*Bioinformatics*, 36(11), 2020, 3299–3306
doi: 10.1093/bioinformatics/btaa172
Advance Access Publication Date: 11 March 2020
Original Paper

OXFORD

Genome analysis

# RobustClone: a robust PCA method for tumor clone and evolution inference from single-cell sequencing data

**Ziwei Chen[1,2], Fuzhou Gong[1,2], Lin Wan[1,2,*] and Liang Ma[3,*]**

[1]NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, [2]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China and [3]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on June 10, 2019; revised on February 10, 2020; editorial decision on March 5, 2020; accepted on March 6, 2020

## Abstract

**Motivation:** Single-cell sequencing (SCS) data provide unprecedented insights into intratumoral heterogeneity. With SCS, we can better characterize clonal genotypes and reconstruct phylogenetic relationships of tumor cells/clones. However, SCS data are often error-prone, making their computational analysis challenging.

**Results:** To infer the clonal evolution in tumor from the error-prone SCS data, we developed an efficient computational framework, termed RobustClone. It recovers the true genotypes of subclones based on the extended robust principal component analysis, a low-rank matrix decomposition method, and reconstructs the subclonal evolutionary tree. RobustClone is a model-free method, which can be applied to both single-cell single nucleotide variation (scSNV) and single-cell copy-number variation (scCNV) data. It is efficient and scalable to large-scale datasets. We conducted a set of systematic evaluations on simulated datasets and demonstrated that RobustClone outperforms state-of-the-art methods in large-scale data both in accuracy and efficiency. We further validated RobustClone on two scSNV and two scCNV datasets and demonstrated that RobustClone could recover genotype matrix and infer the subclonal evolution tree accurately under various scenarios. In particular, RobustClone revealed the spatial progression patterns of subclonal evolution on the large-scale 10X Genomics scCNV breast cancer dataset.

**Availability and implementation:** RobustClone software is available at https://github.com/ucasdp/RobustClone.

**Contact:** lwan@amss.ac.cn or maliang@ioz.ac.cn

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Tumor evolution has been a subject of longstanding discussion (Nowell, 1976). Understanding evolutionary mechanisms underlying cancer progression and characterizing intra-tumor heterogeneity are believed to guide the principles in predicting and controlling cancer progression, metastasis and therapeutic responses (Schwartz and Schäffer, 2017). A tumor is comprised of subpopulations of cells with distinct genotypes called subclones (Lawson *et al.*, 2018). By taking the advantage of high-throughput next-generation sequencing, computational methods based on bulk DNA-sequencing data have been developed to deconvolve subclonal genotypes and/or infer their evolutionary relationships (Deshwar *et al.*, 2015; El-Kebir *et al.*, 2015, 2016, 2018; Jiang *et al.*, 2016; Jiao *et al.*, 2014; Zaccaria *et al.*, 2018; Zare *et al.*, 2014).

The rapid advances in single-cell sequencing (SCS) technology have greatly enhanced the resolution of tumor cell profiling and are expected to better characterize intratumoral heterogeneity (Lawson *et al.*, 2018; Navin, 2014). While pioneer works utilize single-cell copy-number variation (scCNV) profiles to construct tumor cell phylogenies (Navin *et al.*, 2011; Wang *et al.*, 2014), many others work on single-cell single nucleotide variation (scSNV) data with application of traditional phylogenetic methods. Xu *et al.* (2012) and Yu *et al.* (2014) applied distance-based methods, UPGMA or neighbor-joining (Felsenstein, 2004), on kidney cancer and colon cancer. More complex models, such as maximum likelihood or Bayesian (Yang, 2014), have also been applied to infer tumor phylogeny with scSNV data (Eirew *et al.*, 2015; Hughes *et al.*, 2014).

Although promising, current SCS data are known to be error-prone due to technique issues, thereby limiting the direct application of traditional phylogenetic approaches to the data. Four common types of errors are often associated with SCS data: false positive (FP) and false negative (FN) mutations, missing bases (MBs), as well as doublets. FPs and FNs are usually caused by allelic dropout events, a very common problem in SCS in which one or both alleles fail to amplify. The FN rate (FNR) varies from 0.1 to 0.43, as reported in many studies (Gawad *et al.*, 2014; Hou *et al.*, 2012; Wang *et al.*, 2014; Xu *et al.*, 2012). FPs occur on the order of $\sim 10^{-5}$, which

exceeding the somatic mutation rate (Hou *et al.*, 2012; Wang *et al.*, 2014; Xu *et al.*, 2012). MBs may issue from insufficient sequencing coverage. The reported missing rate (MR) can be as high as 58% in SCS data (Gawad *et al.*, 2014; Hou *et al.*, 2012). Another source of error in SCS data is cell doublets, which result from unintended measurement of two or more cells (Roth *et al.*, 2016; Zafar *et al.*, 2017). Cell doublet rates vary greatly depending on the isolation technique. For example, in fluorescence-activated cell sorting, the cell doublet rate is reported <1%, while in oral pipette and droplet encapsulation techniques, the doublet rate is reported from 1% to 10% (Zafar *et al.*, 2017). When these errors occur together in data, the downstream analysis can be greatly biased.

Methods that explicitly account for errors in SCS data, especially scSNV data, have emerged in recent years. SCITE (Jahn *et al.*, 2016) and OncoNEM (Ross and Markowetz, 2016), model the noise of SCS, as well as construction of mutation or subclonal trees based on scSNV. SCG (Roth *et al.*, 2016) also models various technical errors by clustering single cells into subclones with a hierarchical Bayesian model and then inferring the subclonal genotypes. These methods are constructed under the infinite site model (ISM), although some account for loss of heterozygosity (LOH), the presence of recurrent mutations is prohibited. This assumption may often be violated in human tumor evolution, where recurrent events, such as back or parallel mutation, as well as LOH all could happen (Davis and Navin, 2016). SiFit (Zafar *et al.*, 2017), a likelihood-based approach, has employed the finite site model (FSM) of evolution and infers cell phylogeny to account for SCS errors. BEAM (Miura *et al.*, 2018) is a Bayesian method that has no explicit restrictions on mutation model. It improves the quality of single-cell sequences by using the intrinsic evolutionary information in single-cell data in a molecular phylogenetic framework. Among the above-mentioned methods, each has its own merits, as they all perform acceptably well under the present amount of small or moderate data size (e.g. the number of cells $\leq$ 500). In recent years, single-cell techniques are rapidly evolving, therefore, lowering the cost of sequencing (Lan *et al.*, 2017). The size of single-cell samples, and the number of mutations, both in forms of SNV and CNV, which can be used in the analysis, are expected to increase in the very near future (Shapiro *et al.*, 2013). These advances could result in a dramatic increase of computational intensity, especially for likelihood-based or Bayesian-based algorithms.

Recently, the low-rank matrix factorization method, robust principal component analysis (RPCA), for recovering low-dimensional subspace from corrupted data, i.e. the data contaminated by an amount of noise, is being extensively studied (Candes *et al.*, 2011; Hsu *et al.*, 2011; Lin *et al.*, 2011; Vidal *et al.*, 2016). RPCA is a generalization of the standard principal component analysis (PCA) by introducing some robustness. Instead of approximating observation with a low-rank matrix, as in PCA, RPCA approximates observed matrix by decomposing it into the sum of a low-rank matrix and a sparse matrix that models the corrupted variables. The decomposition of RPCA can be implemented by scalable and fast algorithms, such as the Augmented Lagrange Multiplier Method (ALM) (Lin *et al.*, 2011). RPCA can be extended naturally to model corrupted data in the presence of missing entries (Vidal *et al.*, 2016). It has wide applications in fields, such as image processing (Vidal *et al.*, 2016) and bioinformatics [e.g. the imputation of single-cell RNA-sequencing data (Chen *et al.*, 2020)].

Cancer cells within a tumor are often heterogeneous. Nevertheless, these cells usually form into subpopulations (subclones) with nearly or completely identical genetic composition. Therefore, the number of subclones should be in general much less than the number of cells or the number of mutated sites. On the other hand, the observed single-cell genotype matrix (GTM) is often incorporated, besides missing entries, with random noise caused by technical errors. Thus making the GTM recovery problem fits perfectly to the RPCA framework, which has the low-rank plus sparse matrices assumption. In this study, we present RobustClone, a computational framework allowing for the recovery of subclone genotypes based on observed GTM of either scSNV or scCNV data, and reconstructing the subclonal evolutionary tree. RobustClone utilizes the extended RPCA method, which can accommodate GTM with missing entries. Using simulated and real data, we demonstrate the power of RobustClone in recovering real GTM and reconstruction of subclonal evolutionary trees under various scenarios. We also show the efficiency of RobustClone on applications to large-scale data.

## 2 Materials and methods

In this section, we first introduce RPCA and the extended RPCA algorithms, which are used to recover the low-rank subspace from data matrix with corrupted and/or missing entries (Section 2.1); we then describe how the proposed computational framework, termed RobustClone, recovers the true GTM of tumor cells, identifies tumor subclones and reconstructs subclonal evolutionary trees, all based on tumor SCS data (Section 2.2). Finally, we provide details of the evaluation methods (Section 2.3) and the simulated and real data used (Section 2.4).

### 2.1 RPCA and extended RPCA

#### 2.1.1 Robust principal component analysis

As a popular tool to recover low-rank matrix, standard PCA is based on the assumption that all sample points are drawn from the same statistical or geometric model (Vidal *et al.*, 2016). In practice, however, the entries of data matrix can be corrupted by gross errors, making standard PCA less robust to intra-sample outliers (Vidal *et al.*, 2016). In this regard, Candes *et al.* (2011) proposed the RPCA to recover low-rank matrix from data with corrupted entries. The RPCA problem can be solved with the ALM, a fast and scalable algorithm proposed by Lin *et al.* (2011).

Assume that the data matrix $D_{m\times n}$ is generated as the sum of two matrices $D = A_0 + E_0$, where $A_0$ represents a low-rank data matrix, while $E_0$ represents the intra-sample outliers. We further assume that many entries of $D$ remain intact, thereby causing many entries of $E_0$ to be zero. The RPCA problem can be formulated as decomposing matrix $D$ into the sum of a low-rank matrix $A$ and a sparse matrix $E$, satisfying

$$\min_{A,E} \text{rank}(A) + \lambda||E||_0, \quad \text{s.t.} \quad A + E = D, \tag{1}$$

where $||E||_0$ is the number of non-zero entries in $E$, and $\lambda$ is the regularization parameter that balances the two terms. However, the task of recovering the low-rank matrix $A$ and the sparse signal $E$ in Problem (1) is generally NP-hard (Vidal *et al.*, 2016).

In order to efficiently compute and solve this problem, Problem (1) can be convex relaxed, as

$$\min_{A,E} ||A||_* + \lambda||E||_1, \quad \text{s.t.} \quad A + E = D, \tag{2}$$

where $||\cdot||_*$ and $||\cdot||_1$ denote the nuclear norm and the $\ell_1$ norm of matrix, respectively. We call Problem (2) the relaxed version of RPCA. It has been theoretically validated that the relaxed RPCA can decompose $D$ and exactly recover the unknown matrices $A$ and $E$ with a probability of almost one under rather broad conditions (see Vidal *et al.* 2016; Candes *et al.* 2011). The optimal choice of $\lambda$ has been shown to be $\lambda_0 = 1/\sqrt{\max(m,n)}$ (Candes *et al.*, 2011; Vidal *et al.*, 2016).

The constrained optimization problem (2) can be solved by the ALM algorithm proposed by Lin *et al.* (2011), a special case of the alternative direction method of multipliers. It can be applied on the following augmented Lagrangian function:

$$\begin{aligned} L(A,E,\Lambda,\mu) = \quad & ||A||_* + \lambda||E||_1 + \\ & \langle \Lambda, D - A - E\rangle + \frac{\mu}{2}||D - A - E||_F^2, \end{aligned} \tag{3}$$

where $||\cdot||_F$ is the Frobenius norm of matrix, $\Lambda$ is the Lagrange multiplier, $\langle\cdot,\cdot\rangle$ is the inner product and $\lambda$ and $\mu$ are the regulation terms. See details of the algorithm in Supplementary Note 1 and Algorithm S1.

#### 2.1.2 Extended RPCA

We further extend RPCA to the cases where entries in the data matrix are not only corrupted but also incomplete. In order to handle the missing entries, we define a linear operator $\mathcal{P}_\Omega(D)$, which maps the missing/incomplete entries to 0, while keeping the observed entries, as

$$[\mathcal{P}_\Omega(D)]_{ij} = \begin{cases} D_{ij}, & \text{if } (i,j) \in \Omega; \\ 0, & \text{otherwise.} \end{cases}$$

The extended RPCA problem is then formulated as follows (Shang *et al.*, 2014; Vidal *et al.*, 2016; Wright *et al.*, 2012):

$$\min_{A,E} ||A||_* + \lambda||E||_1, \quad \text{s.t.} \quad \mathcal{P}_\Omega(A+E) = \mathcal{P}_\Omega(D), \tag{4}$$

with the intention of recovering the low-rank matrix and the sparse component $(A, E)$ of $D = A + E$ only from the observations $\mathcal{P}_\Omega(D)$. The low-rank and sparse components can also be exactly recovered with high probabilities under conditions similar to those for RPCA (Shang *et al.*, 2014; Vidal *et al.*, 2016; Wright *et al.*, 2012). Shang *et al.* (2014) show that the optimization problem (4) is equivalent to the following constrained optimization problem:

$$\min_{A,E} ||A||_* + \lambda||\mathcal{P}_\Omega(E)||_1 \quad \text{s.t.} \quad A + E = D, \tag{5}$$

which can be solved by applying the ALM algorithm (Supplementary Algorithm S2) with the following augmented Lagrangian function:

$$L(A,E,\Lambda,\mu) = ||A||_* + \lambda||\mathcal{P}_\Omega(E)||_1 + \langle \Lambda, D - A - E\rangle + \frac{\mu}{2}||D - A - E||_F^2. \tag{6}$$

In choices of $\lambda$, we put more weight on $||\mathcal{P}_\Omega(E)||_1$ when the data matrix has higher MR, i.e. $\lambda = \lambda_0(1 + 3 \times \text{MR}) = (1 + 3 \times \text{MR})/\sqrt{\max(m,n)}$. When there are no missing entries (MR = 0), then $\lambda = \lambda_0$, which is the theoretical optimal choice of $\lambda$ for general RPCA.

### 2.2 RobustClone

We next introduce the proposed computational framework, RobustClone, which lays on the foundation of RPCA and extended RPCA. It applies to tumor SCS data to recover the true genotypes of cells, identify subclones and reconstruct subclonal evolution trees.

RobustClone takes input of the observed GTM $Y_{m\times n} = [y_{ij}]_{m\times n}$ from either scSNV or scCNV data, where $y_{ij}$ denotes the genotype of locus $j \in \{1,\ldots,n\}$ of individual cell $i \in \{1,\ldots,m\}$. The value of $y_{ij}$ can be either binary (e.g. $y_{ij} \in \{0,1\}$: '0'-unmutated, '1'-mutated) or ternary (e.g. $y_{ij} \in \{0,1,2\}$, the number of mutant alleles) for scSNV data, and it can be a non-negative integer for scCNV data (e.g. $y_{ij} \in \{0,1,2,\ldots,p\}$, e.g. the number of copies of a DNA fragment, where usually the normal case in the diploid genome has $y_{ij} = 2$). We use 'NA' for entries in the GTM with incomplete or missing values. The flowchart of RobustClone is shown in Figure 1. The main steps for RobustClone are organized in the following subsections.

#### 2.2.1 Recover the true GTM of cells

RobustClone will first recover a matrix $X$, which approximates the underlying true genotypes of subclones of cells, from the original observed matrix $Y$. Since tumor cells are clustered into homogenous subpopulations (subclones) and cells within the same subpopulation are much more similar, with identical genotype or minor/rare variations, than those outside of the clusters. Thus, the underlying low-dimensional structures of cell populations are embedded in the noisy observed matrix $Y$. Therefore, we apply the RPCA method to recover the low-dimensional subspace $X$ from $Y$, as follows:

$$Y = X + E, \tag{7}$$

where $X$ represents a low-rank matrix, which approximates the underlying genotypes of cell subpopulations (subclones), while the sparse matrix $E$ represents the noise in the original data as well as cell-specific variations. For scSNV data with binary values, $E$ contains the noise caused by technical errors, such as FPs, FNs and cell doublets. For scCNV data, $E$ can be the noise generated by DNA sequencing and/or the errors caused during the estimation of copy numbers.

When there are no missing entries in $Y$, RobustClone applies the RPCA model (2). While in cases with missing entries, the extended RPCA model (5) is applied to recover the low-rank matrix $X$ (Fig. 1A).

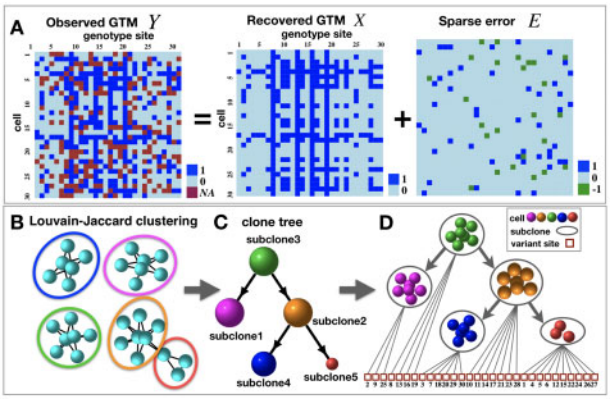


**Fig. 1.** Overview of the computational framework of RobustClone that recovers the true genotypes of cells, identifies subclones and reconstructs subclonal evolution trees using tumor SCS data. (**A**) RobustClone decomposes the observed genotype $Y$ into the sum of the low-rank GTM $X$ and a sparse matrix $E$ by RPCA or the extended RPCA model. (**B**) RobustClone divides the individual cells into clusters, as our identified subclones, by applying the Louvain–Jaccard method on the recovered low-rank GTM $X$. (**C**) RobustClone reconstructs the subclonal evolutionary tree: RobustClone identifies the subclonal tree by finding the MST using Euclidean distance after extracting the consensus subclonal genotypes; the radius sizes of the nodes on the subclonal tree are proportional to the number of cells contained in each subclone. (**D**) The subclonal evolutionary tree that describes the subclonal development of the tumor and the newly mutated genotypes of each subclone from its parent subclone

#### 2.2.2 Identify subclones by Louvain–Jaccard clustering method

Since the imputed and recovered low-rank GTM $X$ may not guarantee an error-free state, we cannot identify subclones simply by aggregating the identical rows of $X$. Instead, RobustClone clusters cells into subclones by applying the Louvain–Jaccard method (Blondel *et al.*, 2008; Levine *et al.*, 2015) on $X$.

The Louvain–Jaccard method is a network-based fast community detection algorithm, which has wide applications in the clustering of single-cell RNA-sequencing data (Shekhar *et al.*, 2016) (Fig. 1B). The community detection algorithm is based on the idea of modularity, as described by Newman and Girvan (2004), the rationale being that nodes within the same community have more connected edges than nodes between communities. In its implementation, the Louvain–Jaccard method first constructs a $k$-nearest neighbor graph of $m$ cells based on Euclidean distance and then clusters cells into subpopulations. The choice of parameter $k$ is empirically dependent on the sample size (number of single cells), and we demonstrate that the results by Louvain–Jaccard algorithm is robust to the choices of $k$ (see Supplementary Note 2 and Figs S11–S13).

The Louvain–Jaccard algorithm does not require prespecification of the number of subclones, and it is fast and scalable that can be applied to large-scale datasets. However, we want to emphasize that RobustClone is not restricted to the Louvain–Jaccard algorithm. Clustering methods, such as hierarchical clustering and $K$-means algorithm, can also be adopted by RobustClone to identify subclones.

#### 2.2.3 Reconstruct subclonal evolution tree

RobustClone reconstructs the subclonal evolution tree of tumor by the following two steps.

First, extracting subclonal genotypes. Since the cells within the same subclone identified by the Louvain–Jaccard method in Section 2.2.2 are homogeneous, with genotypes being identical or almost identical. RobustClone simply extracts the consensus genotypes for each genetic site: (i) for scSNV data, the genotype with highest frequency among cells within the same subclone is selected as the consensus, and (ii) for scCNV data, the median of the copy numbers among the cells from the same subclone is taken. The vector of consensus genotypes is then defined as the subclonal genotype by RobustClone.

Second, reconstructing the subclonal evolutionary tree (Fig. 1C and D). RobustClone calculates the Euclidean distance between each pair of subclones using their subclonal genotypes and then finds the minimum spanning tree (MST) among the subclones based on their distances. To determine the root of the tree, RobustClone selects the subclone that has the shortest Euclidean distance to a given reference/normal sample. We want to note that, it is also possible to construct the cell tree, by applying classic phylogenetic algorithms, with the recovered GTM.

### 2.3 Evaluations

We compare the performance of RobustClone to state-of-the-art methods (Jahn *et al.*, 2016; Miura *et al.*, 2018; Ross and Markowetz, 2016; Roth *et al.*, 2016; Zafar *et al.*, 2017) under various simulated scenarios. The evaluations are based on several metrics that measure different aspects of the goodness of the recovered GTM, including: (i) the FPs + FNs ratios of output GTM to input GTM, (ii) the percentage of correctly imputed MBs and (iii) the error rate of the recovered GTM to the ground truth. Metrics (i) and (ii) were also utilized by Miura *et al.* (2018).

In addition, we evaluate and compare the tree reconstruction error of subclonal trees, which is calculated as the average differences between the shortest pairwise cell distance along the reconstructed and the true subclonal trees (Ross and Markowetz, 2016). When applying tree reconstruction error, cells are represented with subclonal genotypes. RobustClone, OncoNEM, SCG and BEAM have built-in subclonal inference steps. For SiFit and SCITE, we apply *K*-medoids clustering based on the distance of cell along their reconstructed cell lineage tree (SiFit) or mutation tree (SCITE) to cluster cells. The best number of clusters is determined by maximizing silhouette score (Zafar *et al.*, 2017). Each cell within a cluster is assigned with a subclonal genotype that corresponding to its cluster medoid cell (details refer to Supplementary Note 5).

For the simulation datasets with doublets, we calculate the first three evaluations with the cells after removing doublets and measure the tree reconstruction error between the true tree and the inferred tree with doublet cells excluded.

### 2.4 Data

#### 2.4.1 Simulation data

We simulated 360 datasets with various changing factors: number of cells ($m$), number of SNVs ($n$), number of subclones ($s$), FPR ($\alpha$), FNR ($\beta$), MR ($\gamma$) and doublet rate ($\delta$). Amongst, 350 datasets are simulated under ISM, which we divided into 7 groups (referred as G1–G7) with each containing 50 datasets (details refer to Supplementary Note 3). For each group, we set one or two changing factors, while keep the rest fixed. For each setting of factors, 10 replicate datasets are simulated. Unless otherwise noted, we set the default technical errors to FPs ($\alpha = 1\%$), FNs ($\beta = 20\%$), MBs ($\gamma = 20\%$) and doublets ($\delta = 10\%$). G1 are small-scale datasets with doublet rate changing from 0 to 0.4. The number of cells and SNVs are all set to 100 with 5 subclones. G2–G4 are median-scale datasets, with $m$ and $n$ set to 500, and 10 subclones by default. The changing factors are $\alpha$ from $10^{-5}$ to 0.4 for G2, $\beta$ from 0.05 to 0.4 for G3 and $\gamma$ from 0.2 to 0.8 for G4. G5–G7 are large-scale datasets, by default, with cells and SNVs set to 1000, and the number of subclones set to 10. G5 have changing number of subclones from 10 to 50. G6 change the number of cells and subclones simultaneously, with $m$ varies from 100 to 5000 and $s$ varies from 5 to 40. G7 have changing number of SNVs in the range of 100–2000. The detailed settings are provided in Supplementary Table S1.

We also simulated a group of median sized datasets under FSM, which consist of 10 replicate datasets. We used the default setting of median-scale dataset ($m$: 500, $n$: 500, $s$: 10) and added 10% recurrent mutations and 20% LOH in addition to the default setting of technical errors (details refer to Supplementary Note 3).

#### 2.4.2 Real data

We tested RobustClone on four real datasets, including two scSNV and two scCNV datasets. The scSNV datasets were (i) high-grade serous ovarian cancer (HGSOC) data (Mcpherson *et al.*, 2016; Roth *et al.*, 2016) and (ii) essential thrombocythemia (ET) data (Hou *et al.*, 2012). The scCNV datasets were (i) xenograft breast tumor data (hereinafter denoted as SA501X3F) (Campbell *et al.*, 2019; Zahn *et al.*, 2017) and (ii) breast cancer data from 10X Genomics (https://www.10xgenomics.com/solutions/single-cell-cnv/).

## 3 Results

### 3.1 RobustClone recovers GTM with high accuracy and efficiency on simulation datasets

In order to demonstrate the steps of RobustClone, we generated an illustrative dataset with 5 subclones, 1000 cells and 300 SNV sites. The errors in the dataset were set as MR 20%, FPR 15% and FNR 15%, respectively. RobustClone can recover the true GTM with high accuracy and its inferred subclonal tree is highly consistent with the simulated topology (see details in Supplementary Note 6 and Fig. S1).

In more systematic evaluations, we applied RobustClone, together with state-of-the-art methods (e.g. BEAM, SCG, SiFit, SCITE and OncoNEM), on 350 simulated datasets with various settings under the ISM (Section 2 and Supplementary Table S1 and Notes 3 and 4).

First, we show RobustClone has comparable performance to other methods when applied to small- or median-scale datasets with different settings of technical error rates (Supplementary Table S1 and Figs S2 and S3). The small-scale datasets (G1) were designed to have 5 subclones and 100 cells × 100 SNVs with varying doublet rates ($\delta$: 0–0.4). RobustClone is not sensitive to the change of $\delta$. All methods, but OncoNEM performed comparatively well on recovering GTM with the small-scale datasets (G1 in Supplementary Figs S2 and S3). On the tree reconstruction distance, OncoNEM has better performance than SiFit, although it still has in average larger distance compared to other methods. RobustClone, SCG, BEAM and SCITE all have good subclonal tree reconstruction performance.

The median-scale datasets have size of 500 cells × 500 SNVs with 10 subclones. Overall, RobustClone is comparable in performance to BEAM, SCITE and SCG, under varying FPR ($\alpha$: $10^{-5}$–0.4), FNR ($\beta$: 0.05–0.4) and MR ($\gamma$: 0.2–0.8). As OncoNEM failed in computation on these median sized data, we have excluded it in the following comparisons. Among the rest of methods, SiFit results in explicitly higher error rate and tree reconstruction distance as compared to the other four. RobustClone performed best in imputation of MBs under most error settings, except for in the situations of extremely high FP and FN errors (G2–G4 in Supplementary Figs S2 and S3). In fact, only SCITE has relatively robust performance on simulations with extreme rates of FP and FN. Noted that, however, the FPR and FNR were given as known parameters when applying SCITE. Also, the simulated data were generated under ISM, which may be beneficial for methods, such as SCITE, that have model-based design specially for ISM. Nevertheless, methods, such as RobustClone and BEAM, that do not rely on ISM still have better control of error rate under median scale and low FPR and FNR (Supplementary Figs S2 and S3).

Next, we show the scalability and efficiency of RobustClone on large-scale datasets. SiFit has significantly worse outputs in both correction of errors and imputation of missing values on these

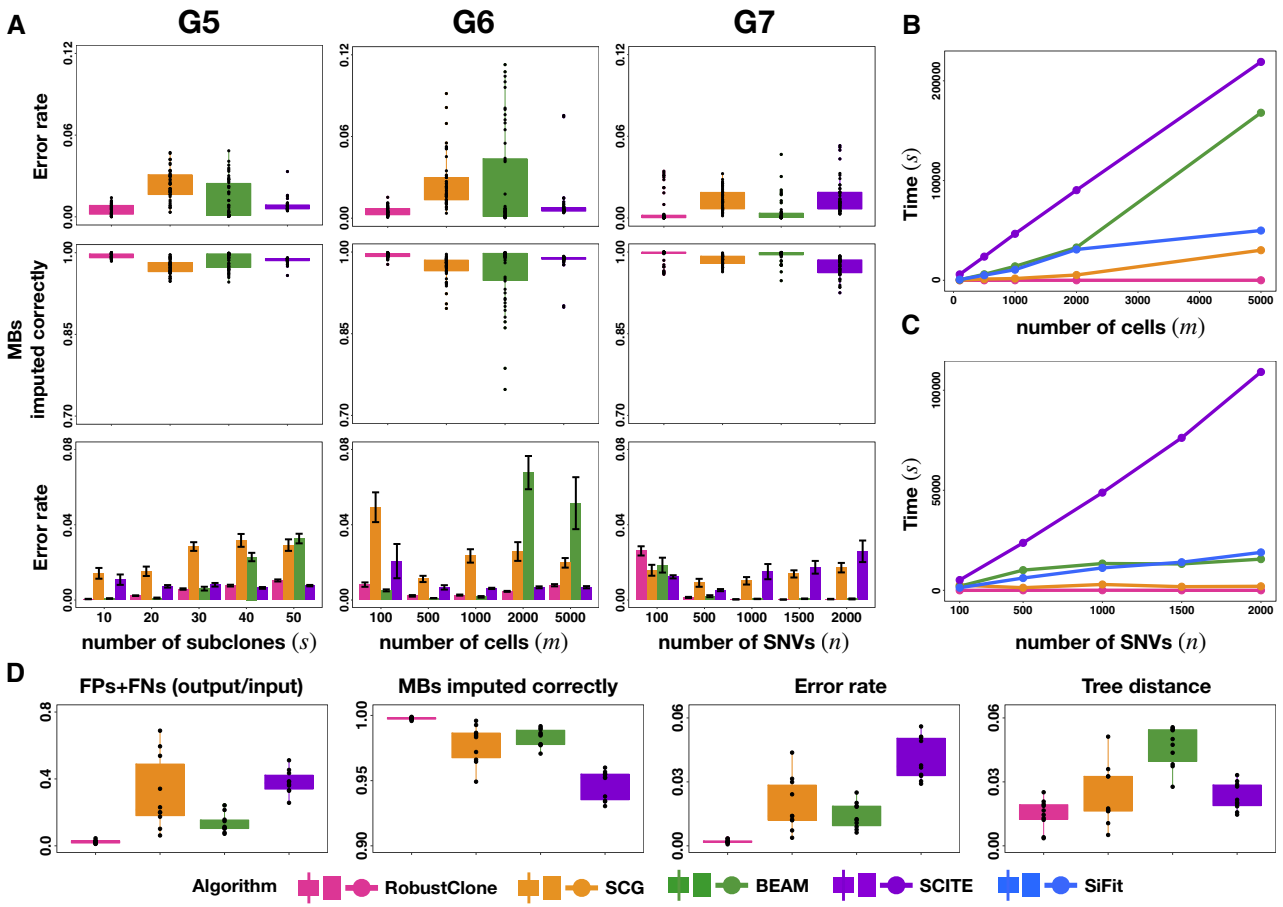


**Fig. 2.** Comparison of accuracy and efficiency among RobustClone, SCG, BEAM, SCITE, SiFit and OncoNEM algorithms on the comparison datasets. (**A**) The percentage of correctly imputed MBs and the error rate of the recovered GTM compared to the ground truth of five algorithms except SiFit on G5–G7. The four evaluations compared with SiFit can be seen in Supplementary Figures S4 and S5. (**B**) The running time of five algorithms on G6. (**C**) The running time of five algorithms on G7. (**D**) The FPs+FNs ratios of output GTM to input GTM, the percentage of correctly imputed MBs, the error rate of the recovered GTM compared to the ground truth and tree distances between the trees inferred by different algorithms and true trees of five algorithms except SiFit on simulation datasets under FSM

large-scale datasets among all applied methods (Supplementary Figs S4 and S5). RobustClone outperformed other methods in general (Fig. 2A top row), especially in imputation of MBs (Fig. 2A, second row). SCG, BEAM and SCITE also performed considerably good overall (Fig. 2A and Supplementary Figs S4 and S5). Datasets in G5 are 1000 cells × 1000 SNVs with the number of subclones varying from 10 to 50. RobustClone, SCG and BEAM all have elevated error rate in the output GTM when increasing the number of subclones. RobustClone and SCITE are less sensitive to the change of the number of subclones. Though SCITE seems to be more robust, RobustClone has much better performance when less subclones present (e.g. $s < 30$) and is only slightly higher in output error rate when more subclones present (e.g. $s > 30$) (Fig. 2A). When we change the number of cells in the input ($m$: 100–5000) coupled with changing number of subclones ($s$: 5–40), while fixing the number of SNVs to 1000, RobustClone and SCITE are less affected by the changes (G6 in Fig. 2A and Supplementary Figs S4 and S5). The output error rate of BEAM seems to be influenced by the increment in both the number of cells and the number of subclones, which results in sharp elevation in cases of 2000 cells (30 subclones) and 5000 cells (40 subclones). Interestingly, only in the cases of 100 cells did SiFit have comparable results to other methods, where its overall output error rates are smaller than SCG and SCITE (Supplementary Fig. S4). When we fix the number of cells to 1000 and increase the number of SNVs from 100 up to 5000, RobustClone has the best performance in all but the 100 SNVs scenario (G7 in Fig. 2A and Supplementary Fig. S4). The performance of BEAM is also outstanding with 500 or more SNVs. Both RobustClone and BEAM have decreased error output following the increment of SNVs. SCG and SCITE, on the other hand, have increased their output error rates as the number of SNVs increases.

Beside better performance on the large-scale datasets, RobustClone also has tremendous advantage on computational efficiency (Fig. 2B and C). RobustClone is the most efficient algorithm, that only takes 38 s for calculation of 5000 cells and takes 14 s for calculation of 2000 SNVs. SCG is the second efficient algorithm, it takes 30 172 s to deal with 5000 cells and 2016 s for calculation of 2000 SNVs. SiFit and BEAM are also time consuming when there is a large number of cells or SNVs in the data. The computational times of SCITE grow exponentially with the increment of cells or SNVs.

In addition to simulations conducted under ISM, we also simulated 10 sets of data ($m \times n$:500 × 500) with recurrent mutation and LOH events (Section 2 and Supplementary Note 3). RobustClone is a model-free method, and it significantly outperforms other methods (Fig. 2D). BEAM is also not restricted to ISM, it ranked second in overall genotype correction, where it has better performance in correcting FP/FN and comparable power in imputation of missing as compared to SCG. However, it has the worst tree reconstruction error, which possibly due to its default setting of using maximum parsimony for tree reconstruction. SCITE and SCG are designed under ISM, which as a result, they have less power in correction of FP/FN. SCITE also underperformed in imputation of MBs and tree reconstruction. Noted that, although SiFit is

applicable to FSM by design, its overall error is still very high under these median size datasets, thus, we excluded its results.

### 3.2 RobustClone works robustly on real data with high MR

We applied RobustClone to a set of real scSNV data with high MR (Hou *et al.*, 2012). The single-cell exome-sequencing data from a sample of JAK2-negative myeloproliferative neoplasm (ET) contains 58 single cells (Hou *et al.*, 2012) with 712 somatic single nucleotide variants. We applied the binarized GTM preprocessed in Ross and Markowetz (2016) as the raw input for RobustClone. The MR of this dataset reaches 58% (Supplementary Fig. S6A). The RobustClone algorithm recovered the GTM by imputation of missing entries and correction of erroneous entries (Supplementary Fig. S6B). The 58 tumor cells were then clustered by RobustClone into 3 subclones, containing 25, 19 and 14 cells, respectively (Supplementary Fig. S6B). With subclone3 identified as the root, RobustClone found an MST in linear topology that connected all three subclones (Supplementary Fig. S6C). This result is consistent with the previous findings in Hou *et al.* (2012), Ross and Markowetz (2016) and Jahn *et al.* (2016).

In addition to scSNV data, RobustClone can also be used to detect copy-number heterogeneity and identify clones with scCNV data. To demonstrate this, we applied RobustClone on copy-number profiles of cells from the passages of a patient-derived primary triple-negative breast cancer xenograft (SA501X3F data), which contains 260 cells and 20 651 genomic bins of copy-number states (Supplementary Fig. S7A). RobustClone recovered a GTM with cells clustered into two subclones, where one subclone consisted of 214 cells (denoted as subcloneA), and the other consisted of 46 cells (denoted as subcloneB) (Supplementary Fig. S7B). The copy-number profiles of the two subclones are shown in Supplementary Figure S7C. It can be seen that the difference in copy numbers between subcloneA and subcloneB is mainly presented on the X chromosome (Supplementary Fig. S7B and C). SubcloneA is completely consistent with the major subclone identified in Zahn *et al.* (2017), Rashid *et al.* (2019) and Campbell *et al.* (2019). If we take into account, the high noise in the data together with the small size of subcloneB and then tune parameter $\lambda$ to panelize more on the sparsity of the error entries ($E$ in Section 2.1), we can identify an extra subclone separate from original subcloneB (Supplementary Fig. S8), which is consistent with the clonal subpopulations identified in Zahn *et al.* (2017). However, since Zahn *et al.* (2017) did not explicitly correct for noise in the GTM before clonal identification, the two derived subpopulations from subcloneB had much uncertainty (Campbell *et al.*, 2019). We believe that the two subclones resulted from the default setting of RobustClone are more robust.

In order to further assess the robustness of RobustClone on the SA501X3F dataset, we randomly performed a 30% dropout of entries in the original GTM (Supplementary Fig. S9A) and reanalyzed it. RobustClone still identified two subclones, and the copy-number profile of cells in different subclones was consistent with the result in Supplementary Figure S7B (Supplementary Fig. S9B). We continued to increase the dropout rate to 50% (Supplementary Fig. S9C). RobustClone found 4 subclones containing 213, 45, 1 and 1 cells, respectively (Supplementary Fig. S9D). The two extra subclones each contain one cell only, which is separate from one of the original major subclones.

These results indicate that RobustClone is highly robust for scSNV and scCNV data with high missing entries.

### 3.3 Recovering scSNV genotype and inferring subclonal tree of HGSOC

We performed RobustClone on a set of HGSOC data (Mcpherson *et al.*, 2016; Roth *et al.*, 2016). The original data matrix contains 420 cells and 43 selected SNV sites with 10.7% missing entries (Fig. 3A). RobustClone efficiently recovered the cellular GTM by imputing the missing values and correcting noisy entries in the observed data (Fig. 3B). Based on the corrected GTM, RobustClone identified five subclones. We labeled them subclone1–subclone5, according to their sizes, consisting of 122, 81, 81, 69 and 67 cells, respectively (Fig. 3C). As subclone4 had the minimum number of mutations, it was assigned as the root subclone. An MST connecting all five subclones was then constructed by RobustClone based on pairwise subclonal distance (Fig. 3C). The newly arisen mutations of each subclone, following the topology of MST, could thus be identified (Supplementary Fig. S10A).

To better understand the composition and the relationship among subclones, we divided the SNV sites into five major blocks. Only subclone4 had some mutations in block5 where site *TP53* mainly presented in heterozygous genotype. The mutations in block5 were carried through all subsequent subclones. These subclones had a high rate of homozygous *TP53* mutant alleles, which is a strong indicator of cancer (Sun *et al.*, 2018). Subclone3 descended from subclone4 and accumulated sparse mutations in all blocks except block1. The mutations in block2 and block4 define the divergence of subclone1 and subclone2 from subclone3. The smallest of the five subclones, subclone5, which carried more mutations in block3, was derived from subclone1.

We compared the subclones identified by RobustClone to the result of SCG, which identified six subclones (clusters) based on the same HGSOC data (Mcpherson *et al.*, 2016; Roth *et al.*, 2016). SCG cluster0 mainly consists of cells in subclone2 and subclone3. The cells in SCG clusters 1, 2, 3 and 5 are mainly distributed in subclone1 and subclone5 (Fig. 3D). Subclone4 contained all cells in SCG cluster4, which was interpreted to be normal cells. Interestingly, heterozygous and/or even homozygous mutations were recovered by SCG in SNV sites corresponding to block1, but only in cluster4. These precancerous mutations were expected to become 'public', or at least be abundant, in subsequent subclones (Opasic *et al.*, 2019; Williams *et al.*, 2016). In comparison, the recovery of GTM by RobustClone with no mutations in block1 in all cells seems to be more reasonable. In addition to the spanning tree of subclones, the subclonal genotypes and/or the corrected GTM could also be used to directly reconstruct the subclonal and/or cellular phylogenies by applying any readily available off-the-shelf methods in classic phylogenetics (Supplementary Fig. S10B and C) (Felsenstein, 2004).

### 3.4 Revealing the spatial heterogeneity of breast cancer with multi-section large-scale scCNV dataset

We applied RobustClone to large-scale breast cancer scCNV data from 10X Genomics. The frozen breast tissues were from three negative ductal carcinoma *in situ* specimens, which were divided into five spatially consecutive parts, denoted as Sections from A to E (Fig. 4A). The raw scCNV-sequencing data were preprocessed by Cell Ranger pipeline, resulting in a 9050 × 55 572 GTM, which included 2061, 2046, 1448, 1665 and 1830 cells from Sections A to E, respectively. Each cell was characterized by the ploidy states of 55 572 20-kb bins over 7 chromosomes, which covered chr3, chr4, chr5, chr6, chr7, chr8 and chr10, respectively.

RobustClone first recovered a low-rank scCNV GTM and then identified 6 subclones with 4234, 1808, 1647, 821, 308 and 232 cells, respectively (Fig. 4C). The copy-number profile (Fig. 4B) of each subclone was obtained by taking the median copy numbers of each bin for cells within the same subclone. As the copy-number profile of subclone1 was consistently diploid (Fig. 4B), it was presumed to consist of normal cells and was, therefore, assigned as the root. RobustClone then found the MST of subclones (Fig. 4C). We used red boxes in Figure 4B to highlight the changes between the copy-number profiles of each subclone and its parent. Subclone4 diverged from subclone1 and had a gain in copy number in nearly all seven chromosomes where chromosomes 5 and 7 changed from diploid to tetraploid and other chromosomes changed to triploid. Subclone6 also derived from subclone1 with copy number loss in chr6 and 10. Subclone4 further differentiated into two major subclones: subclone2, which changed from triploid to tetraploid on chr4 and 8; subclone3, which gained one more ploidy, mainly on chr3, 6 and 10. Subclone5 gained further copy numbers on all chromosomes, except chr4, on the basis of subclone2.

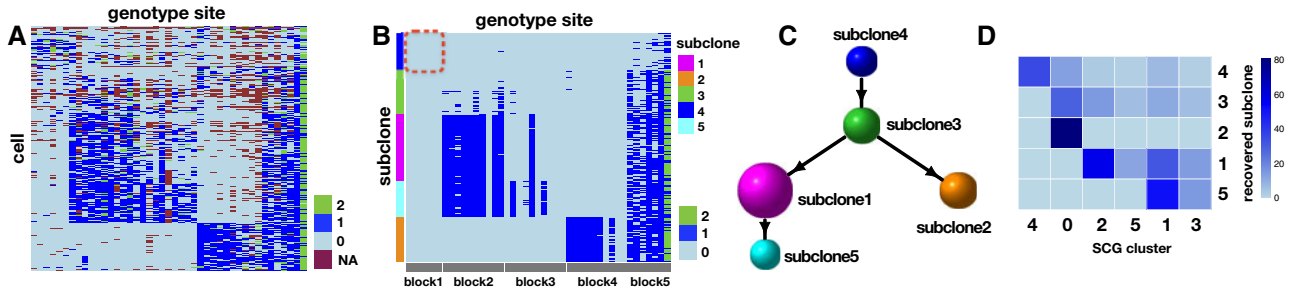


**Fig. 3.** RobustClone reconstructs the subclonal evolution tree on HGSOC data. (**A**) The heatmap of the noisy SNV profile before inference. Each row of the heatmap represents a single cell, and the column represents the genotype sites (GS). (**B**) The heatmap of the GTM recovered by RobustClone. Each row of the heatmap represents a single cell, and the column represents the GS. The GS in the red circle are unmutated, but they were recovered as mutated sites by SCG, as shown in Figure 3B of Roth *et al.* (2016). (**C**) The subclonal evolution tree reconstructed by RobustClone using MST. (**D**) The number of overlapped cells of HGSOC data contained in subclones identified by RobustClone and cells contained in clones identified by SCG

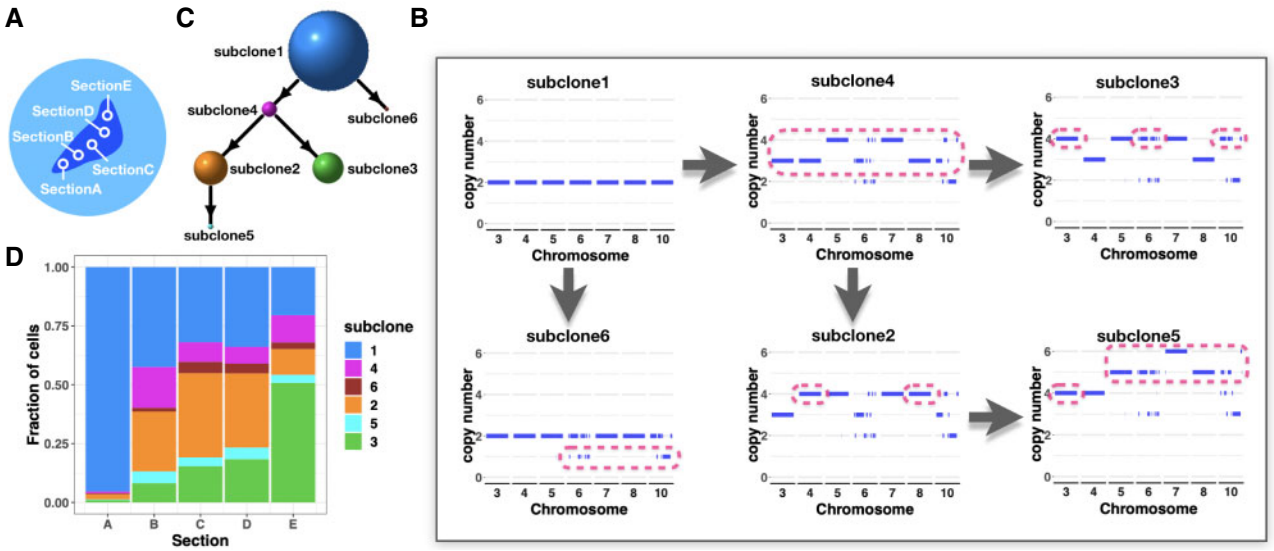


**Fig. 4.** RobustClone reveals a spatial progression pattern of subclones on the large-scale copy-number profile of breast cancer dataset from 10X Genomics. (**A**) Schematic representation of the regional division of frozen breast cancer samples; the samples are from the five connective spatial points of Sections A–E. (**B**) The copy-number profile of subclones recovered by RobustClone. (**C**) The subclonal evolution tree reconstructed by RobustClone. (**D**) Stacked histogram representing subclonal proportions across the five sections

The subclonal composition of the five spatial sections is shown in Figure 4D. Section A is dominated by normal cells of subclone1. Subclone2 occupies the largest proportion of subclones apart from normal cells (subclone1) in the middle sections B, C and D. In contrast, Section E is governed by subclone3. These results reveal the great spatial heterogeneity within tumor.

## 4 Discussion

In this study, we proposed RobustClone, a tool for the robust recovery of noisy scSNV and scCNV data based on the RPCA and extended RPCA. RobustClone is a model-free approach, which achieves high accuracy in the imputation of MBs and correction of FPs and FNs.

Understanding intratumoral heterogeneity and inferring of clonal evolution have long been the subjects of research interests (Schwartz and Schaffer, 2017). Under the error-prone single-cell DNA data, two general ideas can be adopted to infer the relationship of tumor cells or subclones. One is jointly modeling errors and subclonal phylogeny under a Bayesian or likelihood framework. The other is to directly correct the errors in the original single-cell GTM and then construct a subclonal tree with the recovered GTM. BEAM, SCITE and SiFit belong to the first kind, they have good performance when cell numbers or SNV sites are not so large. RobustClone and SCG belong to the latter type. They both exhibit computational efficiency in large dataset, and they may also utilize the state-of-the-art methods in molecular phylogenetics for subclonal tree reconstruction.

RobustClone constructs the subclonal tree by clustering the cells into subclones using Louvain–Jaccard method and then inferring the MST of the subclones. The MST-based approach for constructing subclonal tree has also been adopted by early tumor studies (Gawad *et al.*, 2014; Ross and Markowetz, 2016; Yuan *et al.*, 2015). Compared with classical phylogeny, which depicts hierarchical relations of cells, MST characterizes the genealogical relationship of tumor clones and explicitly reflects the progression of subclones. We show that the MST by RobustClone matches the phylogeny well in Supplementary Figure S10. It is worth noting that in some cases, extant subclones as well as their common ancestors are partially observed or incompletely sampled, which pose challenges to both MST- and phylogeny-based methods. OncoNEM has tried to resolve this problem by inferring unobserved populations that can improve the likelihood of the MST (Ross and Markowetz, 2016). However, its application is limited to small sample size, where for large sample size, its searching algorithm will become too computationally expensive (Ross and Markowetz, 2016). On the other hand, large sample size may greatly reduce the possibility of missing clones. In this study, we mainly focus on recovering the genotypes of cells/clones/subclones from SCS data with high accuracy, especially for large

data size, and users of RobustClone can choose whether to construct MST and/or phylogenetic tree according to their preferences.

Beside point mutation events, which generate SNVs, CNV events also happen commonly in tumors. Unlike other methods that apply only to one type of data, RobustClone can be performed on both scSNV and scCNV data. The application of real data in both cases demonstrated that RobustClone has considerable power in recovering the genotypes of single cells and/or clones, as well as reconstructing cell and/or clone trees. Noted that, the current RobustClone works solely on either scSNV or scCNV, but cannot take both data types together. Our future work will put effort on jointly correcting for errors in scSNV and scCNV together, to gain more accurate information for the inference of cancer evolution.

## Acknowledgements

We thank Prof. Sudhir Kumar, Prof. Weiwei Zhai and Dr Hechuan Yang for useful discussion and comments, and thank Glen Stecher for supports on the Linux version of BEAM. We also acknowledge the anonymous reviewers for their insightful and constructive feedback.

## Funding

This work was supported by the National Key R&D Program of China [2018YFB0704304]; National Natural Science Foundation of China [11571349, 91630314, 81673833 and 11971459]; the Strategic Priority Research Program of CAS [XDB13050000]; National Center for Mathematics and Interdisciplinary Sciences of CAS; LSC of CAS; and the Youth Innovation Promotion Association of CAS.

*Conflict of Interest*: none declared.

## References

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.*, **2008**, P10008.

Campbell,K.R. *et al.* (2019) clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biol.*, **20**, 54.

Candes,E.J. *et al.* (2011) Robust principal component analysis? *J. AMC*, **58**, 1–37.

Chen,C. *et al.* (2020) scRMD: Imputation for single cell RNA-seq data via robust matrix decomposition. *Bioinformatics*, btaa139, 10.1093/bioinformatics/btaa139.

Davis,A. and Navin,N.E. (2016) Computing tumor trees from single cells. *Genome Biol.*, **17**, 113.

Deshwar,A.G. *et al.* (2015) PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.*, **16**, 35.

Eirew,P. *et al.* (2015) Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, **518**, 422–426.

El-Kebir,M. *et al.* (2015) Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, **31**, i62–i70.

El-Kebir,M. *et al.* (2016) Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Syst.*, **3**, 43–53.

El-Kebir,M. *et al.* (2018) Inferring parsimonious migration histories for metastatic cancers. *Nat. Genet.*, **50**, 718–726.

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Gawad,C. *et al.* (2014) Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl. Acad. Sci. USA*, **111**, 17947–17952.

Hou,Y. *et al.* (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, **148**, 873–885.

Hsu,D. *et al.* (2011) Robust matrix decomposition with sparse corruptions. *IEEE Trans. Inf. Theory*, **57**, 7221–7234.

Hughes,A.E.O. *et al.* (2014) Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing. *PLoS Genet.*, **10**, e1004462.

Jahn,K. *et al.* (2016) Tree inference for single-cell data. *Genome Biol.*, **17**, 86.

Jiang,Y. *et al.* (2016) Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci. USA*, **113**, E5528–E5537.

Jiao,W. *et al.* (2014) Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, **15**, 35.

Lan,F. *et al.* (2017) Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.*, **35**, 640–646.

Lawson,D.A. *et al.* (2018) Tumour heterogeneity and metastasis at single-cell resolution. *Nat. Cell Biol.*, **20**, 1349–1360.

Levine,J.H. *et al.* (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, **162**, 184–197.

Lin,Z. *et al.* (2011) The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv: 1009.5055v2*.

Mcpherson,A. *et al.* (2016) Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.*, **48**, 758–767.

Miura,S. *et al.* (2018). Computational enhancement of single-cell sequences for inferring tumor evolution. *Bioinformatics*, **34**, i917–i926.

Navin,N. *et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature*, **472**, 90–94.

Navin,N.E. (2014) Cancer genomics: one cell at a time. *Genome Biol.*, **15**, 452.

Newman,M.E.J. and Girvan,M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E*, **69**, 026113.

Nowell,P. (1976) The clonal evolution of tumor cell populations. *Science*, **194**, 23–28.

Opasic,L. *et al.* (2019) How many samples are needed to infer truly clonal mutations from heterogenous tumours? *BMC Cancer*, **19**, 403.

Rashid,S. *et al.* (2019) Dhaka: variational autoencoder for unmasking tumor heterogeneity from single cell genomic data. *Bioinformatics,* btz095, https://doi.org/10.1093/bioinformatics/btz095.

Ross,E.M. and Markowetz,F. (2016) OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.*, **17**, 69.

Roth,A. *et al.* (2016) Clonal genotype and population structure inference from single-cell tumor sequencing. *Nat. Methods*, **13**, 573–576.

Schwartz,R. and Schäffer,A.A. (2017) The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.*, **18**, 213–229.

Shang,F. *et al.* (2014) Robust principal component analysis with missing data. In: *CIKM '14: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*. pp. 1149–1158. ACM, New York, NY, USA.

Shapiro,E. *et al.* (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, **14**, 618–630.

Shekhar,K. *et al.* (2016) Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, **166**, 1308–1323.

Sun,J.X. *et al.* (2018) A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLoS Comput. Biol.*, **14**, e1005965.

Vidal,R. *et al.* (2016) *Generalized Principal Component Analysis*. Springer, Berlin Heidelberg.

Wang,Y. *et al.* (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, **512**, 155–160.

Williams,M.J. *et al.* (2016) Identification of neutral tumor evolution across cancer types. *Nat. Genet.*, **48**, 238–244.

Wright,J. *et al.* (2012) Compressive principal component pursuit. In: 2012 *IEEE International Symposium on Information Theory Proceedings (ISIT)*. IEEE, Cambridge, MA.

Xu,X. *et al.* (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, **148**, 886–895.

Yang,Z. (2014) *Molecular Evolution: A Statistical Approach*. Oxford University, Oxford.

Yu,C. *et al.* (2014) Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell Res.*, **24**, 701–712.

Yuan,K. *et al.* (2015) Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol.*, **16**, 36.

Zaccaria,S. *et al.* (2018) Phylogenetic copy-number factorization of multiple tumor samples. *J. Comput. Biol.*, **25**, 689–708.

Zafar,H. *et al.* (2017) SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol.*, **18**, 178.

Zahn,H. *et al.* (2017) Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods*, **14**, 167–173.

Zare,H. *et al.* (2014) Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.*, **10**, e1003703.