


METHOD

Open Access



SIEVE: joint inference of single-nucleotide variants and cell phylogeny from single-cell DNA sequencing data

Senbai Kang¹, Nico Borgsmüller^{2,3}, Monica Valecha^{4,5}, Jack Kuipers^{2,3}, Joao M. Alves^{4,5}, Sonia Prado-López^{4,5,6}, Débora Chantada⁷, Niko Beerenwinkel^{2,3}, David Posada^{4,5,8} and Ewa Szczurek^{1*} 

*Correspondence:
szczurek@mimuw.edu.pl

¹ Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland

² Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland

³ SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland

⁴ CINBIO, Universidade de Vigo, 36310 Vigo, Spain

⁵ Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain

⁶ Institute of Solid State Electronics E362, Technische Universität Wien, Vienna, Austria

⁷ Department of Pathology, Hospital Álvaro Cunqueiro, Vigo, Spain

⁸ Department of Biochemistry, Genetics, and Immunology, Universidade de Vigo, 36310 Vigo, Spain

Abstract

We present SIEVE, a statistical method for the joint inference of somatic variants and cell phylogeny under the finite-sites assumption from single-cell DNA sequencing. SIEVE leverages raw read counts for all nucleotides and corrects the acquisition bias of branch lengths. In our simulations, SIEVE outperforms other methods in phylogenetic reconstruction and variant calling accuracy, especially in the inference of homozygous variants. Applying SIEVE to three datasets, one for triple-negative breast (TNBC), and two for colorectal cancer (CRC), we find that double mutant genotypes are rare in CRC but unexpectedly frequent in the TNBC samples.

Keywords: Single-cell DNA sequencing, Statistical phylogenetic models, Cell phylogeny reconstruction, Somatic variant calling, Finite-sites assumption, Acquisition bias correction

Background

Intra-tumour heterogeneity is a consequence of accumulated somatic mutations during tumour evolution [1, 2] and the culprit of acquired resistance and relapse in clinical cancer therapy [3, 4]. Phylogenetic inference is a powerful tool to understand the development of intra-tumour heterogeneity in time and space. Variant allele profiles derived from bulk sequencing data have typically been used to reconstruct the tumour phylogeny at the level of clones [5–9]. More recently, the development of single-cell DNA sequencing (scDNA-seq) [10–12] has enabled single-nucleotide variant (SNV) calling [13–18] and phylogeny reconstruction [15, 19–26] down to the single-cell level.

A statistical phylogenetic model is defined by an instantaneous transition rate matrix, a tree topology and tree branch lengths. Such a model defines a Markov process for the evolution of nucleotides or genotypes [27]. Studying the evolutionary process and estimating important parameters such as the branch lengths using statistical



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

phylogenetic models has a long tradition, benefits from well established theory, and has many applications, such as interpreting temporal cell dynamics [28].

However, compared to statistical phylogenetic models, most methods for phylogeny reconstruction from scDNA-seq operate within a simpler modelling framework. First, although branch lengths are a critical part of a phylogenetic tree and reflect the real evolutionary distances among cells, they are often ignored. Those approaches that do infer branch lengths [22, 26] employ the data from the variant sites and ignore information from *background sites* (that have a wildtype genotype), which may lead to so-called acquisition bias and overestimated branch lengths [29, 30].

Moreover, variant calling and phylogenetic inference are commonly considered independent tasks. Variant calling is typically performed first, and phylogenetic inference is performed on the called variants. However, variant calling, particularly from scDNA-seq data, can be hampered by missing data and low coverage, potentially resulting in wrong calls that could mislead phylogenetic inference. A feasible strategy to alleviate this problem is to integrate tree reconstruction with variant calling [12], where phylogenetic information on cell ancestry is used to obtain more reliable variant calls. Recently developed methods for scDNA-seq data approach this strategy from different perspectives [15, 31]. However, those methods do not operate within the statistical phylogenetic framework, in particular do not infer branch lengths of the tree. Moreover, either they fully follow the infinite-sites assumption (ISA), which is often violated in real datasets [32, 33], or relax this assumption to only a limited extent. As a result, they may miss important events in the evolution of tumours. Thus, methods have not yet been developed which, employing statistical phylogenetic models under the finite-sites assumption (FSA), infer cell phylogeny from raw scDNA-seq data and simultaneously call variants.

To address this, we propose SIEVE (Single-cell EVolution Explorer), a statistical method that exploits raw read counts for all nucleotides from scDNA-seq to reconstruct the cell phylogeny and call variants based on the inferred phylogenetic relations among cells. To our knowledge, SIEVE is the first approach that employs a statistical phylogenetic model following FSA, where branch lengths, measured by the expected number of somatic mutations per site, are corrected for the acquisition bias using the data from the background sites, and simultaneously calls variants and allelic dropout (ADO) states from raw read counts data. SIEVE incorporates solutions tailored for scDNA-seq tumour data. First, it includes a trunk in the tree structure, representing the branch joining the healthy root to the most recent common ancestor (MRCA) of the subpopulation of the analysed cells. As such, the model captures the early, important gene mutations, common for all cells in the trunk. Second, it employs a dedicated probabilistic model of the raw nucleotide read counts at the modelled sites, and discerns between single and double mutations at these sites. Thanks to its flexibility, the model is able to detect 12 different types of genotype transitions, corresponding to nine types of events in evolutionary history. SIEVE is implemented and available as a package of BEAST 2, which allows for benefiting from other packages in this framework. Using simulated data, we assess the performance of our model in comparison to existing methods. To illustrate the functionality of SIEVE, we apply it to datasets from two patients with CRC and one with TNBC.

Results

SIEVE is a statistical method for joint inference of SNVs and cell phylogeny from scDNA-seq data

SIEVE takes as input raw read count data at candidate SNV sites, accounting for the read counts for three alternative nucleotides and the total depth at each site (Fig. 1a)

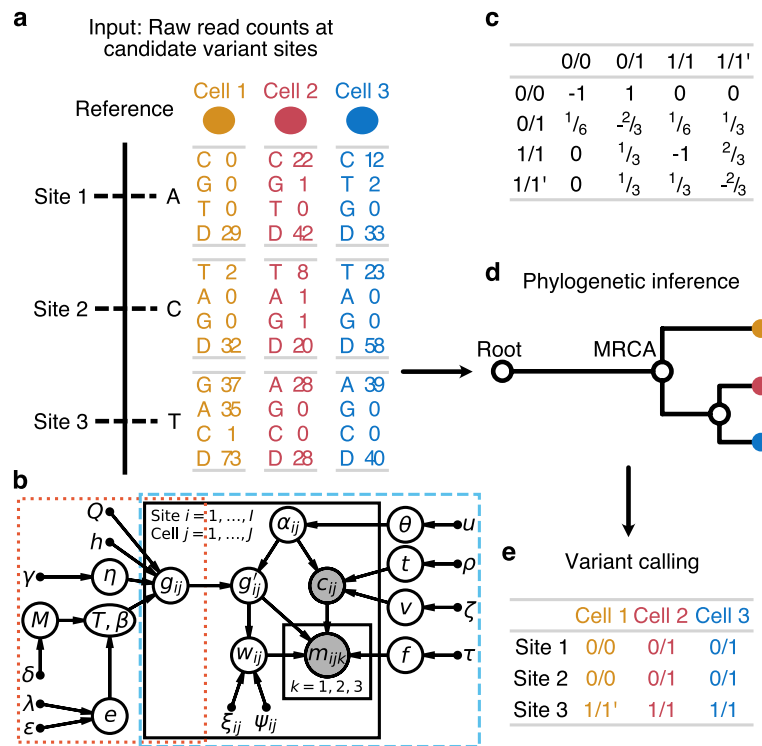


Fig. 1 Overview of the SIEVE model. **a** Input data to SIEVE at candidate SNV sites. For a specific cell at an SNV site, fed to SIEVE are the read counts for all nucleotides: reads of the three alternative nucleotides with values in descending order and the total coverage (denoted by D in **a**). **b** Graphical representation of the SIEVE model. Bridged by g_{ij} , the genotype for site i in cell j , the orange dotted frame encloses the statistical phylogenetic model, and the blue dashed frame highlights the model of raw read counts. Shaded circle nodes represent observed variables, while unshaded circle nodes represent hidden random variables. Small filled circles correspond to fixed hyper parameters. Arrows denote local conditional probability distributions of child nodes given parent nodes. The sequencing coverage c_{ij} follows a negative binomial distribution parameterised by the number of sequenced alleles α_{ij} , the mean of allelic coverage t and the variance of allelic coverage v . α_{ij} is a hidden categorical variable parameterised by ADO rate θ , which has a uniform prior with fixed hyper parameter u . t also has a uniform prior with fixed parameter ρ , while v has an exponential prior parameterised by ζ . The nucleotide read counts m_{ij} given c_{ij} follow a Dirichlet-multinomial distribution parameterised by ADO-affected genotype g'_{ij} , which is a hidden random variable depending on α_{ij} and genotype g_{ij} , effective sequencing error rate f , which has an exponential prior with fixed hyper parameter τ , and overdispersion w_{ij} , which is a hidden categorical variable dependent on g'_{ij} parameterised by fixed parameters ξ_{ij} and ψ_{ij} for each category. g_{ij} is determined by the statistical phylogenetic model parameterised by fixed rate matrix Q , fixed number of categories h as well as shape parameter η with exponential prior for site-wise substitution rates, and tree topology \mathcal{T} along with branch lengths β . \mathcal{T} and β have a coalescent prior with an exponentially growing population parameterised by effective population size M , which has a multiplicative inverse prior, and growth rate e , which has a laplace prior parameterised by λ and ϵ . **c** The transition rate matrix in the statistical phylogenetic model. During an infinitesimal time interval only one change is allowed to occur. **d** The cell phylogeny inferred from the data with SIEVE. Not only is the tree topology crucial, but also the branch lengths. The root represents a normal cell, and the only direct child of the root is the most recent common ancestor (MRCA) of all cells. **e** Variant calling given the inferred cell phylogeny. For further details, see the “Methods” section

and combines a statistical phylogenetic model with a probabilistic graphical model of the read counts, incorporating a Dirichlet Multinomial distribution of the nucleotide counts (Fig. 1b; [Methods](#)). The statistical phylogenetic model allows for acquisition and loss of mutations on both maternal and paternal alleles (Fig. 1c). It considers four possible genotypes, 0/0 (referred to as *wildtype*), 0/1 (*single mutant*), 1/1 (*double mutant*, where the two alternative nucleotides are the same) and 1/1' (*double mutant*, where the two alternative nucleotides are different). With these genotypes, SIEVE is able to discern 12 different types of genotype transitions, which can be categorised into nine types of mutation events, namely single mutation, homozygous coincident double mutation, heterozygous coincident double mutation, single back mutation, double back mutation, homozygous single mutation addition, heterozygous single mutation addition, homozygous substitute single mutation, and heterozygous substitute single mutation (Table 1; [Methods](#)). Based on the inferred tree (Fig. 1d), SIEVE calls the maximum likelihood somatic mutations (Fig. 1e). With these calls and the recognised mutation events on the branches of the tree, we detect parallel evolution in the case when the same event re-occurs on independent branches of the tree. The tree contains a trunk joining the root representing a healthy cell with the most recent common ancestor (MRCA) of the modelled cells, representing the acquisition of clonal mutations at the initial stage of tumour progression. SIEVE leverages the noisy raw read counts to integrate genotype uncertainty into cell phylogeny inference. Benefiting from the inferred cell relationships, SIEVE is able to reliably infer the single-cell genotypes, especially for sites where only few reads are available. SIEVE is implemented as a package of BEAST 2, a flexible and mature framework for statistical phylogenetic modelling [34].

We investigated the performance of SIEVE using simulated data with different means and variances of allelic coverage, reflecting different *coverage qualities* ([Methods](#)). Specifically, we simulated data with low mean and high variance of allelic coverage (low quality), with high mean and medium variance (medium quality), and with high mean

Table 1 12 types of genotype transitions that SIEVE is able to identify, with their interpretation as mutation events. The genotype transitions correspond to possible changes of genotypes on a branch from the parent node to the child node. If any of these events occurs on independent branches of the phylogenetic tree, it is also considered as a parallel evolution event. For detailed explanations of the mutation events, see the “[Methods](#)” section

Genotype transition	Mutation event
0/0 → 0/1	Single mutation
0/0 → 1/1	Homozygous coincident double mutation
0/0 → 1/1'	Heterozygous coincident double mutation
0/1 → 0/0	Single back mutation
1/1 → 0/1	Single back mutation
1/1' → 0/1	Single back mutation
1/1 → 0/0	Double back mutation
1/1' → 0/0	Double back mutation
0/1 → 1/1	Homozygous single mutation addition
0/1 → 1/1'	Heterozygous single mutation addition
1/1' → 1/1	Homozygous substitute single mutation
1/1 → 1/1'	Heterozygous substitute single mutation

and low variance (high quality). Other important dataset characteristics were varied, including the number of cells and mutation rate, which is measured by the number of somatic mutations per site per generation.

SIEVE accurately estimates tree topology and branch lengths

We first evaluated the accuracy of SIEVE in inferring the simulated cell phylogeny with branch lengths using the branch score (BS) distance [35] (Fig. 2a; [Methods](#)). We compared to CellPhy [26] and SiFit [22], which were fed with the variant calls from Monovar [13]. Here, we gave SiFit an advantage of setting the true positive error rate used in the simulation ([Methods](#)). Thanks to the acquisition bias correction, SIEVE reports branch lengths as expected number of somatic mutations per site, while CellPhy and SiFit per SNV site. SCIPHI [15] does not infer branch lengths, hence its BS distance could not be computed. SIEVE consistently outperformed CellPhy and SiFit, regardless of the number of cells, mutation rate and coverage quality. This may be because, in contrast to SIEVE, CellPhy and SiFit do not model raw reads and, importantly for the BS distance, do not correct the inferred branch lengths for acquisition bias. We also found that the BS distance of SIEVE had a negative nonlinear association with the number of background sites (Additional file 1: Fig. S1), explaining the relatively greater differences under higher mutation rates. These results proved the necessity for correcting the acquisition bias with enough background sites to obtain accurate branch lengths.

As the BS distance is dominated by the branch lengths, we further assessed SIEVE's accuracy in inferring the tree structure using the normalised Robinson-Foulds (RF) distance [36]. Compared to CellPhy, SiFit and SCIPHI (Fig. 2b; [Methods](#)), SIEVE was the most robust method to changes of mutation rate, number of cells and coverage quality. When the data hardly contained mutations violating the ISA (mutation rate being 10^{-6} , with less than 0.1% double mutant genotypes and at most 1% SNV sites with parallel mutations), all methods achieved a similar median RF distance (around 0.15–0.3). Since in contrast to SCIPHI, SIEVE, CellPhy and SiFit employ statistical phylogenetic models following FSA, this indicates that models following FSA are also applicable to data evolving under the ISA. SIEVE outperformed CellPhy and SiFit when the number of cells and the mutation rate increased. When the data clearly violated the ISA (mutation rates being 8×10^{-6} and 3×10^{-5} , with 0.02–0.3% and 0.1–1% double mutant genotypes, as well as 2–8% and 10–27% SNV sites with parallel mutations indicative of FSA, respectively), SCIPHI inferred reasonable tree topologies from datasets with a small number of cells (40). However, its performance dramatically dropped with 100 cells, especially when the data was of medium or high coverage quality. The behaviour of SCIPHI might be related to its estimation of ADO rate and single mutant genotype calling in these scenarios.

SIEVE accurately infers parameters in the model of raw read counts

We next investigated the accuracy of parameter estimates, including *effective* sequencing error rate, ADO rate, and wildtype and alternative overdispersion (Additional file 1: Fig. S2; [Methods](#)). Here, the effective sequencing error rate (Additional file 1: Fig. S2a) takes into account both amplification and sequencing error rates in scDNA-seq. Wildtype and alternative overdispersion are parameters in the distribution of

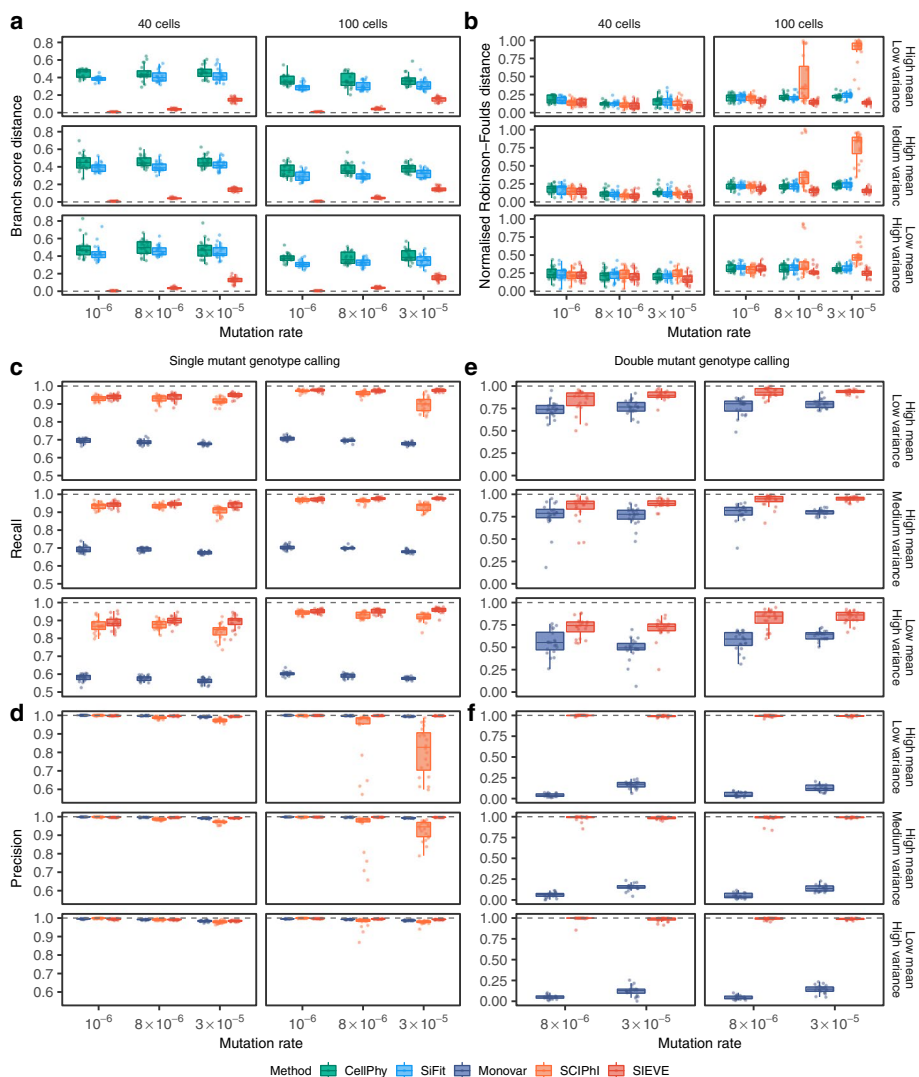


Fig. 2 Benchmarking result of the SIEVE model. Varying are the number of tumour cells, mutation rate and coverage quality. Each simulation is repeated $n = 20$ times with each repetition denoted by coloured dots. The grey dashed lines represent the optimal values of each metric. Box plots comprise medians, boxes covering the interquartile range (IQR), and whiskers extending to 1.5 times the IQR below and above the box. **a, b** Box plots of the tree inference accuracy measured by the BS distance where the branch lengths are taken into account (**a**) and the normalised RF distance where only tree topology is considered (**b**). **c, d** Box plots of the single mutant genotype calling results measured by the fraction of true positives respectively in the ground truth positives, i.e. the sum of true positives and false negatives, (recall, **c**) as well as in the predicted positives, i.e. the sum of true positives and false positives, (precision, **d**). **e, f** Box plots of the double mutant genotype calling results measured by recall (**e**) and precision (**f**), where the variant calling results when mutation rate is 10^{-6} are omitted as very few double mutant genotypes are generated (less than 0.1%)

nucleotide read counts related to different genotypes. The former corresponds to genotype 0/0 and 1/1, while the latter to genotype 0/1 and 1/1'. SIEVE accurately inferred most parameters in all simulated scenarios regardless of the number of cells, mutation rate and coverage quality. Although SIEVE's accuracy of estimating ADO rate slightly decreased with the coverage quality, it still was the best among the competing

methods. For data with medium and high coverage quality, 100 cells and higher mutation rates (8×10^{-6} and 3×10^{-5}), SCIPhI tended to overestimate ADO rates.

SIEVE accurately calls single and double mutations

Next, we assessed SIEVE's performance in calling the single mutant genotype (Fig. 2c, d, Additional file 1: Figs. S3a,b and S4; Methods). As opposed to Monovar, recall for SIEVE and SCIPhI increased with the number of cells but was less sensitive to the coverage quality (Fig. 2c). The recall of SIEVE was higher than that of SCIPhI by 0.16–18.55% and that of Monovar by 28.89–71.74%. Unlike Monovar, both SIEVE and SCIPhI benefit from the information provided by cell phylogenies. We speculate that the advantage of SIEVE over SCIPhI stems from the use of raw read counts for all nucleotides, while SCIPhI only employs the sequencing coverage and the read count of the most prevalent alternative nucleotide.

Moreover, SIEVE and Monovar achieved comparable precision (Fig. 2d) and false positive rates (Additional file 1: Fig. S3a) regardless of the number of cells, mutation rate and coverage quality. However, this did not hold for SCIPhI. By analysing the types of false positives among the predicted single mutant genotypes (Additional file 1: Fig. S4; Methods), we found that SCIPhI tended to miscall wildtype genotypes as single mutant genotype (i.e. 0/0 are called as 0/1) (Additional file 1: Fig. S4a). This occurred with high mutation rates (8×10^{-6} and 3×10^{-5}), especially in scenarios where SCIPhI inferred inaccurate trees (Fig. 2b) and overestimated ADO rates (Additional file 1: Fig. S2b). The reason is twofold. First, the ISA upon which SCIPhI builds naturally limits its application to data following FSA. Second, under these scenarios, SCIPhI tends to mistake sites with no variant support for ADO events, and hence its high ADO rate. SIEVE avoids such mistakes by leveraging a model of sequencing coverage (Methods), thereby accounting for the related overdispersion and correctly estimating the ADO rate. We also noticed that when data clearly violated ISA, both Monovar and SCIPhI miscalled more double mutant genotypes as the single mutant genotype than SIEVE (Additional file 1: Fig. S4b).

We then focused on the results of double mutant genotype calling (Fig. 2e, f, Additional file 1: Fig. S3c,d; Methods), where SCIPhI was excluded as it is unable to call such mutations. The recall of double mutant genotypes for SIEVE and Monovar increased with the number of cells and the coverage quality (Fig. 2e), while SIEVE showed higher recall for such genotypes than Monovar. Moreover, SIEVE outperformed Monovar with high precision (almost 1, Fig. 2f) and low false positive rate (almost 0, Additional file 1: Fig. S3c).

SIEVE accurately calls ADOs for data of adequate coverage quality

We further assessed SIEVE's performance in ADO calling (Additional file 1: Fig. S5), where there are no published methods for us to compare with. When calling ADOs, SIEVE's performance was independent of the number of cells or mutation rate, but highly dependent on the coverage quality. The reason is that SIEVE calls ADOs by inferring the number of sequenced alleles, assuming it is proportional to the observed sequencing coverage (Methods). Consequently, for data with medium and high coverage quality the average F1 score of ADO calling was high (0.86 and 0.93, respectively), whereas for data with low coverage quality, which is typical for current scDNA-seq data,

the ADO calling performance deteriorated, with average F1 score being only 0.10. Since the coverage quality of real data is low, we do not report ADO calling results for all real datasets analysed below (Additional file 1: Table S1).

SIEVE accurately infers cell phylogenies and calls variants in the presence of copy number aberrations (CNAs)

Both SIEVE and two compared methods, CellPhy and SCIPhI, work with the assumption that the genomes of the cells are diploid. SiFit allows deletions, thus considering copy number 1 or 2. Occurrences of CNAs change the copy number for some of the sites. Leaving such sites in the data introduces discrepancy with the assumption, but may give more statistical power for model inference. To investigate the degree to which the performance of SIEVE and other models is affected by CNAs, we considered simulation scenarios where both deletions and amplifications were added, by changing the copy number to any integer from the [0, 10] interval that is different than 2 (Methods). We varied the amount of genomic sites having CNAs in either small or large amount ($\frac{1}{3}$ or $\frac{2}{3}$ of all sites, respectively), and all methods were run both with CNA sites included and excluded from the input data.

The presence of CNAs had very little effect on the performance of inferring the simulated cell phylogeny with branch lengths by all evaluated methods. Indeed, the BS distances obtained by the methods were at a similar level, regardless of the presence of the CNAs and their amount (Additional file 1: Fig. S6a). In contrast, the presence of CNAs worsened the performance of all methods in the task of inferring the topology of phylogeny, as measured by the normalised RF distance. When the CNAs were present in a small amount, the normalised RF distance for all methods was only slightly increased, regardless of the inclusion of the CNA sites or their exclusion from the input data. In the case when the CNAs were present in a large amount, the normalised RF distance increased stronger and the methods visibly benefited from including CNA sites, as they suffered from insufficient information when the CNA sites were excluded (Additional file 1: Fig. S6b).

In terms of inferring the single mutant genotype, the recall and precision of SIEVE and SCIPhI were not affected much by the presence of CNA sites, regardless of their amount and inclusion or exclusion from the data. In contrast, these measures decreased for Monovar, deteriorating most strongly when CNAs were present in large amounts and included in the data (Additional file 1: Fig. S7a,b). The existence of CNA sites had little influence on the false positive rate of SCIPhI, and only slightly increased the false positive rates of Monovar and SIEVE (Additional file 1: Fig. S7c). The F1 scores of SIEVE and SCIPhI were invariant to the CNA sites, while that of Monovar dropped proportionally to the amount of CNAs in the case when they were included in the data (Additional file 1: Fig. S7d). For inferring double mutant genotypes, adding CNAs had very little impact on the performance of both SIEVE and Monovar (Additional file 1: Fig. S8).

Overall, although assuming a diploid genome, SIEVE is robust to the existence of CNA sites in the input data for both inferring cell phylogeny and calling variants. For phylogeny inference using SIEVE it is rather desirable to potentially increase statistical power and include all sites in the data, even if they were affected by CNAs.

SIEVE achieves favourable run times and low memory usage in the default, multi-thread mode

We further evaluated the run times and memory requirements of SIEVE and other approaches (Additional file 1: Fig. S9 and Table S2; [Methods](#)). While SIEVE in single thread mode was not competitive, it achieved stellar run time performance in the default, multi-thread mode. In particular, SIEVE outperformed other Bayesian methods and was similar in run time performance as compared to CellPhy, a model based on maximum likelihood inference and using bootstrap to estimate node support. With the increase of the number of both cells and sites, the run time of SIEVE in the multi-thread mode increased much slower compared to other methods. This indicates that SIEVE is scalable to large number of cells and sites. In terms of memory usage, all methods performed similarly well, except for SiFit, which required tremendous amounts of memory.

SIEVE inferred a phylogenetic tree and called variants for CRC cells

We applied SIEVE to a new single-cell whole genome sequencing (scWGS) dataset, where 28 tumour cells were isolated from three primary tumour biopsies of a patient with CRC (CRC28; see the [“Methods”](#) section). We identified 8470 candidate SNV sites and 1,163,335,103 background sites. To take into account branch-wise substitution rate variation, we employed a relaxed molecular clock model [37] (same for the following datasets; see the [“Methods”](#) section). In the inferred maximum clade credibility (MCC)

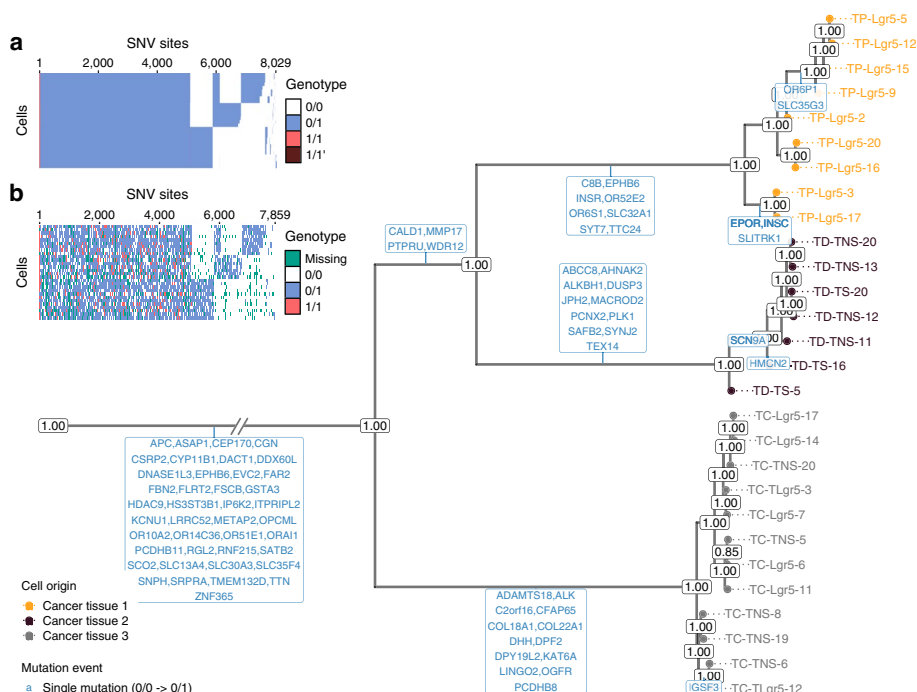


Fig. 3 Results of phylogenetic inference and variant calling for the CRC28 dataset. Shown is SIEVE’s maximum clade credibility tree. The exceptionally long trunk has been folded (marked by slashes). Cells are coloured according to the corresponding biopsies. The numbers at each node represent posterior probabilities (threshold $p > 0.5$). At each branch, genes with non-synonymous mutations are depicted in blue. **a, b** Variant calling heatmap for SIEVE (**a**) and Monovar (**b**). Listed in the legend are the categories of predicted genotypes by each method. Cells in the row are in the same order as that of leaves in the phylogenetic tree

tree (Fig. 3; see Additional file 1: Fig. S10 for the branch lengths), tumour cells grouped into three highly supported clades corresponding to the three biopsies. The average length of the branches was 4.2×10^{-7} . The estimated effective sequencing error and ADO rates were 7.6×10^{-4} and 0.20, respectively.

Among the trees obtained by other methods (Additional file 1: Fig. S11), the tree obtained by CellPhy was the most similar to the one by SIEVE and also the closest in terms of normalised RF and BS distance (Additional file 1: Fig. S12). Although all methods grouped tumour proximal (TP) cells identically as an independent subclone, SCIPhI and SiFit clustered tumour distal (TD) and tumour central (TC) cells distinctly. Both SIEVE and CellPhy agreed that TP and TD cells were closer than TC cells during the evolutionary history. The fact that the different biopsies form well-supported clades exposes a strong geographical clonal structure suggesting regular growth and limited cell migration. From the four compared models, only SIEVE and CellPhy reported node support values, giving clear intuitions about the confidence for each clade.

We mapped non-synonymous mutations to the internal branches (Methods), where only single mutations were found, indicating that the mutational process likely followed the ISA. Many mutations resided on the trunk (clonal mutations), including established CRC driver genes [38, 39], such as *APC*, as well as genes related to the metastatic progression of CRC [40, 41], such as *ASAP1* and *RGL2*. For all mapped genes, SIEVE identified only one type of mutation event, i.e. single mutations that correspond to the switch of the genotype from 0/0 to 0/1. The lack of other mutation events that are possible to identify using our model (see Table 1) indicates that for this sample the model did not detect any violations of the ISA.

SIEVE identified 8029 SNV sites among the candidate SNV sites (Fig. 3a), where most of the genotypes were single mutant and few were double mutant, including 1/1'. The variant calling results of SIEVE and Monovar (Fig. 3b) were overall similar. However, the calls from Monovar were clearly more noisy, with many missing entries and more double mutant genotypes, some of which might be false positives according to the simulation results. The proportion of genotypes called by SIEVE and Monovar were summarised in Additional file 1: Table S3 (same for the following datasets).

SIEVE inferred a phylogenetic tree and called variants for TNBC cells

We then applied SIEVE to a single-cell whole exome sequencing (scWES) dataset [42], containing 16 tumour cells collected from a patient with TNBC (TNBC16; see the "Methods" section). We identified 5912 candidate SNV sites and 152,027,822 background sites. The estimated tree was supported by high posterior probabilities (Fig. 4) with a relatively long trunk and short terminal branches (Additional file 1: Fig. S13). The average branch length was 4.6×10^{-6} . We estimated that the effective sequencing error rate was 8.2×10^{-4} and the ADO rate was 0.05.

SCIPhI and CellPhy returned trees that were similar in structure to the one obtained by SIEVE (Additional file 1: Fig. S14), where the tree inferred by SCIPhI was the closest to that inferred by SIEVE (Additional file 1: Fig. S15a) in terms of normalised RF distance and the one inferred by CellPhy was the closest in terms of the BS distance (Additional file 1: Fig. S15b). Finally, the tree obtained by SiFit was the least similar to all other methods.

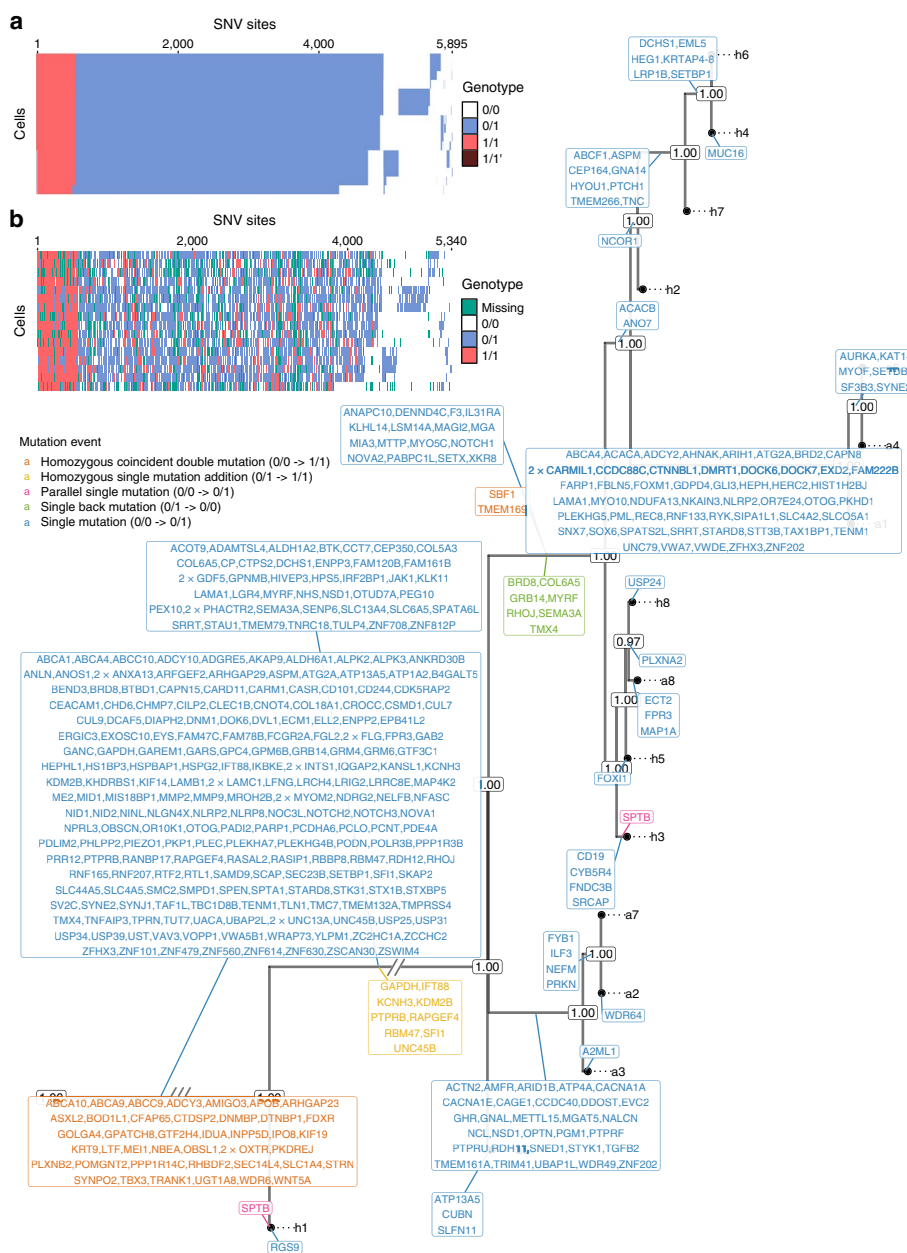


Fig. 4 Results of phylogenetic inference and variant calling for TNBC16 [42] dataset. Shown is SIEVE's maximum clade credibility tree. Two exceptionally long branches are folded with the number of slashes proportional to the branch lengths. Tumour cell names are annotated to the leaves of the tree. The numbers at each node represent the posterior probabilities (threshold $p > 0.5$). At each branch, genes with non-synonymous mutations are depicted in different colours, representing various types of mutation events. **a, b** Variant calling heatmap for SIEVE (**a**) and Monovar (**b**). Listed in the legend are the categories of predicted genotypes by each method. Cells in the row are in the same order as that of leaves in the phylogenetic tree

While for the previous CRC28 dataset the events identified by SIEVE consisted solely of single mutations (transitions from 0/0 to 0/1 genotype), which are typically analysed and often detected by other methods, the TNBC16 dataset is the showcase of SIEVE's ability to detect more diverse types of mutation events. By mapping

non-synonymous mutations to the internal branches, we identified five different types of mutation events (Methods), including several violations of the ISA, such as back mutations and parallel mutations. These, apart from the standard single mutations, included 44 homozygous coincident double mutations (transitions from 0/0 to 1/1 genotype), nine homozygous single mutation additions (from 0/1 to 1/1 genotype), two parallel single mutations (from 0/0 to 0/1 genotype that occurred more than once on the tree), and seven single back mutations (from 0/1 to 0/0 genotype). Demeulemeester et al. [33] suggested that single back mutation events might occur due to retained mutability of the variant allele, thus making it likely to be mutated again. An alternative explanation for single back mutations could be an occurrence of a loss of heterozygosity. Other events violating the ISA might be due to mutational hotspots and hypermutable motifs [33]. As expected, most of the mutations, including single and double mutant genotypes, resided on the trunk, and some of them occurred in genes which were also reported in the original study [42], such as *TBX3*, *NOTCH2*, *NOTCH3* and *SETBP1*. In the original study, the evolutionary tree of SNVs was reconstructed using hierarchical clustering. Unfortunately, clustering is not a phylogenetic method based on shared ancestry, and assumes ultrametricity (perfect clock). In contrast to hierarchical clustering, our approach gives more insights into the evolutionary history of the tumour. In particular, it infers the error rates, categorises the types of the mutation events that occurred, and gives posterior estimates for the nodes (the node supports). The high support values (Fig. 4) indicate that the tree inferred by SIEVE is highly plausible.

SIEVE identified 5,895 SNV sites (Fig. 4a). In contrast to Monovar, SIEVE calls genotypes for all analysed sites, including sites with missing data (Fig. 4b).

SIEVE inferred a phylogenetic tree and called variants for CRC samples mixed with normal cells

Finally, we applied SIEVE to another scWES dataset [43], which consisted of 35 tumour and normal cells as well as 13 adenomatous polyp cells from a patient with CRC (CRC0827 in [43]; referred to as CRC48 below; see the “Methods” section). The tumour cells came from two distinct anatomical locations (cancer tissue 1 and 2). We identified 707 candidate SNV sites as well as 119,486,190 background sites. From the inferred phylogenetic tree (Additional file 1: Figs. S16-S17), we identified two tumour clades matching their anatomical locations and one clade for adenomatous polyp and normal cells. Nine cells collected from the tumour biopsies were clustered outside the tumour clades, suggesting that these were normal cells within the tumour biopsies, which was also pointed out in the original study. The average branch length of the inferred tree was 2.1×10^{-7} . We estimated that the effective sequencing error rate was 8.3×10^{-4} and the ADO rate was 0.10.

Other methods reported distinct trees (Additional file 1: Figs. S18-S21), which might result from the relatively insufficient number of (candidate) variant sites as input. CellPhy, SCIPHi and SiFit were also able to distinguish the same set of normal cells from tumour cells. However, SiFit was unable to group tumour cells into two clades matching their anatomical locations as well as SIEVE.

From the non-synonymous mutations mapped to the branches, we observed unique subclonal mutations, including an established CRC driver mutation, *SYNE1* [39]. In addition to multiple single mutation events, we located two parallel single mutations (*CHD3* and *PLD2*), which evolved independently in adenomatous polyps and in tumour cells. Moreover, a mutated gene, *MLH3*, known being related to DNA mismatch repair [44], was found on the branch leading to the tumour subclone. This might be one of the reasons why this phylogenetic tree demonstrates a strong imbalance of branch lengths, with much longer branches found in the tumour subtree.

The variant calling results of SIEVE shared a similar but less noisy structure to those of Monovar (Additional file 1: Fig. S16a,b). We identified 678 SNV sites in total.

Discussion

Here we present a statistical approach for cell phylogeny inference and variant calling from scDNA-seq data. SIEVE leverages raw read counts to directly reconstruct cell phylogenies and then to reliably call single-cell variants. SIEVE tackles a considerably challenging problem, i.e. the propagation of errors in variant calling to the inference of cell phylogeny, by sharing information between these two tasks. Important characteristics of SIEVE include accounting for the FSA and correction for acquisition bias for tree branch lengths, which prevents from overfitting the phylogenetic model, and, finally, modelling the trunk of the evolutionary tree accommodating the events that are common for all cells.

Inferring mutation status accurately from highly noisy scDNA-seq data remains a demanding problem. A pivotal strength of SIEVE is its characteristic of using genotypes as a bridge between tree inference and variant calling so that these tasks are united. SIEVE is able to reliably differentiate wildtype, single and double mutant genotypes. The benchmarking shows that SIEVE, regarding variant calling, outperforms methods which employ no cell relationships (Monovar) and which, despite accounting for such information, do not include an instantaneous transition rate matrix and branch lengths (SCIPhI). Regarding tree reconstruction, SIEVE is more robust than SCIPhI, which infers phylogenies following ISA from raw scDNA-seq data. It also outperforms methods that rely on variants called by other approaches as a pre-processing step, thereby likely being misled by wrongly inferred variants (CellPhy and SiFit). The high performance of SIEVE can also be attributed to the fact that it is the only model that performs acquisition bias correction, allowing for more accurate branch lengths, and models the distribution of sequencing coverage and accounting for its overdispersion. Finally, SIEVE is also able to reliably call ADOs given data of adequate coverage quality.

Although MCMC is employed in the inference, our results show that SIEVE is an efficient method regarding both run time and memory consumption in the default, multi-thread mode. It also has the potential of favourable scalability to large numbers of cells and sites, where the latter is particularly relevant to the inference of accurate cell phylogenies. Naturally, the more candidate variant sites are available, the more statistical power they confer.

Currently, SIEVE only considers SNVs and assumes a diploid genome. Further improvement could embrace small indels and CNAs to improve phylogenetic inference and variant calling, yet care must be taken to differentiate deletions during evolution

from ADOs. Additionally, SIEVE only allows at most one ADO for each site and cell. Further extension could expand to locus dropout, which directly results in missing data.

We apply SIEVE to real scDNA-seq datasets harnessed from CRC and TNBC. SIEVE calls far fewer double mutant genotypes and gives more reliable mutation assignment than Monovar does, in line with the simulation results. We also notice that SIEVE identifies double mutant genotypes, which is rare in CRC but frequent in TNBC, indicating the noteworthy role such genotypes play in the evolution of different types of cancer. Future studies could be based on the phylogenetic tree and variants inferred by SIEVE to identify somatic mutations potentially related to the resistance and relapse in the clinical therapy of cancer. SIEVE can also be applied to targeted sequencing data, where a user-defined number of background sites could be specified for acquisition bias correction. Moreover, SIEVE's applicability is not restricted to cancer samples, and it can also be used to trace lineages of healthy cells.

In the real data analysis we utilise the relaxed molecular clock model implemented in BEAST 2. This shows one of the advantages of SIEVE being a package of BEAST 2, and the potential of exploiting the functionality of other BEAST 2 packages in our model.

Conclusions

The SIEVE model successfully exploits raw read counts from scDNA-seq data and jointly infers phylogeny and variants. Our comprehensive simulations show that SIEVE can produce reliable cell phylogeny and somatic variants, facilitating the downstream analysis. With the advancement of scDNA-seq technology, we expect the improvement of the coverage quality where the inference of ADO states is reliable. Although we mainly illustrate the application of SIEVE to scDNA-seq data from tumours, it is applicable to studying evolution also in other tissues.

Methods

Single-cell isolation, whole-genome amplification and sequencing

We isolated EpCAM+ cells from one normal and three tumoural regions (TP: tumour proximal; TC: tumour central; TD: tumour distal) from the patient with a BD FACSAria III cytometer. We successfully amplified the genomes of 28 tumour cells and 18 normal cells with Ampli1 (Silicon Biosystems) and built whole-genome sequencing libraries using the KAPA (Kapa Biosystems) library kit. Each library was sequenced at $\approx 6\times$ on an Illumina Novaseq 6000 at the Spanish National Center of Genomic Analysis (CNAG-CR; <https://www.cnag.crg.eu/>). We called this dataset CRC28.

Data preprocessing

For the public TNBC16 [42] and CRC48 [43] datasets, we downloaded the raw sequencing reads from the SRA database in FASTQ format. For the three datasets (CRC28, TNBC16 and CRC48) we trimmed the Illumina adapter sequences using cutadapt (version 1.18) and mapped reads to the 1000G Reference Genome hs37d5 using BWA MEM (version 0.7.17). After de-duplication with Picard (version 2.18.14),

we used GATK (version 3.7.0) for local realignment based on indel calls from the 1000G Phase 1 and the Mills and 1000G gold standard. Subsequently, we recalibrated the base scores using GATK (version 4.0.10) with polymorphisms from dbSNP (build 138) and indels from the 1000G Phase 1. Exact commands used to run the tools are featured in Supplementary Note.

SIEVE model

SIEVE is a statistical approach which combines a statistical phylogenetic model with a probabilistic model of raw read counts. We implement SIEVE under BEAST 2 [34], a popular Bayesian phylogenetic framework that uses Markov Chain Monte Carlo (MCMC) for the estimation of phylogenetic trees and model parameters.

Input data

SIEVE takes as input raw read counts of all four nucleotides at candidate SNV sites (Fig. 1a). Specifically, for cell $j \in \{1, \dots, J\}$ at candidate SNV site $i \in \{1, \dots, I\}$, the input data to SIEVE is in the form of $\mathcal{D}_{ij}^{(1)} = (\mathbf{m}_{ij}, c_{ij})$, where $\mathbf{m}_{ij} = \{m_{ijk} \mid k = 1, 2, 3\}$ corresponds to the read counts of three alternative nucleotides with values in descending order and c_{ij} to the sequencing coverage for cell j and site i . Candidate SNV sites are defined as statistically significant SNVs that could potentially occur in single cells (see the “Candidate site identification” section).

For scWGS and scWES datasets, raw read counts from I' background sites are denoted $\mathcal{D}^{(2)}$. The number of background sites is used to correct acquisition bias (see the “SIEVE likelihood” section). For datasets lacking background information (for instance, from targeted sequencing), SIEVE accepts a user-specified number of background sites only for acquisition bias correction.

Candidate site identification

To identify candidate variant sites, we employ a strategy similar to SCIPHI [15]. Specifically, a likelihood ratio test is conducted for SNV detection, but with a modification enabling to capture sites containing double mutant genotypes. To this end, the Beta-Binomial distribution is fitted with free mean and overdispersion parameters at each site across all cells with non-zero variant read counts, and the corresponding likelihood is denoted L_1 . Next, another constrained Beta-Binomial distribution is fitted using the same set of cells with fixed mean being 0.25 and free overdispersion, whose likelihood is denoted L_0 . As a result, the test statistic $-2 \log \frac{L_0}{L_1}$ asymptotically follows the χ^2 distribution with degrees of freedom being 1. The null hypothesis (H_0) is thus that the mean = 0.25, and the alternative hypothesis (H_1) is that the mean \neq 0.25. A site is classified as candidate variant when the corresponding p -value is larger than 0.05 or the fitted mean is larger than 0.25. This analysis is performed on tumour cells. Normal cells are additionally used to filter out germline mutations. This candidate site identification procedure is implemented in a tool named DataFilter.

The sites identified by DataFilter are referred to as ‘candidate’ since they could sometimes be false discoveries due to technical errors in scDNA-seq. Moreover, the

actual variant calling, i.e. determination of whether the variant occurs in each of the candidate sites in each cell is performed by SIEVE, and not DataFilter. Notably, all other methods that identify evolutionary trees, including CellPhy [26], SiFit [22], or SCIPhI [15], require an input either actual variants in each cell (CellPhy, SiFit) or the candidate variant sites (SCIPhI). The identification of these candidate sites is crucial for model performance, as it limits the number of sites where the variation may occur, which is much smaller compared to the full set of all possible sites.

Statistical phylogenetic model

The statistical phylogenetic model behind SIEVE includes an instantaneous transition rate matrix, which is defined by a continuous-time homogeneous Markov chain. We consider four possible genotypes $G = \{0/0, 0/1, 1/1, 1/1'\}$, where 0, 1, and 1' are used to denote the reference nucleotide, an alternative nucleotide, and a second alternative nucleotide which is different from that denoted by 1, respectively. The fundamental evolutionary events we consider are single mutations and single back mutations. The former happen when 0 mutates to 1, or 1 and 1' mutate to each other, while the latter occur when 1 or 1' mutates to 0. Hence, genotypes 0/0 and 0/1 represent wildtype and single mutant genotypes, respectively, whereas genotype 1/1 and 1/1' represent double mutant genotypes. We intentionally use the non-standard nomenclature of single and double mutants to discern important evolutionary events. In contrast, calling both 0/1 and 1/1' a heterozygous mutation genotype would be more standard and correct, but would not differentiate between the genotype that has only a single allele changed with respect to the reference (0/1) from the genotype that has two alleles changed (1/1'). We only consider unphased genotypes, so we do not differentiate between 0/1 and 1/0 or between 1/1' and 1'/1.

The joint conditional probability of all cells at SNV site i having genotype $g_{ij} \in G, j = 1, \dots, J$ is determined according to the statistical phylogenetic model by

$$P\left(\mathbf{g}_i^{(L)} \mid \mathcal{T}, \boldsymbol{\beta}, Q, h, \eta\right) = \sum_{\mathbf{g}_i^{(A)} \setminus g_{i,2J}} P\left(\mathbf{g}_i^{(L)}, \mathbf{g}_i^{(A)} \setminus g_{i,2J} \mid \mathcal{T}, \boldsymbol{\beta}, Q, h, \eta\right). \quad (1)$$

In Eq. (1), $\boldsymbol{\beta}$ represents the branch lengths measured by the expected number of somatic mutations per site and Q is the instantaneous transition rate matrix of the Markov chain. \mathcal{T} is the rooted binary tree topology, representing the genealogical relations among cells. We specifically require the root of \mathcal{T} to have only one child, representing the most recent common ancestor (MRCA) of all cells. The branch between the root and the MRCA is the trunk of the cell phylogeny. The trunk is one of novelties of our approach, introduced to represent the accumulation of clonal mutations (shared among all cells) in the initial phase of tumour progression. Therefore, with J existing cells, labelled by $\{1, \dots, J\}$, as leaves, \mathcal{T} has J internal hidden ancestor nodes, labelled by $\{J + 1, \dots, 2J\}$, and $2J - 1$ branches, whose lengths are kept in $\boldsymbol{\beta}$. The trunk is essential for \mathcal{T} to assure that the root, labelled by $2J$, represents a normal ancestor cell even if the data only contains tumour cells. Hence the genotype of the root for SNV site i , denoted $g_{i,2J}$, is fixed

to 0/0. $\mathbf{g}_i^{(L)}$ represents the genotypes of J cells as leaves of \mathcal{T} , while $\mathbf{g}_i^{(A)}$ is the genotypes of all ancestor cells as internal nodes of \mathcal{T} . Note that we marginalise the genotypes of the ancestor nodes except for the root. We also consider among-site substitution rate variation following a discrete Gamma distribution with mean equal 1, parameterised by the number of rate categories h and shape η [45]. \mathcal{T}, β, η in Eq. (1) are hidden variables, estimated using MCMC (see the “Posterior and MCMC” section), whereas h is a hyperparameter that is fixed (4 by default). Note that variant calling effectively corresponds to the determination of the values of the variables $\mathbf{g}_i^{(L)}$.

In the transition rate matrix Q (Fig. 1c), each entry denotes a rate from one genotype to another during an infinitesimal time interval Δt . Note that at most one change is allowed to occur in Δt . For instance, the transition of 0/0 moving to 1/1 during Δt is impossible as two single somatic mutations are required; thus, the corresponding transition rate is 0. The transition rate from genotype 0/0 to 0/1 represents the somatic mutation rate and is set to 1. The back mutation rate is measured relatively to the somatic mutation rate and therefore is $1/3$.

With the genotype state space G defined, for a given branch length β , the underlying four-by-four transition probability matrix $R(\beta)$ of the Markov chain is represented using matrix exponentiation of the product of Q and β as $R(\beta) = \exp(Q\beta)$ [27].

Model of raw read counts

The probability of observing the input data \mathcal{D}_{ij} for cell j at site i is factorised as

$$P(\mathcal{D}_{ij}) = P(\mathbf{m}_{ij} | c_{ij})P(c_{ij}), \quad (2)$$

where the first component is the model of nucleotide read counts and the second the model of sequencing coverage.

Model of sequencing coverage After single-cell whole-genome amplification (sc-WGA) some genomic regions are more represented than others. After scDNA-seq, this results in an uneven coverage along the genome, much more than in the case of bulk sequencing. Here, to model the sequencing coverage c in the presence of overdispersion, we employ a negative binomial distribution.

$$P(c | p, r) = \binom{c+r-1}{r-1} p^r (1-p)^c, \quad (3)$$

with parameters p and r . We reparameterise the distribution with $p = \mu/\sigma^2$ and $r = \mu^2/\sigma^2 - \mu$, where μ and σ^2 are the mean and the variance of the distribution of the sequencing coverage c , respectively.

Theoretically, each cell j at site i has its specific μ_{ij} and σ_{ij}^2 parameters, which, however, are impossible to be estimated freely. Hence, we make additional assumptions and pool

the data for better estimates, adapting the approach of [46]. We assume that μ_{ij} and σ_{ij}^2 have the following forms, respectively:

$$\begin{aligned} \mu_{ij} &= \alpha_{ij} t s_j, \\ \sigma_{ij}^2 &= \mu_{ij} + \alpha_{ij}^2 v s_j^2. \end{aligned} \tag{4}$$

In Eq. (4), t is the mean of allelic coverage (the expected coverage per allele) and v is the variance of allelic coverage. We estimate t and v with MCMC (see the “[Posterior and MCMC](#)” section). $\alpha_{ij} \in \{1, 2\}$ is a hidden random variable denoting the number of sequenced alleles for cell j at site i . According to the statistical phylogenetic model, both alleles are expected to be sequenced. However, due to the frequent occurrence of allelic dropout (ADO) during scWGA, there are cases where only one allele is amplified and therefore α_{ij} is 1. Equation (4) reflects the fact that the expected sequencing coverage and its raw variance are proportional to the number of sequenced alleles. Note that inferring the hidden variable α_{ij} corresponds to identifying occurrences of ADO events, and hence the ability of SIEVE to perform ADO calling. We denote the prior distribution of α_{ij}

$$\begin{cases} P(\alpha_{ij} = 1 | \theta) = \theta, & \text{if ADO occurs,} \\ P(\alpha_{ij} = 2 | \theta) = 1 - \theta, & \text{otherwise,} \end{cases} \tag{5}$$

where θ is a parameter corresponding to the the probability of ADO occurs, i.e. the ADO rate, which is estimated using MCMC.

In Eq. (4), s_j is the size factor of cell j which makes sequencing coverage from different cells comparable and is estimated directly from the sequencing coverage using

$$\hat{s}_j = \text{median}_{i:c_{ij} \neq 0} \frac{c_{ij}}{\left(\prod_{\substack{j'=1 \\ c_{ij'} \neq 0}}^{J'} c_{ij'} \right)^{\frac{1}{J'}}}, \tag{6}$$

where J' is the number of cells with non-zero coverage at a site. By taking into account only the non-zero values, the estimate \hat{s}_j is not affected by the missing data, which is prevalent in scDNA-seq.

Table 2 Definition of the distribution of g'_{ij} conditional on g_{ij} and α_{ij}

g'_{ij}	g_{ij}	α_{ij}	$P(g'_{ij} g_{ij}, \alpha_{ij})$
0/0	0/0	2	1
0/-	0/0	1	1
0/1	0/1	2	1
1/1	1/1	2	1
1/-	1/1	1	1
1/ γ'	1/ γ'	2	1
1/-	1/ γ'	1	1
0/-	0/1	1	$\frac{1}{2}$
1/-	0/1	1	$\frac{1}{2}$
Others			0

Model of nucleotide read counts We denote the genotype affected by ADO $g'_{ij} \in G \cup \{0/-, 1/-\}$, where 0/- and 1/- are the results of ADO occurring to g_{ij} . For instance, 0/- is caused either by 0 dropped out from 0/0 or by 1 dropped out from 0/1. Then the probability of g'_{ij} is denoted by

$$P(g'_{ij} | g_{ij}, \alpha_{ij}), \tag{7}$$

which is defined at length in Table 2.

We model the read counts of three alternative nucleotides m_{ij} given the sequencing coverage c_{ij} with a Dirichlet-multinomial distribution as

$$P(m_{ij} | c_{ij}, \mathbf{a}_{ij}) = \frac{F(c_{ij}, \mathbf{a}_{ij0})}{\prod_{k=1:3}^{m_{ijk}>0} F(m_{ijk}, \mathbf{a}_{ijk}) F(c_{ij} - \sum_{k=1}^3 m_{ijk}, \mathbf{a}_{ij4})}, \tag{8}$$

with parameters $\mathbf{a}_{ij} = \{a_{ijk} | k = 1, \dots, 4\}$ and $\mathbf{a}_{ij0} = \sum_{k=1}^4 a_{ijk}$. F is a function in the form of

$$F(x, y) = \begin{cases} xB(y, x), & \text{if } x > 0, \\ 1, & \text{otherwise,} \end{cases} \tag{9}$$

where B is the beta function. Note that $c_{ij} - \sum_{k=1}^3 m_{ijk}$ is the read count of the reference nucleotide.

To improve the interpretation of Eq. (8), we reparameterise it with $\mathbf{a}_{ij} = w_{ij} \mathbf{f}_{ij}$, where $\mathbf{f}_{ij} = \{f_{ijk} | k = 1, \dots, 4\}$, $\sum_{k=1}^4 f_{ijk} = 1$ is a vector of expected frequencies of each nucleotide and w_{ij} represents overdispersion. \mathbf{f}_{ij} are categorical hidden variables dependent on g'_{ij} :

$$\mathbf{f}_{ij} = \begin{cases} \mathbf{f}_1 = \left(\frac{1}{3}f, \frac{1}{3}f, \frac{1}{3}f, 1-f\right), & \text{if } g'_{ij} = 0/0 \text{ or } 0/-, \\ \mathbf{f}_2 = \left(\frac{1}{2} - \frac{1}{3}f, \frac{1}{3}f, \frac{1}{3}f, \frac{1}{2} - \frac{1}{3}f\right), & \text{if } g'_{ij} = 0/1, \\ \mathbf{f}_3 = \left(1-f, \frac{1}{3}f, \frac{1}{3}f, \frac{1}{3}f\right), & \text{if } g'_{ij} = 1/1 \text{ or } 1/-, \\ \mathbf{f}_4 = \left(\frac{1}{2} - \frac{1}{3}f, \frac{1}{2} - \frac{1}{3}f, \frac{1}{3}f, \frac{1}{3}f\right), & \text{if } g'_{ij} = 1/1', \end{cases} \tag{10}$$

where f is the expected frequency of nucleotides whose existence is solely due to technical errors during sequencing. To be specific, f is defined as the effective sequencing error rate including amplification (where a nucleotide is wrongly amplified into another one during scWGA) and sequencing errors.

w_{ij} is also a categorical hidden variable dependent on g'_{ij} :

$$w_{ij} = \begin{cases} w_1, & \text{if } g'_{ij} = 0/0, 0/-, 1/1, \text{ or } 1/-, \\ w_2, & \text{if } g'_{ij} = 0/1 \text{ or } 1/1', \end{cases} \tag{11}$$

where w_1 is wildtype overdispersion and w_2 is alternative overdispersion.

By plugging in Eqs. (10) and (11), (8) is equivalently represented with

$$P(\mathbf{m}_{ij} | c_{ij}, g'_{ij}, f, w_{ij}) = \begin{cases} P_{0/0} = P\left(\mathbf{m}_{ij} \mid c_{ij}, g'_{ij} = 0/0, \mathbf{f}_1, w_1\right), \\ P_{0/-} = P\left(\mathbf{m}_{ij} \mid c_{ij}, g'_{ij} = 0/-, \mathbf{f}_1, w_1\right), \\ P_{0/1} = P\left(\mathbf{m}_{ij} \mid c_{ij}, g'_{ij} = 0/1, \mathbf{f}_2, w_2\right), \\ P_{1/1} = P\left(\mathbf{m}_{ij} \mid c_{ij}, g'_{ij} = 1/1, \mathbf{f}_3, w_1\right), \\ P_{1/-} = P\left(\mathbf{m}_{ij} \mid c_{ij}, g'_{ij} = 1/-, \mathbf{f}_3, w_1\right), \\ P_{1/Y} = P\left(\mathbf{m}_{ij} \mid c_{ij}, g'_{ij} = 1/Y, \mathbf{f}_4, w_2\right). \end{cases} \tag{12}$$

Note that $P_{0/0}$ and $P_{0/-}$ share the same \mathbf{f} and w_1 , showing that the model of nucleotide read counts is not enough to discriminate 0/0 from 0/-, and so do $P_{1/1}$ and $P_{1/-}$. In such cases, incorporating the model of sequencing coverage helps resolve the entanglement.

To understand Eq. (12), first take $P_{0/0}$ as an example. Theoretically, no alternative nucleotides are supposed to exist if no technical errors occur. Thus, any observations of any alternative nucleotides can only result from technical errors, and the expected frequency of the reference nucleotide is accordingly adjusted to $1 - f$. For another example $P_{0/1}$, say the reference nucleotide is A and the alternative nucleotide is C, and both their read count frequencies are supposed to be $1/2$ if no technical errors occur. For the other two alternative nucleotides, G and T, their observations could only result from technical errors, and both their frequencies are $f/3$. Moreover, either A or C may be sequenced as a different nucleotide (each with probability $1/2$). In the former case, the frequency of A decreases by $f/2$. In the latter case, if C is sequenced as A (with probability $f/3$) the frequency of A increases by $1/2 \times f/3$. Overall, the frequency of A decreases by $f/3$, resulting in $1/2 - f/3$.

f, w_1 and w_2 in Eq. (12) are estimated with MCMC.

SIEVE likelihood

We denote the conditional variables in Eq. (1) as $\Theta = \{\mathcal{T}, \boldsymbol{\beta}, Q, h, \eta\}$ and those in the model of raw read counts as $\Phi = \{t, v, \theta, f, w_1, w_2\}$. Given the input data $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$, the log-likelihood of the SIEVE model is

$$\log \mathcal{L}(\Theta, \Phi) = \log \mathcal{L}^{(1)}(\Theta, \Phi) + \log \mathcal{L}^{(2)}(f, w_1), \tag{13}$$

where $\mathcal{L}^{(1)}$ is the tree likelihood corrected for acquisition bias computed from candidate SNV sites in $\mathcal{D}^{(1)}$, while $\mathcal{L}^{(2)}$ is the likelihood computed from background sites in $\mathcal{D}^{(2)}$, referred to as the background likelihood. Equation (13) does not contain $g_{ij}, g'_{ij}, \alpha_{ij}$ since they are marginalised out (see below).

Since we only use data from SNV sites to compute the tree likelihood, the tree branch lengths $\boldsymbol{\beta}$ are prone to be overestimated [29, 30]. The overestimation of $\boldsymbol{\beta}$ due to only using data from SNV sites is called acquisition bias, which is corrected in SIEVE according to [47]:

$$\log \mathcal{L}^{(1)} = \log P\left(\mathcal{D}^{(1)} \mid \Theta, \Phi\right) + I' \log \left(\frac{1}{I} \sum_{i=1}^I C_i\right), \tag{14}$$

where the first component is the uncorrected tree log-likelihood for SNV sites, and C_i in the second component is the likelihood of SNV site i being invariant (see below). The regularisation term $I' \log \left(\frac{1}{I} \sum_{i=1}^I C_i \right)$ renders SIEVE in favour of trees with short branch lengths where $\mathcal{L}^{(1)}$ is large due to the increasing averaged C .

To compute the uncorrected tree log-likelihood, we marginalise out α_{ij} and g'_{ij} :

$$\begin{aligned}
 P(\mathbf{m}_{ij}, c_{ij} | g_{ij}, \Phi) &= P(\mathbf{m}_{ij}, c_{ij} | g_{ij}, f, w_{ij}, t, \nu, \theta) \\
 &= \sum_{\alpha_{ij}, g'_{ij}} P(\mathbf{m}_{ij}, c_{ij}, \alpha_{ij}, g'_{ij} | g_{ij}, f, w_{ij}, t, \nu, \theta) \\
 &= \sum_{\alpha_{ij}, g'_{ij}} P(\mathbf{m}_{ij} | c_{ij}, g'_{ij}, f, w_{ij}) P(g'_{ij} | g_{ij}, \alpha_{ij}) \\
 &\quad \times P(c_{ij} | \alpha_{ij}, t, \nu) P(\alpha_{ij} | \theta) \\
 &= \begin{cases} P_{0/0} \cdot P(c_{ij} | \alpha_{ij} = 2, t, \nu) \cdot (1 - \theta) \\ \quad + P_{0/-} \cdot P(c_{ij} | \alpha_{ij} = 1, t, \nu) \cdot \theta, & \text{if } g_{ij} = 0/0, \\ P_{0/1} \cdot P(c_{ij} | \alpha_{ij} = 2, t, \nu) \cdot (1 - \theta) \\ \quad + \frac{1}{2}(P_{0/-} + P_{1/-}) \cdot P(c_{ij} | \alpha_{ij} = 1, t, \nu) \cdot \theta, & \text{if } g_{ij} = 0/1, \\ P_{1/1} \cdot P(c_{ij} | \alpha_{ij} = 2, t, \nu) \cdot (1 - \theta) \\ \quad + P_{1/-} \cdot P(c_{ij} | \alpha_{ij} = 1, t, \nu) \cdot \theta, & \text{if } g_{ij} = 1/1, \\ P_{1/Y} \cdot P(c_{ij} | \alpha_{ij} = 2, t, \nu) \cdot (1 - \theta) \\ \quad + P_{1/-} \cdot P(c_{ij} | \alpha_{ij} = 1, t, \nu) \cdot \theta, & \text{if } g_{ij} = 1/1', \end{cases} \tag{15}
 \end{aligned}$$

where $P_{0/0}, P_{0/-}, P_{0/1}, P_{1/1}, P_{1/-}, P_{1/Y}$ are defined in Eq. (12) and $P(g'_{ij} | g_{ij}, \alpha_{ij})$ is defined in Eq. (7). In the second line of Eq. (15), the probability is factorised out according to Fig. 1b.

To compute $\log P(\mathcal{D}^{(1)} | \Theta, \Phi)$ in Eq. (14), we assume that the SNV sites evolve independently and identically. By plugging Eqs. (1) and (15), $\log P(\mathcal{D}^{(1)} | \Theta, \Phi)$ is denoted by

$$\begin{aligned}
 \log P(\mathcal{D}^{(1)} | \Theta, \Phi) &= \sum_{i=1}^I \log \sum_{\mathbf{g}_i^{(L)}} P(\mathcal{D}_i^{(1)} | \mathbf{g}_i^{(L)}, \Phi) \sum_{\mathbf{g}_i^{(A)} \setminus g_{i,2J}} P(\mathbf{g}_i^{(L)}, \mathbf{g}_i^{(A)} \setminus g_{i,2J} | \Theta) \\
 &= \sum_{i=1}^I \log \sum_{\mathbf{g}_i^{(L)}} \left[\prod_{j=1}^J P(\mathbf{m}_{ij}, c_{ij} | g_{ij}, \Phi) \right. \\
 &\quad \left. \times \sum_{\mathbf{g}_i^{(A)} \setminus g_{i,2J}} P(\mathbf{g}_i^{(L)}, \mathbf{g}_i^{(A)} \setminus g_{i,2J} | \Theta) \right] \\
 &= \sum_{i=1}^I \sum_{j=1}^J \log \sum_{\mathbf{g}_i^{(L)}, \mathbf{g}_i^{(A)} \setminus g_{i,2J}} \left[P(\mathbf{m}_{ij}, c_{ij} | g_{ij}, \Phi) \right. \\
 &\quad \left. \times P(\mathbf{g}_i^{(L)}, \mathbf{g}_i^{(A)} \setminus g_{i,2J} | \Theta) \right], \tag{16}
 \end{aligned}$$

which is efficiently computed out by Felsenstein’s pruning algorithm [48], with the extension of the model of raw read counts applied on leaves. Specifically, the Felsenstein’s pruning algorithm is applied to an extended tree \mathcal{T} , where additional leaf nodes corresponding to the data are attached at the bottom of \mathcal{T} : for each node corresponding to genotype g_{ij} there is a leaf node added, corresponding to data $(\mathbf{m}_{ij}, c_{ij})$, and the transition

probability between the genotype node and the leaf is given by Eq. (15). For I candidate SNV sites, J cells and K genotype states in G (for SIEVE $K = 4$), the time complexity of Felsenstein’s pruning algorithm is $\mathcal{O}(IJK^2)$.

C_i in Eq. (14) is determined similarly to Eq. (16) by computing the joint probability of observing the data $\mathcal{D}_i^{(1)}$ and $\mathbf{g}_i^{(L)} = 0/0$:

$$\begin{aligned}
 C_i &= P\left(\mathcal{D}_i^{(1)}, \mathbf{g}_i^{(L)} = 0/0 \mid \Theta, \Phi\right) \\
 &= P\left(\mathcal{D}_i^{(1)} \mid \mathbf{g}_i^{(L)} = 0/0, \Phi\right) \sum_{\mathbf{g}_i^{(A)} \setminus g_{i,2J}} P\left(\mathbf{g}_i^{(L)} = 0/0, \mathbf{g}_i^{(A)} \setminus g_{i,2J} \mid \Theta\right) \\
 &= \prod_{j=1}^J P\left(\mathbf{m}_{ij}, c_{ij} \mid g_{ij} = 0/0, \Phi\right) \sum_{\mathbf{g}_i^{(A)} \setminus g_{i,2J}} P\left(\mathbf{g}_i^{(L)} = 0/0, \mathbf{g}_i^{(A)} \setminus g_{i,2J} \mid \Theta\right).
 \end{aligned}
 \tag{17}$$

Formally, to compute the background likelihood, we should account for the fact that the background sites, similarly to the variant sites, also evolve under the phylogenetic model and involve similar computations as above. This, however, would result in a large additional computational burden due to the large number of background sites compared to the variant sites. Thus, to estimate the background log-likelihood efficiently, we make several simplifications and compute it only approximately. First, we assume that across I' background sites each cell has the same genotype 0/0 and both alleles are covered. We further ignore the model of sequencing coverage and the tree log-likelihood in the computations. As a result, by employing an alternative expression of Dirichlet-multinomial distribution $\log \mathcal{L}^{(2)}$ is efficiently obtained as

$$\begin{aligned}
 \log \mathcal{L}^{(2)}(f, w_1) &= \sum_{i=1}^{I'} \sum_{j=1}^J \log P_{0/0} \\
 &= \sum_{i=1}^{I'} \sum_{j=1}^J \log \left[\frac{\Gamma(w_1)\Gamma(c_{ij} + 1)}{\Gamma(c_{ij} + w_1)} \prod_{k=1}^3 \frac{\Gamma(m_{ijk} + \frac{1}{3}fw_1)}{\Gamma(\frac{1}{3}fw_1)\Gamma(m_{ijk} + 1)} \right. \\
 &\quad \left. \times \frac{\Gamma(c_{ij} - \sum_{k=1}^3 m_{ijk} + (1-f)w_1)}{\Gamma((1-f)w_1)\Gamma(c_{ij} - \sum_{k=1}^3 m_{ijk} + 1)} \right] \\
 &= I'J \left[\log \Gamma(w_1) - 3 \log \Gamma\left(\frac{1}{3}fw_1\right) - \log \Gamma((1-f)w_1) \right] \\
 &\quad + \sum_{c=1}^{\max(c_{ij})} N_c (\log \Gamma(c + 1) - \log \Gamma(c + w_1)) \\
 &\quad + \sum_{k=1}^3 \sum_{m_k=1}^{\max(m_{ijk})} N_{m_k} \left(\log \Gamma\left(m_k + \frac{1}{3}fw_1\right) - \log \Gamma(m_k + 1) \right) \\
 &\quad + \sum_{c-\sum_{k=1}^3 m_k=1}^{\max(c_{ij}-\sum_{k=1}^3 m_{ijk})} N_{c-\sum_{k=1}^3 m_k} \left(\log \Gamma\left(c - \sum_{k=1}^3 m_k + (1-f)w_1\right) \right. \\
 &\quad \left. - \log \Gamma\left(c - \sum_{k=1}^3 m_k + 1\right) \right),
 \end{aligned}
 \tag{18}$$

where $P_{0/0}$ is defined in Eq. (12). N_c , N_{m_k} for $k = 1, 2, 3$ and $N_{c - \sum_{k=1}^3 m_k}$ represent, across I' background sites and J cells, the unique occurrences of sequencing coverage c , of alternative nucleotide read counts m_1, m_2, m_3 , and of reference nucleotide read counts $c - \sum_{k=1}^3 m_k$, respectively. In Eq. (18), some items, namely $\log \Gamma(c + 1)$, $-\log \Gamma(m_k + 1)$ for $k = 1, 2, 3$, and $-\log \Gamma(c - \sum_{k=1}^3 m_k + 1)$, only depends on the data, which remain constants during MCMC. Therefore, they are ignored in the computation of background likelihood. It is clear that the background likelihood helps estimate f and w_1 .

The time complexity of Eq. (18) is $\mathcal{O}(c)$ with c being the number of unique values of sequencing coverage across all cells and background sites. Since IJK^2 is usually much larger than c , the overall time complexity of model likelihood is $\mathcal{O}(IJK^2)$.

Priors

To define priors for model parameters and for the tree coalescent, we employ the prior distributions defined in BEAST 2. We impose on \mathcal{T} and β in Eq. (1) a prior distribution following the Kingman coalescent process with an exponentially growing population. The tree prior is parameterised by scaled population size M and exponential growth rate q , and is denoted by

$$P(\mathcal{T}, \beta | M, e), \quad (19)$$

whose analytical form is defined in [49]. M and e are hidden random variables and are estimated using MCMC. Note that, by default, M represents the number of time units, e.g. the number of years, and the mutation rate is measured by the number of mutations per time unit per site. Their product results in the unit of branch length, i.e. the number of mutations per site. Since scDNA-seq data usually does not contain temporal information as a result of collecting samples at the same time, it is impossible to differentiate M from the mutation rate. However, if the mutation rate is known, one could alternatively estimate a time-calibrated cell phylogeny.

As prior distributions, we assign to M

$$P(M | \delta) = \frac{1}{\delta}, \quad (20)$$

where δ is the current proposed value of M . Note that this is supposed to be normalised to define a proper probability distribution, but this form is sufficient to define a proper posterior (see the “Posterior and MCMC” section).

For e , we choose

$$e | \lambda, \epsilon \sim \text{Laplace}(\lambda, \epsilon), \quad (21)$$

where we choose mean $\lambda = 10^{-3}$ and scale $\epsilon = 30.7$ (default in the BEAST 2 software). We choose an exponential distribution as the prior for η in Eq. (1):

$$\eta | \gamma \sim \exp(\gamma), \quad (22)$$

where $\gamma = 1$.

For the model of sequencing coverage described in Eqs. (3) and (4), we set the prior for t within a large range of values with

$$t | \rho \sim \text{Uniform}(0, \rho), \tag{23}$$

where $\rho = 1000$, and the prior for v with

$$v | \zeta \sim \exp(\zeta), \tag{24}$$

where $\zeta = 25$. In terms of θ in Eq. (5), it also has a uniform prior:

$$\theta | u \sim \text{Uniform}(0, u), \tag{25}$$

where $u = 1$.

For the model of nucleotide read counts described in Eqs. (10) to (12), we choose an exponential prior for f :

$$f | \tau \sim \exp(\tau), \tag{26}$$

where $\tau = 0.025$, and a log normal prior for both w_1 and w_2 :

$$\begin{aligned} w_1 | \xi_1, \psi_1 &\sim \text{Log-Normal}(\xi_1, \psi_1), \\ w_2 | \xi_2, \psi_2 &\sim \text{Log-Normal}(\xi_2, \psi_2), \end{aligned} \tag{27}$$

where we choose for w_1 the log-transformed mean $\xi_1 = 3.9$ (150 for untransformed) and the standard deviation $\psi_1 = 1.5$, and for w_2 the log-transformed mean $\xi_2 = 0.9$ (10 for untransformed) and the standard deviation $\psi_2 = 1.7$. Specifically, the mean is log-transformed using

$$\xi_{\text{transformed}} = \log(\xi_{\text{untransformed}}) - \frac{\psi^2}{2}.$$

These specific values reflect our belief that w_1 is greater than w_2 , and are chosen in such a way that both distributions cover a large range of possible values for w_1 and w_2 .

Posterior and MCMC

With the model likelihood and priors defined, the posterior distribution of the unknown parameters is

$$\begin{aligned} P\left(\mathcal{T}, \boldsymbol{\beta}, M, e, \eta, t, v, \theta, f, w_1, w_2 \mid \mathcal{D}^{(1)}, \mathcal{D}^{(2)}\right) &= \frac{1}{Z} P\left(\mathcal{D}^{(1)}, \mathcal{D}^{(2)} \mid \mathcal{T}, \boldsymbol{\beta}, \eta, t, v, \theta, f, w_1, w_2\right) \\ &\times P(\mathcal{T}, \boldsymbol{\beta} \mid M, e) P(M \mid \delta) P(e \mid \lambda, \epsilon) P(\eta \mid \gamma) \\ &\times P(t \mid \rho) P(v \mid \zeta) P(\theta \mid u) P(f \mid \tau) \\ &\times P(w_1 \mid \xi_1, \psi_1) P(w_2 \mid \xi_2, \psi_2), \end{aligned} \tag{28}$$

where Z is a normalisation constant, representing the probability of the observed data.

Since the posterior distribution does not have a closed-form analytical formula, we employ the MCMC algorithm with Metropolis-Hastings kernel to sample from the posterior distribution in Eq. (28). Given the current state of the parameters q , we propose a new state q^* according to proposal distributions $P(q^* | q)$ that assure the reversibility and ergodicity of the Markov chain. With one parameter changed a time, q^* is accepted with probability

$$\min \left\{ 1, \frac{P(T^*, \beta^*, M^*, e^*, \eta^*, t^*, v^*, \theta^*, f^*, w_1^*, w_2^* | \mathcal{D}^{(1)}, \mathcal{D}^{(2)})P(q | q^*)}{P(T, \beta, M, e, \eta, t, v, \theta, f, w_1, w_2 | \mathcal{D}^{(1)}, \mathcal{D}^{(2)})P(q^* | q)} \right\}, \quad (29)$$

where the normalisation constant Z cancels out after plugging in Eq. (28).

For sampling the structure of the cell phylogeny, we take advantage of proposal distributions implemented in the BEAST 2 software [49] and modify them to make sure they are compatible with our tree topology, so that the sampled trees are binary and contain a trunk. Specifically, the tree branch lengths are changed by scaling the heights of the internal nodes. For tree topological exploration, we use the Wilson-Balding move to perform subtree pruning and regrafting. Specifically, a random node and half of its subtree is pruned and reattached to a random branch not belonging to the moved subtree. A subtree-slide move is also used, where a random node and half of its subtree slides either upwards or downwards along branches and cross at least one node. Both those two moves include changes to the lengths of some branches. The final type of move swaps two randomly selected subtrees.

For sampling unknown parameters, we perform either scaling operations or random Gaussian walks.

SIEVE runs with a two-stage sampling strategy. In the first stage the acquisition bias correction is switched off and all parameters are explored, while in the second stage the acquisition bias correction is turned on and parameters not affecting branch lengths are fixed with their estimates from the previous stage. This two-stage strategy proved to yield more accurate parameter and tree estimates than a strategy where both parameters and tree would be explored at once, with the acquisition bias correction enabled. Additionally, the initial tree in the second stage is set to the tree summarised from the first stage.

Variant calling, ADO calling, maximum likelihood gene annotation and mutation event classification

During the sampling process $\mathbf{g}_i^{(L)}$, $\mathbf{g}_i^{(A)}$, g'_{ij} and α_{ij} (Eqs. (1), (15) and (16)) are hidden variables that are marginalised out. Therefore, to obtain estimates of these hidden variables, we infer their maximum likelihood configuration with the max-sum algorithm [50], using the maximum clade credibility tree [51] and parameters estimated from the MCMC posterior samples.

To be specific, by determining the maximum likelihood genotypes of the leaves ($\mathbf{g}_i^{(L)}$), we are able to call variants. By inferring the maximum likelihood g'_{ij} and α_{ij} , the ADO state is determined. Moreover, by computing the maximum likelihood genotypes of the internal nodes ($\mathbf{g}_i^{(A)}$), SIEVE maps mutations to specific tree branches.

Mutation events are classified into different categories based on the corresponding genotype transitions (see Table 1). The single mutation ($0/0 \rightarrow 0/1$) happens when an allele of the wildtype is mutated. The homozygous coincident double mutation ($0/0 \rightarrow 1/1$) refers to the case when both alleles of the wildtype are mutated to the same alternative nucleotide, while the heterozygous coincident double mutation ($0/0 \rightarrow 1/1$) refers to the case when both alleles of the wildtype are mutated to different alternative nucleotides. The single back mutation ($0/1 \rightarrow 0/0$, $1/1 \rightarrow 0/1$ and $1/1 \rightarrow 0/1$) happens when a mutated allele mutates back to the reference nucleotide, while the double

back mutation ($1/1 \rightarrow 0/0$ and $1/1' \rightarrow 0/0$) happens when both mutated alleles mutate back to the reference nucleotide. The homozygous single mutation addition ($0/1 \rightarrow 1/1$) refers to the case when the unmutated allele of the single mutant genotype mutates to the same alternative nucleotide as the mutated allele, while for the heterozygous single mutation addition ($0/1 \rightarrow 1/1'$) the unmutated allele mutates to an alternative nucleotide different from the mutated allele. For the homozygous substitute single mutation ($1/1' \rightarrow 1/1$), one of the mutated alleles mutates to the same alternative nucleotide as the other mutated allele, while for the heterozygous substitute single mutation ($1/1 \rightarrow 1/1'$) one of the mutated alleles mutates to another alternative nucleotide.

Summary of model assumptions

Taken together, SIEVE makes several assumptions about the evolutionary process behind the observed single cell data. First, the model assumes that the genome is diploid. This assumption stands behind most of our model equations. In order not to violate this model assumption, one should pre-process the data to exclude non-diploid regions. On the other hand, this comes with the cost of excluding sites in these regions. Leaving such sites introduces discrepancy with the assumption, but might give more statistical power for model inference. Thus, we leave this decision of excluding copy number altered regions as a preprocessing step to the user.

Another important assumption, made by most methods for phylogenetic reconstruction, is that the sites are independently affected by the mutational process. This assumption is key to computational performance, as it allows to factorise the model likelihood across the sites.

One more assumption made behind SIEVE is that the phylogenetic tree has a trunk, which connects a healthy cell as the root and its only child as the MRCA of all cells in the data. When there are only tumour cells in the data, the MRCA represents the first tumour cell founding the tumour tissue, and since many clonal mutations accumulate during the foundation process of tumour, the trunk is expected to be long. When both healthy and tumour cells are available, the MRCA is also a healthy cell, and since only very few, if any, mutations accumulate between two healthy cells, the trunk is expected to be short. The incorporation of the trunk comes in handy in practice not only because it can help to identify normal cells mixed with tumour cells, but also because an out-group is not needed to root the tree.

Finally, SIEVE follows the finite sites assumption (FSA), which is both more general and more plausible than the infinite sites assumption (ISA). Events violating the ISA are expected biologically and probabilistically [32, 33]. It is important to note that per definition, SIEVE and other models that follow the FSA are well suited to model both cases (when ISA is violated and not). More specifically, the ISA is a special case of the FSA, so the models that follow the FSA also account for the ISA.

Summary of evolutionary features accounted for by the model

In contrast to other models, SIEVE is able to identify 12 types of genotype transitions, corresponding to nine types of mutation events (Table 1). Moreover, when such events affecting the same site are detected on more than one branch, our model is able to detect parallel evolution. This is because SIEVE considers four genotype states ($0/0$, $0/1$, $1/1$,

1/1') and is based on the underlying Markov process model that follows the FSA. Among those nine mutation events, only one of them, namely the single mutation, corresponding to the transition from genotype state 0/0 to 0/1, is accounted for by models that follow the ISA. Moreover, SiFit, which follows the FSA but has a restricted genotype state space compared to SIEVE, is also unable to identify all 12 genotype transitions that are detectable by SIEVE.

Moreover, SIEVE's another feature is its compatibility with molecular clock models implemented in BEAST 2, including the strict, relaxed and random local molecular clock model [52, 53]. The use of these models opens the door for the estimation of divergence times (event timing) and substitution rates using sound statistical models.

Importantly, we separate these features from model assumptions, as these are properties that SIEVE supports in an unforced manner. For instance, SIEVE is able to identify 12 genotype transitions, but not all of them are necessarily to appear on the tree.

ScDNA-seq data simulator

In order to benchmark the performance of SIEVE against those of other published methods, we simulated scDNA-seq data by modifying CellCoal [54] (commit 594e063). In contrast to CellCoal, the sequencing coverage is generated according to Eqs. (3) to (6). Given the sequencing coverage, read counts are simulated with a Multinomial distribution including errors. Input configuration follows the one described for CellCoal [54].

The simulator mimics both the biological evolution and the sequencing process. We first generated a binary genealogical cell lineage tree following the coalescent process assuming a strict molecular clock and created a reference genome where each site was initialised by the reference genotype with one of the four nucleotides. With a specific mutation rate, each site was evolved independently along the tree according to a rate matrix which contains ten diploid genotypes encoded with nucleotide pairs (Additional file 1: Table S4). The rate matrix allows mutations and back mutations, where the probability of the latter is $\frac{1}{3}$ of the former. All simulated sites for which at least one cell has a non-reference genotype are considered as true SNV sites. Next, we added at most one ADO to cell j at site i according to the ADO rate. If ADO happens, the number of sequenced alleles α_{ij} drops from two to one. We recorded the true ADO states across cells for the SNV sites. Size factors for cells in Eq. (4) were sampled from a normal distribution (mean = 1.2, variance = 0.2). Using the negative binomial distribution, we simulated the sequencing coverage with given t and ν . Based on the ADO-affected genotype and sequencing coverage, the read count for each nucleotide was simulated using a Multinomial distribution with a given amplification error rate and sequencing error rate.

Simulation design

We designed simulations to compare multiple methods in different aspects. The benchmarking framework was built using Snakemake [55].

Simulations only considering SNVs

We assumed that the tumour cell samples belonged to an exponentially growing population (growth rate = 10^{-4}) with an effective population size of 10^4 . The number of tumour cells was chosen to be either 40 or 100. We selected three mutation rates: 10^{-6} , 8×10^{-6} and

3×10^{-5} . For different mutation rates, different total number of sites were chosen to result in around 1000 SNV sites for 100 cells (1.3×10^5 sites for 10^{-6} , 2×10^4 sites for 8×10^{-6} , and 6.5×10^3 sites for 3×10^{-5}), as well as between 250 and 1000 SNV sites for 40 cells (8×10^4 sites for 10^{-6} , 2×10^4 sites for 8×10^{-6} and 5×10^3 sites for 3×10^{-5}). Additionally, we varied t and ν in Eqs. (3) and (4) to simulate different coverage qualities. For high quality data, we chose high mean ($t = 20$) and low variance ($\nu = 2$) of allelic coverage. For medium quality data, we chose high mean ($t = 20$) and medium variance ($\nu = 10$). For low quality data, we chose low mean ($t = 5$) and high variance ($\nu = 20$), which was specifically created to mimic the CRC28 dataset.

Other important parameters in the simulation were fixed as follows: in Eq. (5) $\theta = 0.163$, in Eq. (12) $w_1 = 100$ and $w_2 = 2.5$, and both amplification error rate and sequencing error rate were 10^{-3} , which resulted in the effective sequencing error rate $f \approx 2 \times 10^{-3}$ in Eq. (12).

We designed in total 18 simulation scenarios, each repeated 20 times.

Simulations considering both SNVs and CNAs

To add CNAs, we selected a set of datasets generated as described above, using the following parameters: 40 cells, medium mutation rate (8×10^{-6}) and medium coverage quality ($t = 20, \nu = 10$). Two levels of CNA prevalence were simulated: around $1/3$ or $2/3$ of all genomic sites. A site could contain CNAs occurring at an early or at a late stage during the evolutionary process with equal probabilities, and the corresponding number of CNAs was sampled in $\{0, 1, 3, \dots, 10\}$. For a site containing early stage CNAs, the probability of a cell carrying such events was sampled uniformly from the $[2/3, 1]$ interval, while for late stage CNAs the probability was sampled from the $(0, 1/3]$ interval. If a site in a cell was sampled to be affected by CNAs, a specific allele was selected for CNA with probability 0.5. To this end, if the sampled CNA value was 0, the read counts for the site and the cell was simply set to 0. Otherwise, we directly manipulated the simulated read counts of the chosen allele by multiplying the CNA value minus one, where the one CNA copy was retained for the other unchosen allele.

The simulated datasets after adding CNAs were stored in two versions: with or without genomic sites containing CNAs, both of which were used as input for all methods.

It is important to note that in these simulations, the CNAs were added independently of the phylogenetic structure. It is thus expected that we were simulating the most pessimistic scenario, as CNAs introducing bias in the data in the same way for phylogenetically related cells could in fact help with better phylogeny reconstruction.

Measurement of cell phylogeny accuracy and quality of variant calling

To assess the accuracy of the cell phylogeny reconstruction considering branch lengths, we computed the BS distance from the inferred tree to the true tree [35]. For any two trees, this difference is computed as:

$$d_{BS} = \sqrt{\sum_i \left(l_{1i}^{(s)} - l_{2i}^{(s)} \right)^2 + \sum_i \left(l_{1i}^{(u)} \right)^2 + \sum_i \left(l_{2i}^{(u)} \right)^2}. \quad (30)$$

where $l_{ji}^{(s)}$ represents the length of a branch shared by both trees, and $l_{ji}^{(u)}$ represents the length of a branch i that is unique for tree j .

To assess the accuracy of the cell phylogeny reconstruction ignoring branch lengths we used the normalised RF distance [36]:

$$d_{RF} = \frac{n_1^{(u)} + n_2^{(u)}}{n_1 + n_2}, \quad (31)$$

where n_j denotes the total number of branches in tree j , while $n_j^{(u)}$ represents the number branches exclusive of tree j .

Thus, BS distance and normalised RF distance values equal to 0 indicate a perfect tree reconstruction. For SIEVE and SiFit, we compute both normalised RF distance and BS distance in the rooted tree mode. For CellPhy, we compute these metrics in the unrooted tree mode as it infers an unrooted tree from data only containing tumour cells. Since SCIPHI reports a rooted tree without branch lengths, we can only compute the normalised RF distance. BS distance and normalised RF distance values were computed using the R package phangorn [56].

To evaluate the variant calling and ADO calling results, we computed precision, recall, F1 score and false positive rate (FPR). For variant calling, we separately compared the performance in calling the single mutant genotype and double mutant genotypes. In particular, when we evaluated the accuracy of single mutant genotype calling, any identification of double mutant genotypes whose true genotype is single mutant genotype was counted as a false negative. Moreover, we analysed two different types of false positives in single mutant genotype calling. The first type corresponds to single mutant calls for sites where the true genotype is a wildtype genotype. The second type are single mutant calls for sites where the true genotype is a double mutant.

For SIEVE and Monovar, we computed the recall, precision, F1 score, and FPR for single mutant genotype calling and double mutant genotype calling. For SCIPHI, we only computed metrics for single mutant genotype calling as it does not call double mutant genotypes. Moreover, we evaluated the accuracy of calling ADO states only for SIEVE, as it is the only method that is able to call them.

Configurations of methods

For Monovar (commit 68fbb68), we used the true values of θ and f as priors for false negative rate and false positive rate and default values for other options.

For SCIPHI (commit 34975f7), we ran it with default options and 5×10^5 iterations.

To run CellPhy (commit 832f6c2) and SiFit (commit 9dc3774), we fed the required data with variants called by Monovar. For CellPhy, we piped the data in VCF format and initialised the tree search with three parsimonious trees. We instructed the tool to use a built-in rate matrix with ten genotypes (GT10), a stationary nucleotide frequency distribution learned from the data (FO), an error model applied to the leaves (E), and the Gamma model of site-wise substitution rate variation (G). For SiFit, we fed the input data as a ternary matrix and used the true values of θ and f as the prior for false negative rate and the estimated false positive rate, respectively. We ran it with 2×10^5 iterations.

On the simulated data, we ran SIEVE with a strict molecular clock model for 2×10^6 and 1.5×10^6 iterations for the first and the second sampling stage, respectively. On the real datasets, we used a log-normal relaxed molecular clock model to take into consideration branch-wise substitution rate variation. To achieve better mixed Markov chains, we employed an optimised relaxed clock model in [37] instead of the default one in BEAST 2.

Since more parameters are added when using the relaxed molecular clock model, we ran the analysis with 3×10^6 iterations for the first stage and 2.5×10^6 iterations for the second, respectively. Note that the parameters introduced by the relaxed molecular clock model are also explored in the second sampling stage. The SNVs were then annotated using Annotvar (version 2020 Jun. 08) [57]. In the main text, the tree was plotted using ggtree [58] and the genotype heatmap was plotted using ComplexHeatmap [59].

Run time analysis

Repeated five times, we used a simulation scenario with the following parameters for run time analysis: medium mutation rate (8×10^{-6}) and medium coverage quality ($t = 20, \nu = 10$). SiFit and SCIPHi were run in the default, single-thread mode, while CellPhy and SIEVE were run in both single- and multi-thread mode, where different numbers of threads were provided to achieve their highest efficiency. SiFit, SCIPHi and the two stages of SIEVE were run for 10^6 iterations, respectively. With bootstrap applied, CellPhy was run with the default setting (a maximum of 1000 replicates with a possible early-stopping). This analysis was performed on a server with 64 cores (AMD Ryzen Threadripper 3990X 64-Core Processor) and 256 GB memory.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-022-02813-9>.

Additional file 1: Supplementary Figs. S1-S21, Tables S1-S4, and Note.

Additional file 2: Review history.

Acknowledgements

We thank Dr. Timothy Vaughan for valuable instructions on package development for BEAST 2.

Peer review information

Stephanie McClelland and Veronique van den Berghe were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 2.

Authors' contributions

S.K. and E.S. conceived the SIEVE model — with input and feedback from J.K., N.B.E. and D.P. S.K. implemented the model, performed all model performance analysis and generated all figures. S.P.L., D.C. and D.P. performed the CRC28 scDNA-seq experiment. N.B.O., M.V. and J.M.A. processed the scDNA-seq datasets. S.K. and E.S. wrote the manuscript with critical comments and input from all the co-authors. E.S. supervised the study. The authors read and approved the final manuscript.

Authors' Twitter handles

Twitter handles: @senbai_kang (Senbai Kang); @cbg_ethz (Nico Borgsmüller, Jack Kuipers, and Niko Beerenwinkel); @linkmonica (Monica Valecha); @JMFAlves (Joao M. Alves); @dposada_ (David Posada); @ewa_szczurek (Ewa Szczurek).

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 766030. E.S. acknowledges the support from the Polish National Science Centre SONATA BIS grant No. 2020/38/E/NZ2/00305. D.P. was supported by the European Research Council (ERC-617457-PHYLOCANCER), the Spanish Ministry of Science and Innovation (PID2019-106247GB-I00), and Xunta de Galicia.

Availability of data and materials

Raw single-cell whole-genome sequencing data from CRC28 have been deposited at the National Center for Biotechnology Information (NCBI) as BioProject PRJNA896550 [60]. We have additionally analysed two published single-cell datasets ([42, 43]). Raw sequencing data for these datasets are available from the Sequence Read Archive (SRA, <https://www.ncbi.nlm.nih.gov/sra>) database under accession codes SRA053195 (TNBC16) and SRP067815 (CRC48).

SIEVE is implemented in Java and is accessible at <https://github.com/szczurek-lab/SIEVE> [61]. DataFilter for selecting candidate variant sites is available at <https://github.com/szczurek-lab/DataFilter> [62]. The simulator is hosted at https://github.com/szczurek-lab/SIEVE_simulator [63], and the reproducible benchmarking framework is available at https://github.com/szczurek-lab/SIEVE_benchmark_pipeline [64]. The scripts for generating all figures in this paper are hosted at https://github.com/szczurek-lab/SIEVE_analysis [65]. All aforementioned code are freely accessible under a GNU General Public License v3.0 license. The source code versions used in the manuscript can be downloaded from the Zenodo repository [66–70].

Declarations

Ethics approval and consent to participate

We obtained fresh frozen primary tumour and normal tissues from a single colorectal cancer patient (CRC28) stored at the Galicia Sur Health Research Institute (IISGS) Biobank, member of the Spanish National Biobank Network (N^o B.0000802). This study was approved by a local Ethical and Scientific Committee (CAE Galicia 2014/015).

Consent for publication

Not applicable.

Competing interests

Other projects in the research lab of E.S. are co-funded by Merck Healthcare KGaA.

Received: 23 April 2022 Accepted: 8 November 2022

Published online: 30 November 2022

References

- Greaves M. Evolutionary Determinants of Cancer. *Cancer Discov.* 2015;5(8):806–20. <https://doi.org/10.1158/2159-8290.CD-15-0439>.
- Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell.* 2021;184(8):2239–2254.e39. <https://www.sciencedirect.com/science/article/pii/S0092867421002944>.
- McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell.* 2017;168(4):613–628. <https://www.sciencedirect.com/science/article/pii/S0092867417300661>.
- Marusyk A, Janiszewska M, Polyak K. Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. *Cancer Cell.* 2020;37(4):471–484. <https://www.sciencedirect.com/science/article/pii/S1535610820301471>.
- Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, Moch H, et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun.* 2012;3(1):1–8.
- Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature.* 2012;486(7403):395–9.
- Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods.* 2014;11(4):396–8.
- Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* 2014;24(11):1881–93.
- Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* 2015;16(1):1–20.
- Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. *Nature.* 2011;472(7341):90–4.
- Navin NE. The first five years of single-cell cancer genomics and beyond. *Genome Res.* 2015;25(10):1499–507.
- Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. *Genome Biol.* 2020;21(1):1–35.
- Zafar H, Wang Y, Nakhleh L, Navin N, Chen K. Monovar: single-nucleotide variant detection in single cells. *Nat Methods.* 2016;13(6):505–7.
- Dong X, Zhang L, Milholland B, Lee M, Maslov AY, Wang T, et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods.* 2017;14(5):491–3.
- Singer J, Kuipers J, Jahn K, Beerenwinkel N. Single-cell mutation identification via phylogenetic inference. *Nat Commun.* 2018;9(1):5144. <https://doi.org/10.1038/s41467-018-07627-7>.
- Luquette LJ, Bohrsen CL, Sherman MA, Park PJ. Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. *Nat Commun.* 2019;10(1):1–14.
- Bohrsen CL, Barton AR, Lodato MA, Rodin RE, Luquette LJ, Viswanadham VV, et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nat Genet.* 2019;51(4):749–54.
- Lähnemann D, Köster J, Fischer U, Borkhardt A, McHardy AC, Schönhuth A. Accurate and scalable variant calling from single cell DNA sequencing data with ProSolo. *Nat Commun.* 2021;12(1):1–11.
- Yuan K, Sakoparnig T, Markowitz F, Beerenwinkel N. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol.* 2015;16(1):1–16.

20. Ross EM, Markowitz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.* 2016;17(1):1–14.
21. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. *Genome Biol.* 2016;17(1):1–17.
22. Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol.* 2017;18(1):1–20.
23. Malikić S, Jahn K, Kuipers J, Sahinalp SC, Beerenwinkel N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature Commun.* 2019;10(1):1–12.
24. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics.* 2019;35(21):4453–5.
25. Zafar H, Navin N, Chen K, Nakhleh L. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Res.* 2019;29(11):1847–59.
26. Kozlov A, Alves JM, Stamatakis A, Posada D. Cell Phy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. *Genome Biol.* 2022;23(1):1–30.
27. Felsenstein J. *Inferring phylogenies*, vol. 2. Sunderland: Sinauer Associates; 2004.
28. Stadler T, Pybus OG, Stumpf MP. Phylodynamics for cell biologists. *Science.* 2021;371(6526):eaa6266.
29. Lewis PO. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology.* 2001;50(6):913–25. <https://doi.org/10.1080/106351501753462876>.
30. Leaché AD, Banbury BL, Felsenstein J, de Oca AnM, Stamatakis A. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. *Syst Biol.* 2015;64(6):1032–47. <https://doi.org/10.1093/sysbio/syv053>.
31. Kuipers J, Singer J, Beerenwinkel N. Single-cell mutation calling and phylogenetic tree reconstruction with loss and recurrence. *Bioinformatics.* 2022;btac577. <https://doi.org/10.1093/bioinformatics/btac577>.
32. Kuipers J, Jahn K, Raphael BJ, Beerenwinkel N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.* 2017;27(11):1885–94.
33. Demeulemeester J, Dentre SC, Gerstung M, Van Loo P. Biallelic mutations in cancer genomes reveal local mutational determinants. *Nat Genet.* 2022;54(2):128–133. <https://doi.org/10.1038/s41588-021-01005-8>.
34. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput Biol.* 2019;15(4):1–28. <https://doi.org/10.1371/journal.pcbi.1006650>.
35. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution.* 1994;11(3):459–68. <https://doi.org/10.1093/oxfordjournals.molbev.a040126>.
36. Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math Biosci.* 1981;53(1):131–147. <https://www.sciencedirect.com/science/article/pii/0025556481900432>.
37. Douglas J, Zhang R, Bouckaert R. Adaptive dating and fast proposals: Revisiting the phylogenetic relaxed clock model. *PLOS Comput Biol.* 2021;17(2):1–30. <https://doi.org/10.1371/journal.pcbi.1008322>.
38. Huang D, Sun W, Zhou Y, Li P, Chen F, Chen H, et al. Mutations of key driver genes in colorectal cancer progression and metastasis. *Cancer Metastasis Rev.* 2018;37(1):173–87.
39. Raskov H, Søby JH, Troelsen J, Bojesen RD, Gögenur I. Driver gene mutations and epigenetics in colorectal cancer. *Ann Surg.* 2020;271(1):75–85.
40. Müller T, Stein U, Poletti A, Garzia L, Rothley M, Plaumann D, et al. ASAP1 promotes tumor cell motility and invasiveness, stimulates metastasis formation in vivo, and correlates with poor survival in colorectal cancer patients. *Oncogene.* 2010;29(16):2393–403. <https://doi.org/10.1038/onc.2010.6>.
41. Sun MS, Yuan LT, Kuei CH, Lin HY, Chen YL, Chiu HW, et al. RGL2 Drives the Metastatic Progression of Colorectal Cancer via Preventing the Protein Degradation of β -Catenin and KRAS. *Cancers.* 2021;13(8). <https://doi.org/10.3390/cancers13081763>.
42. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature.* 2014;512(7513):155–60. <https://doi.org/10.1038/nature13600>.
43. Wu H, Zhang X, Hu Z, Hou Q, Zhang H, Li Y, et al. Evolution and heterogeneity of non-hereditary colorectal cancer revealed by single-cell exome sequencing. *Oncogene.* 2017;36(20):2857–67.
44. D'Andrea AD. 4 - DNA Repair Pathways and Human Cancer. In: Mendelsohn J, Gray JW, Howley PM, Israel MA, Thompson CB, editors. *The Molecular Basis of Cancer (Fourth Edition)*. fourth edition ed. Philadelphia: W.B. Saunders; 2015. p. 47–66.e2. <https://www.sciencedirect.com/science/article/pii/B9781455740666000044>.
45. Yang Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol.* 1996;11(9):367–372. <https://www.sciencedirect.com/science/article/pii/0169534796100410>.
46. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* 2010;11(10):R106. <https://doi.org/10.1186/gb-2010-11-10-r106>.
47. Felsenstein J. Phylogenies from restriction sites: a maximum-likelihood approach. *Evolution.* 1992;46(1):159–73.
48. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol.* 1981;17(6):368–76. <https://doi.org/10.1007/BF01734359>.
49. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics.* 2002;161(3):1307–1320. <https://www.genetics.org/content/161/3/1307>.
50. Bishop CM, Nasrabadi NM. *Pattern recognition and machine learning*. New York, US: Springer; 2006.
51. O'Reilly JE, Donoghue PC. The efficacy of consensus tree methods for summarizing phylogenetic relationships from a posterior sample of trees estimated from morphological data. *Syst Biol.* 2018;67(2):354–62.
52. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed Phylogenetics and Dating with Confidence. *PLOS Biol.* 2006;4(5):null. <https://doi.org/10.1371/journal.pbio.0040088>.
53. Drummond AJ, Suchard MA. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* 2010;8(1):114. <https://doi.org/10.1186/1741-7007-8-114>.
54. Posada D. Cell Coal: Coalescent Simulation of Single-Cell Sequencing Samples. *Mol Biol Evol.* 2020;37(5):1535–42. <https://doi.org/10.1093/molbev/msaa025>.

55. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520–2. <https://doi.org/10.1093/bioinformatics/bts480>.
56. Schliep K, Potts AJ, Morrison DA, Grimm GW. Intertwining phylogenetic trees and networks. *Methods Ecol Evol*. 2017;8(10):1212–1220. <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12760>.
57. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164–e164. <https://doi.org/10.1093/nar/gkq603>.
58. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*. 2017;8(1):28–36. <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12628>.
59. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32(18):2847–9. <https://doi.org/10.1093/bioinformatics/btw313>.
60. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. SIEVE: joint inference of single-nucleotide variants and cell phylogeny from single-cell DNA sequencing data. *Datasets*. Bioproject; 2022. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA896550>. Accessed 1 Nov 2022.
61. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. SIEVE: joint inference of single-nucleotide variants and cell phylogeny from single-cell DNA sequencing data. *GitHub*; 2022. <https://github.com/szczurek-lab/SIEVE>.
62. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. SIEVE: joint inference of single-nucleotide variants and cell phylogeny from single-cell DNA sequencing data. *GitHub*; 2022. <https://github.com/szczurek-lab/DataFilter>.
63. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. SIEVE: joint inference of single-nucleotide variants and cell phylogeny from single-cell DNA sequencing data. *GitHub*; 2022. https://github.com/szczurek-lab/SIEVE_simulator.
64. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. SIEVE: joint inference of single-nucleotide variants and cell phylogeny from single-cell DNA sequencing data. *GitHub*; 2022. https://github.com/szczurek-lab/SIEVE_benchmark_pipeline.
65. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. SIEVE: joint inference of single-nucleotide variants and cell phylogeny from single-cell DNA sequencing data. *GitHub*; 2022. https://github.com/szczurek-lab/SIEVE_analysis.
66. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. SIEVE v0.15.6. *Zenodo*; 2022. <https://doi.org/10.5281/zenodo.7270031>.
67. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. DataFilter v0.1.0. *Zenodo*; 2022. <https://doi.org/10.5281/zenodo.7270015>.
68. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. SIEVE_simulator v1.3.0. *Zenodo*; 2022. <https://doi.org/10.5281/zenodo.7270021>.
69. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. SIEVE_benchmark_pipeline v0.1.0. *Zenodo*; 2022. <https://doi.org/10.5281/zenodo.7270025>.
70. Kang S, Borgsmüller N, Valecha M, Kuipers J, Alves JM, Prado-López S, et al. SIEVE_analysis v0.1.0. *Zenodo*; 2022. <https://doi.org/10.5281/zenodo.7270027>.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

