Computational and Structural Biotechnology Journal 20 (2022) 2978–2985

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini review

# Somatic variant calling from single-cell DNA sequencing data

Monica Valecha [a,b], David Posada [a,b,c,*]

[a] CINBIO, Universidade de Vigo, 36310 Vigo, Spain
[b] Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Spain
[c] Department of Biochemistry, Genetics, and Immunology, Universidade de Vigo, 36310 Vigo, Spain

ARTICLE INFO

*Article history:*
Received 1 April 2022
Received in revised form 6 June 2022
Accepted 6 June 2022
Available online 14 June 2022

*Keywords:*
Variant calling
Single-cell genomics
Allele dropout
Amplification error
Somatic variants

ABSTRACT

Single-cell sequencing has gained popularity in recent years. Despite its numerous applications, single-cell DNA sequencing data is highly error-prone due to technical biases arising from uneven sequencing coverage, allelic dropout, and amplification error. With these artifacts, the identification of somatic genomic variants becomes a challenging task, and over the years, several methods have been developed explicitly for this type of data. Single-cell variant callers implement distinct strategies, make different use of the data, and typically result in many discordant calls when applied to real data. Here, we review current approaches for single-cell variant calling, emphasizing single nucleotide variants. We highlight their potential benefits and shortcomings to help users choose a suitable tool for their data at hand.

**Contents**

## 1. Introduction

In recent years, single-cell sequencing studies have gained momentum due to their capacity to disentangle biological differences in apparently homogeneous tissues [1–6]. While the single-cell field has logically focused on single-cell RNA sequencing (scRNA-seq) [7], due to its ability to unveil functional variation directly, single-cell DNA sequencing (scDNA-seq) [8] has also been

*Abbreviations:* scDNA-seq, single-cell DNA sequencing; scRNA-seq, single-cell RNA sequencing; scMethyl-seq, single-cell Methylation sequencing; scATAC-seq, single-cell sequencing assay for transposase-accessible chromatin; scHi-C, single-cell Hi-C sequencing; scWGA, single-cell whole-genome amplification; ADO, allelic dropout; LDO, locus dropout; SNV, single nucleotide variant; SV, structural variant; CNV, copy number variant; Indel, short insertion or deletion; hSNP, heterozygous single-nucleotide polymorphism; VAF, variant allele frequency.

[*] Corresponding author at: CINBIO, Universidade de Vigo, 36310 Vigo, Spain.
*E-mail address:* dposada@uvigo.es (D. Posada).

https://doi.org/10.1016/j.csbj.2022.06.013
2001-0370/

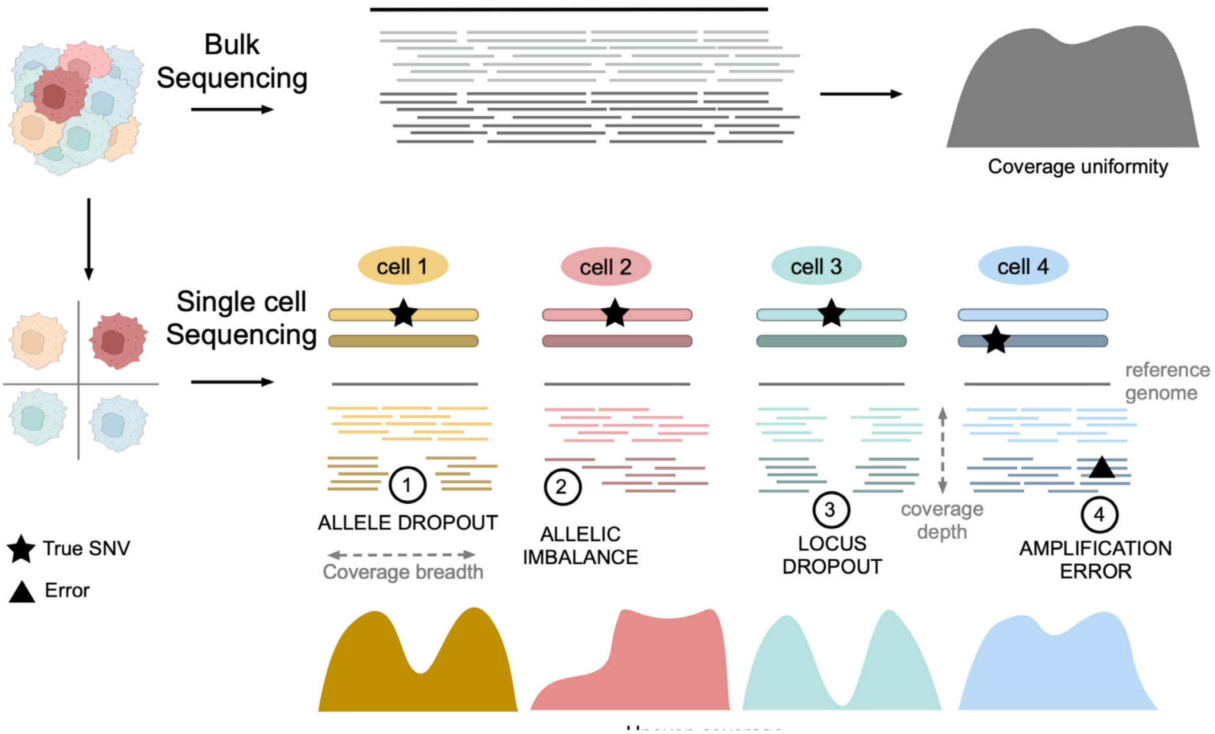

**Fig. 1.** Single-cell whole-genome amplification biases. Technical biases arising during single-cell whole genome amplification can be detected from the sequencing reads, like allele dropout (ADO) (1), allelic imbalance (AI) (2), locus dropout (LDO) (3), and amplification errors (4). Coverage breadth and depth are more heterogeneous for single-cell compared to bulk sequencing.

beneficial in understanding the role of somatic mutations in development, aging, and disease, particularly in cancer [9–19]. Nowadays, single-cell omic sequencing goes beyond the study of genomic variants or expression levels and includes the analysis of methylation (e.g., scMethyl-seq [20]) and chromatin architecture with techniques like scATAC-seq [21] and scHi-C [22]. These single-cell sequencing technologies are used in research fields like cancer, microbiology, neurology, development, or reproduction biology [23–26]. This review will focus on the bioinformatic methods developed to identify genomic alterations from scDNA-seq data, particularly single-nucleotide variants (SNVs). Excellent reviews already exist for the detection of copy-number variants (CNVs) from scDNA-seq [27] and also for the identification of SNVs from scRNA-seq data [28], so we will not pursue these topics further.

### *1.1. Errors in single-cell DNA sequencing data*

A single cell typically contains a limited amount of DNA, around 6–7 pg in the case of human cells. Because of this, most single-cell protocols require a whole-genome amplification step (scWGA) to produce enough DNA for sequencing [8,29]. After scWGA, some regions in the original genome become overrepresented, while others are underrepresented or never amplified. When the single-cell libraries are sequenced, the coverage along the genome is very heterogeneous. Often, maternal or paternal alleles are disproportionally represented (i.e., allelic imbalance or bias; AI) or absent (i.e., allelic dropout; ADO) in the sequencing reads. Furthermore, errors in the DNA can occur during cell lysis, DNA extraction, library preparation, or scWGA, leading to a number of spurious nucleotide changes in the sequencing reads [8,30]. scDNA-seq data are therefore very noisy. This technical noise can result in missing or wrong genotype calls during single-cell variant calling, both false positives and false negatives (Fig. 1).

### *1.2. Modeling of single-cell DNA sequencing errors*

To deal with single-cell noise, multiple variant callers have been proposed specifically for scDNA-seq data (described below). These tools can call a variety of genomic variants, including single-nucleotide variants (SNVs), copy-number variants (CNVs), small insertion or deletions (indels), and structural variants (SVs). Calling genomic variants from scRNA-seq data is less common, but some tools have been developed for this purpose [31]. Below we review and discuss the technical properties and usability of different scDNA-seq callers. As an excellent review is in place for scDNA-seq CNV detection [27], we concentrate on scDNA-seq SNV calling. But before exploring the different scDNA-seq variant callers available, we believe it is worth describing how these methods typically deal with the technical errors resulting from the scWGA step. The basic model for scDNA-seq error assumes only two states for the genotypes, mutated (1) or not (0), and that false positive and false negative scWGA errors occur at rates α and β, respectively.

If for a given cell and locus, P (A | G), is the probability of the amplified genotype (A) given the true genotype (G), then:

$$\begin{aligned} &P(0|0) = 1 - \alpha \\ &P(0|1) = \beta \\ &P(1|0) = \alpha \\ &P(1|1) = 1 - \beta \end{aligned} \tag{1}$$

While the probabilities above correspond to the genotypes, the input data for all variant callers are the sequencing reads. Read counts at a locus are typically modeled using a Beta-Binomial distribution, as for bulk data [32]. Therefore, the probability of amplified genotype $A$ given the observed number of reference ($r$) and alternate read ($a$) counts at a given locus and cell is:

$$\begin{aligned} P(r,a|A=0) &= \binom{r+a}{a}\varepsilon^a(1-\varepsilon)^r \\ P(r,a|A=1) &= \binom{r+a}{a}\mu^a(1-\mu)^r \end{aligned} \quad (2)$$

where ε and μ are Beta distributed variables for the probability of drawing an alternate read. Note that ε represents the sequencing error. Finally, the joint probability of the observed read counts and the amplified genotype given the true genotype can be computed by multiplying the probabilities above:

$$\begin{aligned} P(r,a,A=0|G=0) &= \binom{r+a}{a}\varepsilon^a(1-\varepsilon)^r(1-\alpha) \\ P(r,a,A=0|G=1) &= \binom{r+a}{a}\varepsilon^a(1-\varepsilon)^r\beta \\ P(r,a,A=1|G=0) &= \binom{r+a}{a}\mu^a(1-\mu)^r\alpha \\ P(r,a,A=1|G=1) &= \binom{r+a}{a}\mu^a(1-\mu)^r(1-\beta) \end{aligned} \quad (3)$$

Then, for $n$ cells and $m$ loci, the likelihood function for the probability of the true genotype $G$ given the read counts is:

$$L(G) = \prod_{i=1}^{n}\prod_{j=1}^{m} P(r_{ij}, a_{ij}|G_{ij}) \quad (4)$$

Statistical methods can then be applied to calculate the maximum likelihood genotypes or their posterior distribution given the observed reads. Similar models have been described or extended by different people [33–39].

## 2. Single-cell variant callers

### 2.1. SNV scDNA-seq callers

We identified ten tools specifically designed for calling SNVs from scDNA-seq data. They adopt different methodological strategies (Table 1), have distinct capabilities (Table 2), possess specific technical features (Table 3), and are freely accessible from public repositories (Table 4). These tools assume that data and errors at different loci are independent and that the SNVs are biallelic and located in diploid regions. The input data are then mapped sequencing reads (BAM format) or the read counts with their base quality scores (mpileup format). While the different tools specify a minimum number of reads per site by default, these are not strict requirements and can be changed.

Monovar [38], the first single-cell variant caller, uses mapped reads from multiple cells to compute the posterior probability of a locus containing at least one alternate allele. In doing so, it calculates the likelihood of the heterozygote and homozygote genotypes, accounting for false-positive errors and ADO, via a dynamic programming algorithm. After assigning each cell the genotype with the highest posterior probability, an optional consensus filter retains only variants called in two or more cells.

SCcaller [40] computes the likelihood of being a heterozygous or homozygous SNV, or an artifact, for a set of candidate variant loci. Then it uses a likelihood-ratio test to distinguish real SNVs from artifacts, whose null distribution accounts for AI, sequencing depth, and quality. Here, the level of AI is estimated independently for each candidate SNV using a kernel smoothing that considers the degree of bias in the read distribution from neighboring heterozygous germline single-nucleotide polymorphisms (hSNPs). SCcaller only calls variants and does not distinguish missing data from an unmutated state.

SCIΦ [39] was the first caller that jointly inferred the evolutionary relationships between cells and the cell genotypes. SCIΦ first identifies candidate SNVs based on the posterior probability of observing one or more mutated cells at each locus. For this, it models the read count distribution considering amplification errors and ADO. The candidate SNVs are then used to estimate the underlying cell phylogeny and model parameters using Markov Chain Monte Carlo (MCMC), considering the zygosity. In a final step, mutations are assigned to the individual cells sampling from the posterior distribution approximated by the MCMC. SCIΦN [41] relaxes the infinite site assumption (see below) of SCIΦ, allowing for mutational recurrence and loss.

LiRA [42] uses read-phasing information from physically-linked hSNPs to distinguish real SNVs from scWGA artifacts. In the set of "spanning" reads that cover the positions of both an SNV and an hSNP, true SNVs will appear together with either the alternate or the reference hSNP allele. In contrast, false SNVs will appear both with the alternate and the reference hSNP allele (Fig. 2). ADO will be detected here when all reads carry either the alternative or the reference hSNP allele. LiRA initially identifies candidate SNVs using the Genome Analysis Toolkit (GATK) [43], jointly for every cell and a matched healthy bulk, requiring no reads for the alternate allele in the bulk sample and at least one read for the alternate allele in the single cell. hSNPs are taken from dbSNP [44]. While this linked-hSNP strategy reduces the number of SNVs that can be identified, it should produce very precise calls. As SCcaller, LIRA only calls mutated genotypes. To maintain the estimated FDR at a tolerable level, LiRA estimates the minimum composite coverage threshold by which FDR ≤10%.

As LiRA, Conbase [45] uses phasing information from hSNPs to correct for errors and allelic dropout, but unlike it, looks for haplo-

**Table 1**
Methodological strategies and assumptions of scDNA-seq variant callers. Tools can use all cells simultaneously (joint calling) or do mutation calling cell by cell (marginal calling). Some callers use phylogenetic information or follow the infinite-sites assumption. The allelic imbalance and the amplification error are assumed to be constant across the genome (global) or not (local). Some tools use linked hSNPs to identify errors.

| | Calling strategy | Phylogeny | Infinite sites assumption | Allelic Imbalance/dropout | Amplification error | Linked hSNPs |
|---|---|---|---|---|---|---|
| Monovar | joint | no | no | global | global | no |
| SCcaller | marginal | no | no | local | global | no |
| SCIΦ | joint | yes | yes | global | global | no |
| LiRA | marginal | no | no | local | local | yes |
| Conbase | joint | no | no | local | local | yes |
| SCAN-SNV | joint | no | no | local | local | no |
| scVILP | joint | yes | yes | global | global | no |
| ProSolo | marginal | no | no | local | local | no |
| SCIΦN | joint | yes | no | global | global | no |
| Phylovar | joint | yes | yes | global | global | no |

**Table 2**
Capabilities of scDNA-seq variant callers. All callers identify somatic variants, whereas some also include germline variants and indels in their output. Some callers can also give homozygous mutant genotypes and impute missing genotypes. Most callers can call singletons (mutations that appear just in one cell), and SCAN-SNV also detects doublets (pairs of cells erroneously treated as a single cell).

| | Germline calls | Somatic calls | Indels | Homozygous mutations | Genotype imputation | Call singletons | Detects doublets |
|---|---|---|---|---|---|---|---|
| Monovar | yes | yes | no | yes | no | yes | no |
| SCcaller | no | yes | yes | yes | no | yes | no |
| SCIΦ | no | yes | no | no | yes | yes | no |
| LiRA | no | yes | no | no | no | yes | no |
| Conbase | no | yes | no | yes | no | no | no |
| SCAN-SNV | no | yes | no | yes | no | yes | yes |
| scVILP | yes | yes | no | no | yes | yes | no |
| ProSolo | yes | yes | no | yes | yes | yes | no |
| SCIΦN | no | yes | no | no | yes | yes | no |
| Phylovar | yes | yes | no | no | no | yes | no |

**Table 3**
Technical features of scDNA-seq variant callers. The input formats can be BAM (https://samtools.github.io/hts-specs/SAMv1.pdf) or mpileup (http://www.htslib.org/doc/samtools-mpileup.html). All tools require a reference human genome, and some of them also need a set of candidate SNVs and SNPs, normal/tumor bulk samples, or a dbSNP file (https://www.ncbi.nlm.nih.gov/snp). The output format is VCF (https://samtools.github.io/hts-specs/VCFv4.2.pdf), its binary counterpart BCF, TSV (tab-separated values), or RDA (R data file).

| | Input format | Other input files | Bulk sample | dbSNP | Output | Computer language |
|---|---|---|---|---|---|---|
| Monovar | BAM | Ref. genome | no | no | VCF | Python |
| SCcaller | BAM | Ref. genome | normal | yes | VCF | Python |
| SCIΦ | Mpileup | Ref. genome | normal | no | VCF | C++ |
| LiRA | BAM | Ref. genome, candidate SNVs | normal | yes | VCF | Python, R |
| Conbase | BAM | Ref. genome, SNPs | normal | no | TSV | Python |
| SCAN-SNV | BAM | Ref. genome | normal | yes | RDA | Python, R |
| scVILP | Mpileup | Ref. genome | no | no | VCF | Python, C++ |
| ProSolo | BAM | Ref. genome, candidate SNVs | tumor | no | BCF | Python |
| SCIΦN | Mpileup | Ref. genome | normal | no | VCF | C++ |
| Phylovar | Mpileup | Ref. genome | no | no | VCF | Python |

**Table 4**
References and URLs for scDNA-seq variant callers.

| | Reference | URL |
|---|---|---|
| Monovar | Zafar et al., 2016 | https://bitbucket.org/hamimzafar/monovar/ |
| SCcaller | Dong et al., 2017 | https://github.com/biosinodx/SCcaller |
| SCIΦ | Singer et al., 2018 | https://github.com/cbg-ethz/SCIPhI |
| LiRA | Bohrson et al., 2019 | https://github.com/parklab/LiRA |
| Conbase | Hård et al., 2019 | https://github.com/conbase/conbase |
| SCAN-SNV | Luquette et al., 2019 | https://github.com/parklab/scan-snv |
| scVILP | Edrisi et al., 2019 | https://github.com/mae6/scVILP |
| ProSolo | Lähnemann et al., 2021 | https://github.com/ProSolo/prosolo |
| SCIΦN | Kuipers et al., 2022 | https://github.com/cbg-ethz/SCIPhIN |
| Phylovar | Edrisi et al., 2022 | https://github.com/NakhlehLab/Phylovar |

type concordance across all cells, distinguishes missing data from unmutated genotypes, only calls SNVs present in at least two cells, and considers SNPs other than dbSNP, augmenting the proportion of phasable SNVs slightly. Conbase starts by identifying SNVs linked to hSNPs where multiple cells show support for the same alternative allele. Then it considers the different combinations ("tuple pairs") of SNV and linked hSNP alleles in the reads across all cells, using adjustable frequency thresholds for the tuple pairs to call the single-cell genotypes.

SCAN-SNV [46] implements a genome-wide spatial model of AI that leverages the variant allele frequency (VAF) at a large set of phased hSNPs. SCAN-SNV first uses GATK HaplotypeCaller on single-cell and bulk sequencing data to generate a list of candidate variant sites. Then, hSNPs in the bulk data are phased with SHAPEIT [47] and used to train the AI model. SCAN-SNV identifies true SNVs by requiring the VAF of candidate SNVs to match the estimated (balanced or imbalanced) local VAF and be inconsistent with typical pre-amplification and early-amplification artifacts. This method also includes a false discovery rate (FDR) tuning strategy based on the stringency of the p-value thresholds for candidate SNVs with low VAFs.

scVILP [33] implements the joint inference of single-cell SNVs and the cell phylogeny as a combinatorial optimization problem. Using statistical models for single-cell errors and sequencing coverage, it tries to identify the set of single-cell genotypes that maximizes the probability of the observed read counts while enforcing the infinite-sites assumption (i.e., SNVs occur once along a "perfect" cell phylogeny). scVILP can impute genotypes at loci with missing data (i.e., without read counts). It requires quite a bit of memory and is more suited for target sequencing data.

ProSolo [48] uses a probabilistic model that considers the specific biases of multiple-displacement amplified (MDA) [49] scWGA in a site-specific manner, following a mechanistic model of amplification bias trained on empirical data [see [50]] and assessing amplification errors upon a bulk sample from which the single cells are supposed to stem. Furthermore, Prosolo can impute genotypes using the bulk sample, calculate the posterior probability of ADO at a particular site, and flexibly control the FDR.

Phylovar [51] is a likelihood-based method for joint variant calling and phylogenetic inference, similar to SCIΦ and scVILP, but that has been specifically designed to scale well with large data sets with thousands of SNVs. Phylovar first identifies candidate SNVS using SCIΦ's LRT. Afterward, a hill-climbing search algorithm is used to maximize the probability of the observed read counts given the cell phylogeny, the placement of mutations, and the single-cell error rates.

### *2.1.1. Calling strategies*

*2.1.1.1. Joint vs. marginal calling.* Single-cell SNV callers could be broadly classified into two groups depending on whether they per-

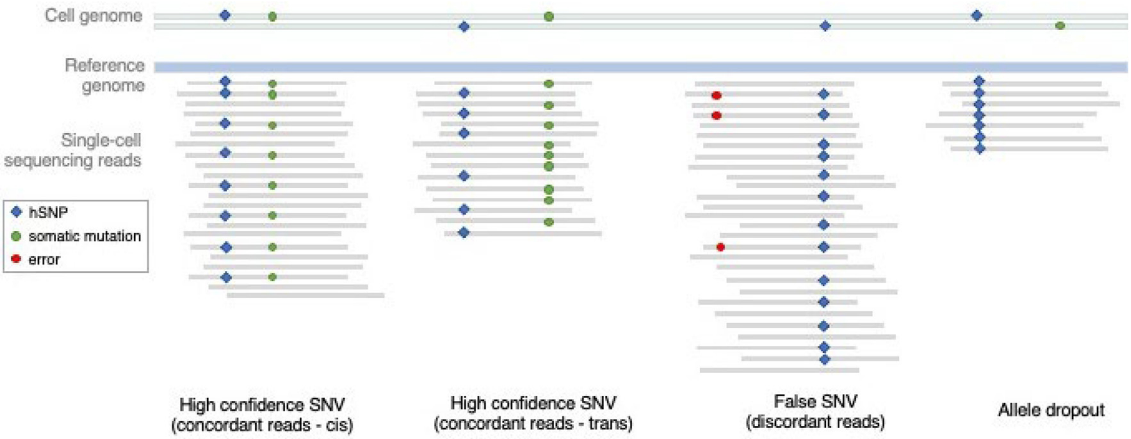
Cell genome
Reference genome
Single-cell sequencing reads
◆ hSNP
● somatic mutation
● error
High confidence SNV (concordant reads - cis)
High confidence SNV (concordant reads - trans)
False SNV (discordant reads)
Allele dropout

**Fig. 2.** SNV assessment with linked hSNPs. Variant alleles at SNVs and linked hSNPs should appear consistently in the same reads if they occur in the same chromosome in the original cell genome (*cis* configuration). On the contrary, they should appear consistently in different reads if they occur in different chromosomes in the original cell genome (*trans* configuration). During cell lysis or scWGA in *cis* with the hSNP variant alleles, errors will appear exclusively on a fraction of the reads that carry the hSNP alternate allele. In contrast, errors in *trans* with the hSNP variant alleles will appear exclusively on a fraction of the reads that do not carry the hSNP alternate allele. ADO becomes evident when all or none of the linked reads carry the hSNP alternate allele.

form joint or marginal variant calling. During joint calling, the information from all cells is considered at once, while in marginal calling, each cell is analyzed in turn. Monovar, Conbase, SCIΦ, SCIΦN, SCAN-SNV, scVILP, and Phylovar can perform joint variant calling, while SCcaller, LiRA, and ProSolo carry out marginal calling.

*2.1.1.2. Use of phylogenetic information.* Most joint calling tools like SCIΦ, SCIΦN, scVILP, and Phylovar leverage the phylogenetic information contained in single-cell genomes to call variants. The key idea is that closely related cells are expected to share the same genotype more often than cells distant in the phylogeny.

*2.1.1.3. Infinite-sites assumption.* Given the low mutation rate typically assumed for somatic cells, the callers that use phylogenetic information (SCIΦ, scVILP, and Phylovar, but not SCIΦN) follow the so-called infinite-sites assumption (ISA), by which a mutation is supposed to occur only once at a given genomic site. Methods that do not use phylogenetic information do not constrain the number of mutations at a given locus and therefore implicitly allow for violations of the ISA. Recent studies suggest that, at least in cancer, the ISA might not hold universally [52,53].

*2.1.1.4. Allelic imbalance and allelic dropout.* AI and ADO are fundamental biases introduced by the scWGA step that can severely confound variant calling from single cells. Different methods use distinct approaches to deal with this. Monovar, SCIΦ, scVILP, SCIΦN, and PhyloVar assume or estimate a global ADO or false-negative rates like the one described in Eq. (1) above, which applies to all loci and cells.

SCcaller and SCAN-SNV estimate how AI varies across the genome using neighboring hSNPs. In the absence of allelic imbalance, the VAF at hSNPs should be 0.5, becoming 0 or 1 in the case of ADO. LiRA and Conbase also take advantage of hSNPs but focus on those located in the same reads as the candidate SNVs. ADO is detected when all reads, or none, contain the variant allele at the linked hSNPs (Fig. 2). Read-backed phasing allows for more reliable identification of singletons –SNVs seen only in one cell– but restricts the SNV call set to sites linked to an hSNP.

To model AI, ProSolo leverages a beta-binomial mixture model from Lodato et al. [50], whose parameters can be learned from empirical data. Calls are improved using an unamplified bulk sample from the same cell population as the single cells.

*2.1.1.5. Amplification error.* As in the case of ADO, the different methods deal distinctively with errors arising during cell isolation, lysis, or scWGA. Monovar, SCcaller, SCIΦ, scVILP, SCIΦN, and Phylovar include a global error or false positive rate parameter across loci and cells (see Eq. (1)).

LiRA and Conbase use linked hSNPs to identify potential false positives at candidate SNVs. SCAN-SNV detects false positives as candidate SNVs that do not match the estimated AI estimated for that region or have a VAF consistent with pre-amplification or early amplification artifacts. ProSolo assumes a locus-specific false positive error rate.

*2.1.1.6. Genotype imputation.* Furthermore, due to the remarkable coverage heterogeneity in scDNA-seq data, often, some genotypes are not called. Some single-cell callers can impute the missing genotypes, either using a matched bulk sample, like ProSolo, or phylogenetic information, according to the genotypes assigned to the internal nodes of the cell phylogeny, like SCIΦ and SCIΦN.

*2.1.1.7. Post-calling filters.* Distinguishing errors from singletons (i.e., mutations seen only in one cell) is not easy unless one relies on specific strategies, like linked reads. Therefore, some tools implement a consensus post-calling filter that keeps only those mutations seen in at least two cells. Conbase and Monovar apply a consensus filter during variant calling, which can be deactivated in the case of Monovar. Indeed, researchers can write their own filters, and other tools exist to improve the quality of the single-cell genotypes [54].

### 2.2. Performance comparison

#### 2.2.1. Statistical performance

The statistical performance of the different single-cell somatic variant callers has been benchmarked every time a new method was introduced. These studies typically compare the callers in terms of precision (the fraction of the called variants that are true), false discovery rate (the fraction of the called variants that are wrong), or recall (the fraction of true variants that are identified).

In general, these comparisons consider subsets of the available methods, and in most cases, the method introduced by the authors performs better than the competitors, throwing inconsistent results among different studies. According to these benchmarks, single-cell variant callers have higher precision than recall, which

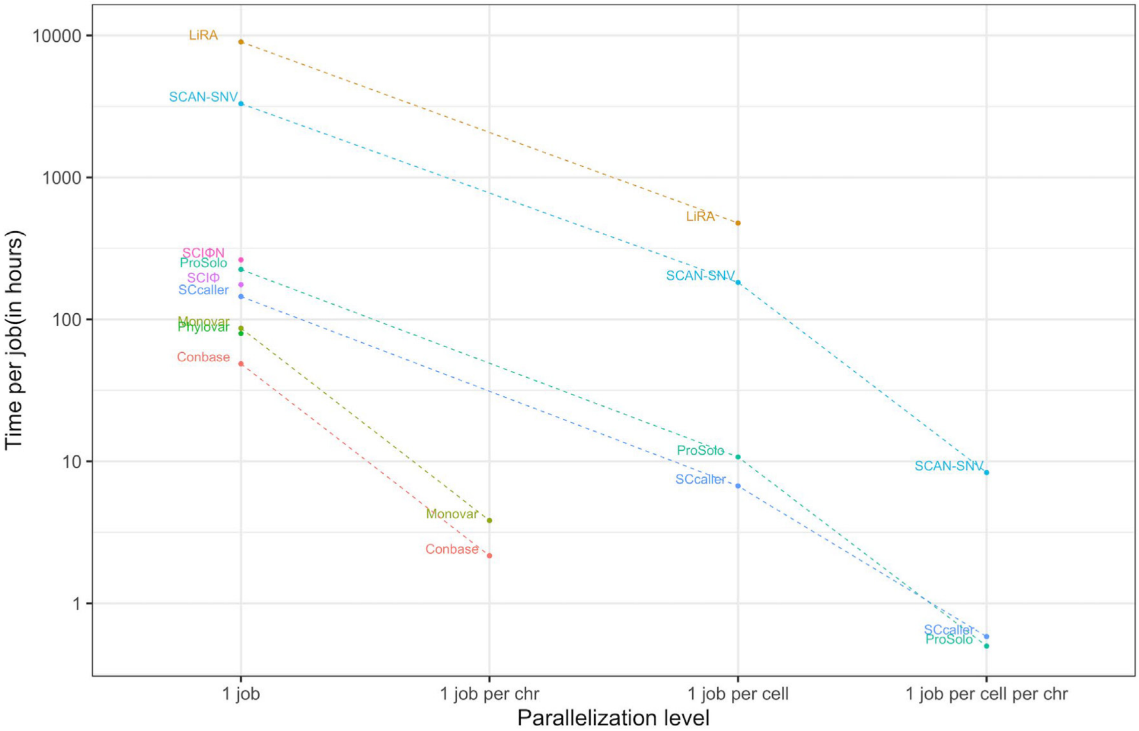

**Fig. 3.** Runtime for single-cell SNV callers. Plot showing run times for scDNA-seq variant callers on a dataset with 24 single-cell whole-genomes. Colors highlight distinct callers. The X-axis represents four different job-splitting strategies (note that different tools have different capabilities in this regard). The Y-axis is in log-scale and represents the maximum number of hours required by a given tool.

the exception being Monovar, which typically results in a lot of calls, with a good recall at the cost of lower precision [39,40,42,45,46,48]. SCcaller is more precise than Monovar, and their relative recall depends on the specific study [40,45,46,48]. On the other hand, SCIΦ has, in general, a good recall (further improved in SCIΦN) and performs similarly to scVILP and Phylovar [39,48,51]. LiRA and Conbase are very precise and can have a good recall when referring to SNVs linked to hSNP [42,45]. However, they cannot identify unlinked SNVs, so their genome-wide recall is very low. SCAN-SNV also shows high precision but at the price of a lower recall rate [46]. ProSolo is the most recent single-cell variant caller. According to its developers, ProSolo has better recall and precision than Monovar, SCIΦ, SCcaller, and SCAN_SNV [48].

### 2.2.2. Speed benchmark

We tested the relative speed of the different callers (Fig. 3). For this, we computed the time necessary to call SNVs from a scDNA-seq (WGS 6X) dataset produced in our lab, consisting of 24 single-cell whole-genomes from a colorectal cancer patient (CRC24 in [37]). We excluded scVILP from this benchmark as it can only work with targeted data. Different callers use distinct parallelization strategies. Phylogeny-based callers like SCIΦ, SCIΦN, and Phylovar can only be run as a single job analyzing all cells simultaneously. In contrast, joint callers like Monovar and Conbase can be run chromosome-wise, reducing the runtime by 10-fold. Under these two computational strategies (single job and chromosome-wise), Conbase was the fastest tool, implying approximately 50 and 3 h, respectively. SCcaller and ProSolo can be parallelized at two different levels, cell-wise and per cell per chromosome. The latter reduced the runtime dramatically, so these two tools finished their analysis in less than one hour, which was the minimum runtime for this dataset. SCAN-SNV as a single job implied thousands of hours, but when the tasks were split by chromosome per cell, the runtime was reduced to approximately 8 h. LiRA cannot work chromosome-wise, as it leverages genome-wide somatic mutation rates, but it can work one cell at a time. Still, under both strategies it was the slowest method.

## 3. Conclusions

Accurate single-cell variant calling is critical for recognizing somatic genomic heterogeneity at the ultimate level of resolution, identifying variants in rare cell populations, and for downstream analyses related to cell biology, development, and somatic evolution. The analysis of scDNA-seq implies numerous challenges derived from technical biases occurring during cell processing and scWGA. Protocols that bypass the scWGA step exist [55,56], but they are typically custom-made and not very portable.

Most single-cell variant callers are very precise, and their recall varies across different scenarios and simulation studies. All the benchmarks have been exclusively carried out when presenting a new single-cell variant caller. Therefore, they are prone to a self-assessment trap [55], with the authors' method being the best under the different scenarios simulated. Comprehensive, third-party benchmarking studies like those carried out for bulk variant calling [57,58] are still lacking for single-cell variant callers. The computational speed of the different tools depends heavily on the possibility of simultaneously running multiple jobs (per cell and per chromosome). SCcaller and Prosolo, or Conbase if only a single job is possible, were the fastest tools for the exemplar dataset we selected for benchmarking.

Choosing the appropriate single-cell variant caller is not easy and depends on the particular question at hand. If the interest is in detecting very reliable but not necessarily many mutations, linked-read strategies should offer a reasonable precision rate. However, if the interest is in describing diversity or in performing genome-wide level analyses, other strategies should offer a better recall. One might want to use results from the intersection of multiple callers [50]. Still, using different algorithms and distinct underlying assumptions will most likely reduce the number of variants finally called.

## CRediT authorship contribution statement

**Monica Valecha:** Conceptualization, Formal analysis, Investigation, Data curation, Writing – original draft. **David Posada:** Conceptualization, Methodology, Resources, Writing – original draft, Supervision, Project administration, Funding acquisition.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We appreciate the comments from two anonymous reviewers. This work was supported by the European Research Council (ERC-617457- PHYLOCANCER) awarded to D.P., Spanish Ministry of Science and Innovation - MICINN (PID2019-106247GB-I00 awarded to D.P.), and by the European Union's Horizon 2020 Research and Innovation Program under the Marie Sklodowska-Curie grant agreement No. 766030.

## References

[1] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 2013;14:618–30.
[2] Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. Mol Cell 2015;58:598–609.
[3] Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. Nat Rev Cancer 2017;17:557–69.
[4] Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature 2017;541:331.
[5] Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet 2016;17:175–88.
[6] Macaulay IC, Voet T. Single cell genomics: advances and future perspectives. PLoS Genet 2014;10:e1004126.
[7] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 2015;16:133–45.
[8] Navin NE. Cancer genomics: one cell at a time. Genome Biol 2014;15:452.
[9] Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z, et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proc Natl Acad Sci U S A 2013;110:21083–8.
[10] Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell 2012;148:886–95.
[11] Li Y, Xu X, Song L, Hou Y, Li Z, Tsang S, et al. Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. GigaScience 2012;1:12.
[12] Miles LA, Bowman RL, Merlinsky TR, Csete IS, Ooi AT, Durruthy-Durruthy R, et al. Single-cell mutation analysis of clonal evolution in myeloid malignancies. Nature 2020;587:477–82.
[13] Su F, Zhang W, Zhang D, Zhang Y, Pang C, Huang Y, et al. Spatial intratumor genomic heterogeneity within localized prostate cancer revealed by single-nucleus sequencing. Eur Urol 2018;74:551–9.
[14] Leung ML, Davis A, Gao R, Casasent A, Wang Y, Sei E, et al. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome Res 2017;27:1287–99.
[15] Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature 2011;472:90–4.
[16] Nam AS, Chaligne R, Landau DA. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. Nat Rev Genet 2021;22:3–18.
[17] Lim B, Lin Y, Navin N. Advancing cancer research and medicine with single-cell genomics. Cancer Cell 2020;37:456–70.
[18] Marioni JC, Arendt D. How single-cell genomics is changing evolutionary and developmental biology. Annu Rev Cell Dev Biol 2017;33:537–53.
[19] Wiedmeier JE, Noel P, Lin W, Von Hoff DD, Han H. Single-cell sequencing in precision medicine. Precis Med Cancer Ther 2019:237–52.
[20] Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods 2016;13:229–32.
[21] Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 2015;523:486–90.
[22] Nagano T, Lubling Y, Várnai C, Dudley C, Leung W, Baran Y, et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature 2017;547:61–7.
[23] Tang X, Huang Y, Lei J, Luo H, Zhu X. The single-cell sequencing: new developments and medical applications. Cell Biosci 2019;9:53.
[24] Wagner DE, Klein AM. Lineage tracing meets single-cell omics: opportunities and challenges. Nat Rev Genet 2020;21:410–27.
[25] Evrony GD, Hinch AG, Luo C. Applications of single-cell DNA sequencing. Annu Rev Genomics Hum Genet 2021;22:171–97.
[26] Kaster A-K, Sobol MS. Microbial single-cell omics: the crux of the matter. Appl Microbiol Biotechnol 2020;104:8209–20.
[27] Mallory XF, Edrisi M, Navin N, Nakhleh L. Methods for copy number aberration detection from single-cell DNA-sequencing data. Genome Biol 2020;21:208.
[28] Liu F, Zhang Y, Zhang L, Li Z, Fang Q, Gao R, et al. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. Genome Biol 2019;20:242.
[29] 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;21:31.
[30] de Bourcy CFA, De Vlaminck I, Kanbar JN, Wang J, Gawad C, Quake SR. A quantitative comparison of single-cell whole genome amplification methods. PLoS ONE 2014;9:e105585.
[31] Gonzalez Castro LN, Tirosh I, Suvà ML. Decoding cancer biology one cell at a time. Cancer Discov 2021;11:960–70.
[32] 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;3:811.
[33] Edrisi M, Zafar H, Nakhleh L. A combinatorial approach for single-cell variant detection via phylogenetic inference. bioRxiv 2019. https://doi.org/10.1101/693960.
[34] Kim KI, Simon R. Using single cell sequencing data to model the evolutionary history of a tumor. BMC Bioinf 2014;15:27.
[35] Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. Genome Biol 2016;17:86.
[36] 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;18:178.
[37] 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;23:37.
[38] Zafar H, Wang Y, Nakhleh L, Navin N, Chen K. Monovar: single-nucleotide variant detection in single cells. Nat Methods 2016;13:505–7.
[39] Singer J, Kuipers J, Jahn K, Beerenwinkel N. Single-cell mutation identification via phylogenetic inference. Nat Commun 2018;9:5144.
[40] 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;14:491–3.
[41] Kuipers J, Singer J, Beerenwinkel N. Single-cell mutation calling and phylogenetic tree reconstruction with loss and recurrence 2022. https://doi.org/10.1101/2022.01.28.478229.
[42] Bohrson 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;51:749–54.
[43] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.
[44] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–11.
[45] Hård J, Al Hakim E, Kindblom M, Björklund ÅK, Sennblad B, Demirci I, et al. Conbase: a software for unsupervised discovery of clonal somatic mutations in single cells through read phasing. Genome Biol 2019;20:68.
[46] Luquette LJ, Bohrson CL, Sherman MA, Park PJ. Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. Nat Commun 2019;10:3908.
[47] Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods 2011;9:179–81.
[48] 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;12:6744.
[49] Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A 2002;99:5261–6.
[50] Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science 2015;350:94–8.
[51] Edrisi M, Valecha MV, Chowdary SBV, Robledo S, Ogilvie HA, Posada D, et al. Phylovar: Towards scalable phylogeny-aware inference of single-nucleotide variations from single-cell DNA sequencing data. bioRxiv 2022:2022.01.16.476509. https://doi.org/10.1101/2022.01.16.476509.
[52] Demeulemeester J, Dentro SC, Gerstung M, Van Loo P. Biallelic mutations in cancer genomes reveal local mutational determinants. Nat Genet 2022:1–6.
[53] 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;27:1885–94.

[54] Miura S, Huuki LA, Buturla T, Vu T, Gomez K, Kumar S. Computational enhancement of single-cell sequences for inferring tumor evolution. Bioinformatics 2018;34:i917–26.

[55] Zahn H, Steif A, Laks E, Eirew P, VanInsberghe M, Shah SP, et al. Scalable whole-genome single-cell library preparation without preamplification. Nat Methods 2017;14:167–73.

[56] Xi L, Belyaev A, Spurgeon S, Wang X, Gong H, Aboukhalil R, et al. New library construction method for single-cell genomes. PLoS ONE 2017;12:e0181163.

[57] Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods 2015;12:623–30.

[58] Mangul S, Martin LS, Hill BL, Lam AK-M, Distler MG, Zelikovsky A, et al. Systematic benchmarking of omics computational tools. Nat Commun 2019;10:1393.