marek Cmero, Jonas Demeulemeester, Steven Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C. Boutros, David D. Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhi, S. Cenik Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quid D. Morris, Paul T. Spellman, David C. Wedge, and Peter Van Loo. The evolutionary history of 2,658 cancers. *Nature* 2020 578:7793, 578(7793):122–128, 2 2020. ISSN 1476-4687. doi: 10.1038/s41586-019-1907-7.
- Adriana Salcedo, Maxime Tarabichi, Shadielle Melijah G. Espiritu, Amit G. Deshwar, Matei David, Nathan M. Wilson, Stefan Drento, Jeff A. Wintersinger, Lydia Y. Liu, Minjeong Ko, Srinivasan Sivanandan, Hongjiu Zhang, Kaiyi Zhu, Tai-Hsien Ou Yang, John M. Chilton, Alex Buchanan, Christopher M. Lalansingh, Christine P'ng, Catalina V. Anghel, Imaad Umar, Bryan Lo, William Zou, Jared T. Simpson, Joshua M. Stuart, Dimitris Anastassiou, Yuanfang Guan, Adam D. Ewing, Kyle Elliott, David C. Wedge, Quid Morris, Peter Van Loo, and Paul C. Boutros. A community effort to create standards for evaluating tumor subclonal reconstruction. *Nature Biotechnology* 2020 38:1, 38(1):97–107, 1 2020. ISSN 1546-1696. doi: 10.1038/s41587-019-0364-z.
- Johannes Smolander, Sofia Khan, Kalaimathy Singaravelu, Leni Kauko, Riikka J. Lund, Asta Laiho, and Laura L. Elo. Evaluation of tools for identifying large copy number variations from ultra-low-coverage whole-genome sequencing data. *BMC Genomics* 2021 22:1, 22(1):1–15, 5 2021. ISSN 1471-2164. doi: 10.1186/s12864-021-07686-Z.
- Jos B Poell, Matias Mendeville, Daoud Sie, Arjen Brink, Ruud H Brakenhoff, and Bauke Ylstra. ACE: absolute copy number estimation from low-coverage whole-genome sequencing data. *Bioinformatics*, 35(16):2847–2849, 8 2019. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTY1055.
- Yao Xiao, Xueqing Wang, Hongjiu Zhang, Peter J. Ulitz, Hongyang Li, and Yuanfang Guan. BulkClone is a probabilistic tool for deconvoluting tumor heterogeneity in fast-sequencing samples. *Nature Communications* 2020 11:1, 11(1):1–11, 9 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18169-2.
- Yukie Kashima, Yoshitaka Sakamoto, Keiya Kaneko, Masahide Seki, Yutaka Suzuki, and Ayako Suzuki. Single-cell sequencing techniques from individual to multiomics analyses. *Experimental & Molecular Medicine*, 52(9), 9 2020. ISSN 1226-3613. doi: 10.1038/s12276-020-00499-2.
- Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev, and Bradley E. Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 6 2014. ISSN 0036-8075. doi: 10.1126/SCIENCE.1254257.
- Itay Tirosh and Mario L. Suvà. Deciphering Human Tumor Biology by Single-Cell Expression Profiling. <https://doi.org/10.1146/annurev-cancerbio-030518-055609>, 3(1):151–166, 3 2019. doi: 10.1146/ANNUREV-CANCERBIO-030518-055609.
- Mario L. Suvà and Itay Tirosh. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Molecular Cell*, 75(1):7–12, 7 2019. ISSN 1097-2765. doi: 10.1016/j.molcel.2019.05.003.
- Xian F. Mallory, Mohammadamin Edrisi, Nicholas Navin, and Luay Nakhleh. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biology*, 21(1):208, 12 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02119-8.
- Sohrab P. Shah, Xiang Xuan, Ron J. DeLeeuw, Mehrnosh Khojasteh, Wan L. Lam, Raymond Ng, and Kevin P. Murphy. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, 22(14):e431–e439, 7 2006. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTL238.
- Daniel Lai, Gavin Ha, and Sohrab Shah. HMMcopy: Copy number prediction with correction for GC and mappability bias for HTS data, 2019.
- Gro Nilsen, Knut Liestøl, Peter Van Loo, Hans Kristian Moen Vollan, Marianne B Eide, Oscar M Rueda, Suet-Feung Chin, Roslin Russell, Lars O Baumbusch, Carlos Caldas, Anne-Lise Borresen-Dale, and Ole Christian Lingjærde. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* 2012 13:1, 13(1):1–16, 11 2012. ISSN 1471-2164. doi: 10.1186/1471-2164-13-591.
- Björn Bakker, Aaron Taudt, Mirjam E. Belderbos, David Porubsky, Diana C.J. J. Spierings, Tristan V. de Jong, Nancy Halsema, Hinke G. Kazemier, Karina Hoekstra-Wakker, Allan Bradley, Eveline S.J.M. J. M. de Bont, Anke van den Berg, Victor Guryev, Peter M. Lansdorp, Maria Colomé-Tatché, and Floris Fojter. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biology*, 17(1):115, 12 2016. ISSN 1474760X. doi: 10.1186/s13059-016-0971-7.
- Aaron Sebastian Taudt. *Hidden Markov models for the analysis of next-generation sequencing data*. PhD thesis, University of Groningen, Groningen, 10 2018.
- Ruli Gao, Alexander Davis, Thomas O McDonald, Emi Sei, Xiuqing Shi, Yong Wang, Pei-Ching Tsai, Anna Casasent, Jill Waters, Hong Zhang, Funda Meric-Bernstam, Franziska Michor, and Nicholas E Navin. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature Genetics*, 48(10):1119–1130, 10 2016. ISSN 1061-4036. doi: 10.1038/ng.3641.
- Tyler Garvin, Robert Aboukhalil, Jude Kendall, Timour Baslan, Gurinder S Atwal, James Hicks, Michael Wigler, and Michael C Schatz. Interactive analysis and assessment of single-cell copy-number variations. *Nature Methods*, 12(11):1058–1060, 11 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3578.
- A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 10 2004. ISSN 1465-4644. doi: 10.1093/biostatistics/kxh008.
- Jeremy J. Shen and Nancy R. Zhang. Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. <https://doi.org/10.1214/11-AOAS517>, 6(2):476–496, 6 2012. ISSN 1932-6157. doi: 10.1214/11-AOAS517.
- Xuefeng Wang, Hao Chen, and Nancy R Zhang. DNA copy number profiling using single-cell sequencing. *Briefings in Bioinformatics*, 19(5):731–736, 9 2018. ISSN 1467-5463. doi: 10.1093/BIB/BBX004.
- Yuchao Jiang, Derek A. Oldridge, Sharon J. Diskin, and Nancy R. Zhang. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Research*, 43(6):e39–e39, 3 2015. ISSN 0305-1048. doi: 10.1093/NAR/GKU1363.
- Nancy R. Zhang, David O. Siegmund, Hanlie Ji, and Jun Z. Li. Detecting simultaneous change-points in multiple sequences. *Biometrika*, 97(3):631–645, 9 2010. ISSN 0006-3444. doi: 10.1093/BIOMET/ASQ025.
- Nancy R Zhang and David O Siegmund. Model selection for high dimensional multi-sequence change-point problems. *Statistica Sinica*, 22:1507–1538, 2012. doi: 10.5707/ss.2010.257.
- Ruijin Wang, Dan Yu Lin, and Yuchao Jiang. SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing. *Cell Systems*, 10(5):445–452, 5 2020. ISSN 2405-4712. doi: 10.1016/j.cels.2020.03.005.
- Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, Richard Durbin, and Alkes L Price. Reference-based phasing using the Haplotypes Reference Consortium panel. *Nature Genetics* 2016 48:11, 48(11):1443–1448, 10 2016. ISSN 1546-1718. doi: 10.1038/ng.3679.
- Simone Zaccaria and Benjamin J. Raphael. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nature Biotechnology*, 39(2):207–214, 2 2021. ISSN 15461696. doi: 10.1038/s41587-020-0661-6.
- Jack Kuipers, Katharina Jahn, Benjamin J. Raphael, and Niko Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Research*, 27(11):1885–1894, 11 2017. ISSN 1088-9051. doi: 10.1101/GR.220707.117.
- Jack Kuipers, Mustafa Anil Tuncel, Pedro Ferreira, Katharina Jahn, and Niko Beerenwinkel. Single-cell copy number calling and event history reconstruction. *bioRxiv*, page 2020.04.28.065755, 4 2020. doi: 10.1101/2020.04.28.065755.
- Joseph Felsenstein. Journal of Molecular Evolution Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *J Mol Evol*, 17:368–376, 1981.
- Ziheng Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular biology and evolution*, 10(6):1396–1401, 1993. ISSN 0737-4038. doi: 10.1093/OXFORDJOURNALS.MOLBEV.A040082.
- Ziheng Yang. Journal of Molecular Evolution Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. *J Mol Evol*, 39:306–314, 1994.
- Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McDoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W. Richard McCombie, James Hicks, and Michael Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 4 2011. ISSN 0028-0836. doi: 10.1038/nature09807.
- Zaibing Li, Xiao Zhang, Chenxin Hou, Yuqing Zhou, Junli Chen, Haoyang Cai, Yifeng Ye, Jiping Liu, and Ning Huang. Comprehensive identification and characterization of somatic copy number alterations in triple-negative breast cancer. *International Journal of Oncology*, 56(2):522–530, 2 2020. ISSN 1019-6439. doi: 10.3892/IJO.2019.4950.
- Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17(1):10–12, 5 2011. doi: http://dx.doi.org/10.14806/ej.17.1.200.
- Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15):2114–20, 8 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu170.
- R. Schmieder and R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 3 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr026.
- The Broad Institute. Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF, 2021.
- Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 4 2012. ISSN 1548-7091. doi: 10.1038/nmeth.1923.
- Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth,

- Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, 8 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp352.
41. Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2, 3 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq033.
  42. Ian Dunham, Anshul Kundaje, Shelley F Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Fretz, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shores, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Pouya Kheradpour, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C.J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A.L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Peter J. Good, Elise A. Feingold, Gregory E. Crawford, Job Dekker, Laura Elitski, Peggy J. Farnham, Morgan C. Giddings, Thomas R. Gingeras, Roderic Guigó, Timothy J. Hubbard, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Matthew L. Eaton, Alex Dobin, Andrea Tanzer, Julien Lagarde, Wei Lin, Chenghai Xue, Brian A. Williams, Chris Zaleski, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakraborty, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Daniel Robyr, Xiaolan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaian Wang, Yoshihide Hayashizaki, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Yijun Ruan, Piero Carninci, Cricket A. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Linda L. Grasfeder, Paul G. Giresi, Anna Battenhouse, Nathan C. Sheffield, Kimberly A. Showers, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniel, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Vishwanath R. Iyer, Meizhen Zheng, Ping Wang, Jason Gertz, Jost Vielmetter, E. Christopher Partridge, Katherine E. Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M. Bowling, Michael Anaya, Marie K. Cross, Michael A. Muratet, Kimberly M. Newberry, Kenneth McCue, Amy S. Nesmith, Katherine I. Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Stephanie L. Parker, Sreeram Balasubramanian, Nicholas S. Davis, Sarah K. Meadows, Tracy Eggleston, J. Scott Newberry, Shawn E. Levy, Devin M. Absher, Wing H. Wong, Matthew J. Blow, Axel Visel, Len A. Pennachio, Hanna M. Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Giovanni Bussotti, Claire Davidson, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David A. Hendrix, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Jing Leng, Michael F. Lin, Jane Loveland, Zhi Lu, Deepa Manthavadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sponer, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Electra Tapanari, Michael L. Tress, Marijke J. Van Baren, Stefan Washietl, Laurens Wilming, Amonida Zadiisa, Zhengdong Zhang, Michael Brent, David Haussler, Alfonso Valencia, Nick Addelman, Roger P. Alexander, Raymond K. Auerbach, Suganthi Balasubramanian, Keith Bettinger, Nitin Bhardwaj, Alan P. Boyle, Alina R. Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D. Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Sushma Iyengar, Victor X. Jin, Konrad J. Karczewski, Maya Kawsowski, Phil Lacroute, Hugo Lam, Nathan Lamarre-Vincent, Jin Lian, Marianne Lindahl-Allen, Renqiang Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Ximeng J. Mu, Henriette O'Geen, Zhengqing Ouyang, Dorrelyn Patacsil, Debasish Raha, Lucia Ramirez, Brian Reed, Minyi Shi, Teri Slifer, Heather Witt, Linfeng Wu, Xiaolin Xu, Koon Kiu Yan, Xinqiong Yang, Kevin Struhl, Sherman M. Weissman, Luiz O. Penalva, Subhradip Kar-makar, Raj R. Bharnvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Alec Victorson, Thomas Auer, Lazaro Centanin, Michael Eichenlaub, Franziska Gruhl, Stephan Heermann, Burkhard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate Wittbrodt, Jochen Wittbrodt, Gaurav Jain, Gayathri Balasundaram, Daniel L. Bates, Rachel Byron, Theresa K. Canfield, Morgan J. Diegel, Douglas Dunn, Abigail K. Ebersol, Tristan Frum, Kavita Garg, Erica Gist, R. Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Audra K. Johnson, Ericka M. Johnson, Tattiana V. Kutavina, Kristen Lee, Dimitra Lotakis, Matthew T. Maurano, Shane J. Neph, Fiedencio V. Neri, Eric D. Nguyen, Hongzhu Qu, Alex P. Reynolds, Vaughn Roach, Eric Rynes, Minerva E. Sanchez, Richard S. Sandstrom, Anthony O. Shafer, Andrew B. Stergachis, Sean Thomas, Benjamin Vernot, Jeff Vierstra, Shinyong Vong, Hao Wang, Molly A. Weaver, Yongqi Yan, Miaohua Zhang, Joshua M. Akey, Michael Bender, Michael O. Dorschner, Mark Groudine, Michael J. MacCoss, Patrick Navas, George Stamatoyannopoulos, Kathryn Beal, Alvis Brazma, Paul Flicek, Nathan Johnson, Margus Lukk, Nicholas M. Luscombe, Daniel Sobral, Juan M. Vaquerizas, Serafim
  - Batzoglou, Arend Sidow, Nadine Hussami, Sofia Kyriazopoulou-Panagiotopoulou, Max W. Libbrecht, Marc A. Schaub, Webb Miller, Peter J. Bickel, Balazs Banfai, Nathan P. Boley, Haiyan Huang, Jingyi Jessica Li, William Stafford Noble, Jeffrey A. Birmes, Orion J. Buske, Avinash D. Sahu, Peter V. Kharchenko, Peter J. Park, Dannon Baker, James Taylor, and Lucas Lochovsky. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012 489:7414, 489(7414):57–74, 9 2012. ISSN 1476-4687. doi: 10.1038/nature11247.
  43. Thomas Derrien, Jordi Estellé, Santiago Marco Sola, David G. Knowles, Emanuele Rainieri, Roderic Guigó, and Paolo Ribeca. Fast Computation and Applications of Genome Mappability. *PLOS ONE*, 7(1):e30377, 1 2012. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0030377.