Bioinformatics, 34, 2018, i917–i926
doi: 10.1093/bioinformatics/bty571
ECCB 2018

Computational enhancement of single-cell sequences for inferring tumor evolution

Sayaka Miura[1,2], Louise A. Huuki[1,2], Tiffany Buturla[1,2], Tracy Vu[1,2], Karen Gomez[1,2] and Sudhir Kumar[1,2,3,*]

[1]Institute for Genomics and Evolutionary Medicine and [2]Department of Biology, Temple University, Philadelphia, PA 19122, USA and [3]Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Jeddah, Saudi Arabia

*To whom correspondence should be addressed.

Abstract

Motivation: Tumor sequencing has entered an exciting phase with the advent of single-cell techniques that are revolutionizing the assessment of single nucleotide variation (SNV) at the highest cellular resolution. However, state-of-the-art single-cell sequencing technologies produce data with many missing bases (MBs) and incorrect base designations that lead to false-positive (FP) and false-negative (FN) detection of somatic mutations. While computational methods are available to make biological inferences in the presence of these errors, the accuracy of the imputed MBs and corrected FPs and FNs remains unknown.

Results: Using computer simulated datasets, we assessed the robustness performance of four existing methods (OncoNEM, SCG, SCITE and SiFit) and one new method (BEAM). BEAM is a Bayesian evolution-aware method that improves the quality of single-cell sequences by using the intrinsic evolutionary information in the single-cell data in a molecular phylogenetic framework. Overall, BEAM and SCITE performed the best. Most of the methods imputed MBs with high accuracy, but effective detection and correction of FPs and FNs is a challenge, especially for small datasets. Analysis of an empirical dataset shows that computational methods can improve both the quality of tumor single-cell sequences and their utility for biological inference. In conclusion, tumor cells descend from pre-existing cells, which creates evolutionary continuity in single-cell sequencing datasets. This information enables BEAM and other methods to correctly impute missing data and incorrect base assignments, but correction of FPs and FNs remains challenging when the number of SNVs sampled is small relative to the number of cells sequenced.

Availability and implementation: BEAM is available on the web at https://github.com/SayakaMiura/BEAM.

Contact: s.kumar@temple.edu

1 Introduction

Tumor sequencing is yielding critical insights into somatic drivers of tumorigenesis and clonal structure of heterogeneous tumors (Brastianos *et al.*, 2015; Gawad *et al.*, 2014; Gundem *et al.*, 2015; McFadden *et al.*, 2014; Nassar *et al.*, 2015; Navin *et al.*, 2011; Nik-Zainal *et al.*, 2012; Sanborn *et al.*, 2015; Xue *et al.*, 2017; Yachida *et al.*, 2010; Zhao *et al.*, 2016). The rapid advancement of single-cell sequencing technologies has made it possible to profile somatic mutations carried by individual cells (Eirew *et al.*, 2015; Francis *et al.*, 2014; Gawad *et al.*, 2014, 2016; Huang *et al.*, 2015; Hughes *et al.*, 2014; Navin, 2014, 2015; Paguirigan *et al.*, 2015; Shapiro *et al.*, 2013; Van Loo and Voet. 2014; Yu *et al.*, 2014; Zafar *et al.*, 2016). Many studies have performed single-cell sequencing on tumors to identify clones and their evolutionary relationships (Eirew *et al.*, 2015; Gawad *et al.*, 2014; Hou *et al.*, 2012; Jan *et al.*, 2012; Li *et al.*, 2012; Melchor *et al.*, 2014; Navin, 2015; Potter *et al.*, 2013; Xu *et al.*, 2012; Yu *et al.*, 2014). Thus, single-cell sequencing will be instrumental in revealing the genetic changes that occur during cancer progression, which is a prerequisite for clone identification and the inference of evolutionary relationships among cells and relative timing of mutation events. But, the utility of current single-cell sequencing technologies is limited by many technical issues (Gawad *et al.*, 2016; Navin, 2014, 2015; Ning *et al.*, 2014; Wang and Navin, 2015). For example, the low physical coverage of some genomic regions and

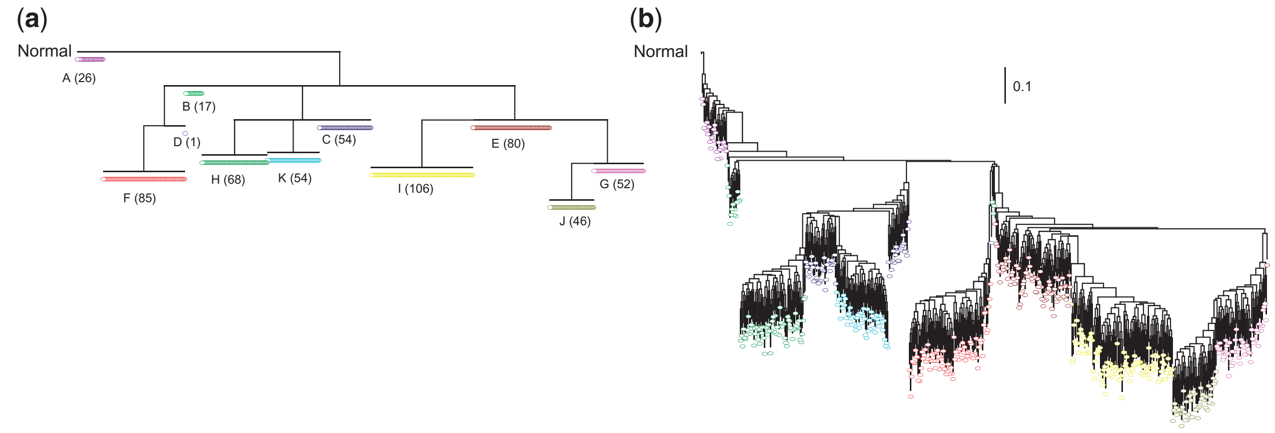

Fig. 1. Impact of missing data and sequencing errors on the inferred cell phylogeny. (**a**) The true evolutionary tree of 1000 cells distributed among tumor clones A–K (shown in different colors); the number of cells sampled for each clone is in parentheses. (**b**) Cell phylogeny inferred using simulated single-cell sequences in which 500 SNVs were sampled. Roth *et al.* (2016) software and parameter settings were used to generate the data with 20% missing bases, 28% false-positives and 7% false-negatives. A maximum likelihood method for phylogenetic analysis of SNV data (Stamatakis, 2014) was used to infer the cell phylogeny. Branch lengths are drawn to scale (number of SNVs/site). The inferred cell phylogeny shows greater sequence divergence than the true phylogeny due to the influence of many false-positive mutations

positions prevents unambiguous assignment of a nucleotide base to those positions, known as 'missing bases' (MBs). Allelic dropout (ADO) events cause false-negatives (FNs) when mutant alleles are present but not amplified. Infidelity of amplification can cause false-positives (FPs) when errors during initial amplification are inherited to subsequent molecules and a 'mutation' is identified that was not present in the sampled cell. Sometimes a single-cell cannot be completely separated from other cells, which results in the sequencing of multiple cells together. FP rates (3×10^{-5}–7×10^{-5} per homozygous wild-type positions) and ADO rates (0.2–0.4 per heterozygous site) can exceed the rate of occurrence of true mutations (Ross and Markowetz 2016). MBs also occur at frequencies as high as 58% (Hou *et al.*, 2012) in single-cell data sequences (Gawad *et al.*, 2016). All of these problems result in inaccurate single-cell sequences even when high sequencing coverage has been achieved.

Many new methods have been developed to compensate for these issues and allow reliable inference from single-cell sequence datasets. For example, OncoNEM (Ross and Markowetz, 2016) and BitPhylogeny (Yuan *et al.*, 2015) identify clones and their evolutionary relationship (i.e. clone phylogeny). SiFit infers cell phylogeny with the consideration of sequencing errors (Zafar *et al.*, 2017). SCITE (Jahn *et al.*, 2016) and Kim and Simon (2014) methods are designed to infer the order of mutations over time from the single-cell sequences. SCG is designed to deal with the issue of multi-cell sequencing when inferring clone sequences (Roth *et al.*, 2016).

These methods produce corrected single-cell sequences, but they do not report their performance in imputing MBs correctly and reducing FPs and FNs. The primary focus of these current methods has been to improve the quality of biological inferences from error-containing single-cell sequencing data. Consequently, the absolute and relative performance of current methods for reducing the error present in single-cell sequences is not known. Significant improvement in the quality of single-cell sequences will enable use of a large number of sophisticated methods in molecular phylogenetics (Nei and Kumar, 2000) for inferring the evolutionary history of clones, reconstructing ancestral clones, identifying early and late occurring driver mutations, and characterizing inter- and intra-tumor heterogeneity. These standard approaches cannot currently be used for tumor single-cell data, because they are not robust to the presence of high levels of sequence error (Zafar *et al.*, 2017). For example, a widely-used maximum likelihood method (Stamatakis, 2014) produces a cell phylogeny (Fig. 1b) from simulated single-cell sequence data (with MBs, FPs and FNs) that is clearly very different from the true tree (Fig. 1a). In addition to various inconsistencies in the evolutionary relationships, the branch lengths leading to the tips of the phylogeny are extensively overestimated, because all the cells of a clone (same color) are actually identical (Fig. 1a). Consequently, the inferred cell phylogeny shows much greater evolutionary depth, resulting in inflated estimates of tumor heterogeneity and incorrect mapping of mutations. Therefore, single-cell sequences require correction before use in downstream biological analysis.

In this article, we present the performance of four existing methods (OncoNEM, SCG, SCITE and SiFit) in correctly imputing MBs and reducing the numbers of FPs and FNs. We excluded methods that did not produce single-cell sequences, e.g. BitPhylogeny (Yuan *et al.*, 2015) and the method of Kim and Simon (2014). In addition, we propose and test a new method, Bayesian Evolution-Aware Method (BEAM), which employs molecular phylogenetics and a Bayesian prediction framework to improve the quality of single-cell sequences (see Materials and methods). Our testing focused on computer simulated datasets, as knowledge of the true single-cell sequences enables direct assessment of the performance of computational methods (Ross and Markowetz, 2016; Roth *et al.*, 2016). We also analyzed one empirical dataset (Li *et al.*, 2012) to gauge the utility of computational approaches in a real-world scenario and the concordance of the inferences produced.

In the following, we present information on the simulated data used in our evaluation of methods, followed by a description of the BEAM approach and the assumptions, parameters, and accuracy measures used. We then present results from our analyses of simulated and empirical data discuss the patterns observed.

2 Materials and methods

2.1 Generation of datasets by computer simulations

Roth *et al.* datasets (R1000 × 50 and R100 × 50 datasets): We used the simulator and parameter settings described by Roth *et al.* (2016) to produce 240 datasets. This simulator first generates a clone phylogeny and then the clone genotypes. A new model phylogeny is

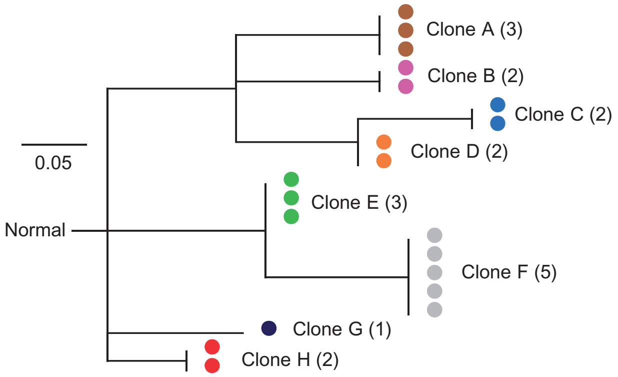

Fig. 2. An example cell phylogeny used in computer simulations. In this phylogeny, there are eight distinct clones (A–H), each represented by 1–5 cells. This phylogeny was used to generate one dataset by Ross and Markowetz (2016) simulator

generated for every dataset (e.g. Fig. 1a for 1000 cells and 10 distinct clones). To generate a clone phylogeny, new clones are created by accumulating mutations (a mutation rate set to 0.1 per site) until all SNV loci (50) are created. The simulator uses an infinite sites assumption, so no mutations override each other, and introduces loss of heterozygosity at a rate of 0.2 per site in which heterozygous mutants are changed into homozygous mutants without allowing any loss of mutations. A clone genotype is assigned to each cell by sampling from a categorical distribution (clonal prevalence), which is generated from a symmetric Dirichlet distribution with the parameter value of 1.

Datasets with 100 and 1000 cells were produced by Roth *et al.* (2016) in which doublet single-cell sequencing was simulated by sampling two clone genotypes at different rates: 5%, 10, 20 and 40% of the cells. The simulator generated allelic count data with ADO by sampling from the empirical distribution of SNV frequencies specified in Roth *et al.* (2016). The depth of coverage at each locus was chosen from a Poisson distribution with a mean of 1000 reads. The number of variant reads was sampled from a Binomial distribution with the parameter selected from an empirical distribution and the depth of coverage sampled from a Poisson distribution. To determine if an allele was present or absent, the Binomial exact test was performed and a p-value threshold of 10^{-6} was used. In the resulting data, 3–51% of the observed mutant alleles were FPs and 2–20% of the observed homozygous wild-type alleles were FNs. We randomly assigned a 'missing' value (MB) to 20% of the bases.

Ross and Markowetz datasets (M10 × 50 – M50 × 300 datasets): This collection of 690 datasets was generated using the Ross and Markowetz (2016) simulator and parameter settings. In their approach, clone phylogenies were first generated by iteratively adding a branch with a node to an existing node that was randomly chosen from a growing phylogeny (1, 5, 10 and 20 clones). Unobserved clones were then introduced by removing clones that had at least two descendant clones (0, 1, 2, 3 and 4 unobserved clones). A new clone phylogeny was generated for each dataset, and each cell is assigned to a clone with a probability corresponding to its size (10, 20, 30 and 50 cells). Figure 2 shows an example phylogeny of 20 cells used to simulate clone evolution for generating M datasets (M10 × 50 – M50 × 300 datasets). Along the clone phylogeny, true clone genotypes were generated by assigning mutations with a uniform probability (50, 100, 200 and 300 SNVs). Observed genotypes were derived from true genotypes by introducing MBs (10, 20, 30 and 40% of SNVs), FPs (10^{-5}, 5, 10, 20 and 30% of mutant alleles) and FNs (5, 10, 20 and 30% of wild-type alleles).

2.2 Accuracy measurements

We recorded the numbers of missing bases (MBs), false-positives (FPs) and false-negatives (FNs) in the simulated single-cell sequence datasets. The total number of correct positions (with no MB, FP, or FN) were aggregated and divided by the product of the number of cells and the number of SNVs. This quantity is referred to as the initial sequence quality (Q_0), which is the same as the mean Hamming distance between the true and the inferred single-cell sequence. After the sequence data was subjected to computational analysis by BEAM, OncoNEM, SCG, SCITE and SiFit, the sequence quality was reassessed (Q_{BEAM}, Q_{OncoNEM}, Q_{SCG}, Q_{SCITE} and Q_{SiFi}, respectively).

While the positions containing MBs, FPs and FNs are known for simulated data, no such information exists in the analysis of empirical data and the computational methods must be applied to all the positions. Therefore, we also compared the total numbers of MBs, FPs and FNs before and after the application of a computational method to a dataset. As the number of MBs in the simulated data was larger than FPs or FNs, and thus the accuracy of MB corrections may dominate this measure.

2.3 New method evaluated (BEAM)

The new Bayesian evolution-aware method (BEAM) uses classical molecular evolutionary phylogenetics to impute missing data and detect base assignment errors in the single-cell sequencing data. It is based on the premise that significant evolutionary information is present in the initial cell sequences regardless of base assignment errors. For example, cells from the same clone show a strong tendency to occur in close proximity in the initial cell phylogeny (Fig. 1b), as seen by the location of cells marked by the same color in the true tree (Fig. 1a), despite the presence of a large number of MBs, FPs and FNs in the simulated sequence data. BEAM uses this intrinsic evolutionary information and computes a Bayesian posterior probability (PP) of observing all possible alleles at each SNV position in each single-cell sequence, as described below.

For brevity, we explain BEAM using an example dataset that was generated using the cell phylogeny in Figure 2. It consists of 20 single-cells from eight distinct clones and 200 SNVs. The simulated sequence dataset contained 800 MBs, 429 FPs and 106 FNs (Fig. 3a). For this data, we first infer a cell phylogeny from the observed single-cell sequences by using a maximum likelihood method specifically suited for phylogenetic analysis of SNV data (Stamatakis 2014) (Fig. 3a). This approach does not require the infinite sites assumption, i.e. mutations are allowed to be lost and they may occur at the same genomic position in different cells, which is different from the principle applied in OncoNEM, SiFit and SCITE (Jahn *et al.*, 2016; Ross and Markowetz, 2016; Zafar *et al.*, 2017).

In this example, cells from the same clones (the same color) generally cluster together, but identical cells of a clone can show extensive observed sequence divergence (e.g. brown cells in Figs 2 and 3a). Given this initial cell phylogeny and the initial cell sequences, we estimate PP of each possible base assignment at each position in a cell sequence following equation 5 in Liu *et al.* (2016). In brief, consider a set of four sequences and their evolutionary tree (Fig. 4). In this tree, x_1 to x_3 represent the nucleotides at a given position in the tumor cell sequence; x_4 is the wild-type base from the normal cell sequence. To estimate PP of a base assignment at x_1, nucleotides at the tip nodes are represented by the vector with the constraint that x_1 is unknown, that is, $\boldsymbol{x} = (\text{'?'}, x_2, x_3, x_4)$. Also, let $\boldsymbol{y} = (y_1, y_2)$ represent the vector of nucleotides y_1 and y_2 at the two ancestral

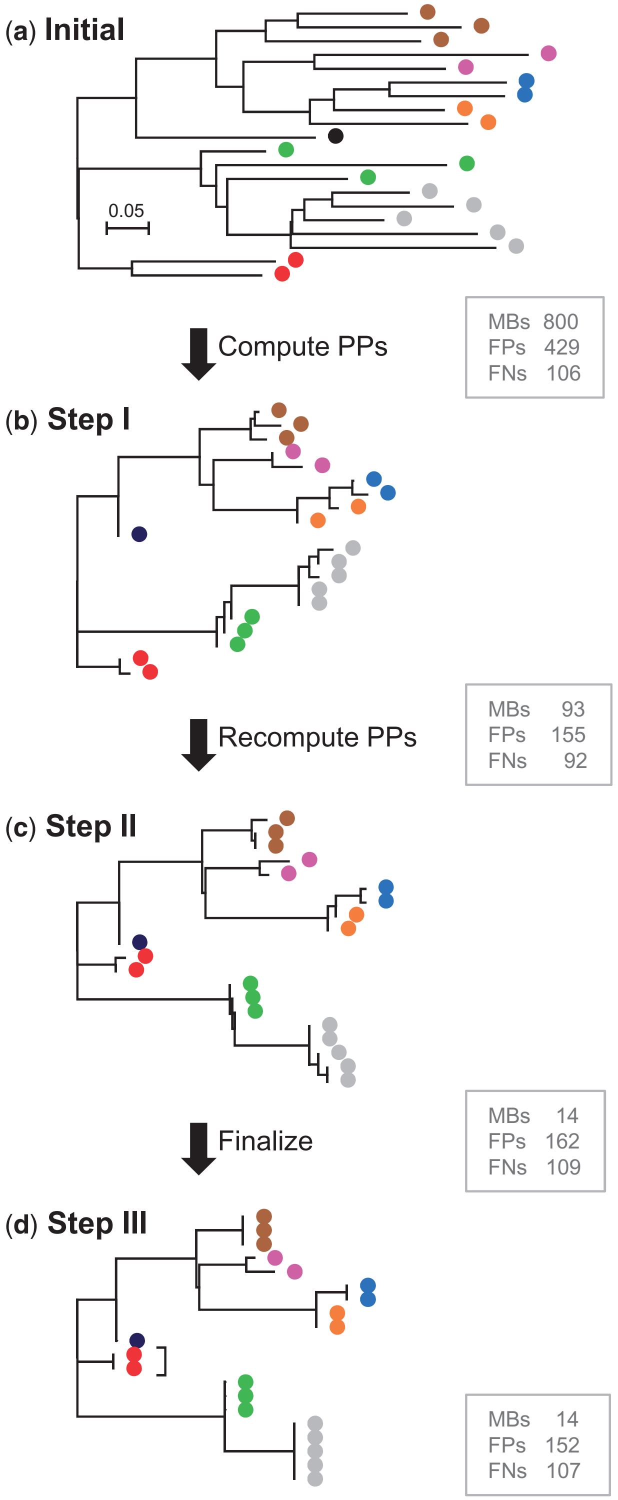

Fig. 3. An overview of the BEAM approach. (**a**) The initial cell phylogeny, along with branch lengths, derived from single-cell sequence data simulated using the model tree in Figure 2. The Ross and Markowetz (2016) simulator and parameter settings were used. The number of MBs, FPs and FNs in the initial sequence data are shown in the box. (**b**) Cell phylogeny produced from the data in which MBs were imputed and FPs and FNs were corrected. The remaining MBs, FPs and FNs are shown. (**c**) Improved cell phylogeny after recomputing PPs by using the phylogeny in b. (**d**) The final cell phylogeny produced by BEAM along with the remaining MBs, FPs and FNs. The topology and branch lengths are very similar to those in the model tree shown in Figure 2

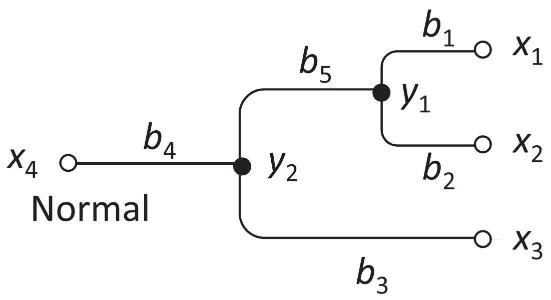

Fig. 4. A simple tree used in explaining PP computation. x_i is a base in the single-cell sequence, y_i is a base in the ancestral sequence and b_i refers to a branch length

nodes in this tree. Then, the Bayesian posterior probability for a set of nucleotides (x_1, $\boldsymbol{y}$) can be computed as follows (Liu *et al.*, 2016):

$$f(x_1, \boldsymbol{y}|\boldsymbol{x}; \boldsymbol{b}) = f(x_1) \times f(\boldsymbol{x}|x_1, \boldsymbol{y}; \boldsymbol{b})/f(\boldsymbol{x}; \boldsymbol{b}) \qquad (1)$$

Here, $\boldsymbol{b}$ is the vector of branch lengths $b_1, \ldots, b_5$. The prior, $f(x_1)$, is the product of the probability of change from its ancestral base y_1 and the frequency of the base y_1 in the sequence data. The likelihood, $f(\boldsymbol{x}|x_1, \boldsymbol{y}; \boldsymbol{b})$, is computed by assuming x_1 is unknown, while the denominator is computed for all possible combinations of x_1, y_1 and y_2. Then, the PP of a nucleotide at x_1 is the sum of probabilities of $f(x_1, \boldsymbol{y}|\boldsymbol{x}; \boldsymbol{b})$ for all possible combinations of y_1 and y_2. The PP of base A at x_1, for example, is given by:

$$\mathrm{PP}\ (x_1 = \mathrm{A}) = \Sigma_{y;x1=\mathrm{A}} f(\mathrm{A}) \times f(\boldsymbol{x}|x_1, \mathrm{y}; \boldsymbol{b})/f(\boldsymbol{x}; \boldsymbol{b}) \qquad (2)$$

We compute PP for each possible base assignment at each position (i) in the single-cell sequence of interest, including those for which a base call is available because they could be FPs or FNs. If there are m single-cell sequences and n SNV positions in the dataset, $m \times n$ computations are performed. We assigned the base with $\mathrm{PP} > 0.7$ to the position of interest. When all of the bases show $\mathrm{PP} < 0.7$ at a position, we assign the base or missing value originally observed at that position. To select this cut-off, we compared Qs between various PP cut-offs (0.6–0.9) on a simulated data and identified PP with the best performance. We found that the use of higher PP cut-offs will increase the number of MBs, whereas a lower PP cut-off will increase the number of FPs and FNs. So, the use of 0.7 as an intermediate cut-off seems to be the best choice. We used this PP cut-off value to analyze all the datasets.

In our example, BEAM improves the single cell sequences by imputing base identities at MB positions by reducing these from 800 to 93, reducing FPs from 429 to 155, and reducing FNs from 106 to 92. Because of these improvements, the inferred cell phylogeny becomes more accurate (Fig. 3b). At the same time, the erroneously long tip branches in the initial phylogeny are shortened via elimination of FPs (Fig. 3a).

In the next step, we use the new cell phylogeny (Fig. 3b) and recompute PPs by applying Equations 1 and 2. Now, MBs decrease from 93 to 14 and FPs and FNs increased slightly (155 to 162 and 92 to 109, respectively; Fig. 3c). The cell phylogeny inferred using these single-cell sequences looks similar to the true phylogeny (Fig. 2). No further iterations are needed after this step, and the last step is to annotate clones. Clone delineations can be done by using the bootstrap procedure to assess the robustness of the branching patterns (Felsenstein, 1985; Nei and Kumar, 2000), merging cells by iterative clustering of nodes along branches (Ross and Markowetz, 2016), or by using a k-medoids clustering approach on a distance matrix that is obtained from the latest cell phylogeny (Zafar *et al.*, 2017). BEAM provides the user with an option to assign identical sequences to all the cells of a clone by erasing potentially spurious mutations. In our simulations, we found that it was best to assign identical clone sequences to cells that were connected with effectively zero branch lengths in the phylogeny (default: <2% of SNVs; Fig. 3d). In this case, BEAM will finalize sequences by computing PPs of each possible base assignment and comparing the average PPs between potential bases in order to assign the base with the highest average PP to all the cells from the same clone. Ultimately, BEAM produces refined single-cell sequences as well as the cell phylogeny.

2.4 Options used for analyzing data

We used default or recommended parameters to perform SCG, SCITE, OncoNEM and BEAM analyses. SCG (Roth *et al.*, 2016) was performed with the doublet option. When the status of mutations (presence or absence) within a predicted clone genotype had <0.95 probability, we assigned missing values to those positions. To assign a clone for each cell, SCG additionally computed the probability of having a predicted clone genotype for each cell. Thus, we assigned the predicted clone genotype with the highest probability. When none of predicted clone genotypes had a high probability (>0.01) for a cell, no predicted clone genotype was assigned to the cell. Cells lacking clone genotype assignments were removed, and datasets consisting entirely of cells lacking clone genotype assignments were removed from accuracy considerations as we considered that SCG failed to correct sequences for these datasets.

OncoNEM (Ross and Markowetz, 2016) analyses were performed using true rates of false positive and false negative base assignment errors in the input files (observed sequences). Maximum Bayes factor for which a smaller model was preferred was 10, and the model search stopped when the best scoring tree stabilized for at least 200 iterations. Mutant nucleotides were assigned when the probability of observing the mutation at a given position was >0.95 and missing bases were assigned when the probability of observing the mutant nucleotide was between 0.05 and 0.95. SCITE (Jahn *et al.*, 2016) analyses were performed by giving true rates of false-positive and false-negative detections of mutations. The desired number of repetitions of the MCMC was 1, and the desired chain length of each MCMC repetition was 900 000. Often, mutant nucleotides were not assigned to any cells. When mutant nucleotides were not assigned to any cells, those sites were assigned with wild-type bases. When multiple possible cell phylogenies for a dataset were produced, we used the option to marginalize out the alternatives. We distinguished heterozygous and homozygous mutations for R1000 × 50 and R100 × 50 datasets. When multiple cell sequences were inferred for a single cell, we replaced inconsistent base assignments with MBs. For SiFit (Zafar *et al.*, 2017) analyses, we also input true rates of false positive and false negative detection of mutations. The number of iterations run for each restart was 10 000. Cell genotypes were inferred by inferring the order of mutations along the cell phylogeny predicted by SiFit. LOH rate and deletion rate were set to zero.

2.5 Empirical data

We analyzed a single-cell sequencing dataset from muscle-invasive bladder tumors (Li *et al.*, 2012), which has been analyzed previously in other articles proposing new methods (Ross and Markowetz, 2016; Zafar *et al.*, 2017). We obtained sequenced reads in FASTQ format by using SRA toolkit (v2.8.1) (Leinonen *et al.*, 2011). We mapped these sequenced reads to the human genome sequence (hg18 from UCSC database; https://genome.ucsc.edu/) by using the Burrows-Wheeler alignment tool (BWA v0.7.12) (Li and Durbin, 2009) with *aln* options. Samtools (v1.3.1) (Li *et al.*, 2009) was used to remove reads with low mapping quality (≤40) when creating BAM files, which were sorted by chromosome coordinate. This initial data processing follows the protocol described in Zafar *et al.* (2016).

We then used Monovar (Zafar *et al.*, 2016) to call mutations. We performed *mpileup* in Samtools with the options presented in the instructions for Monovar (i.e. minimum base quality was zero). Monovar analysis was performed with the default or recommended options, i.e. offset for prior probability of false-positive error was 0.002; offset for prior probability of allelic drop out was 0.2; threshold for variant calling was 0.05; and the number of threads used in multiprocessing was 2.

For downstream computational analysis, we selected SNVs in coding regions that were identified by Zafar *et al.* (2016). Nucleotides identified among the majority of normal cells were assigned as wild-types, and the other bases found in tumor cells were assigned as mutants. Our analyses did not distinguish homozygous and heterozygous mutations. Following Zafar *et al.* (2016), we assigned missing values to positions with coverage depth less than 6×, in addition to positions where Monovar did not predict a genotype. Lastly, we removed one cell that contained a very large number of missing bases.

3 Results

3.1 Analysis of simulated large datasets

We first present results from the analysis of 120 large datasets that consisted of 1000 cells each with 50 SNVs (R1000 × 50). The initial sequence quality (Q_0) of these datasets ranged from 65 to 75% (Fig. 5a), which was caused by 20% MBs, 7–51% FPs and 2–16% FNs. Four of the five methods were able to handle these large datasets and produced refined single-cell sequences of much higher quality than the input (Fig. 5a). The average Q_{BEAM} was 89% and varied from 80–95%. The performance was the best when the input sequences contained the fewest errors and the worst when the input sequence error was the largest. SCITE performed similarly well ($Q_{\mathrm{SCITE}} = 90\%$) and showed very similar trends describing the relationship between input and output quality. While SCG also worked well ($Q_{\mathrm{SCG}} = 89\%$), it showed much greater variability in performance. SiFit showed the smallest improvements ($Q_{\mathrm{SiFit}} = 80\%$). Overall, the sequence quality improved for all datasets after a computational method was used.

We found that the correct imputation of MBs was the primary reason for the improvements observed for R1000 × 50 datasets. The fraction of MBs found in the single-cell sequences decreased, on average, by 100, 100 and 96% after the application of BEAM, SCG and SCITE, respectively, and most of the missing data were correctly imputed (88, 88 and 87%, respectively) (Fig. 5b). Although SiFit imputed all of the MBs without producing new MBs, SiFit correctly imputed a much smaller fraction of MBs than the other methods (70% of MBs were correctly imputed). Interestingly, however, no method showed a significant ability to reduce the total number FPs and FNs in this data (Fig. 5c), as the numbers of FPs and FNs in the output were greater than in the input (output/input ratio >1.0). Of all the methods, SiFit showed the worst average ratio (2.18), which can happen because the correction procedure has to be applied to all the bases in the input sequence data as the positions with FP and FN

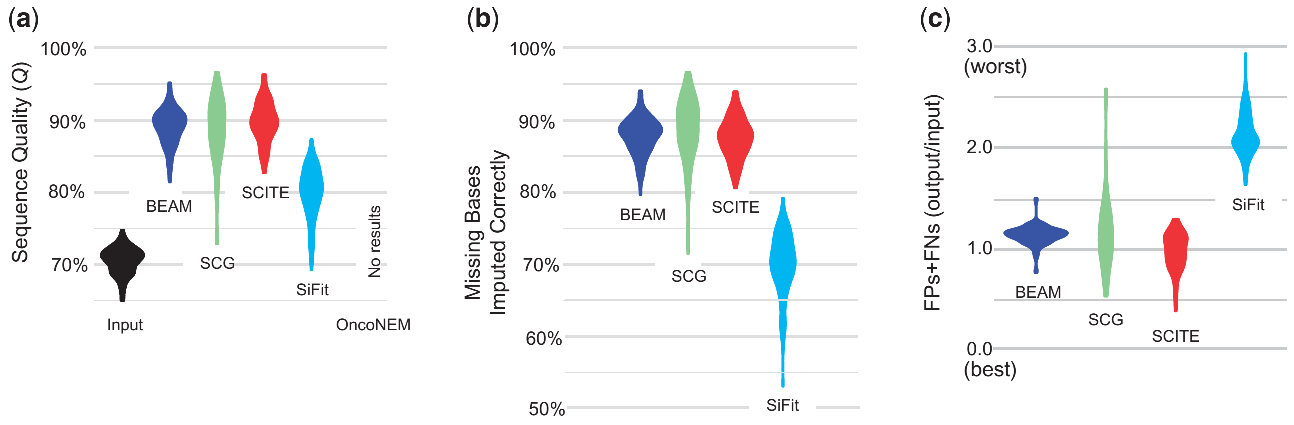

Fig. 5. Improvement in the single-cell sequences realized by using computational methods for processing larger datasets (R1000 × 50). (**a**) Sequence quality of data input to computational methods and of data output from four computational methods. (**b**) Proportions of missing bases imputed correctly by computational methods. (**c**) The ratio of the number of FPs and FNs in the output and the input sequences (output/input)

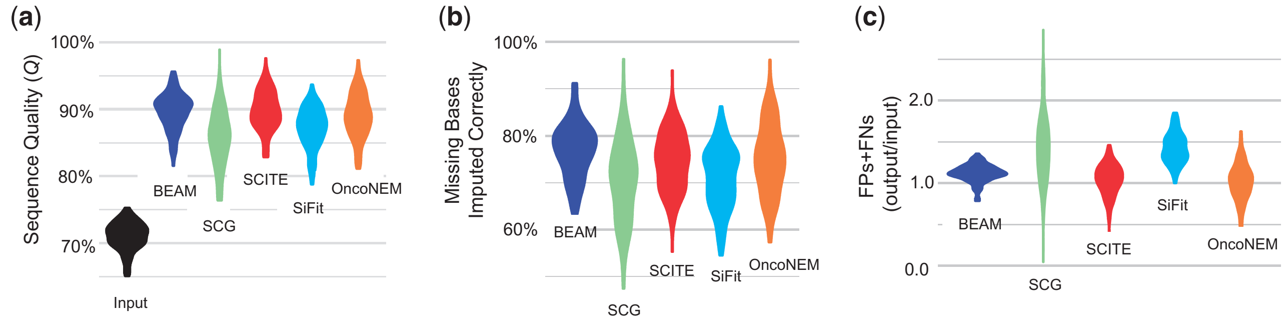

Fig. 6. Improvement in the single-cell sequences for medium sized datasets (R100 × 50). (**a**) Sequence quality of the data input to five computational methods and of output from computational processing. (**b**) Proportions of missing bases imputed correctly by computational methods. (**c**) The ratio of the number of FPs and FNs in output and input sequences (output/input)

are not known in the real world data analysis. This causes the creation of many new false positives and false negatives.

Next, we tested the accuracy of computational methods for datasets in which the number of sequenced cells was reduced to 100 (R100 × 50). Again, all methods successfully improved single-cell sequences with very similar output sequence quality (Fig. 6a). Outcomes were similar to those observed for R1000 × 50 datasets, except that SiFit performed much better and OncoNEM produced results. As with the larger dataset (R1000 × 50), the correct imputation of MBs was the primary improvement observed (Fig. 6b), and all of the methods produced sequences in which the numbers of FPs and FNs were similar to or much larger than the error in the input sequence data (Fig. 6c).

In both R1000 × 50 and R100 × 50 datasets, we observed that the identification of FPs and/or FNs was less effective than the imputation of MBs for all the methods (Figs 5c and 6c). We hypothesized that, unlike the imputation of MBs, detection and correction of FPs and FNs was very sensitive to the available cell relationship information that can be gleaned from the initial error-prone single-cell sequencing data. We tested this hypothesis by increasing the number of SNVs used to 500 (R1000 × 500), while keeping the number of errors per SNV the same as in R1000 × 50 dataset. We applied BEAM and SCG to the new collection of datasets and found that the output/input ratio of FPs and FNs for BEAM became much less than 1 (Fig. 7a). That is, BEAM was able to produce sequences with ∼40% fewer FPs and FNs. This improvement over R1000 × 50 datasets is explained by the fact that the initial cell relationships derived using the input data in BEAM are more accurate when the number of SNVs analyzed is large (compare phylogenies in Figs 7c and 1b). This improvement enables the Bayesian analysis to generate better predictions, because the phylogenetic prior is closer to the truth.

The performance of SCG did not improve, and still created more FP and FN errors in the output (Fig. 7b). This may be because SCG first clusters cells to identify clones and then assigns clone genotypes. This procedure does not appear to benefit from increased evolutionary information in the dataset. Additional analyses support this reasoning, as BEAM, OncoNEM, SCITE and SiFit were able to detect and correct many FPs and FNs when the number of SNVs sampled was large in relation to the number of cells sampled.

3.2 Analysis of smaller datasets

Next, we analyzed 690 datasets that were generated by Ross and Markowetz (2016) (M datasets). All contained fewer cells than the R1000 datasets (largest) and R100 datasets (medium-sized), but the M datasets contained the same or larger numbers of SNVs (50, 100, 200 and 300 SNVs) compared to the R datasets. All five methods produced results for all datasets, in which the initial quality of cell sequences was between 65 and 75%. BEAM, SCITE and SiFit increased the quality of the sequences to an average 92% (Fig. 8a), but OncoNEM (82%) was less accurate and SCG performed poorly (65%). OncoNEM and SCG did not impute missing data as accurately as BEAM, SCITE and SiFit (Fig. 8b).

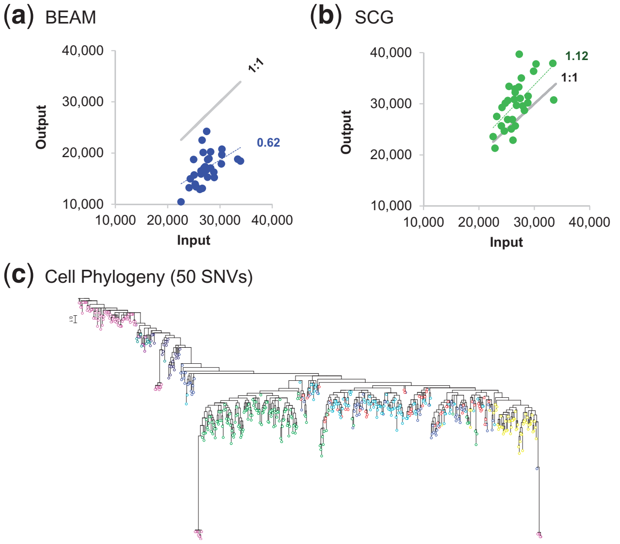

Fig. 7. Detection and correction of FP and FNs in datasets containing 1000 cell sequences with SNVs. Scatter graphs show the relationships of the number of the FPs and FNs in the input data (x-axis) and the output data (y-axis) produced by (**a**) BEAM and (**b**) SCG. Linear regression slopes through the origin are 0.62 and 1.12, respectively. (**c**) The initial cell phylogeny when the number of SNVs is small (50), which is very different from the true cell phylogeny and the cell phylogeny derived using 500 SNVs (Fig. 1b)

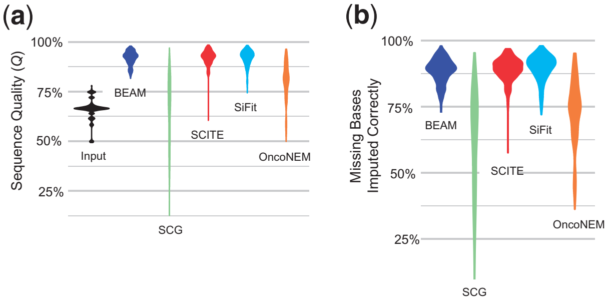

Fig. 8. Improvement in the single-cell sequences for datasets containing a small number of cells (M). **(a)** Sequence quality of the data input to computational methods and of the data output from all five methods. (**b**) Proportions of missing bases imputed correctly. These are aggregate patterns from the analysis of 690 datasets; see Figures 9 and 10 for more detailed results

All methods, except SCG, decreased the numbers of FPs and FNs in the analysis of M datasets (Fig. 9), because the ratio of the number of SNVs to the number of cells was greater than that for R datasets. In fact, increasing the number of SNVs provides a proportional increase in the performance of BEAM, OncoNEM and SiFit. Patterns observed for M datasets (Fig. 9) confirm the results for R datasets with 500 SNVs (Fig. 7): BEAM becomes more accurate with larger numbers of SNVs and SCG's performance does not improve. OncoNEM performed the best in correcting FPs and FNs in M datasets, but achieves this at the expense of producing many MBs (Fig. 8).

As expected, the quality of the inferred sequences became higher as the number of SNVs increased (Fig. 10a). The quality of the inferred sequences was higher when the number of cells was high (Fig. 10b), because all methods utilize similar cell sequences (or clusters of SNVs) to impute MBs and correct FPs and FNs. For example, the PP calculation is affected by the base assignments of neighboring cells, which will become more accurate when a base assignment is supported by larger number of cells from the same clone. Overall, BEAM, SCITE and SiFit provided the most robust results when the number of cells was small, and OncoNEM and SCG were greatly impacted if multiple cells per clone were not sampled. Lastly, the quality of the output sequences was a direct function of the fraction of MBs (Fig. 10c), FPs (Fig. 10d) and FNs (Fig. 10e).

Before proceeding with an analysis of empirical datasets, it is important to note that we did not introduce any loss of mutant alleles in our simulations, e.g. loss of heterozygosity (LOH) and the loss of genomic segments. Such mutations will negatively impact the performance of most methods, as the evolutionary relatedness of sequences will be disturbed by such losses. Therefore, such positions should be detected and removed before applying these computational methods. Also, we found that a high rate of doublet sequencing did not adversely impact the performance of any of the computational methods (Fig. 10f). This result is consistent with the finding in Jahn *et al.* (2016), who reported that doublet sequencing did not lead to lower accuracy in biological inference, which was a major motivation for the development of SCG (Roth *et al.*, 2016).

3.3 Analysis of an empirical dataset

We applied all five methods to a previously published dataset of a muscle-invasive bladder tumor (Li *et al.*, 2012). This dataset contained 55 cells with 84 SNVs in protein coding regions. The cell

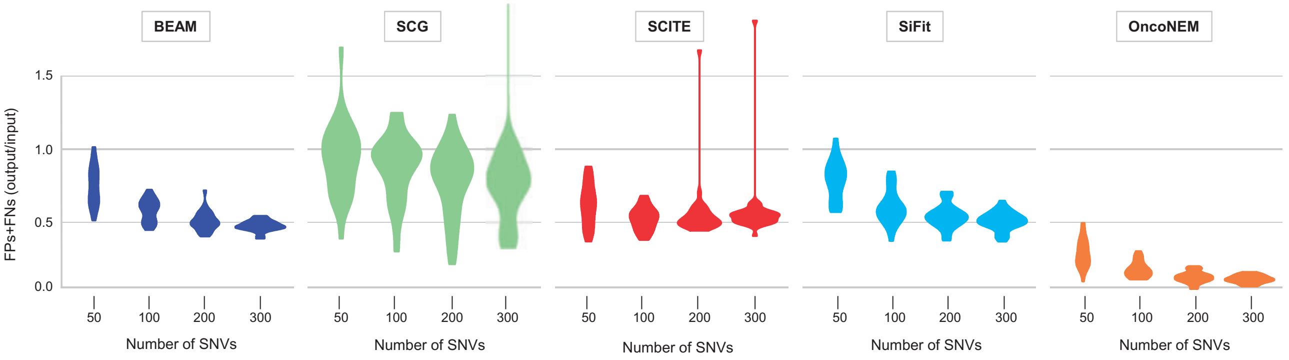

Fig. 9. Improving FPs and FNs in the single-cell sequences (M20). The ratio of FP and FN base assignments in the output and the input sequences (output/input) for BEAM, SCG, SCITE, SiFit and OncoNEM. For SCG, the trend was truncated at 2.0 for simplicity

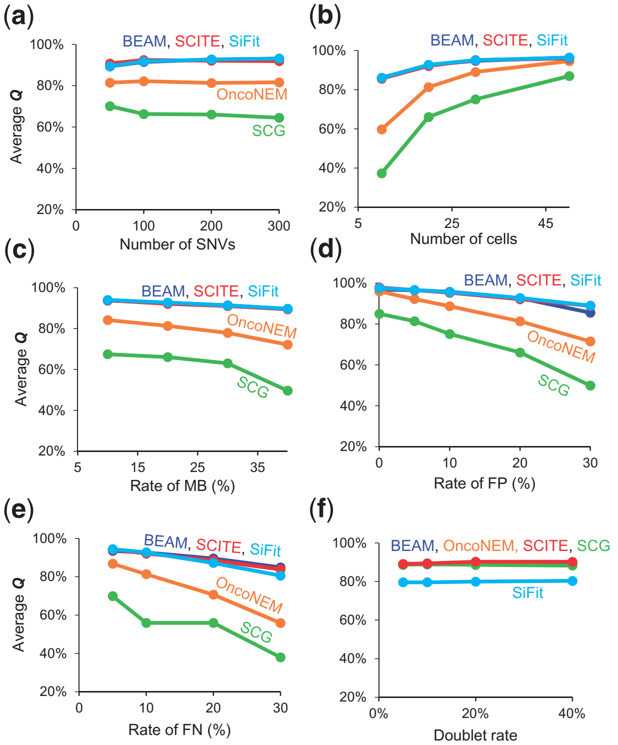

Fig. 10. The relationship of the output sequence quality with the various simulation parameters. (**a**) Number of SNVs in the data, (**b**) number of cells sequenced and fractions of (**c**) missing bases, MBs, (**d**) false positives, FPs and (**e**) false negatives, FNs in M datasets. (**f**) Effect of increasing doublet sequencing rates on the quality of the output sequencing for R datasets

phylogeny before and after the application of BEAM is shown in Figure 11a and b, respectively. As observed from computer simulated datasets, the phylogeny based on the initial single-cell sequences shows high diversity among cells, with no clear demarcation of clones; note that we colored tips in Figure 11a based on BEAM's clone predictions in Figure 11b. Following the application of BEAM, the cell phylogeny shows distinct clonal structure with 11 different tumor clones (Fig. 11b).

Li *et al.* (2012) suggested that all of the tumor cells were derived from a single ancestral cell. This conclusion is supported by BEAM, as the inferred cell phylogeny showed that all of the tumor cells had a common ancestral cell. Li *et al.* (2012) reported that a set of 15 cells, which they identified as a single clone, arose early in the tumor's evolution. BEAM found that 14 of these cells belonged to a group of early-emerging clones (gray, green and pink clones in Fig. 11b), and one cell was a part of a later arising (yellow) clone. Two BEAM-identified clones (purple and red) are the closest relatives and were contained in the second clone in Li *et al.* (2012). Cells from the third clone in Li *et al.* (2012) are divided among 5 closely-related clones by BEAM. Therefore, the cell phylogeny produced by BEAM provides a more detailed clonal structure, which remains consistent with the clonal structure (three major clones) presented by Li *et al.* (2012).

The initial cell phylogeny (Fig. 11a) was transformed into the final cell phylogeny (Fig. 11b) by BEAM, because the final single-cell sequences contained only 6 positions with missing values. This is a major improvement, as the original dataset contained 666 positions with missing values. In imputing MBs, BEAM assigned a mutant base to 337 positions and the normal base to 323 positions. 63% of the mutant base and 74% of the wild-type base assignments made by BEAM were also suggested by all other methods (Fig. 11c and d, respectively). In fact, more than 93% of BEAM's assignments for MBs were shared with at least one other computational tool. Therefore, extensive consensus exists among methods. We also examined whether BEAM's base assignments for MBs were supported by the read count data. As mentioned in an earlier section, Monovar does not assign a base when there are only a few reads. So, we tested whether the wild type base assignments by BEAM were supported by higher read counts than other bases at the examined positions. This was indeed the case, the read counts of wild type alleles assignments averaged 7 times higher. Assignment of mutant alleles to MBs were also supported by 22% larger read counts than the wild type alleles. Therefore, read count data generally supports the assignments made by BEAM.

In addition to imputing almost all MBs in the observed sequences (99% of MBs), BEAM corrected wrong mutant and wild-type base assignments. In these predictions, 105 mutation calls were found to be false positives, 65% of which were also detected by at least two other methods. That is, they were consensus assignments (≥3 out of 5 methods). Also, 183 wild-types were detected to be false negatives, i.e. they should have been mutant assignments. 88% of these were consensus assignments, becase at least two other methods suggested the same base assignment. Therefore, we expect that the use of these computational methods will enable better biological conclusions.

4 Discussion

We have reported that computational methods are generally capable of imputing missing bases with high accuracy and, thus, can improve the quality of the tumor single cell sequences. In particular, BEAM, SCG and SCITE performed well in imputing missing bases for datasets with a large number of cells. Our results also confirm Ross and Markowetz (2016)'s conclusion regarding the accuracy of SCG for datasets representing a large number of cells but containing few SNVs. However, we have newly reported that the gain in accuracy is due to correct imputation of missing data, and that SCG does not perform well in correcting FPs and FNs. In fact, all methods require a large number of SNVs to detect and correct FPs and FNs. And, as mentioned earlier, no other methods provided information on their performance in accurately imputing MBs and correcting FPs and FNs, so our results provide the knowledge of potential errors that each method may produce in actual empirical data analyses, which will be a guide to analyze the inferences of these methods for practitioners.

We have shown that, when the number of SNVs sampled is large, many methods also show good performance in detecting and correcting false positive and false negative mutation assignments. In our analyses, BEAM, SCITE and SiFit methods performed very well for datasets containing a small number of cells, both for small and large number of SNVs. These three methods employ molecular phylogenetics, but BEAM is based exclusively on classical molecular phylogenetic methods and applies a Bayesian framework to impute MBs and correct FPs and FNs by using the initial cell sequence phylogeny as a prior. In contrast, SCITE and SiFit employ a model that account for sequencing error rates in the process of inferring the evolutionary tree of cells, which also results in improved cell sequences (Jahn *et al.*, 2016; Zafar *et al.*, 2017). SiFit produces refined cell sequences by inferring the order of mutations along the inferred maximum likelihood cell phylogeny with given error rates, whereas SCITE uses a Bayesian method to search for a cell phylogeny that intrinsically maps mutations along branches in the phylogeny. On the other hand, BEAM does not determine the order of mutations and considers all possible ancestral base assignments to refine cell sequences.

As BEAM also uses a Bayesian method, we consider BEAM and SCITE to be in the same class of methods, with the difference that

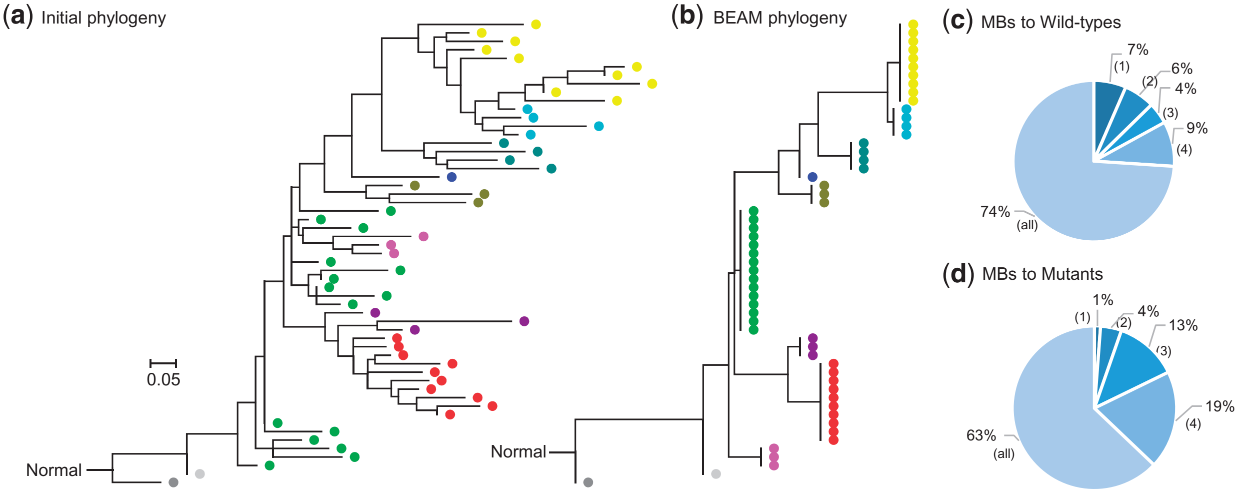

Fig. 11. Analysis of an empirical dataset. (**a**) The cell phylogeny produced by using the initial single-cell sequences from Li *et al.* (2012). (**b**) The cell phylogeny produced by BEAM after imputing MBs and correcting FPs and FNs. The proportions of imputed MBs that were assigned (**c**) a wild-type base and (**d**) a mutant base are shown, with different slices indicating the number of methods (in parentheses) that made the same assignment, including BEAM

SCITE needs to model sequencing error rates, but BEAM does not require this modelling. Also, BEAM decouples the inference of cell phylogenies from the corrections of observed single cell sequences, as compared to joint inference by SCITE. Interestingly, both of them show comparable results, which are among the best. That is, BEAM obviates the need to apply the same sequencing error rate model throughout the tree, which may be preferable because mutational patterns change in cells based on their evolved state (Frank and Nowak, 2004), potentially resulting in heterogeneity of error rate models among clones. An advantage of BEAM over SCITE may be that BEAM performs slightly better when the number of SNVs is large (Fig. 9), because the initial cell phylogeny derived using observed single-cell sequences is expected to be more accurate for datasets with a large number of SNVs.

Our results show that methods that incorporate the cell phylogeny are more powerful than others, especially when the number of cells per clone is small and the number of SNVs is large. This is because BEAM, SiFit and SCITE perform better than OncoNEM, which aims to infer clone phylogeny, and SCG, which employs a mixture model that identifies groups of cells with shared clone genotypes. Because tumor cells descend from pre-existing cells, there is evolutionary continuity in cell sequencing datasets, which enables computational methods to correctly impute missing data and make correct base assignments. We find that molecular evolutionary methods that have been successfully applied for species and strain phylogenetics for decades serve as a strong foundation for phylogenetic approaches with greater power to impute missing data and refine cell sequences for small datasets.

Acknowledgements

We thank Heather Rowe and Chivonne Matthews for comments and editorial support. We also thank Zachary Hanson-Hart for technical support.

Funding

Grants from Temple University and from the National Institutes of Health to S.K. (LM012487) and S.M. (LM012758) provided support for this research.

Conflict of Interest: none declared.

References

Brastianos,P.K. *et al.* (2015) Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov.*, **5**, 1164–1177.

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

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Francis,J.M. *et al.* (2014) EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov.*, **4**, 956–971.

Frank,S.A. and Nowak,M.A. (2004) Problems of somatic mutation and cancer. *Bioessays*, **26**, 291–299.

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.

Gawad,C. *et al.* (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.*, **17**, 175–188.

Gundem,G. *et al.* (2015) The evolutionary history of lethal metastatic prostate cancer. *Nature*, **520**, 353–357.

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

Huang,L. *et al.* (2015) Single-Cell Whole-Genome Amplification and Sequencing: methodology and Applications. *Annu. Rev. Genomics Hum. Genet.*, **16**, 79–102.

Hughes,A. *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.

Jan,M. *et al.* (2012) Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl. Med.*, **4**, 149ra118.

Kim,K.I. and Simon,R. (2014) Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics*, **15**, 27.

Leinonen,R. *et al.* (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,Y. *et al.* (2012) Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *Gigascience*, **1**, 12.

Liu,L. *et al.* (2016) A molecular evolutionary reference for the human variome. *Mol. Biol. Evol.*, **33**, 245–254.

McFadden,D.G. *et al.* (2014) Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell*, **156**, 1298–1311.

Melchor,L. *et al.* (2014) Single-cell genetic analysis reveals the composition of initiating clones and phylogenetic patterns of branching and parallel evolution in myeloma. *Leukemia*, **28**, 1705–1715.

Nassar,D. *et al.* (2015) Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nat. Med.*, **21**, 946–954.

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

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

Navin,N.E. (2015) The first five years of single-cell cancer genomics and beyond. *Genome Res.*, **25**, 1499–1507.

Nei,M. and Kumar,S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford, New York.

Nik-Zainal,S. *et al.* (2012) The life history of 21 breast cancers. *Cell*, **149**, 994–1007.

Ning,L. *et al.* (2014) Current challenges in the bioinformatics of single cell genomics. *Front. Oncol.*, **4**, 7.

Paguirigan,A.L. *et al.* (2015) Single-cell genotyping demonstrates complex clonal diversity in acute myeloid leukemia. *Sci. Transl. Med.*, **7**, 281re2.

Potter,N.E. *et al.* (2013) Single cell mutational profiling and clonal phylogeny in cancer. *Genome Res.*, **23**, 2115–2125.

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.

Sanborn,J.Z. *et al.* (2015) Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc. Natl. Acad. Sci. USA*, **112**, 10995–11000.

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

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Van Loo,P. and Voet,T. (2014) Single cell analysis of cancer genomes. *Curr. Opin. Genet. Dev.*, **24**, 82–91.

Wang,Y. and Navin,N.E. (2015) Advances and applications of single-cell sequencing technologies. *Mol. Cell.*, **58**, 598–609.

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

Xue,Y. *et al.* (2017) An approach to suppress the evolution of resistance in BRAFV600E-mutant cancer. *Nat. Med.*, **23**, 929–937.

Yachida,S. *et al.* (2010) Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, **467**, 1114–1117.

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.

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

Zafar,H. *et al.* (2016) Monovar: single-nucleotide variant detection in single cells. *Nat. Methods*, **13**, 505–507.

Zhao,Z.M. *et al.* (2016) Early and multiple origins of metastatic lineages within primary tumors. *Proc. Natl. Acad. Sci. USA*, **113**, 2140–2145.