Bootstrap confidence for molecular evolutionary estimates from tumor bulk sequencing data

Jared Huzar1, Madelyn Shenoy1, Maxwell D. Sanderford1, Sudhir Kumar1,2,3 and Sayaka Miura1,2*

1Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, United States, 2Department of Biology, Temple University, Philadelphia, PA, United States, 3Center for Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia

Bulk sequencing is commonly used to characterize the genetic diversity of cancer cell populations in tumors and the evolutionary relationships of cancer clones. However, bulk sequencing produces aggregate information on nucleotide variants and their sample frequencies, necessitating computational methods to predict distinct clone sequences and their frequencies within a sample. Interestingly, no methods are available to measure the statistical confidence in the variants assigned to inferred clones. We introduce a bootstrap resampling approach that combines clone prediction and statistical confidence calculation for every variant assignment. Analysis of computer-simulated datasets showed the bootstrap approach to work well in assessing the reliability of predicted clones as well downstream inferences using the predicted clones (e.g., mapping metastatic migration paths). We found that only a fraction of inferences have good bootstrap support, which means that many inferences are tentative for real data. Using the bootstrap approach, we analyzed empirical datasets from metastatic cancers and placed bootstrap confidence on the estimated number of mutations involved in cell migration events. We found that the numbers of driver mutations involved in metastatic cell migration events sourced from primary tumors are similar to those where metastatic tumors are the source of new metastases. So, mutations with driver potential seem to keep arising during metastasis. The bootstrap approach developed in this study is implemented in software available at https://github.com/SayakaMiura/CloneFinderPlus.

KEYWORDS
tumor evolution, bootstrap, bulk sequencing, metastasis, driver mutation

1 Introduction

Tumors are characterized by a mixture of cell populations in which many distinct clones arise due to somatic mutations (Gerlinger et al., 2012; Nik-Zainal et al., 2012; de Bruin et al., 2014; Zhao et al., 2016). These clones may increase in frequency during tumor progression, and they may spread to other locations resulting in metastasis (Gerlinger et al., 2012; Nik-Zainal et al., 2012; de Bruin et al., 2014; Zhao et al., 2016). Genetic variation in tumors is commonly profiled by bulk sequencing of tumor samples. In bulk sequencing, many cells in the sample are sequenced together to produce somatic variants and their population frequencies. This information informs the degree of genetic heterogeneity in tumors, but not the number of distinct clones present or the sequences of these clones. Knowledge of individual clone sequences is necessary to reconstruct the evolutionary relationship of tumor cells, the dynamics of mutational processes, and the history of metastatic cell migrations.
For this reason, several methods for analyzing bulk sequencing data are available (Beerenwinkel et al., 2015; Miura et al., 2020). Some methods are designed to identify clusters of genetic variants with similar variant allele frequencies (VAFs) indicative of their co-presence in the same genotype (Roth et al., 2014; Malikic et al., 2015; Popic et al., 2015; El-Kebir et al., 2018; Xiao et al., 2020). This strategy is commonly used when bulk sequencing data from only a single sample is available. More accurate clone predictions can be achieved when multiple tumor samples are sequenced from a patient, which enables the inference of clone genotypes and their evolutionary relationships (Murugaesu et al., 2015; Hao et al., 2016; Harbst et al., 2016; Reiter et al., 2017; Martinez et al., 2018; Miura et al., 2018; Hu et al., 2019).

Inferred clone sequences from bulk sequencing data are estimates. However, none of the current clone prediction methods provide an assessment of the uncertainty associated with these estimates. Uncertainty in clone inferences should occur because they are based on the similarities of VAFs that are calculated from observed sequencing reads with and without variants. Especially when the number of reads is small, the variance of VAF can be large (Figure 1). Thus, single nucleotide variants (SNVs) with small read counts are expected to affect the accuracy of clone prediction more strongly than those with large read counts. Unfortunately, all current methods comparing VAFs ignore this variance and simply present inferred clones and variation assignments without presenting the assignment variance. Here, we suggest using a bootstrap resampling approach to overcome this shortcoming. We have implemented this idea for use with the CloneFinder method (Miura et al., 2018) to demonstrate the usefulness of the bootstrap resampling in assessing the uncertainty of clone inference and embracing it in the downstream analysis such as the mapping of metastatic migration histories. We apply the bootstrap approach to analyze an empirical dataset, which yields insights into driver mutations and metastasis migrations.

2 Materials and methods

2.1 Bootstrap approach for tumor evolution estimates

Our bootstrap approach samples sequencing reads with and without variants at genomic positions (Figures 2A, B). The total number of reads sampled at a given position remains the same as in the original dataset, but reads are sampled with replacement. Since the same read can be sampled multiple times, some reads and even variants will be missing at that position in the bootstrap replicate dataset. Reads are resampled for each position, and a bootstrap replicate dataset is generated for each tumor sample. A pseudo-multi-tumor dataset is then generated by combining bootstrap replicates of tumor samples, and this pseudo-multi-tumor dataset is analyzed using the desired clone prediction method (e.g., CloneFinder) to infer clones (Figure 2C). Similarly, more pseudo-multi-tumor datasets are generated, and clones are inferred in many bootstrap replicates.

The bootstrap approach can be used to build consensus clone sequences. Predicted clones from all replicates are pooled together and (nearly) identical clone genotypes (for a user-supplied SNV count cutoff) are grouped (Figure 2D). Each clone group is then represented by a consensus clone and the proportion of bootstrap replicates in which a clone appears is the bootstrap support for detecting that consensus clone. To construct a consensus clone sequence, a base reconstructed in a greater proportion of bootstrap clones than the desired threshold is selected for each variant.
historical, where paths with the highest frequency are connected with each other first, followed by others with lower frequencies until the minimum desired bootstrap support requirement is met (Figure 2F).

2.2 Advanced CloneFinder (CloneFinder+)

We implemented and tested the bootstrap approach in our CloneFinder method which is known to perform well for inferring clones using bulk-sequencing datasets from multi-tumor samples (Miura et al., 2018; Miura et al., 2020). We also took this opportunity to advance CloneFinder by adding a step to preprocess the sequencing datasets by FastClone (Xiao et al., 2020) software that clusters SNVs based on VAFs. This allows CloneFinder to start with a larger collection of tumor genotypes than the original approach. Thus, the advanced CloneFinder, CloneFinder+, begins with the analysis of VAFs using FastClone (Figure 3A). For a tumor sample, FastClone clusters SNVs based on their VAF similarities and predicts relationships of SNV clusters, i.e., ancestor-descendant, sibling, or monoclonal (Figure 3B). CloneFinder+ constructs candidate clone sequences by accumulating all predicted SNVs from the root cluster to a target cluster. These clones are the candidate clones for a given tumor sample. This analysis is done for every tumor sample individually, and candidate clones are inferred for each tumor sample. All candidate clones from all the tumor samples are then pooled and duplicate clones are removed. Also, potentially spurious candidate clone sequences are filtered, e.g., those with many ambiguous base assignments (Supplementary Note for details). Lastly, a candidate clone sequence matrix, M, is constructed (Figure 3C). This is a binary matrix, where M[i,j] = 0/1 represents the absence/presence of a SNV at the jth variant in the ith candidate clone.

Next, we apply CloneFinder to infer clone genotypes (Miura et al., 2018) (Figures 3C–G). A phylogeny of candidate clones is first built using the matrix M. This phylogeny is used to identify missing ancestral clones that have persisted without being replaced by their descendant clones. Ancestral clone sequences are inferred at internal nodes of the phylogeny, and all unique ancestral clones are added to the collection of candidate clones, M (Figure 3E). Then, the presence of these candidate clones in the bulk-sequencing dataset is assessed using VAFs. A matrix of VAFs, V, is constructed, where the number of rows is equal to the number of tumor samples and the number of columns is equal to the number of SNVs (Figure 3D). Next, tumor frequencies of all candidate clones are estimated through regression analysis. Here, V, M, and f, a two-dimensional matrix of estimated clone frequencies of the tumor samples, have the following relationship,

\[ \frac{1}{2f} \times M = V. \]  

(1)

This relationship is valid only when variants are not affected by copy number alterations (CNAs). Thus, variants that are affected by CNAs need to be excluded from the dataset; or VAFs should be adjusted before the analysis using estimated cancer cell fraction (CCF), i.e., VAF is CCF divided by two. When an estimated clone frequency within a given tumor sample is lower than the desired threshold, that clone is assumed to be absent from the tumor sample. Candidate clones predicted to be absent from all tumor samples are removed.
Next, CloneFinder+ decomposes potential hybrid sequences using the strategy in CloneFinder (Miura et al., 2018) (Figure 3F). Briefly, SNVs with the same expected VAF are subclassified into two groups based on their similarity to observed VAFs. Alternate candidate clone sequences are constructed by combining SNV groups. All possible combinations of SNV groups are added to the M and are evaluated using the regression analysis (Eq. 1). All combinations estimated to be present in appreciable frequency are retained. Accordingly, M is updated, and this process of searching ancestral and hybrid clones is repeated until no new clone sequences are added. Lastly, CloneFinder+ finalizes the predicted clone sequences, where variants not assigned to any clone sequences are added to a clone based on their VAF similarities (Figure 3G).

2.3 Assembly of computer-simulated data

We obtained previously simulated datasets from https://github.com/raphael-group/machina. These datasets were generated by modeling the evolution of primary and metastatic tumors (El-Kebir et al., 2018). In this simulation, metastatic tumors were founded by one or more than one cancer cells that migrated from another tumor site (primary or another metastatic tumor). Clones were defined as a group of cells with the same sequences. Thus, a new clone could arise even from a single mutation. Each clone phylogeny was unique, and the number of clones was 6–26. Each dataset contained 9–99 SNVs. We excluded five datasets that contained tumor samples with only one variant because this is unrealistic in empirical data. In total, we analyzed 75 datasets.

2.4 Assembly of empirical data

We obtained metastatic cancer datasets from Zhao et al. (2016), who performed bulk sequencing (exome sequencing) and identified SNVs. For each SNV, these datasets contained total sequencing read counts and counts of reads with altered bases (SNVs). In total, we obtained 40 datasets with three to eight tumors. For each SNV, we predicted if it was a driver mutation using the Cancer Genome Interpreter (CGI) web tool (http://www.cancergenomeinterpreter.org), which uses the OncodriveMUT method (Tamborero et al., 2018). We also used the CRAVAT
web tool (http://www.cravat.us), which performs the CHASM prediction (Carter et al., 2009; Douville et al., 2013). Driver mutations were predicted without specifying a cancer type.

To map mutations at branches of the phylogeny, we analyzed predicted clone sequences and reconstructed ancestral clone sequences using MEGA (Kumar et al., 2012; Tamura et al., 2021).

2.5 Data analysis with the bootstrap approach

Each simulated and empirical dataset was analyzed using CloneFinder+, and the reliability of the inferences was assessed using the bootstrap approach. In the bootstrap analysis, we generated 30 replicates because using more (100) replicates essentially produced the same result (Supplementary Figure S1). In the CloneFinder+ analysis, we clustered variants without giving the tumor purity (the step of FastClone analysis) and used variants with at least 50 reference read counts and two mutant read counts to assess the quality of candidate clones. Note that CloneFinder+ does not require the value of tumor purity and the maximum number of clones to be inferred. During the analysis, candidate clone genotypes with <1% clone frequencies for all tumor samples were discarded. We grouped identical bootstrap clones to derive consensus clone sequences while allowing at most one base assignment difference. We selected a base in >90% of bootstrap clones for each variant position.

We inferred cell migration history using the PathFinder method (Kumar et al., 2020), and the reliability was assessed using the bootstrap approach. PathFinder was performed by providing the correct primary tumor sites, sequences of CloneFinder+ clones that were predicted with >5% clone frequencies, and tumor sites that contained each clone. Note that PathFinder does not require the value of tumor purity.

2.6 Accuracy measurements

To evaluate the accuracy in inferring correct clones, we paired each simulated clone sequence with its most similar inferred clone sequence. We allowed an inferred clone to be paired with more than one simulated clone. We counted the number of sequence differences between inferred and simulated clones paired. We calculated the average when more than one inferred clone was paired with a given simulated clone, which was divided by the sequence length to estimate genotype error (GE) for a given simulated clone.

To evaluate the accuracy of inferred migration history, we counted the number of migration paths that were correctly inferred, those not identified, and incorrect paths following a previous study (Kumar et al., 2020).
3 Results

3.1 Bootstrap confidence for predicted clones

The bootstrap approach is widely employed in molecular evolutionary and phylogenetic analyses to estimate variances and confidence limits (Efron and Tibshirani, 1994; Nei and Kumar, 2000). Our bootstrap approach samples sequencing reads. So, VAFs will be perturbed, with greater perturbations experienced by VAFs computed from a small number of reads (Figure 1). Since the number of reads is often highly variable among SNVs in empirical data [e.g., 8–1,436 for ATP401 from Zhao et al. (2016)], the variances of their VAFs are also variable. We first show the GEs of inferred clones and their bootstrap supports.

We found that the median GE of the clones with higher bootstrap supports (>50%) was zero (Figure 4A). Therefore, clones with >50% bootstrap support can be considered to be reliably inferred. We also found that many bootstrap consensus clones (77%) received low bootstrap supports (≤10%), and these clones contained many genotype errors (GE >5%; Figure 4A). Interestingly, these clones had a wide range of GEs, even overlapping with those from higher bootstrap supports.
This pattern was consistent with previous studies (Miura et al., 2020) indicating that inferred clone sequences were slightly more accurate for data with monoclonal seeding, i.e., lower intra-tumor heterogeneity of tumors from a primary or a metastatic tumor. We classified the datasets into (1) those with metastatic tumors that received a single seeding clone from another tumor site, i.e., monoclonal seeding, and (2) those with more than one seeding event, i.e., polyclonal seeding. Thus, intra-tumor heterogeneity of tumors from polyclonal seeding is higher, while inter-tumor heterogeneity among tumor sites is smaller than those with monoclonal seeding.

We found that the bootstrap approach performed well on both types of datasets, as inferred clones with greater bootstrap confidence tended to have more accurately inferred clone sequences (lower GE) (Figures 4B, C). We also found that GE tended to be slightly better for datasets with monoclonal seeding, indicating that inferred clone sequences were slightly more accurate for data with monoclonal seeding, i.e., lower intra-tumor heterogeneity with higher-inter tumor heterogeneity. This pattern was consistent with previous studies (Miura et al., 2020).

We also tested the impact of the number of tumors on the performance of the bootstrap approach. These simulated datasets were generated by modeling the evolution of primary and metastatic tumors, where metastatic tumors were seeded by clones that migrated from either a primary or a metastatic tumor. We classified the datasets into (1) those with metastatic tumors that received a single seeding clone from another tumor site, i.e., monoclonal seeding, and (2) those with more than one seeding event, i.e., polyclonal seeding. Thus, intra-tumor heterogeneity of tumors from polyclonal seeding is higher, while inter-tumor heterogeneity among tumor sites is smaller than those with monoclonal seeding.

We found that the bootstrap approach performed well on both types of datasets, as inferred clones with greater bootstrap confidence tended to have more accurately inferred clone sequences (lower GE) (Figures 4B, C). We also found that GE tended to be slightly better for datasets with monoclonal seeding, indicating that inferred clone sequences were slightly more accurate for data with monoclonal seeding, i.e., lower intra-tumor heterogeneity with higher-inter tumor heterogeneity. This pattern was consistent with previous studies (Miura et al., 2020).

We also tested the impact of the number of tumors on the performance of the bootstrap approach. These simulated datasets were generated by modeling the evolution of primary and metastatic tumors, where metastatic tumors were seeded by clones that migrated from either a primary or a metastatic tumor. We classified the datasets into (1) those with metastatic tumors that received a single seeding clone from another tumor site, i.e., monoclonal seeding, and (2) those with more than one seeding event, i.e., polyclonal seeding. Thus, intra-tumor heterogeneity of tumors from polyclonal seeding is higher, while inter-tumor heterogeneity among tumor sites is smaller than those with monoclonal seeding.

We found that the bootstrap approach performed well on both types of datasets, as inferred clones with greater bootstrap confidence tended to have more accurately inferred clone sequences (lower GE) (Figures 4B, C). We also found that GE tended to be slightly better for datasets with monoclonal seeding, indicating that inferred clone sequences were slightly more accurate for data with monoclonal seeding, i.e., lower intra-tumor heterogeneity with higher-inter tumor heterogeneity. This pattern was consistent with previous studies (Miura et al., 2020).

We next tested if different scenarios of the cell migration history affected the performance of the bootstrap approach. These simulated datasets were generated by modeling the evolution of primary and metastatic tumors, where metastatic tumors were seeded by clones that migrated from either a primary or a metastatic tumor. We classified the datasets into (1) those with metastatic tumors that received a single seeding clone from another tumor site, i.e., monoclonal seeding, and (2) those with more than one seeding event, i.e., polyclonal seeding. Thus, intra-tumor heterogeneity of tumors from polyclonal seeding is higher, while inter-tumor heterogeneity among tumor sites is smaller than those with monoclonal seeding.

We found that the bootstrap approach performed well on both types of datasets, as inferred clones with greater bootstrap confidence tended to have more accurately inferred clone sequences (lower GE) (Figures 4B, C). We also found that GE tended to be slightly better for datasets with monoclonal seeding, indicating that inferred clone sequences were slightly more accurate for data with monoclonal seeding, i.e., lower intra-tumor heterogeneity with higher-inter tumor heterogeneity. This pattern was consistent with previous studies (Miura et al., 2020).

We also tested the impact of the number of tumors on the performance of the bootstrap approach. These simulated datasets were generated by modeling the evolution of primary and metastatic tumors, where metastatic tumors were seeded by clones that migrated from either a primary or a metastatic tumor. We classified the datasets into (1) those with metastatic tumors that received a single seeding clone from another tumor site, i.e., monoclonal seeding, and (2) those with more than one seeding event, i.e., polyclonal seeding. Thus, intra-tumor heterogeneity of tumors from polyclonal seeding is higher, while inter-tumor heterogeneity among tumor sites is smaller than those with monoclonal seeding.

We found that the bootstrap approach performed well on both types of datasets, as inferred clones with greater bootstrap confidence tended to have more accurately inferred clone sequences (lower GE) (Figures 4B, C). We also found that GE tended to be slightly better for datasets with monoclonal seeding, indicating that inferred clone sequences were slightly more accurate for data with monoclonal seeding, i.e., lower intra-tumor heterogeneity with higher-inter tumor heterogeneity. This pattern was consistent with previous studies (Miura et al., 2020).

Next, we tested the performance of the bootstrap approach using 40 empirical datasets from various cancer types. As observed for the simulation study, most of the bootstrap consensus clones had low bootstrap supports (<10%), and few clones were identified with good (>50%) bootstrap supports (Figure 5A). We did not find a clear association between cancer types and bootstrap support values, as the distribution of bootstrap values was similar among different cancer types (Figure 5B). We also did not observe an association between tumor mutation burden (number of mutations) and bootstrap values for any cancer types (Figures 5C–F). Therefore, the performance of the bootstrap approach was not affected by either cancer type or tumor mutation burden.

3.3 Bootstrap confidence for inferred cell migration histories

Since the bootstrap approach performed well to place a confidence limit on inferred clones, we next tested if the bootstrap approach is also useful to assess the reliability of a downstream inference of predicted clones. As an example of a downstream analysis of predicted clones, we inferred metastatic cell migration histories using the same simulated datasets.

We found that correct paths often had a high bootstrap support (a median bootstrap support = 86.5%), while incorrect paths tended to have low bootstrap supports (a median of 10%) (Figure 6). However, as observed in the analysis of inferred clone sequences, a few correct paths were not well supported. Actually, bootstrap support for correct paths varied considerably, indicating that these paths were not repeatedly found in many bootstrap replicate datasets. Thus, some migration paths are difficult to reconstruct, which is consistent with previous findings (Kumar et al., 2020). Overall, these results indicated that the reliability of inferred migration path from predicted clones could be assessed using the bootstrap approach.

3.4 Patterns of metastatic cell migrations and driver mutation occurrences

Since point estimates of clones and cell migration paths have limited accuracy, placing bootstrap confidence on inferences is essential in real empirical data analysis. As an example, we illustrate empirical data analysis for revealing the patterns of metastatic cell migrations and driver mutation occurrences. In this analysis, we excluded eight datasets without primary tumors because the information on the primary tumor site was necessary for the inference of migration history.
First, we show the result of a lung cancer patient with three metastatic tumors in the heart, liver, and GEJ (ATP401 patient). The inferred clone phylogeny using CloneFinder+ (without bootstrap reliability assessment) indicated that clone C1 originated from the root clone (the most recent common ancestor of all clones), and a metastatic tumor in the liver contained this clone (Figure 7A). The PathFinder analysis (without bootstrap reliability assessment) using these predicted clones and phylogeny produced by CloneFinder+ clones without the bootstrap assessment. The numbers of drivers and total mutations are shown for each migration path. The primary and metastatic tumors are shown in blue and red boxes, respectively. (C) Bootstrap consensus migration history. The number along a path is bootstrap support (%). Dotted arrows indicate paths with <40% bootstrap support. (D) Driver mutation count and (E) driver mutation rates were compared between the paths originating from primary and metastatic tumors. The p values were computed using t-test. ATP401 patient was used. CGI was used for driver mutation prediction.

FIGURE 7
Clone phylogeny and metastatic cell migration history of a lung cancer patient. (A) Inferred clone phylogeny using CloneFinder+ without the bootstrap assessment. Grey circles represent tumor clones, and their predicted tumor sites (>0% clone frequency) are shown within boxes below the clone IDs. Tumor sites shown at internal nodes are predicted sites by PathFinder. Letters along branches are the branch ID, and branches are colored based on predicted tumor sites. All mutations are mapped at branches of the phylogeny through ancestral sequence reconstruction. When a cell migration event is inferred at a branch, the number of drivers and total mutations are shown. (B) Inferred cell migration history by PathFinder using CloneFinder+ clones without the bootstrap assessment. The primary and metastatic tumors are shown in blue and red boxes, respectively. (C) Bootstrap consensus migration history. The number along a path is bootstrap support (%). Dotted arrows indicate paths with <40% bootstrap support. (D) Driver mutation count and (E) driver mutation rates were compared between the paths originating from primary and metastatic tumors. The p values were computed using t-test. ATP401 patient was used. CGI was used for driver mutation prediction.
those from the primary tumor than those from a metastatic tumor ($p < 0.01$; t-test) (Figure 7D). Since migration paths with a larger number of associated mutations may simply result in a larger number of driver mutations, we normalized the count of driver mutation, i.e., we computed the driver mutation rate by dividing it by the total number of mutations for a path. We excluded paths without any associated mutations. Similarly, the driver mutation rate was significantly greater for those from the primary tumor than those from a metastatic tumor ($p < 0.01$ by t-test) (Figure 7E). Therefore, driver mutations occurred more frequently at migrations from the primary tumor than those from metastatic tumors for this patient.

To test if most of the patients similarly had higher driver mutation rates for migration paths from the primary tumor than metastatic tumors, we analyzed 32 datasets of metastatic cancer patients. Similar to the ATP401 patient, many paths were from metastatic tumors (Figure 8A), indicating that migration events from metastatic tumors were not rare, consistent with previous studies (Kumar et al., 2020; Chroni et al., 2022). We found that only six patients showed significant differences in driver mutation rates between those from primary tumors and from metastatic tumors ($p < 0.01$ by t-test for both driver prediction methods), and all of them had higher driver mutation rates for paths from the primary tumors, which was similar to the ATP401 patient (Figure 8B). However, most of the patients (26) did not show a significant difference in driver mutation rates between migration paths from primary and metastatic tumors ($p > 0.01$ by t-test for at least one driver prediction method). Therefore, the numbers of driver mutations were often not significantly different between metastatic cell migration events sourced from primary tumors and those from metastatic tumors.
4 Discussion

In this study, we showed the potential of bootstrap resampling procedures to place confidence limits on estimates obtained from tumor sequencing data. We found that incorrect inferences tended to receive low bootstrap support. Overall, the bootstrap approach performed well to distinguish spurious inferences.

Although the primary usage of the bootstrap approach is to place a confidence limit on inferred clones and downstream analysis (e.g., cell migration inferences), consensus clone sequences and consensus cell migration history can be also built by aggregating all bootstrap inferences. It is important to note that consensus inferences are not expected to be extensively more accurate than the point estimates, because the analysis of bootstrap replicates should not repeatedly produce correct inferences that are not found in the point estimate (Figure 9; Supplementary Figure S2).

In this study, we also developed CloneFinder+ by advancing CloneFinder, which now additionally analyzes the similarity of VAFs to cluster SNVs. The accuracy of CloneFinder+ was slightly better than the original version of CloneFinder (Figure 9). In conclusion, we showed that the bootstrap approach performed well to place a confidence limit on tumor evolution inference. The bootstrap approach can be coupled with any clone prediction method. Therefore, it will be useful for method developers to add a bootstrapping option.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

SM conceived the idea and supervised the project. JH and SM acquired and processed data. JH and SM developed methods and performed computational experiments. JH, SM, and MDS implemented the method. SM and MS performed empirical data analysis. All authors discussed the results. JH, SM, and SK wrote the manuscript. All authors read and approved the final manuscript.

Funding

This research was supported by the National Institutes of Health to SK (LM013385-03) and SM (LM014005-02).

Acknowledgments

We thank Dr. Antonia Chroni, Tracy Vu, Vivian Aly, Jiyeong Choi, and Nandita Nagendra for their technical support. Publication of this article was funded in part by the Temple University Libraries Open Access Publishing Fund.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2023.1090730/full#supplementary-material

References


Harbst, K., Lauss, M., Cirenajwis, H., Isaksson, K., Rosengren, F., Törngren, T., et al. (2016). Multiregion whole-exome sequencing uncovers the genetic evolution and


