SECEDO: SNV-based subclone detection using ultra-low coverage single-cell DNA sequencing

Hana Rozhoňová[1,2,3,†], Daniel Danciu[1,4,†], Stefan Stark[1,3,4], Gunnar Rätsch[1,3,4,5,*], André Kahles[1,3,4,*], and Kjong-Van Lehmann[1,3,4,6,7*]

[1] Biomedical Informatics Group, Department of Computer Science, ETH Zurich, Zurich, Switzerland
[2] Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland
[3] Swiss Institute of Bioinformatics, Lausanne, Switzerland
[4] Biomedical Informatics Research, University Hospital Zurich, Zurich, Switzerland
[5] Department of Biology, ETH Zurich, Zurich, Switzerland
[6] Center for Integrated Oncology, University Hospital Cologne, Cologne, Germany
[7] Joint Research Center for Computational Biomedicine, RWTH Aachen University Medical School, Aachen, Germany
[†] Joint first authors
[*] Joint corresponding authors

Abstract. Recently developed single-cell DNA sequencing technologies enable whole-genome, amplification-free sequencing of thousands of cells at the cost of ultra-low coverage of the sequenced data ($<$ 0.05x per cell), which mostly limits their usage to the identification of copy number alterations (CNAs) in multi-megabase segments. Aside from CNA-based subclone detection, single-nucleotide variant (SNV)-based subclone detection may contribute to a more comprehensive view on intra-tumor heterogeneity. Due to the low coverage of the data, the identification of SNVs is only possible when superimposing the sequenced genomes of hundreds of genetically similar cells. Here we present Single Cell Data Tumor Clusterer (SECEDO, lat. 'to separate'), a new method to cluster tumor cells based solely on SNVs, inferred on ultra-low coverage single-cell DNA sequencing data. The core aspects of the method are an efficient Bayesian filtering of relevant loci and the exploitation of read overlaps and phasing information. We applied SECEDO to a synthetic dataset simulating 7,250 cells and eight tumor subclones from a single patient, and were able to accurately reconstruct the clonal composition, detecting 92.11% of the somatic SNVs, with the smallest clusters representing only 6.9% of the total population. When applied to four real single-cell sequencing datasets from a breast cancer patient, SECEDO was able to recover the major clonal composition in each dataset at the original sequencing depth of 0.03x per cell, an 8-fold improvement relative to the state of the art. Variant calling on the resulting clusters recovered more than twice as many SNVs with double the allelic ratio compared to calling on all cells together, demonstrating the utility of SECEDO.
SECEDO is implemented in C++ and is publicly available at https://github.com/ratschlab/secedo.

Keywords: single-cell sequencing · intra-tumor heterogeneity · clustering

1 Introduction

Somatic single-nucleotide variants (SNVs) are commonly associated with cancer progression and growth [39]. The recent development of single-cell DNA sequencing technologies [12] offer the ability to study somatic SNVs at a single-cell level, providing much more detailed information about tumor composition and phylogeny than traditional bulk sequencing [18,33]. However, several technical obstacles decrease the interpretability of the data obtained using these technologies. In particular, most of the current single-cell DNA sequencing technologies require a whole-genome amplification step, which introduces artifacts such as DNA-amplification errors, allelic drop-out, imbalanced amplification, etc. [12]. Several approaches [4,10,16,19,27,38,44] have been proposed to detect SNVs based on such data.

Approaches that do not require whole-genome amplification have been developed to overcome issues related to amplification [20,33]. A prominent example of such technologies is 10X Genomics' Chromium Single Cell CNV Solution[1]. This technology allows the sequencing of hundreds to thousands of cells in parallel, albeit with only extremely low sequencing coverage (<0.05x per cell). Hence, its use has been limited to the inference of copy number variations (CNVs) and alterations (CNAs) [1,11,41,42]. The attempts to also use these data for the identification of tumor subclones based solely on SNVs have so far failed to provide a solution that would be able to recover the clonal composition at the original sequencing depth [32]; in particular, the algorithm of [32] requires a minimum coverage of $\geq$ 0.2x per cell, roughly four times more than what is currently achievable using the 10X Genomics technology [1,41].

In this work, we propose SECEDO, a novel algorithm for clustering cells based on SNVs using single-cell sequencing data with ultra-low coverage. Using an extensive set of simulated data, as well as four real data sets, we show that SECEDO is able to correctly identify tumor subclones in data sets with per-cell coverage as low as 0.03x, improving the current state of the art by a factor of eight and thus rendering the algorithm applicable to currently available 10X Genomics single-cell data. We also provide an efficient C++ implementation of SECEDO, which is able to quickly cluster sequencing data from thousands of cells while running on commodity machines.

2 Methods

Overview

Due to the extremely low coverage of the data (< 0.05x per cell), deciding whether two cells have identical or distinct genotypes is a difficult problem. Most loci are covered, if at all, by only one read (**Supplementary Figure S1**). This makes it difficult, if not impossible, to interpret an observed mismatch when comparing data from two cells. The mismatch could be caused by an actual somatic SNV, by a sequencing error, or by a heterozygous locus that was sequenced in different phase in the two cells. The situation is further complicated by the fact that both heterozygous loci and sequencing errors occur with higher frequency than somatic SNVs in cancer: the frequency of heterozygous loci in a typical human genome is about 10^{-3} [5,30] and sequencing errors (using the Illumina technology) happen with a rate of about 10^{-3} [9,17,29,35], while the prevalence of somatic SNVs in cancer is typically reported to be between 10^{-9} and 10^{-3} [2,23]. Hence, it is crucial to jointly leverage the information from all cells at the same time.

The pivotal blocks in the SECEDO pipeline are (1) the Bayesian filtering strategy for efficient identification of relevant loci and (2) the derivation of a global cell-to-cell similarity matrix utilizing both the structure of reads and the haplotype phasing, which proves to be more informative than considering only one locus at a time.

SECEDO first performs a filtering step, in which it examines the pooled sequenced data for each locus and uses a Bayesian strategy to eliminate loci that are unlikely to carry a somatic SNV (**Figure 1**). The filtering step drastically increases the signal-to-noise ratio by reducing the number of loci by 3 to 4 orders of magnitude (depending on the coverage), while only eliminating approximately half of the loci that carry a somatic SNV. Moreover, the eliminated mutated loci typically have low coverage or high error rate and would not be very useful for clustering. In the second step, SECEDO builds a cell-to-cell similarity matrix based only on read-pairs containing the filtered loci, using a probabilistic model that takes into account

[1] https://www.10xgenomics.com/products/single-cell-cnv/

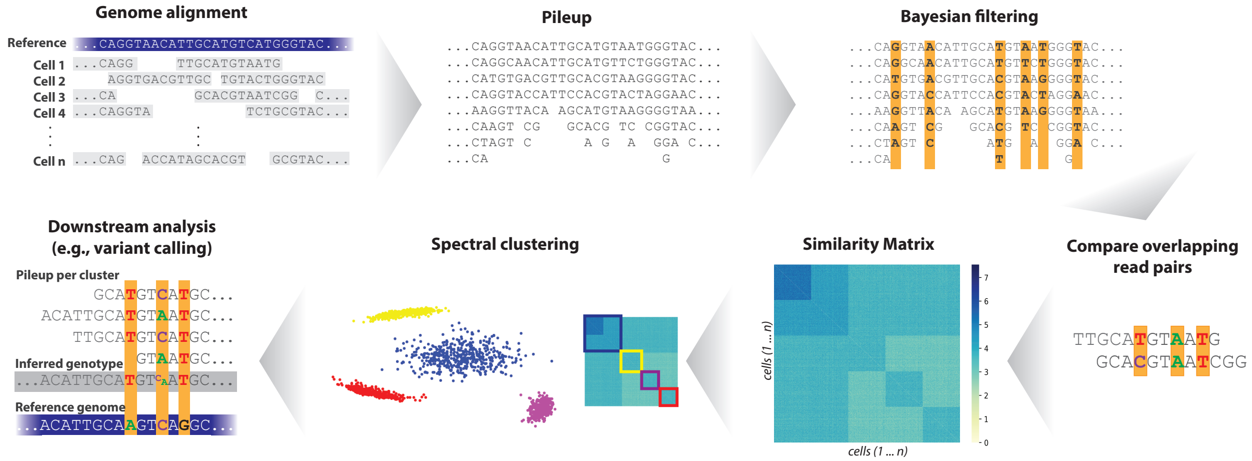

Fig. 1. The SECEDO pipeline. After sequencing, reads are piled up per locus and a Bayesian filter eliminates loci that are unlikely to carry a somatic SNV. For each pair of reads, SECEDO compares the filtered loci and updates the likelihoods of having the same genotype and of having different genotypes for the corresponding cells. The similarity matrix, computed as described in Methods, is then used to cluster the cells into 2 to 4 groups (the number of groups depends on the data and is determined automatically by SECEDO) using spectral clustering. The algorithm is then recursively applied to each cluster until a termination criterion is reached.

the probability of sequencing errors, the frequency of SNVs, the filtering performance, and, crucially, the structure of the reads, i.e. the fact that the whole read was sampled from the same haplotype. In the third step of the pipeline, we use spectral clustering to divide the cells into two or more groups. At this point, we reduced the problem to an instance of the well-studied community detection problem [36], so spectral clustering is a natural choice. Optionally, the results of spectral clustering can be further refined in a fourth step using the Expectation-maximization algorithm [7]. The whole pipeline is then repeated for each of the resulting subclusters. The process is stopped if (1) there is no evidence for the presence of at least two clusters in the similarity matrix, or (2) the clusters are deemed too small. Downstream analysis, for instance, variant calling, can then be performed by pooling sequencing data from all cells in one cluster based on the results of SECEDO to create a pseudo-bulk sample.

Filtering uninformative loci

Consideration of all genomic loci is not desirable when performing the clustering and variant calling, since most positions are not informative for clonal deconvolution. The most informative loci with respect to the clustering of the cells are the loci carrying somatic SNVs since they provide (1) information on assignment of cells to clusters and (2) information on haplotype phasing (due to loss/gain of heterozygosity). To a lesser extent, this is also true for germline heterozygous loci since they provide information on haplotype phasing. In other words, loci at which all the cells have the same homozygous genotype do not provide any information relevant to the task of dividing the cells into genetically homogeneous groups, so they can be excluded from downstream analysis.

Due to the low sequencing coverage, it is generally not possible to reliably assign genotypes to individual cells. However, we identify loci of interest by using the *pooled data* across all the cells to approximate posterior probabilities that the cells have the same genotype. Consider for example a specific locus at which all cells have genotype AA. Assuming sequencing errors happen independently with probability θ and are unbiased (i.e. all types of substitutions are equally probable), the fraction of As in the pooled data is in expectation $(1-\theta)$ and the fraction of all other bases is $\theta/3$. A locus with a significantly different proportion of observed bases indicates that there may be two (or more) different genotypes contributing to the observed data. In particular, we compute the posterior probability that all cells at the locus share the same homoyzgous genotype using an approximate Bayesian procedure. If this posterior is lower than a chosen threshold K, the locus is marked as ‘informative’.

Formally, let C_1, C_2, C_3, C_4 be the bases sorted from the most to the least frequent in the pooled data at the given position, c_1, c_2, c_3, c_4 the corresponding counts ($c_1 \geq c_2 \geq c_3 \geq c_4$), c the total coverage ($c = c_1 + c_2 + c_3 + c_4$). Next, let M be an indicator random variable that is 1 if all cells in the sample have the same homozygous genotype and 0 otherwise. Applying Bayes rule, we can compute $P(M \mid c_1, c_2, c_3, c_4)$ as:

$$P(M \mid c_1, c_2, c_3, c_4) = \frac{P(c_1, c_2, c_3, c_4 \mid M)P(M)}{P(c_1, c_2, c_3, c_4)} \tag{1}$$

We compute or approximate the individual terms as follows:

- $P(M)$ can be estimated from literature: the prevalence of somatic SNVs in cancer lies between 10^{-9} and 10^{-3} [2,23]; the frequency of heterozygous sites in a typical human genome lies between ca 0.04 and 0.11% [5,30]. In order to be conservative, we choose the largest probability ($\approx 10^{-3}$) in both cases, resulting in $P(M) \approx 1 - 2 \cdot 10^{-3} = 0.998$.
- $P(c_1, c_2, c_3, c_4 \mid M)$, is equal to

$$P(c_1, c_2, c_3, c_4 \mid M) = \sum_{g \in \mathcal{G}} P(c_1, c_2, c_3, c_4 \mid \text{genotype of all cells is } g)P(g), \tag{2}$$

  where $\mathcal{G} = \{\text{AA, CC, GG, TT}\}$ is the set of all possible homozygous genotypes.
  The probability of observing data (c_1, c_2, c_3, c_4) given that the genotype of all cells is g ($g = C_iC_i$) has a multinomial distribution with c trials and event probabilities equal to $\left(1 - \theta, \frac{\theta}{3}, \frac{\theta}{3}, \frac{\theta}{3}\right)$:

$$P(c_1, c_2, c_3, c_4 \mid \text{genotype of all cells is } g) = \frac{c!}{c_1!c_2!c_3!c_4!}(1-\theta)^{c_i}\left(\frac{\theta}{3}\right)^{c-c_i}$$

  Assuming the error rate θ is small, the result of the equation above is negligible for any c_i that is not close to c. As a consequence, if the prior $P(g)$ is approximately the same for all genotypes, we can approximate the sum in **Equation 2** with the largest term:

$$P(c_1, c_2, c_3, c_4 \mid M) \approx \max_{g \in \mathcal{G}} P(c_1, c_2, c_3, c_4 \mid \text{genotype of all cells is } g)P(g). \tag{3}$$

- Computing $P(c_1, c_2, c_3, c_4)$ is intractable, as it would involve summing over all possible combinations of the cells' genotypes. We instead approximate the evidence by

$$P(c_1, c_2, c_3, c_4) \approx \frac{c!}{c_1!c_2!c_3!c_4!}\left[p_{hom}(1-\theta)^{c_1}\left(\frac{\theta}{3}\right)^{c_2+c_3+c_4} + p_{het}\left(\frac{1}{2} - \frac{\theta}{3}\right)^{c_1+c_2}\left(\frac{\theta}{3}\right)^{c_3+c_4}\right.$$
$$\left. + p_{hom}p_{mut}\left(\frac{3}{4} - \frac{\theta}{3}\right)^{c_1}\left(\frac{1}{4}\right)^{c_2}\left(\frac{\theta}{3}\right)^{c_3+c_4} + p_{het}p_{mut}\frac{c!}{c_1!c_2!c_3!c_4!}\left(\frac{1}{2} - \frac{\theta}{3}\right)^{c_1}\left(\frac{1}{4}\right)^{c_2+c_3}\left(\frac{\theta}{3}\right)^{c_4}\right]$$

  where $p_{hom}, p_{het}, p_{mut}$ represent the probability of a locus being homozygous, heterozygous and mutated, respectively. The first summation term estimates $P(c_1, c_2, c_3, c_4)$ for a homozygous locus, the second term assumes a heterozygous locus, the third term corresponds to a homozygous locus that suffered a somatic mutation, and the last term to a heterozygous locus with a somatic mutation. See **Supplemental Material S1** for a more detailed derivation.

We then include the locus into the subset of informative positions if $P(M \mid c_1, c_2, c_3, c_4) \leq K$ for a suitable constant K (see **Supplemental Material S2** and **Supplementary Table S1**).

Filtering heterozygous loci is similar. Here, let $P(M' \mid c_1, c_2, c_3, c_4)$ be the probability that all cells have the same *heterozygous* genotype. The individual terms in **Equation 1** are identical except that the event probabilities for the multinomial distribution are $\left(\frac{1}{2} - \frac{\theta}{3}, \frac{1}{2} - \frac{\theta}{3}, \frac{\theta}{3}, \frac{\theta}{3}\right)$. However, since heterozygous loci are three orders of magnitude fewer than homozygous loci [5,30] in addition to potentially being useful in haplotype phasing, in practice it is sufficient to use a simple heuristic: we denote the locus as informative if $c_1 > 1.5 \cdot c_2$, where c_1 and c_2 are the most frequent and the second most frequent bases at that locus, respectively (the expectation is that at a heterozygous locus c_1 and c_2 are close to each other). In addition, we reject all loci for which $c_1 + c_2 + c_3 < 5$.

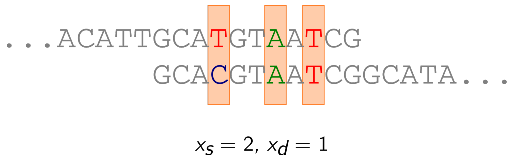

Fig. 2. Illustration of an overlap between two reads. The orange shaded positions are the positions chosen as informative. In this example, length of the overlap is 3, the number of positions where the bases are the same, x_s, is 2 and the number of positions where they are different, x_d, is 1. For our purposes, an overlap is fully described by the tuple (x_s, x_d).

The final set of informative loci then includes those positions that were marked as informative by both filtering steps (i.e. filtering of both homozygous and heterozygous loci). In practice, sequencing artefacts may lead to loci with unusually high coverage. For this reason, we also eliminate any loci with coverage more than two standard deviations away from the expected coverage. In addition, we also eliminate loci where $c - c_1 < 5$.

Cell-to-cell similarities

We define the similarity $s(i, j)$ of cells i and j as the log-odds of the probability that cells i and j have the same genotype and the probability that they have different genotypes, given the corresponding sets of reads. Each of the two probabilities is then approximated as a product of probabilities of individual *overlaps* of two reads, one read from cell i and one read from cell j (**Figure 2**). Formally:

$$s(i,j) = \log\left(\frac{P\left[C(i) = C(j) \mid r_i, r_j, h, \epsilon\right]}{P\left[C(i) \neq C(j) \mid r_i, r_j, h, \epsilon\right]}\right) = \log\left(\frac{P\left[r_i, r_j \mid C(i) = C(j), h, \epsilon\right]}{P\left[r_i, r_j \mid C(i) \neq C(j), h, \epsilon\right]}\right) \tag{4}$$

$$\approx \sum_{k,l} \log\left(\frac{P\left[x_s(r_i^k, r_j^l), x_d(r_i^k, r_j^l) \mid C(i) = C(j), h, \epsilon\right]}{P\left[x_s(r_i^k, r_j^l), x_d(r_i^k, r_j^l) \mid C(i) \neq C(j), h, \epsilon\right]}\right), \tag{5}$$

where r_i is the set of reads from cell i, r_i^k is the k-th read from cell i, $x_s(p, q)$ and $x_d(p, q)$ the number of matches and mismatches, respectively, between reads p and q, $C(i)$ is the (true) cluster assignment of cell i, ϵ is the proportion of SNVs in the set of informative positions and h the proportion of homozygous loci in the set of informative positions (see below). We assume that observing cells with the same genotype and with different genotypes has the same prior probability. (Notice that decomposing the probabilities in **Equation 4** over pairs of reads is indeed only an approximation. In particular, the decomposition in **Equation 5** would only be precise if no two reads coming from one cell were overlapping; in the opposite case, the probabilities of read pairs containing one of these overlapping reads are non-independent. However, since the per-cell coverage is so low (**Supplementary Figure S1**), the number of such non-independent pairs is negligible.)

Notice that by decomposing the probabilities over the overlaps of reads we gain information not only on the number of matches and mismatches between the two reads (i.e. information on potential differences between the two cells), but also information on haplotype phasing. Moreover, it also allows us to put more weight on longer (and hence supposedly more informative) overlaps. For example, a long overlap with only matches is an indication that the two cells might have the same genotype. A long overlap with only mismatches, on the other hand, is not a strong indication towards the cells being from different clusters – another likely scenario is that the two reads were sampled from different haplotypes and we just observe a row of heterozygous loci in different phase. As a result, overlaps with a combination of matches and mismatches are the ones most strongly suggesting the 'different genotypes' case (**Supplementary Figure S2**). We also show, using simulated data, that considering the number of matches and mismatches in the whole overlap of two reads provides strictly more information than considering each locus independently (**Supplementary Figure S3**).

Below we give details on the computation of **Equation 5**.

Assumptions We make the following simplifying assumptions:

1. There are only two kinds of cells; the prevalence of differences between their genomes is μ.
2. All cells are diploid.
3. The somatic SNVs are with equal probability of type AA+AB (a homozygous site in cluster 1, heterozygous in cluster 2) and AB+AA (a heterozygous site in cluster 1, homozygous in cluster 2); in particular, we do not take into account homozygous SNVs (AA+BB) or mutations that involve more than two different alleles (e.g. AB+AC) [6].
4. The mutated and germline heterozygous loci are distributed randomly and independently in the genome; in particular, they do not tend to cluster together.
5. Sequencing errors happen independently at each sequenced base with probability θ and are unbiased (i.e. all types of substitutions are equally likely).
6. We run the clustering on the set of informative positions; in the set there are loci of three types:
   - loci that carry a somatic SNV,
   - loci that have the same homozygous genotype in all cells, and were not discarded during the previous step only because of sequencing errors, and
   - loci that have the same heterozygous genotype in all cells, and were not discarded during the previous step e.g. because of preferential amplification.

   We denote the frequency of the homozygous loci in the data set by h and the frequency of the mutated loci by ϵ; the frequency of the germline heterozygous loci is then $(1 - h - \epsilon)$.

Parameters The algorithm has three parameters: h, the fraction of the homozygous loci in the set of selected positions, ϵ, the fraction of the mutated loci in the set, and θ, the error rate. In our analyses, we used $h = 0.5$, $\epsilon = 0.01$, and $\theta = 0.05$ (the θ parameter has higher value than the usually reported sequencing error rate, because the set of informative positions is enriched in positions carrying sequencing errors).

Computing the probabilities of overlaps We define:

- $P_{s,s}$, the probability that sequencing of two bases of the same kind results again in two bases of the same kind:
$$P_{s,s} = (1-\theta)^2 + \frac{\theta^2}{3} \tag{6}$$
(both bases are sequenced without error, or both are misread to the same base),
- $P_{s,d}$, the probability that sequencing of two bases of the same kind results in bases that differ from each other:
$$P_{s,d} = 1 - P_{s,s}, \tag{7}$$
- $P_{d,s}$, the probability that two different bases are read as the same:
$$P_{d,s} = 2 \cdot (1-\theta) \cdot \frac{\theta}{3} + \frac{2\theta^2}{9} \tag{8}$$
(one of the two bases is misread to the other one, or both are misread to the same base),
- $P_{d,d}$ the probability that two different bases are sequenced as different:
$$P_{d,d} = 1 - P_{d,s}. \tag{9}$$

The probability of observing x_s matches and x_d mismatches in an overlap of length $x_s + x_d$, assuming cells i and j have the same genotype, is now:

$$P\left[x_s, x_d \mid C(i) = C(j), h, \epsilon\right] = \binom{x_s + x_d}{x_s} \sum_{k=0}^{x_s} \sum_{l=0}^{x_d} \binom{x_s}{k} \binom{x_d}{l} \cdot \overbrace{\left(1 - h - \frac{\epsilon}{2}\right)^{k+l} \left(\frac{1}{2}\left(P_{s,s}^{k} \cdot P_{s,d}^{l} + P_{d,s}^{k} \cdot P_{d,d}^{l}\right)\right)^{\delta(k+l)}}^{\text{heterozygous positions}} \cdot \overbrace{\left(h + \frac{\epsilon}{2}\right)^{(x_s + x_d - k - l)} \cdot P_{s,s}^{(x_s - k)} \cdot P_{s,d}^{(x_d - l)}}^{\text{homozygous positions}}, \tag{10}$$

where $\delta(x)$ is a function defined as 0, if $x = 0$, and 1, otherwise. In the formula we sum over all possible combinations of $(k + l)$ heterozygous loci and $(x_s + x_d - k - l)$ homozygous loci; k of the heterozygous loci result in a match, the remaining l in a mismatch.

The probability of observing x_s matches and x_d mismatches assuming cells i and j are in different clusters is:

$$P\left[x_s, x_d \mid C(i) \neq C(j), h, \epsilon\right] = \binom{x_s + x_d}{x_s} \cdot \sum_{k=0}^{x_s} \sum_{p=0}^{x_s-k} \sum_{l=0}^{x_d} \sum_{q=0}^{x_d-l} \frac{x_s!}{k!p!(x_s-k-p)!} \cdot \frac{x_d!}{l!q!(x_d-l-q)!}$$

$$\cdot \overbrace{(1-h-\epsilon)^{k+l} \left(\frac{1}{2}\left(P_{s,s}^{k} \cdot P_{s,d}^{l} + P_{d,s}^{k} \cdot P_{d,d}^{l}\right)\right)^{\delta(k+l)}}^{\text{heterozygous positions}} \cdot \overbrace{h^{(p+q)} \cdot P_{s,s}^{p} P_{s,d}^{q}}^{\text{homozygous positions}}$$

$$\cdot \overbrace{\left(\frac{\epsilon}{2}\right)^{(x_s+x_d-k-l-p-q)} \cdot \left(P_{s,s} + P_{d,s}\right)^{(x_s-k-p)} \cdot \left(P_{d,d} + P_{s,d}\right)^{(x_d-l-q)}}^{\text{mutated positions}} \tag{11}$$

Here k denotes the number of heterozygous positions giving rise to a match, l the number of heterozygous positions giving rise to a mismatch, p the number of positions with the same homozygous genotype in both types of cells that give rise to a match and q the number of these positions that result in a mismatch.

Clustering

We first normalize the computed similarity matrix by making sure all elements are positive: $S^* = -S + \min_{i,j} s(i,j)$. The cells are then clustered using a slight variation on spectral clustering [34] as follows. We compute the symmetric normalized Laplacian $\mathcal{L} = I - D^{-\frac{1}{2}} S^* D^{-\frac{1}{2}}$ and determine its first k (we used $k = 6$ in all experiments in this paper) eigenvectors, corresponding to the k smallest eigenvalues. We then cluster into 1,2,3 or 4 clusters using k-means [3,26], computing the inertia values i_1, i_2, i_3, i_4 for each of the four options and the inertia gaps $g_k = i_k - i_{k-1}, k = 2, 3, 4$, and define $g_1 := 0$. The final number of clusters is $\max_{k=2,3,4} \{k \mid g_k > 0.75 g_{k-1}\}$.

One important aspect of clustering is the stopping criterion, i.e. the decision whether a specific group of cells should be divided into subclusters or not. We suggest a (to the best of our knowledge new) heuristic approach to automatically decide if the computed normalized similarity matrix S^* indicates that there are two (or more) different clusters of cells. We fit a Gaussian mixture model with 1, 2, 3 or 4 components to the smallest k eigenvectors of S^* and compare their likelihood using the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). If the model with only one component is preferred by AIC/BIC over the models with 2, 3 or 4 components we do not split the data further. We further do not accept the split if the resulting subclone has too few cells (we used 500 in our experiments). We also require that the mean within-cluster coverage is at least 9, the lowest coverage sufficient for a reliable variant call (see **Supplemental Material S3**).

3 Results

SECEDO recovers tumor subclones with average precision of 97% on simulated data

In order to test the performance of our method, we simulated a dataset consisting of 7,250 cells divided into 9 groups of various sizes: one group of healthy cells and 8 groups of tumor cells. The genome of the healthy cells was created using Varsim [31] based on the GRCh38 human reference genome. Common variants from dbSNP [37] (3,000,000 single-nucleotide polymorphisms, 100,000 small insertions, 100,000 small deletions, 50,000 multi-nucleotide polymorphisms, 50,000 complex variants) were added to the genome. The genome of the tumor cells was built by adding 2,500 to 20,000 of both coding and non-coding SNVs (subclonal SNV fraction of 3%-27%, [8]), randomly chosen from the COSMIC v94 (Catalogue Of Somatic Mutations In Cancer) database [40], to the parent genome, in addition to 250 small insertions, 250 small deletions, 200 multi-nucleotide variants and 200 complex variants (**Figure 3**). Paired-end reads, with each mate of length

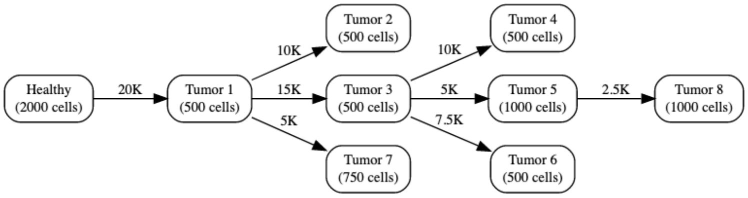

Fig. 3. Theoretical phylogenetic tree of the synthetic dataset, comprising 7,250 cells and 9 unequally sized subclones. Edge labels indicate the number of additional SNVs in each subclone relative to the parent. Node labels indicate the number of cells in each subclone. The mean per-cell coverage is 0.05x.

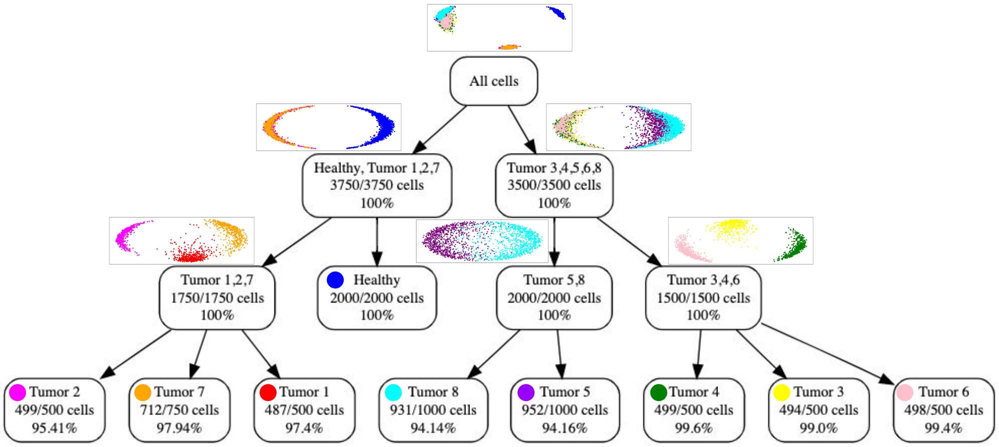

Fig. 4. Clustering a synthetic dataset with 9 unequally sized subclones totaling 7,250 cells. Each node represents one recursive SECEDO clustering step. The percentage at the bottom of each node indicates the clustering precision (correctly clustered cells relative to total cells in cluster). The scatter plots above parent nodes depict the 2nd and 3rd eigenvectors of the similarity matrix Laplacian. For leaf nodes SECEDO correctly determined that further clustering is not desirable.

100 bp, were simulated using ART [15] at an average coverage of 0.05x per cell and with the error profile of Illumina HiSeq 2000 machines. The reads were then aligned using Bowtie 2 [21] and filtered using Samtools [24] to select for reads mapped only in proper pair, non-duplicate and only primary alignments.

For efficiency reasons, we build the pileup files used by the Bayesian filtering using our own implementation rather than existing tools that are not optimized for use on thousands of cells simultaneously (e.g. Samtools, which currently does not offer a multi-threaded pileup creation). The pileup creation, distributed on 23 commodity machines (one for each chromosome) using 20 threads each, takes about 70 minutes (down from 72 hours when using Samtools' pileup creation on the same machines). We ran SECEDO on the resulting pileup files on an Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz using 20 threads and 32GB of RAM. The filtering, clustering and VCF generation took 21 minutes. For the top level clustering, the filtering step kept about 1 in 16,000 loci. Somewhat counter-intuitively, the number of filtered loci approximately doubled at each level as we traveled down the clustering tree. This is due to the fact that the discriminative power of the Bayesian filtering degrades as the mean pooled coverage decreases (from 248 at the root to 20 at the leaves), such that a larger proportion of loci that are not relevant are let through. SECEDO was able to recover all 9 subclones with an average precision of 97.45% (see **Figure 4**). Note that SECEDO is not attempting to reconstruct the evolutionary history of the tumor, but merely trying to efficiently find a grouping of cells that reflect the current subclonal structure and enable downstream tasks like variant calling. Therefore, the clustering tree reconstructed by SECEDO does not reflect the actual developmental process that gave rise to the given population of cancer cells; indeed, the SECEDO clustering tree (see **Figure 4**) differs from the true phylogenetic tree of the population (see **Figure 3**).

In order to show the potential of the resulting clusters for somatic variant calling, we identified the most likely genotype of each cluster using a simple MAQ-based approach [25] (**Supplemental Material S3**) and

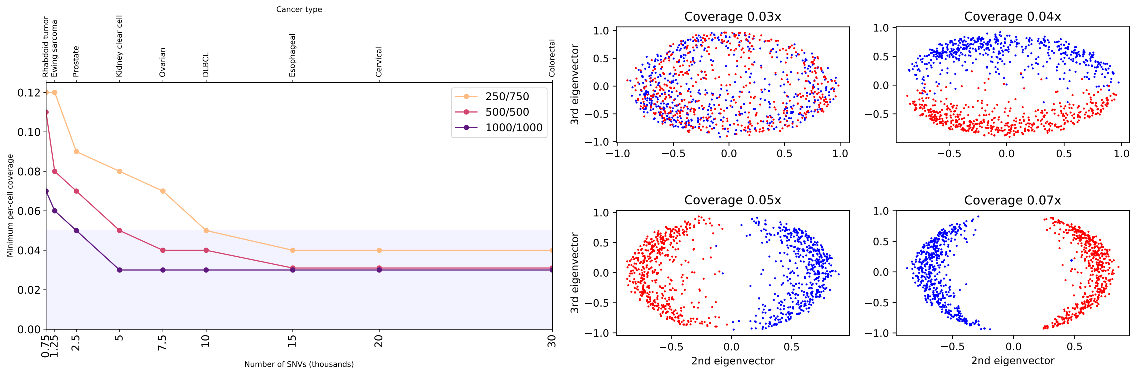

Fig. 5. Analysis of the minimal requirements under which SECEDO can be applied to a given dataset. **Left:** Minimum required coverage for successful clustering of sub-clones differing in the given number of SNVs, in three scenarios. The yellow and red lines show the required average per-cell coverage for clustering 1,000 cells, with a ($^1/_4$, $^3/_4$) split, and an equal ($^1/_2$, $^1/_2$) split, respectively. The bottom line shows the required per-cell coverage for clustering 2,000 with an equal split. The shaded area marks the coverage currently achievable in practice. The top labels indicate the cancer type with mean mutation rate closest to the given SNV density (cancer mutation rates according to [23]). DLBCL = Diffuse large B-cell lymphoma. **Right:** Scatter plots of the 2nd and 3rd eigenvectors of the similarity matrix Laplacian for increasing coverage values (1,000 cells, 500 in each group; 5,000 SNVs). The higher the coverage, the clearer the separation between the two clusters.

generated VCF files for each cluster against the GRCh38 human reference genome. Similarly to other variant callers that remove germline variants [6], we then removed the ground-truth variants that were present in the healthy cells and compared the remaining SNVs against the ground truth SNVs provided by Varsim for each cluster. SECEDO was able to detect 92.11% of the somatic SNVs (vs. 77.79% when calling variants on the unclustered cells) with a 52.41% average precision (see **Supplementary Table S2**).

SECEDO is able to correctly group cells starting at 0.05x coverage and 500 cells per cluster

One practical question of crucial importance is how to determine if, given a dataset, SECEDO will be able to correctly cluster the cells for meaningful downstream processing. To answer this question, we conducted a series of experiments to determine the conditions under which SECEDO can successfully be applied to a given dataset. There are three cluster attributes that affect SECEDO's ability to separate cell clusters: (a) the number of cells, (b) the average per-cell coverage, and (c) the number of SNVs in which the clones differ. In order to test the interplay of these three cluster attributes, we devised a series of synthetic datasets, each consisting of 1,000 cells belonging to two groups. The sizes of the two groups were either equal (i.e. 500 cells in each group) or in ratio 1:3 (i.e. one cluster consisted of 250 cells and the other one of 750 cells). Then, for a given number of SNVs and given sizes of clusters, we gradually lowered the per-cell coverage until the algorithm was unable to cluster the cells correctly. The genome creation, reads simulation, and alignment were done as described in the previous section. For most parameter configurations, the currently achievable per-cell coverage of 0.05x is sufficient for SECEDO to correctly cluster the cells (see **Figure 5**, left). Since SECEDO is able to discriminate between balanced clusters of 500 cells that differ in as little as 5,000 SNVs (equivalent to an SNV prevalence of ca $1.67 \cdot 10^{-6}$), the method can be applied to a wide variety of cancers, starting from those with very high mutation rates, such as melanoma (mean prevalence of somatic SNVs ca 10^{-5}) down to pancreatic and breast cancer (mean prevalence of somatic SNVs ca 10^{-6}) [2,23]. Note that there is a relationship between tumor mutational burden and SECEDO's ability to distinguish subclones. SECEDO is able to identify complex subclonal structures (such as in **Figure 3**) in cancers with high mutational burden (e.g. melanoma), whereas in cancers with lower mutational burden (e.g. pancreatic and breast cancer) only major clones could be identified, as shown in the next section.

An important thing to note is that SECEDO's discriminative power goes beyond 5,000 SNVs when increasing the pooled coverage (e.g. by increasing the number of sequenced cells): **Figure 4** shows that SECEDO was able to accurately isolate cells in Tumor 8, even though they only differ in 2,500 SNVs from

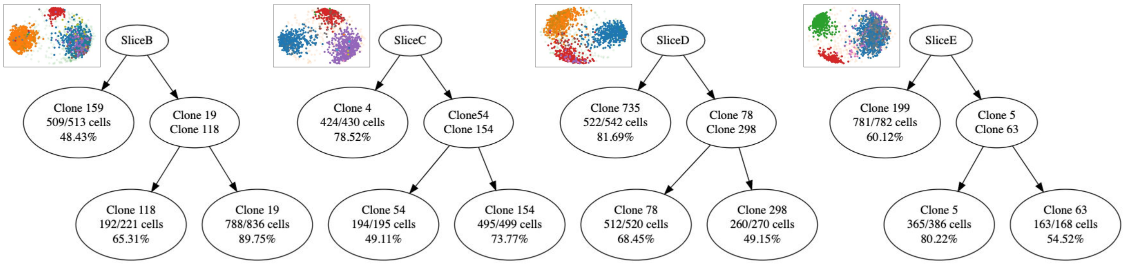

Fig. 6. Clustering of the four tumor sections in the 10x Genomics ductal carcinoma dataset. The first row in each node denotes the cluster name; for consistency, we used the same cluster numbering as [42][3]. The second row denotes the number of cells recovered by SECEDO vs the total number of cells as identified by [42]. The last row denotes the precision of the clustering, i.e. the percentage of cells in the SECEDO cluster that match the originally reported cluster. The lower precision values are due to the fact that cells categorized by [42] as ”None” based on the CNV signature are assigned a category by SECEDO based on the genomic signature.

their parent. While **Figure 5** indicates that at a coverage of 0.05x a minimum of 5,000 SNVs are required to distinguish the subclones, this result was obtained for a pooled coverage of ≈ 41, which is nearly half the pooled coverage for Tumors 5 and 8 (≈ 76). As expected, the discriminative power of SECEDO also increases with the per-cell coverage (see **Figure 5**, right), since the per-cell coverage acts as a multiplying factor for the pooled coverage.

SECEDO recovers dominant subclones in a breast cancer dataset

In order to test the performance of SECEDO on real data, we downloaded a publicly available 10X Genomics single-cell DNA sequencing dataset[2] sequenced using an Illumina NovaSeq 6000 System. The dataset contains five tumor sections (labeled A to E) of a triple negative ductal carcinoma, each section containing roughly 2,000 cells [1]. The mean per-cell coverage in the data set is 0.03x, with individual coverages ranging from 0.006x to 0.086x. Using CNV profiling [42], three dominant clones were identified in each of the sections, except for section A, which only has one dominant clone and was thus not included in our analysis.

We applied SECEDO to the four datasets corresponding to sections B,C,D, and E. The filtering step reduced the number of loci in each tumor section to roughly 1,000,000 bp (ca 0.03% of the original size); the average pooled coverage across the ≈ 2,000 cells in each dataset ranged from 45 to 55. SECEDO was able to correctly recover the three dominant clones in each of the four tumor sections. The clustering results match with high accuracy (96.68% recall, 66.59% precision) those in [42] (**Figure 6**). The scatter plots of the 2nd and 3rd eigenvector of the similarity matrix confirm that each tumor section consists of three highly separable clusters.

We then called SNVs on each subclone of Slice B independently, and on the entire slice. In order to call SNVs, we created a Panel of Normals from the cells categorized as normal by [42] based on the CNV profile (Clone19 in the left-most tree of **Figure 6**). We ran MuTect 1.1.4 [6] with the default settings, using dbSNP [37] and Cosmic v94 [40] as priors. The number of distinct SNVs in the two tumor subclones is more than double the number of variants that were called when pooling all cells together (**Figure 7**, left). The histogram of the allelic factor for the sublconal and global SNVS shows a significant shift to the right for the subclonal SNVs, an indication that the clustering correctly identified and separated genetically similar cells, enabling the detection of twice as many SNVs at twice the allelic ratio (**Figure 7**, right).

4 Discussion

We introduced SECEDO, a method that is able to correctly identify subclones in single-cell sequencing datasets with coverage as low as 0.03x per cell (providing an 8-fold improvement on the state of the art [32]),

[2] https://www.10xgenomics.com/resources/datasets/
[3] Available at https://github.com/raphael-group/chisel-data/

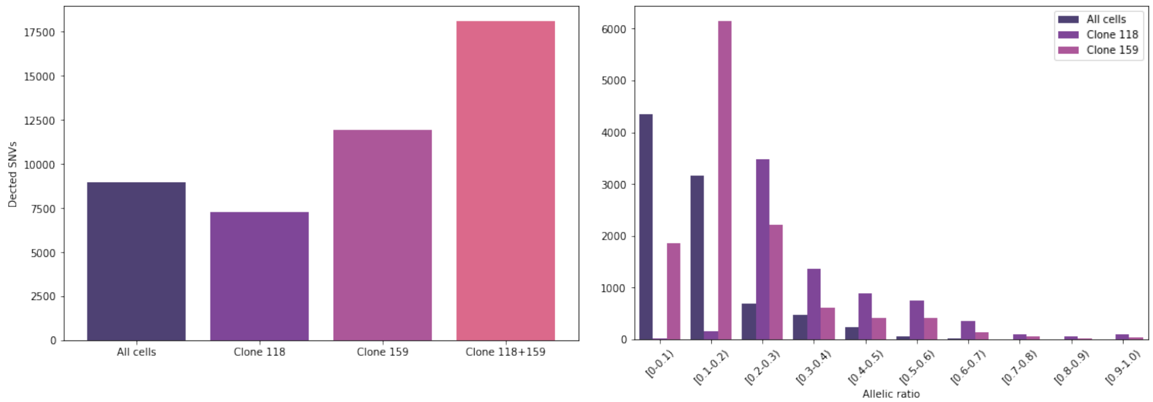

Fig. 7. Left: SNVs detected on Slice B of the breast cancer dataset by running Mutect with the default settings. The middle columns show the number of SNVs reported when running Mutect on each of the two cancer subclones separately, as detected by SECEDO. The last column shows the number of SNVs reported by Mutect when merging the two cancer subclones, while the significantly lower value in the first column represents the number of SNVs reported by Mutect when run against all cells in the tumor slice (including the healthy cells). **Right:** Histogram of the allelic ratio for the SNVs detected by applying MuTect to the two tumor subclones and to all cells in slice B of the breast cancer dataset. The shift to the right in the allelic ratio indicates that the clustering increased the tumor purity.

as demonstrated in four breast cancer datasets (**Figure 6**). It is thus readily applicable to the single-cell DNA sequencing data generated with the 10X Genomics technology. SECEDO is driven by two main key ideas, which are the efficient filtering of uninformative positions as well as taking into account information from overlapping reads. SECEDO performs the clustering with no additional information and, unlike the state-of-the-art method [32], does not require a normal sample for identification of potential mutations. We provide an efficient, well-tested, ready-to-use C++ implementation of SECEDO, which uses established data formats for both input and output, and can thus be easily incorporated into existing bioinformatics pipelines.

We demonstrated SECEDO's applicability to currently available single-cell sequencing data and find that SECEDO correctly clustered cells on a series of synthetic and four breast cancer datasets. CNA frequencies and patterns vary significantly across cancer types [14,43], similarly to SNV frequency. Since SECEDO does not use copy-number information to cluster cells, it can infer sub-clones even in cancer types where CNAs do not vary or where the frequency of CNAs is generally low (e.g. pancreatic neuroendocrine tumors [8]). It is also notable that not all CNAs affect the SNV profile of a cell. Thus, CNA-based clustering may lead to suboptimal grouping of cells, e.g. from a variant calling perspective. SECEDO is able to group cells with similar SNV profiles irrespective of their CNA profiles. This can lead to improvements in the precision and accuracy of the variant calling. Using the clusters identified by SECEDO, we were able to recover 92.11% of the SNVs present in the synthetic data set using a simple variant caller. On Slice B of the breast cancer data set, the number and the confidence of the called SNVs more than doubled after clustering using SECEDO, compared to calling variants on the entire slice.

While SECEDO enables accurate cell-clustering and variant calling, there are a number of areas for future improvement. First, SECEDO currently only uses single-nucleotide substitutions to cluster cells, which are known to be the most common type of mutations in adult and childhood cancers [13,22,28]. We expect that the clustering accuracy could be further improved if e.g. short insertions and deletions were additionally used. Second, the smallest subclones that SECEDO was able to detect had ≈200 cells. However, as technology inevitably improves and the sequencing coverage increases, SECEDO's resolution and variant calling quality will also proportionally increase.

We hope that SECEDO will facilitate new types of analyses and form the basis for future methodological development in the field of cancer research and treatment outcome prognosis.

Acknowledgements

We would like to thank Ximena Bonilla for her constructive feedback on the manuscript. This work was supported by ETH core funding to G.R. (funding D.D., S.S., A.K., KV.L.).

References

1. 10X Genomics: Application note: Assessing tumor heterogeneity with single cell CNV. https://pages.10xgenomics.com/rs/446-PBO-704/images/10x_AN026_SCCNV_Assessing_Tumor%20Heterogeneity_digital.pdf (2018)
2. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al.: Signatures of mutational processes in human cancer. Nature **500**(7463), 415–421 (2013). https://doi.org/10.1038/nature12477
3. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. Technical Report 2006-13, Stanford InfoLab (2006), http://ilpubs.stanford.edu:8090/778/
4. Bohrson, C.L., Barton, A.R., Lodato, M.A., Rodin, R.E., Luquette, L.J., Viswanadham, V.V., Gulhan, D.C., Cortés-Ciriano, I., Sherman, M.A., Kwon, M., et al.: Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nature Genetics **51**(4), 749–754 (2019). https://doi.org/10.1038/s41588-019-0366-2
5. Bryc, K., Patterson, N., Reich, D.: A novel approach to estimating heterozygosity from low-coverage genome sequence. Genetics **195**(2), 553–561 (2013). https://doi.org/10.1534/genetics.113.154500, https://www.genetics.org/content/195/2/553
6. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G.: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology **31**(3), 213–219 (2013). https://doi.org/10.1038/nbt.2514
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39**(1), 1–38 (1977), http://www.jstor.org/stable/2984875
8. Dentro, S.C., Leshchiner, I., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A.G., Yu, K., Rubanova, Y., Macintyre, G., Demeulemeester, J., . . . , Van Loo, P.: Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell **184**(8), 2239–2254.e39 (2021). https://doi.org/https://doi.org/10.1016/j.cell.2021.03.009, https://www.sciencedirect.com/science/article/pii/S0092867421002944
9. Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H.: Substantial biases in ultra-short read data sets from high-throughput dna sequencing. Nucleic Acids Research **36**(16), e105 (2008). https://doi.org/10.1093/nar/gkn425
10. Dong, X., Zhang, L., Milholland, B., Lee, M., Maslov, A.Y., Wang, T., Vijg, J.: Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nature Methods **14**(5), 491–493 (2017). https://doi.org/10.1038/nmeth.4227
11. Durante, M.A., Rodriguez, D.A., Kurtenbach, S., Kuznetsov, J.N., Sanchez, M.I., Decatur, C.L., Snyder, H., Feun, L.G., Livingstone, A.S., Harbour, J.W.: Single-cell analysis reveals new evolutionary complexity in uveal melanoma. Nature Communications **11**(1), 496 (2020)
12. Gawad, C., Koh, W., Quake, S.R.: Single-cell genome sequencing: current state of the science. Nature Reviews Genetics **17**(3), 175–188 (2016). https://doi.org/10.1038/nrg.2015.16
13. Gröbner, S.N., Worst, B.C., Weischenfeldt, J., Buchhalter, I., Kleinheinz, K., Rudneva, V.A., Johann, P.D., Balasubramanian, G.P., Segura-Wang, M., Brabetz, S., et al.: The landscape of genomic alterations across childhood cancers. Nature **555**(7696), 321–327 (2018)
14. Harbers, L., Agostini, F., Nicos, M., Poddighe, D., Bienko, M., Crosetto, N.: Somatic copy number alterations in human cancers: An analysis of publicly available data from the cancer genome atlas. Frontiers in oncology p. 2877 (2021)
15. Huang, W., Li, L., Myers, J.R., Marth, G.T.: ART: a next-generation sequencing read simulator. Bioinformatics **28**(4), 593–594 (2011). https://doi.org/10.1093/bioinformatics/btr708, https://doi.org/10.1093/bioinformatics/btr708
16. Hård, J., Al Hakim, E., Kindblom, M., Björklund, Å.K., Sennblad, B., Demirci, I., Paterlini, M., Reu, P., Borgström, E., Ståhl, P.L., et al.: Conbase: a software for unsupervised discovery of clonal somatic mutations in single cells through read phasing. Genome Biology **20**(1), 68 (2019). https://doi.org/10.1186/s13059-019-1673-8
17. Kelley, D.R., Schatz, M.C., Salzberg, S.L.: Quake: quality-aware detection and correction of sequencing errors. Genome Biology **11**(11), R116 (2010). https://doi.org/10.1186/gb-2010-11-11-r116
18. Kuipers, J., Jahn, K., Beerenwinkel, N.: Advances in understanding tumour evolution through single-cell sequencing. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer **1867**(2), 127–138 (2017). https://doi.org/https://doi.org/10.1016/j.bbcan.2017.02.001, https://www.sciencedirect.com/science/article/pii/S0304419X17300392

19. Lähnemann, D., Köster, J., Fischer, U., Borkhardt, A., McHardy, A.C., Schönhuth, A.: ProSolo: Accurate variant calling from single cell DNA sequencing data. bioRxiv (2020). https://doi.org/10.1101/2020.04.27.064071, https://www.biorxiv.org/content/early/2020/04/28/2020.04.27.064071
20. Laks, E., McPherson, A., Zahn, H., Lai, D., Steif, A., Brimhall, J., Biele, J., Wang, B., Masud, T., Ting, J., et al.: Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. Cell **179**(5), 1207–1221.e22 (2019). https://doi.org/https://doi.org/10.1016/j.cell.2019.10.026, https://www.sciencedirect.com/science/article/pii/S0092867419311766
21. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nature Methods **9**, 357–359 (2012)
22. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., Getz, G.: Discovery and saturation analysis of cancer genes across 21 tumour types. Nature **505**(7484), 495–501 (2014)
23. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al.: Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature **499**(7457), 214–218 (2013). https://doi.org/10.1038/nature12213, https://doi.org/10.1038/nature12213
24. Li, H.: A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**(21), 2987–2993 (2011). https://doi.org/10.1093/bioinformatics/btr509, https://doi.org/10.1093/bioinformatics/btr509
25. Li, H., Ruan, J., Durbin, R.: Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research **18**, 1851–1858 (2008)
26. Lloyd, S.: Least squares quantization in PCM. IEEE Transactions on Information Theory **28**(2), 129–137 (1982)
27. Luquette, L.J., Bohrson, C.L., Sherman, M.A., Park, P.J.: Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. Nature Communications **10**(1), 3908 (2019). https://doi.org/10.1038/s41467-019-11857-8
28. Ma, X., Liu, Y., Liu, Y., Alexandrov, L.B., Edmonson, M.N., Gawad, C., Zhou, X., Li, Y., Rusch, M.C., Easton, J., et al.: Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature **555**(7696), 371–376 (2018)
29. May, A., Abeln, S., Buijs, M.J., Heringa, J., Crielaard, W., Brandt, B.W.: NGS-eval: NGS Error analysis and novel sequence VAriant detection tooL. Nucleic Acids Research **43**(W1), W301–W305 (2015). https://doi.org/10.1093/nar/gkv346, https://doi.org/10.1093/nar/gkv346
30. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al.: A high-coverage genome sequence from an archaic denisovan individual. Science **338**(6104), 222–226 (2012). https://doi.org/10.1126/science.1224344
31. Mu, J.C., Mohiyuddin, M., Li, J., Bani Asadi, N., Gerstein, M.B., Abyzov, A., Wong, W.H., Lam, H.Y.: VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. Bioinformatics **31**(9), 1469–1471 (2014)
32. Myers, M.A., Zaccaria, S., Raphael, B.J.: Identifying tumor clones in sparse single-cell mutation data. Bioinformatics **36**(Supplement_1), i186–i193 (2020). https://doi.org/10.1093/bioinformatics/btaa449, https://doi.org/10.1093/bioinformatics/btaa449
33. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al.: Tumour evolution inferred by single-cell sequencing. Nature **472**(7341), 90–94 (2011)
34. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. p. 849–856. NIPS'01, MIT Press, Cambridge, MA, USA (2001)
35. Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J.L., Mayer, G.: Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports **8**(1), 10950 (2018). https://doi.org/10.1038/s41598-018-29325-6
36. Porter, M.A., Onnela, J., Mucha, P.J.: Communities in networks. Notices of the American Mathematical Society **56**(9), 1082–1097 (2009)
37. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K.: dbSNP: the NCBI database of genetic variation. Nucleic Acids Research **29**(1), 308–311 (2001). https://doi.org/10.1093/nar/29.1.308
38. Singer, J., Kuipers, J., Jahn, K., Beerenwinkel, N.: Single-cell mutation identification via phylogenetic inference. Nature Communications **9**(1), 5144 (2018). https://doi.org/10.1038/s41467-018-07627-7
39. Stratton, M.R., Campbell, P.J., Futreal, P.A.: The cancer genome. Nature **458**(7239), 719–724 (2009). https://doi.org/10.1038/nature07943
40. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al.: COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research **47**(D1), D941–D947 (2018). https://doi.org/10.1093/nar/gky1015

41. Velazquez-Villarreal, E.I., Maheshwari, S., Sorenson, J., Fiddes, I.T., Kumar, V., Yin, Y., Webb, M.G., Catalanotti, C., Grigorova, M., Edwards, P.A., et al.: Single-cell sequencing of genomic dna resolves sub-clonal heterogeneity in a melanoma cell line. Communications Biology **3**(1), 318 (2020). https://doi.org/10.1038/s42003-020-1044-8
42. Zaccaria, S., Raphael, B.J.: Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. Nature Biotechnology **39**(2), 207–214 (2021). https://doi.org/10.1038/s41587-020-0661-6
43. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.Z., Wala, J., Mermel, C.H., et al.: Pan-cancer patterns of somatic copy number alteration. Nature genetics **45**(10), 1134–1140 (2013)
44. Zafar, H., Wang, Y., Nakhleh, L., Navin, N., Chen, K.: Monovar: single-nucleotide variant detection in single cells. Nature Methods **13**(6), 505–507 (2016). https://doi.org/10.1038/nmeth.3835