Bioinformatics, 2019, 1–9
doi: 10.1093/bioinformatics/btz737
Advance Access Publication Date: 4 October 2019
Original Paper

Phylogenetics

Meltos: multi-sample tumor phylogeny reconstruction for structural variants

Camir Ricketts[1,2,†], Daniel Seidman[1,†], Victoria Popic[3], Fereydoun Hormozdiari[4], Serafim Batzoglou[3] and Iman Hajirasouliha[2,*]

[1]Tri-Institutional Training Program in Computational Biology & Medicine, New York, NY 10065, [2]Department of Physiology and Biophysics, Institute for Computational Biomedicine, Englander Institute for Precision Medicine, The Meyer Cancer Center, Weill Cornell Medicine of Cornell University, New York, NY 10021, [3]Department of Computer Science, Stanford University, Stanford, CA 94305 and [4]Department of Biochemistry and Molecular Medicine, MIND Institute and Genome Center, University of California, Davis, CA 95616, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that in their opinion, the first two authors should be regarded as joint First Authors.
Associate Editor: Alfonso Valencia

Received on December 15, 2018; revised on August 10, 2019; editorial decision on September 19, 2019; accepted on September 25, 2019

Abstract

Motivation: We propose Meltos, a novel computational framework to address the challenging problem of building tumor phylogeny trees using somatic structural variants (SVs) among multiple samples. Meltos leverages the tumor phylogeny tree built on somatic single nucleotide variants (SNVs) to identify high confidence SVs and produce a comprehensive tumor lineage tree, using a novel optimization formulation. While we do not assume the evolutionary progression of SVs is necessarily the same as SNVs, we show that a tumor phylogeny tree using high-quality somatic SNVs can act as a guide for calling and assigning somatic SVs on a tree. Meltos utilizes multiple genomic read signals for potential SV breakpoints in whole genome sequencing data and proposes a probabilistic formulation for estimating variant allele fractions (VAFs) of SV events.

Results: In order to assess the ability of Meltos to correctly refine SNV trees with SV information, we tested Meltos on two simulated datasets with five genomes in both. We also assessed Meltos on two real cancer datasets. We tested Meltos on multiple samples from a liposarcoma tumor and on a multi-sample breast cancer data (Yates *et al.*, 2015), where the authors provide validated structural variation events together with deep, targeted sequencing for a collection of somatic SNVs. We show Meltos has the ability to place high confidence validated SV calls on a refined tumor phylogeny tree. We also showed the flexibility of Meltos to either estimate VAFs directly from genomic data or to use copy number corrected estimates.

Availability and implementation: Meltos is available at https://github.com/ih-lab/Meltos.

Contact: imh2003@med.cornell.edu

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

1 Introduction

While many investigations into cancer-driving genomic variants focus on somatic single-nucleotide variants (SNVs) due to their relative ease of identification, somatic structural variants (SVs) have also been shown to be a driving force behind common cancers such as triple-negative breast (Kawazu *et al.*, 2017), small-cell lung (Govindan *et al.*, 2012), neuroblastoma (Pugh *et al.*, 2013), high-grade serous ovarian (Network, 2011), esophageal (Cheng *et al.*, 2016) and castration-resistant prostate cancers (Viswanathan *et al.*, 2018).

Unfortunately, accurate SV detection is a challenging task, especially in cancer datasets, where it is additionally confounded by *tumor heterogeneity*. Spatial heterogeneity in tumor samples often results in the need for multiple biopsies to more accurately characterize the diversity of somatic mutations. Indeed recent studies have shown that this heterogeneity can lead to ineffective patient treatments (Bozic and Nowak, 2014; Diaz *et al.*, 2012). For these reasons, it is imperative that any genome-wide analysis carried out on cancer samples be done with a holistic view of the tumor evolutionary history and with special attention given to the possibility of *subclonal* somatic variations. This has inspired efforts in presenting both SNVs and SVs as joint contributors to variation (Easton *et al.*, 2017). Both SNVs and SVs can contribute to malignant transformation of tumors and there have been examples of there being a high correlation between SVs and SNVs. Both classes of variation have also been shown to sometimes arise from

the same underlying mechanism such as homologous recombination deficiency in breast and ovarian cancer (Funnell *et al.*, 2018). These studies have suggested that incorporating both classes of variants can lead to more accurate results. Therefore, as personalized medicine begins to drive the shifting paradigm in cancer therapy, identifying actionable driver mutations in a robust way is highly important. Thus, leveraging multi-sample whole genome sequencing (WGS) to also identify somatic SVs as part of the whole spectrum of somatic mutations, is increasingly becoming valuable. Especially when considering that single cell sequencing presents the ideal potential for assessing heterogeneity but the sparsity of the data, allelic dropout and costs remain a significant limitation of this approach (Yuan *et al.*, 2017). Current tools that aim to cluster SNVs and SVs highlight the need for an approach that captures phylogenetic information and provide insights into joint intra-clonal evolution of SNVs and SVs (Cmero *et al.*, 2017; Easton *et al.*, 2017; Eaton *et al.*, 2018).

Additionally, characterizing early and late somatic mutations only present in subclones is integral to informing potential successful combination therapies for tumors and avoiding the selection of resistant subclones (Morrissy *et al.*, 2016; Wang *et al.*, 2016). The ability to detect mutations and subclones that occur at a low prevalence and frequency is also a significant issue in downstream variant analysis. While standard short-read sequencing platforms revolutionized our capacity to sequence whole genomes of tumor samples in recent years, identifying low prevalence somatic variations is still a very challenging task. This is particularly the case for structural variation discovery because of the fundamental limitations of standard short-reads (Alkan *et al.*, 2011). As a result, call sets produced by the state-of-the-art SV discovery methods such as (Hormozdiari *et al.*, 2009; Layer *et al.*, 2014; Rausch *et al.*, 2012) and their successors suffer from a high false discovery rate and predictions from different tools often do not agree with each other even in germline genomes (Alkan *et al.*, 2011). Indeed the somatic SV discovery problem in cancer genomes is more complex and challenging in the presence of tumor heterogeneity and complex rearrangements.

While the phylogenetic relationship between somatic SNVs has been utilized to study the sub-clonal structure of heterogeneous tumors, this same relationship among SVs has not been studied to date due to several additional challenges. In particular: (a) difficultly in detecting SVs and estimating their variant allele fractions (VAFs) (b) presence of simultaneous catastrophic events such as chromothripsis, as well as dramatic copy number changes or loss of heterozygosity (c) poor detection signals in low-coverage WGS.

Indeed in recent years, several algorithms for automatic reconstruction of cancer phylogeny trees has been developed by focusing on SNVs (Deshwar *et al.*, 2015; Donmez *et al.*, 2017; El-Kebir *et al.*, 2015; Hajirasouliha *et al.*, 2014; Jiao *et al.*, 2014; Malikic *et al.*, 2015; Marass *et al.*, 2016; Popic *et al.*, 2015; Satas and Raphael, 2017; Yuan *et al.*, 2015; Zare *et al.*, 2014). While a limited number of clonality studies on CNVs and SVs in tumor samples exist, they all either do not consider the structural changes (i.e. just use changes in copy number profiles) or do not utilize cancer phylogenetic relationships (Deshwar *et al.*, 2015; Eaton *et al.*, 2018; Ha *et al.*, 2014; McPherson *et al.*, 2017; Oesper *et al.*, 2014; Roth *et al.*, 2014).

Cancer phylogeny trees, which will also be referred to as lineage trees throughout this paper, are a means of organizing a series of mutations, in this case both SNVs and SVs, into a data structure that represents an order in which the mutations occurred within the cell lines of the samples in question. Each node in the tree contains a series of mutations that share the same presence profile across the samples, which means that these mutations were called in the exact same set of samples, along with similar variant allele frequencies.

These nodes are connected by directional edges, and the edges can only be formed if the nodes involved follow these set of rules that we refer to as evolutionary constraints and are mathematically presented in Equations (9), (10) and (11):

- A predecessor mutation cannot be present in a smaller subset of these samples.
- A mutation cannot have a VAF higher than that of its predecessor mutation (except due to CNVs).
- The sum of the VAFs of mutations cannot exceed the VAF of a common predecessor mutation.

In this study, we employ algorithmic and heuristic innovations to tackle the problem of reconstructing cancer phylogeny trees using SVs and address several challenging issues connected to this problem. In particular, we aim to utilize lineage trees built more reliably from somatic SNVs to learn about the evolution of SVs in a multiple-sample scenario. We also aim to use this a means of helping to reduce the current false discovery rate of existing somatic SV callers, especially for subclonal and relatively low-frequency SVs. This indeed allows us to advance our understanding of the entire landscape of somatic variations present within heterogeneous tumors. We attempt to cluster both somatic SNVs and SVs to explain the evolution of somatic variation and identify subpopulations of mutations that may be informative when designing targeted treatment for tumors. We do these developments using multi-region WGS data together with deep, targeted sequencing data when available, allowing us to more accurately assess the diverse genetic profile of tumors and provide more precise insights than those concluded using a single sample from the primary tumor.

In this work, we explore the hypothesis that many somatic SV events (e.g. midsize deletions, inversions, short interspersed nuclear elements insertions), similar to somatic SNVs, are the result of clonal evolution in cancer samples and correspond to tumor phylogeny. Thus, if we build the tumor lineage tree using high-quality SNVs, where potentially detection, VAF estimation and tree reconstruction is much easier, the tree can act as a guide for SV assignments. The approach of integrating SNV and SV signatures has been shown to be beneficial by other groups as well (Funnell *et al.*, 2018).

To that end, we present Meltos, a novel approach to estimate the variant allele frequency of somatic SVs from multi-region WGS signals and producing the most likely tumor lineage tree containing SNVs and SVs based on the data. Our probabilistic framework allows us to assess multiple types of signals taken from the data simultaneously and more accurately calculate the VAF of SV events while also allowing for the use of VAF and cancer cell fraction (CCF) estimates taken from other tools. Meltos uses a novel framework to assign SV clusters on the branches of the given lineage tree, while augmenting the tree topology if needed to allow for cases of independent evolution of SVs.

We tested Meltos on *simulated* datasets to ensure that Meltos is able to place SV events at the appropriate location in a lineage tree provided the existence of the ground truth. We also assessed the VAF estimation method through this process and used SVClone (Cmero *et al.*, 2017) generated VAFs as well to investigate Meltos assignments using corrected VAF estimates. To ensure the utility of our method in real data scenarios, we applied it to multi-region breast tumor samples obtained from patient PD9770 in a study by Yates *et al.* (Yates *et al.*, 2015). Tumor regions and matched normal samples were sequenced and SV calls were further validated by associated copy number changes and used as a measure of true positive and false positive calls in this individual (Section 3.2). We also applied *Meltos* on a multi-sample liposarcoma dataset collected and sequenced at Stanford University School of Medicine that is believed to have undergone chromothripsis. This dataset consists of seven spatially distinct samples on which variant calling was done.

2 Materials and methods

2.1 Overview

The Meltos algorithm is designed to augment a lineage tree that was built using SNV data with SVs from the same dataset, in order to leverage the higher confidence lineage tree made from SNVs as evidence for the presence of the SVs in question. To do this, the algorithm takes as input the BAM file of spatially distinct samples from which it extracts necessary read counts, as well as sets of SV calls and SNV calls from the same dataset, along with the preconstructed lineage tree on those same SNVs. It then filters the input for confidence based on read depth, and calculates variant allele

frequencies (VAFs) from the read data. From there, the distribution of SV and SNV VAFs are matched and clustered together. SVs that have similar VAFs to clusters of SNVs are assigned to the lineage tree where their presence profile allows. Then, clusters of leftover SVs are made into new potential nodes for the tree, and evolutionary constraints (Section 1) are used to see if there are any viable locations within the tree to insert the new nodes. This assignment process is summarized in Supplementary Figure S3. The underlying assumption being that there is a true underlying clonal structure of the tumor, of which a tree with only SNVs is just a subset. Therefore, the result of this approach is a more complete lineage tree with both SNVs and SVs represented within the phylogeny. Supplementary Figure S1 demonstrates the general workflow of our method.

2.2 Input and preprocessing

Given a collection of WGS samples, ideally obtained from several spatially distinct regions and a matched normal sample, an SV caller is used to identify a list of potential calls among these samples. For example, we can use an enhanced version of TARDIS (Soylev *et al.*, 2017), a standalone SV discovery tool capable of clustering signals for a wide range of SV event types, or LUMPY (Layer *et al.*, 2014) another widely-used SV calling tool. TARDIS is based on our previously published tools Variation Hunter/CommonLAW (Hormozdiari *et al.*, 2011), and can leverage multiple samples simultaneously. In addition to discordant read-pairs, it integrates split-read and read-depth signals to further improve the precision of SV calls.

Note that while we utilized TARDIS, other SV callers can be also used in this step. Some other notable SV callers with specific focus on somatic events include NovoBreak, a local assembly based tool (Chong *et al.*, 2017), or Weaver (Li *et al.*, 2016), an allele-specific quantifier of structural variations.

Somatic single nucleotide variants (SNVs) and their associated variant allele frequencies (VAF) are also obtained from the same dataset. A common strategy for obtaining a high-quality set of SNVs is performing extra targeted deep sequencing on the SNVs sites initially detected using WGS (Ding *et al.*, 2012). When additional deep sequence data are available, the VAF of SNVs can be estimated more robustly and with higher precision. Given a high quality set of SNVs and their estimated VAFs across different samples, we then build an accurate SNV tumor lineage tree where nodes correspond to clusters of SNVs. This step can be done using available methods such as LICHeE (Popic *et al.*, 2015). A list of potential SV events and their associated WGS signals together with a lineage tree built from SNVs, is the final input to our method, Meltos. A script for preparing input for Meltos in the correct format is also provided in Meltos repository.

2.3 VAF estimation for SVs

Meltos is designed as a multi-purpose tool. It first estimates the variant allele frequencies of SVs using the given information from whole-genome signals that are combined within a maximum likelihood formulation. The signals are obtained by quantification of reads after realignment in the region of a given SV event. Discordant read pairs, concordant read pairs, breakpoint read depth and split read information are all considered for an accurate estimation of the VAF of each SV.

This implementation extends the number of signals used for VAF calculation in order to provide better estimates than what we have seen with other tools such as Breakdown (Fan *et al.*, 2014). Breakdown uses maximum likelihood estimation approach but does not utilize breakpoint read depth in its estimate. Moreover, the utility of Breakdown is unfortunately limited to only large-scale deletion events and not other types of SVs.

In each sample i, we first estimate the variant allele frequency, $v_{i,j}$ of the candidate SV j from the whole genome sequence signals. We use a maximum likelihood approach, by first calculating the probability of seeing the SV j, given the observed number of supporting discordant paired-end reads $d_{i,j}$, the average observed depth of coverage $n_{i,j}(a)$ and $n_{i,j}(b)$, the observed number of concordant reads $c_{i,j}$ and the observed number of split reads $s_{i,\,j}$. We obtain $n_{i,j}(a)$ and $n_{i,j}(b)$ by the alignment of concordant reads over breakpoints of the candidate SV [similar to the method used in (Sindi *et al.*, 2012)]. Note that, for simplicity, we just consider somatic SV events in regions where polyploidy was not reported. We also expect that the vast majority of somatic events are indeed heterozygous. While this approach is still a simplified model (e.g. it does not account for multi-breakpoint overlapping events or complex SVs such as chromothripsis), we demonstrate its utility as a first step.

Let $P_{i,j,k}$ be the probability that the candidate SV j has a variant allele frequency equal to $k(0 \leq k \leq 0.5)$ in a given sample i. Assuming the well-defined Poisson distribution model form whole-genome sequencing data, we extend the approaches presented in (Fan *et al.*, 2014; Sindi *et al.*, 2012), in particular to utilize more signals taken from the data such that:

$$P_{i,j,k} = P(n_{i,j}(a)|k) \cdot P(n_{i,j}(b)|k) \cdot P(s_{i,j}|k) \cdot P(d_{i,j}|k) \cdot P(c_{i,j}|k) \quad (1)$$

where $n_{i,j}(a)$ and $n_{i,j}(b)$ are the number of reads covering left and right breakpoints of the candidate SV, respectively, $s_{i,j}$ is the number of split reads supporting the variant, $d_{i,j}$ is the number of discordantly mapped read pairs and $c_{i,j}$ is the number of normal read pairs in the candidate SV interval. Counts can be modeled using a Poisson distribution, across multiple sequencing libraries for each sample, so that the likelihood function for SV j in sample i having VAF k can be described as:

$$\text{Pois}(\lambda_n; n(a)) \cdot \text{Pois}(\lambda_n; n(b)) \cdot \text{Pois}(\lambda_s; s) \cdot \text{Pois}(\lambda_d; d) \cdot \text{Pois}(\lambda_c; c) \quad (2)$$

Each lambda parameter is estimated separately using the following functions, where t is the total number of reads, ℓ_{avg} is the average fragment length, k is the variant allele frequency, r is read length, g is the haploid genome length and w is the window size. It is important to note that GC correction is also implemented where necessary and the deepTools (Ramrez *et al.*, 2016) software package is used to create GC-corrected bams.

$$\lambda_n = \frac{t}{g} \cdot r \cdot (1-k) \quad (3)$$

$$\lambda_s = \frac{t}{g} \cdot 2r \cdot k \quad (4)$$

$$\lambda_d = \frac{t}{g} \cdot (\ell_{\text{avg}} - 2r) \cdot k \quad (5)$$

$$\lambda_c = \frac{t}{g} \cdot w \cdot (1-k) \quad (6)$$

Using a discrete list of possible values of k, we identify the value of k which maximizes the likelihood of what we observe in the data.

Once we have estimated the VAFs for the input SVs, we determine the mean and SD in each sample over all input SVs. We then do the same for the SNVs used to construct the lineage tree. We map the distribution of SNV VAFs onto the distribution of the SV VAFs by replacing the mean and SD of the SNV VAFs with that of the SVs. This is accomplished by treating each sample overall as a Gaussian distribution for SNVs and SVs. We modify the values of the SV VAFs so that they correspond to a value from the SNV distribution for the same sample with the same number of SDs from the mean as the SV had in its own distribution. We do this in order to make comparisons between the distribution of VAFs from SVs and SNVs possible, as the equations for each of the two types of VAFs are quite different.

$$\text{VAF}'_{i,\text{SV}} = \mu_{i,\text{SNV}} + \frac{(\text{VAF}_{i,\text{SV}} - \mu_{i,\text{SV}})}{\sigma_{i,\text{SV}}} \cdot \sigma_{i,\text{SNV}} \quad (7)$$

Where i is a particular sample from the input, μ represents the mean of the VAF values for sample i, either from the SVs or SNVs and σ is the SD.

2.4 SV assignment algorithm

We present our SV assignment technique in multiple steps below:

Step 1. Creation of presence profiles from SV VAFs

First, we create presence profiles for each SV in the input based on the estimated VAF for each SV in each sample. This is done by comparing each of the VAFs to a pair of experimentally determined thresholds. If the VAF in question is below the lower threshold, we assign an absence in that sample and if the VAF in question is above the upper threshold, we assign a presence in that sample. If it falls between the two, we assign an ambiguity to that presence profile, which is resolved in a later step. The set of all non-ambiguous SVs will be referred to as R and the set of all ambiguous SVs as A. These thresholds are tunable parameters and protect against miscalls by SV callers. These profiles define the subset of samples a variant occurs in where the underlying assumption is that a variant present is more samples is more likely to have occurred before those in less samples. The parameters govern the binary profile assigned to an SV and therefore can have an important effect on which clusters/nodes it is assigned to. We expect that it captures most of the topology of the true underlying evolutionary tree. Parameter values that result in too large of a range then subclonal SVs can be interpreted as clonal due to being falsely called as positive in some samples.

Step 2. Initial clustering

We take the VAFs from the centroids of the nodes in the input SNV lineage tree, pool them with the VAFs from R and we bin all of these values based on matching presence profiles. We cluster the VAFs within each of these bins with a Gaussian EM algorithm. Each of these clusters now has a corresponding Gaussian distribution as a bi-product of the clustering process, which is used to make future comparisons.

Step 3. Resolution of ambiguous SV profiles

We take the SVs from A and the SVs from R that formed a cluster with a number of members below a user-given threshold, call this union of sets U and we compare the SVs in U to the remaining clusters made from R, call this set of clusters C, using a heuristic based on the Gaussian distribution of the cluster from C, which approximates the probability that the distribution of that cluster could have produced the VAFs from the member of U in question: for a comparison between an SV $x \in U$ and an existing cluster $n \in C$ with sample set S:

$$\Delta(x,n) = \prod_{\forall i \in S} \frac{1}{2\pi n.\sigma_i^2} e^{\frac{-(x_i - n.\mu_i)^2}{2n.\sigma_i^2}} \tag{8}$$

If this probability is high enough, we modify the presence profile of x to match the profile of n. If there is no cluster for which x is sufficiently probable, x is omitted from the rest of the process and is not assigned to the final tree, as we have insufficient evidence to place it into the phylogeny.

Step 4. SV assignment to nodes

With the new presence profiles for SVs in place, we repeat the clustering process from Step 2, but the remaining SVs into ambiguous and non-ambiguous are not separated. We then iterate through the clusters produced by this method, and find the clusters containing centroids from the SNV lineage tree. All SVs that share a cluster with one of those centroids are considered sufficiently probable to have been produced by the same Gaussian distribution as the VAFs from the SNVs. We therefore conclude that they belong in the same node, and we place the SVs into the lineage tree in the node corresponding to that centroid. We then take the remaining clusters that did not contain one of the original tree's centroids, and we create tree nodes for all such clusters which have more members than the user-given threshold from Step 2. SVs still in clusters below this threshold are not considered for addition to the tree, and are omitted from further analysis.

Step 5. Determining possible node additions

We convert the remaining clusters into lineage tree nodes, preserving their mean and variance from the Gaussian process, as well as all SV assignments from the cluster. We then collect all of these nodes into a set, P.

Step 6. Producing valid tree modifications

$\forall p \in P$, we then find all the locations in the tree at where p could potentially be added to the tree by assigning p as a child of a node already present in the lineage tree, and reassigning a subset of that node's children as p's children by breaking edges in the existing tree and creating new edges. Let's call the total set of all edges needing to be broken and needing to be made to make a change to the tree E.

We confirm whether or not each of the edges that would be formed by enacting E meets the same evolutionary constraints used to create the original lineage tree. Suppose u and v are two possible nodes, one of which is already present in lineage tree T and the other is being investigated as a possible addition. The edge $u \rightarrow v$ is only possible if the two nodes satisfy the following three conditions $\forall i$:

$$u.\overline{\text{VAF}_i} \geq v.\overline{\text{VAF}_i} - \epsilon_{uv} \tag{9}$$

$$\text{if } u.\overline{\text{VAF}_i} = 0, \quad \text{then } v.\overline{\text{VAF}_i} = 0 \tag{10}$$

$$\sum_{c\, s.t.\, (u \rightarrow c) \in T} c.\overline{\text{VAF}_i} \leq u.\overline{\text{VAF}_i} + \epsilon \tag{11}$$

Where ϵ and ϵ_{uv} are experimentally determined allowed error parameters. If all such edges to be created in E meet all three constraints, we add that set of tree modifications to a set N.

Step 7. Assignment of heuristic values

$\forall E \in N$ we assign a value based on a summation of a heuristic for each edge made by enacting E, and a heuristic for each edge broken by enacting E. For a comparison between 2 nodes n and m with sample set S:

$$\Delta(m,n) = \prod_{i \in S} \frac{1}{2\pi n.\sigma_i^2} e^{\frac{-(m.\mu_i - n.\mu_i)^2}{2n.\sigma_i^2}} \prod_{i \in S} \frac{1}{2\pi m.\sigma_i^2} e^{\frac{-(n.\mu_i - m.\mu_i)^2}{2m.\sigma_i^2}} \tag{12}$$

This heuristic is based on the probability of the Gaussian distribution of each node on the edge producing the centroid of the other node. When considering all the edges that are made or broken in E, we create a heuristic value via the following formula:

$$h = \sum_{u,v\, s.t.\, (u \rightarrow v) \in M} \Delta(u,v) - \sum_{u,v\, s.t.\, (u \rightarrow v) \in B} \Delta(u,v) \tag{13}$$

where M is the set of edges to be made in the current change, and B is the set of edges to be removed from the tree in the current change. The values of the newly created edges in E are added to the total, while the values of edges broken in E are subtracted from the total. By attempting to maximize this heuristic, we are trying to make the changes to the tree which result in creating the most probable edges while breaking the least probable edges.

Step 8. Addition of new nodes

We find the set $E \in N$ with the maximum value generated by Equation (13), and now we make the modifications to the tree listed in it. We then remove all E in N that contained the node which was just added to the tree. Finally, we return to Step 6. It is important to note here that SV-only nodes are allowed by Meltos to be added to a tree in various locations as shown in Supplementary Figure S2. SV-only nodes contain, as the name suggests, no SNVs and only SVs that cluster together and cannot be placed in an existing SNV node.

3 Results

3.1 Experimental simulated data

3.1.1 Simulated dataset 1—SV assignment

We sought to assess Meltos' ability to make SV assignments using simulated genomes containing SNVs and SVs. We developed a new method based on the POMEGRANATE (https://github.com/viq854/pomegranate) algorithm that would allow us to carry out the initial cell population simulations with both SNVs and SVs. We utilize this new approach in our pipeline to generate simulated trees.

The algorithm creates a germline cell population containing no explicit mutations, and treats this as the root of the phylogenetic tree being generated. In a series of iterations, the algorithm chooses with a given probability whether or not a cell line mutates, thereby producing a child branch in the tree, or dies and is removed from the simulation. We create a baseline tree of cell populations containing simulated SNVs (Table 1) in this fashion. We then choose a subsample of cell populations in this tree to serve as the sample set for our later analyses, and we introduce known deletions given in the input into those samples. These deletion breakpoints were taken from TARDIS calls on a breast cancer dataset and used to propagate SVs down the tree from the initial cell population into any offspring cell populations, thus keeping the tree consistent with its own mutations. The mutations in these sample simulations, as well as the tree structure of the simulated phylogenetic tree, serves as the truth set for our later analyses.

We take those cell populations with lists of mutations, and produce a fasta file based on the hg19 reference genome, modifying variants at sites dictated by the simulation for SNVs and applying deletions according to the specifications in the input set of SVs by omitting sequence from the fasta. We use this process to create five simulated single-cell-line samples from the same lineage with an average of 33 SVs per sample (Table 1). We then use DWGSIM (https://github.com/nh13/DWGSIM) to generate reads with an average coverage of 50x from the fasta sequences modified with our simulated mutations. A fraction of reads cover each variant produced by the simulator and the resulting VAFs are also affected by variance in simulated coverage.

Table 1. Simulation statistics

Sample no.	# of SNVs	# of SVs placed	# of SVs recovered	% recovered
1	280	37	28	76
2	322	46	38	83
3	265	34	28	82
4	278	37	31	84
5	315	45	38	84

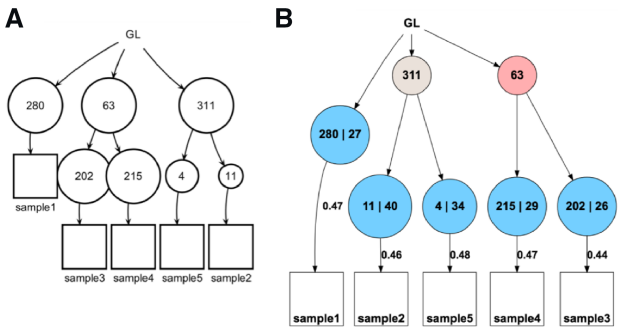

Fig. 1. (**A**) Representation of topology of the true tree for the five simulated genomes. (**B**) Meltos assignments of SVs in the simulated genomes. Nodes with numbers separated by a 'l' show the number of SNVs on the left and the SVs on the right and represent the only previously existing nodes affected by Meltos. GL—germline healthy cell

Finally, we align the generated reads to the reference genome using bwa (Li and Durbin, 2009) and create bam files using samtools (Li *et al.*, 2009). These serve as the input to TARDIS (Soylev *et al.*, 2017), which produces the SV calls we use to test our own algorithm.

We then compare the mutations successfully placed into nodes in Meltos output with the original truth set, and check to see if the tree structure is consistent. We then count the number of SVs that appeared in corresponding correct locations in the final tree based on their presence in our simulated truth set.

Meltos was able to successfully assign 156 of 185 SVs (84%) in simulated genomes (Fig. 1) to nodes that correspond to the nodes those mutations belonged to in the true tree from the simulation.

Initially, 199 deletions were generated in the five simulated genomes but 185 of the SVs were recovered TARDIS SV calling. These 185 were then used as input into the Meltos pipeline. Further refinement of the VAF estimation may be necessary in order to assign the 29 SVs that Meltos was not able to assign to the tree. CNVs were not simulated in this dataset so does not lead to variability in estimates.

It is important that Meltos is able to place a majority of SVs that fit within the evolutionary context of the tumor in the correct nodes. As a novel approach to this new problem, Meltos is clearly able to place a large number of SVs in the tree. One limitation of Meltos' approach is that these assignments are dependent on the accuracy of VAF estimates. It is important that Meltos is able to not only infer VAFs but also use VAF estimates from other tools. Variant allele frequency estimation for SVs is still a largely unexplored problem, SVClone (Cmero *et al.*, 2017) is one of the few tools outside of Meltos to tackle SV VAF estimation. SVClone (Cmero *et al.*, 2017) is an available tool which estimates variant allele frequency for SVs using a different approach. SVClone infers and clusters CCFs of SV breakpoints and is able to take into account purity, ploidy and copy-number information. It uses a decision tree to infer SV directionality and counts supporting and non-supporting reads utilizing a linear adjustment factor for purity where necessary. These counts are then used in calculating a VAF using a Bayesian Dirichlet Process mixture model that is implemented using Markov-Chain Monte-Carlo. This provided the opportunity to explore Meltos assignments with VAFs corrected for associated copy-number.

SVClone estimated VAFs were obtained for simulated SVs and used for assignment. Meltos was able to assign 167 of 185 SVs (90%). Therefore with both approaches Meltos is able to capture a vast majority of the SVs and their appropriate relation to the SNVs in the tree. Given that VAF estimation for SVs is still prone to incurring error, having the ability to fit this large majority of the SVs is a promising step.

3.1.2 Simulated dataset 2—Joint CNV and SV assignment

Using a similar simulation architecture to the previous, we simulate a set of 11 genomes with POMEGRANATE. We do this by creating a phylogenetic tree with random SNV mutations, and then inserting deletions and duplications into the tree based on positions and lengths taken from known SVs in an input dataset, as previously described in Section 3.1.1. The 11 genomes come from 5 randomly selected regions of that simulated phylogenetic tree, and each of the genomes in these regions is collected together into a simulated tumor sample. These sampled genomes work as approximations of related cell populations found in different regions of a tumor as the tumor's phylogeny progresses, and are meant to serve as a proxy for tumor samples of a real analysis. We created fasta files for each of these 11 genomes, sampled reads from each and collected the reads together in a ratio that matches the expected cell fraction ratio for each of the tumor samples, with greater than half of the reads of each sample taken from a mutation free fasta file reference genome to help simulate the heterozygosity of each of the tumor mutations. The result of this is the 5 genomes represented in the final simulated tree (Fig. 2).

We added functionality to POMEGRANATE so that it could simulate sequence amplification events to represent DNA-gain SCNVs and performed a series of analyses using multiple combinations of events called by TARDIS from the above simulated samples

as input to Meltos. This was done to observe Meltos' ability to place SVs in the presence of others types of alterations.

The ability of Meltos to refine the SNV tree with CNVs indicating gain or loss of DNA is a unique approach when compared with tools such as CANOPY (Jiang *et al.*, 2016) and SPRUCE (El-Kebir *et al.*, 2016). Especially when considering the complexity of the simulated tree and the number of events Meltos is able to place onto the tree. While both CANOPY and SPRUCE utilize less restrictive models at this point, Meltos is able to achieve >50% assignments (Table 2) in a highly complex phylogenetic context utilizing Gaussian clustering of estimated input VAFs of both SNVs and CNVs (Fig. 2).

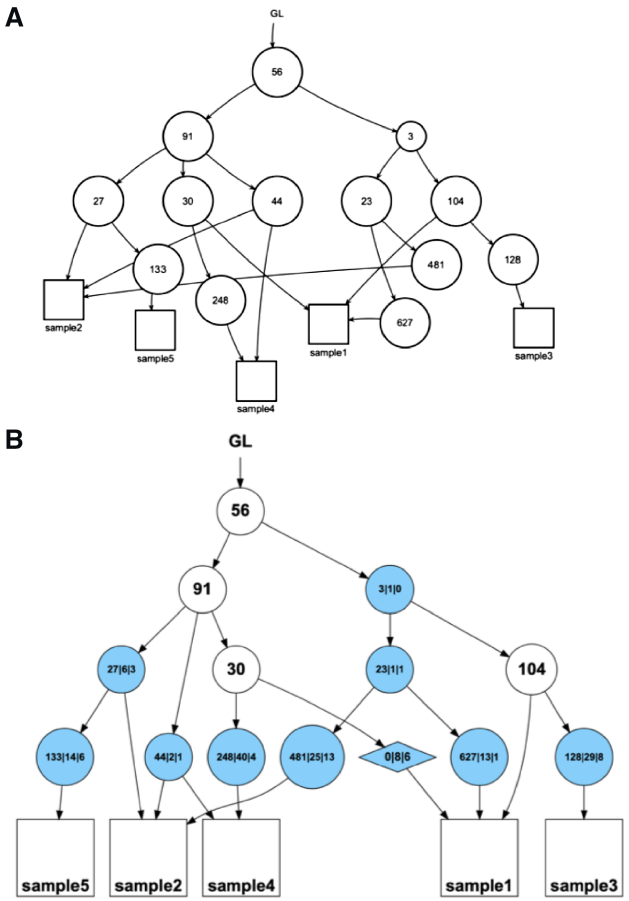

Fig. 2. (**A**) Representation of topology of the SNV tree for the five simulated genomes. (**B**) Meltos assignments of SVs including copy number variants in the simulated genomes. Nodes with numbers separated by 'l' contain SNVs (number on the left), SVs (middle) and DNA-gain CNVs (right). GL—germline healthy cell, Diamond node—SV only node which was added by Meltos and contains no SNVs. Corresponds to the mixed (profiled) run in Table 2 and Supplementary Table S1

Table 2 shows the results of our combinations of simulations. We experimented both with letting MELTOS itself handle its VAF estimates for mutations, and with SVClone performing that analysis instead and giving those VAFs to MELTOS instead. The tests consist of runs on deletions as input, runs with large duplication events as DNA-gain copy number alterations (min size of 1000 bp), and runs with combinations of both classes.

Meltos is dependent on the higher-confidence lineage tree given to it as input by way of LICHeE, and in cases where mutations may be misplaced due the inaccuracies in the SNV tree prior, it can cause other clusters from meeting their necessary phylogenetic constraints and decreases the ability of Meltos to make assignments. Supplementary Table S1 shows the contents of nodes that MELTOS created through clustering, but failed to place into the tree in Step 6 of the MELTOS pipeline due to their being no possible location where they meet all three perfect phylogeny constraints.

While combining the two classes of events saw a decrease in assignment percentage (Table 2), combining the two together resulted in a larger number of mutations being regarded as significant, and more SVs and duplications were placed into nodes. Some of these nodes were not assigned to tree, but with some additional refinement of the initial ssnv tree scaffold, it is possible that Meltos assignment of the combined events would be benefited as with MELTOS's clustering technique, all mutations provide additional signal to help support the validity of other, lower confidence mutations with similar VAFs.

3.2 Experimental real data

We evaluated Meltos on two different real datasets: (a) Three multi-region breast cancer whole genome samples from patient PD9770 in a large study (Yates *et al.*, 2015) were obtained, along with matched normal (approximately 40× sequence coverage). (b) Seven Illumina sequenced multi-region whole genomes taken from a chromothriptic liposarcoma and their matched normal sample at 35× coverage (Spies *et al.*, 2017), generated by our collaborators at Stanford University School of Medicine.

TARDIS was used to identify SVs within each sample and the matched normal. The candidate SVs found in samples from patient PD9770 consisted of an average of **2412 deletions, 141 inversions and 2206 mobile element insertions (MEIs)** among the tumor samples and 2181 deletions, 111 inversions and 943 MEIs in the matched normal (Supplementary Table S2). The candidate SV calls for the liposarcoma contained an average of **2128 deletions, 172 inversions and 5141 MEIs** among the tumor samples (Supplementary Table S3). In all cases, calls were filtered using a threshold of at least four pieces of evidence supporting the variant. As has come to be common among SV callers, the results can vary significantly depending on the caller and the signals they use. For instance, LUMPY identified 1703 deletions not previously characterized in the 1000 Genomes project and not called by TARDIS (Supplementary Fig. S6). Conversely, TARDIS identified 199 deletions not in either dataset. This trend continues in all other samples, exemplified by Supplementary Figure S6B and C and this varying degree of sensitivity allows for false positive calls to cause issues with accuracy of call sets.

Table 2. Meltos mutation assignments

Simulation test	Total tardis calls	True SV no.	# of deletions placed	# of duplications placed	% SVs recovered (%)	False positives
Deletions (no SVClone)	326	290	162	0	56	15
Deletions (profiled)	263	219	138	0	63	13
Duplications (profiled)	95	69	0	43	62	0
Mixed (profiled)	358	310	117	40	51	25

Notes: Table of how many of the simulated deletions and duplications (amplifications) were assigned to the tree during each of the tests run during the second Meltos simulation. 'Deletions' tests were exclusively run with simulated deletions. 'Duplications' tests were exclusively run with simulated copy number amplifications. 'Mixed' runs contained both deletions and duplication events. Rows with 'No SVClone' allowed Meltos to calculate its own VAF values. 'Profiled' runs utilized SVClone with the appropriate presence profiles.

Table 3. SV tree assignments–PD9770

Deletion size	Existing node	Valid SV-only node	Invalid SV-only node
50–500 bp	6	2	1
501–5000 bp	5	5	1
>5000 bp	8	15	3

Meltos is primarily devised to identify the evolutionary context for somatic SVs. This is done not only to identify genetic differences between subpopulations within the tumor but also presents a possible technique through which falsely called somatic events can be identified through their lack of fit within the established evolutionary context of the tumor based on our assumptions.

In samples obtained from PD9770 (Yates *et al.*, 2015), 160 high-quality somatic SNVs were detected. To build a representative tumor lineage for this tumor, we first used LICHeE (Popic *et al.*, 2015) on this data. These somatic SNVs clustered into five nodes: node 1 with 43 SNVs, node 2 with 51 SNVs, node 3 with 18 SNVs, node 4 with 33 SNVs and node 5 with 15 SNVs (Fig. 3). The Meltos workflow was then applied to utilize this tree as an evolutionary prior (Supplementary Fig. S1).

3.2.1 SV assignment in patient PD9770

3.2.1.1 Applying the Meltos workflow. Yates *et al.* (Yates *et al.*, 2015) validated 86 deletions using breakpoint associated copy number changes within the WGS data. These regions were compared with somatic deletions identified by TARDIS within the tumor samples (samples A, C and D). Sixty-three deletions called by TARDIS were found to be present in these validated regions and represented a true somatic call set.

Meltos was then applied to this call set in order to assess its ability to assign these SVs (min size: 102 bp, max size: 31 kb) to the inferred lineage tree inferred from SNVs (Fig. 3A). Meltos estimated the variant allele frequency of each SV in each sample it was found to be present. A majority of these deletions were found to have VAFs $\leq$ 0.5 (Fig. 4). This aligns with the assumption that most somatic deletions tumors are indeed *heterozygous*. The Meltos algorithm was able to successfully place 41 of the 63 validated deletions onto the tree alongside SNVs and represents a unique step towards explaining SVs and SNVs together (Table 3). Nineteen SVs were assigned to existing nodes while twenty-two SVs were placed into two SV-only nodes that represent new additions to the tree. This approach is also meant to discern false positive calls because these calls are less likely to fit within the evolutionary context provided by the lineage tree. Yates *et al.* were able to identify 63 deletion regions in their own calls that were unvalidated when considering copy number changes. These were considered false positive calls. Using this set of unvalidated calls along with visual inspection of the deletions in the Integrative Genomics Viewer (http://software.broadinstitute.org/software/igv/), we also identified 14 TARDIS calls deemed false positives.

These 14 calls were combined with the 63 true positive calls and provided as input to Meltos. This run once again yielded true positive assignments of **41/63** (**65**%) (Table 1) while only **2/14** (**14**%) false positive were assigned to the lineage tree. Both false positive clustered within SV only nodes and were not assigned to a previously existing node on the tree. By comparison, 14 other TARDIS calls were randomly selected and used to replace the false positives as input and Meltos was able to place 7/14 (50%) onto the tree. This highlights the ability of Meltos to assign true events onto the tree and signifies an important step towards comprehensively assessing the variety of mutations within a tumor and order in which they arose. As Meltos goes through further development, it will play an important role in assessing not only the quality of SV calls but how we can use this to explore the clonality of the tumor. Based on Figure 3, we were able to identify 12 SVs that emerged within the tumor genome with similar presence and VAF as 51 SNVs and 7 additional SVs which were acquired afterwards.

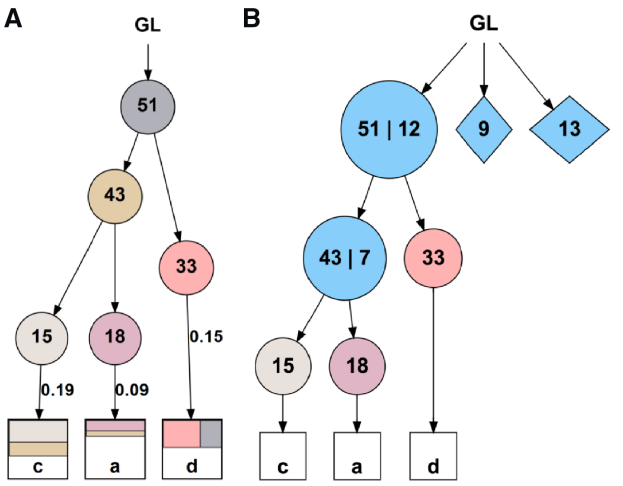

Fig. 3. (**A**) LICHeE lineage tree built from somatic SNVs used as a prior for our novel SV placement algorithm. (**B**) Meltos assignments of SVs to tree. Nodes with values separated by 'l' show the SNVs on the left and SVs on the right. *Some new nodes not linked to their associated samples to simplify diagram.* GL—germline healthy cell, Diamond node—SV only node which was added by Meltos and contains no SNVs

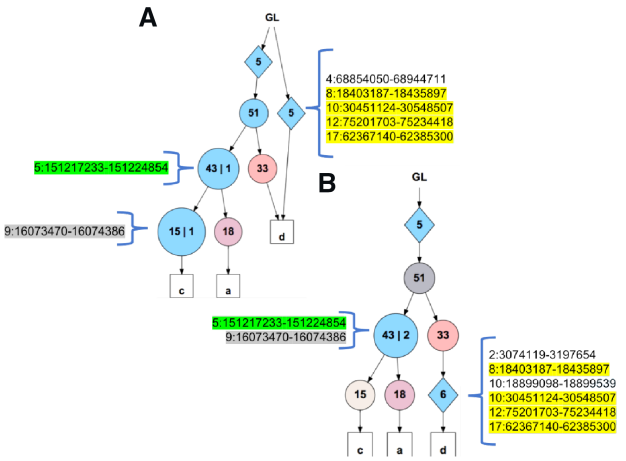

Fig. 4. (**A**) Assignment of SVs using Meltos VAF estimates. (**B**) Assignment of SVs using SVClone VAF estimate. Nodes with values separated by 'l' show the SNVs on the left and SVs on the right. *Some new nodes not linked to their associated samples to simplify diagram.* GL—germline healthy cell, Diamond node—SV only node which was added by Meltos and contains no SNVs

3.2.1.2 Exploring VAF estimation using SVClone. For each dataset, we estimate VAF using not only the VAF algorithm developed for Meltos but also using SVClone so as to explore the approaches to VAF estimation. This also allowed us to compare SV assignment to the tree using these sets of estimates.

In order to correct estimates for not only copy number but also tumor purity and ploidy for the Breast Cancer dataset, we ran TitanCNA (Yates and Campbell, 2012) in order to identify the optimal model that describes the ploidy of each multi-region sample within PD9770 tumor and also copy number segments to correct VAF estimates and reduce error (Table 4).

There is a positive correlation between Meltos and SVClone VAFs in each case although estimates do vary between both approaches. The spearman correlation coefficient between each set of estimates are as follows: PD9770A—0.9825753, PD9770C—0.8520243, PD9770D–0.8739152. The positive correlation signifies that SVs for which SVClone estimates a higher VAF, Meltos is likely

Table 4. TitanCNA estimates

Sample	Purity	Ploidy
PD9770A	0.4976	3.01
PD9770C	0.4952	3.059
PD9770D	0.3807	3.223

to estimate a higher VAF for such an SV as well and vice versa. In each case, Pearson correlation coefficients were lower—0.7497379, 0.4679937, 0.5722435—indicating that the difference between estimates were not consistent in value.

These results were not particularly surprising as the purpose of introducing SVClone into the pipeline presents an opportunity to correct VAFs for copy number and tumor purity before using Meltos to then do tree assignments of SVs onto the SNV tumor lineage tree.

To further investigate Meltos' ability to cluster SVs using copy number corrected SVClone VAFs, we selected 14 SVs with consistent estimated presence profiles across both Meltos and SVClone. Meltos assigned 12 of the 14 SVs (85%) to the tree when using Meltos estimated VAFs while 13 of 14 SVs (93%) were able to be placed onto the SNV lineage tree when using SVClone VAFs (Fig. 4). Meltos is only able to place an SV into the tree once it passes phylogenetic constraints. Using SVClone consistently allowed for SVs to fit within the lineage tree at lower levels where previously Meltos estimated VAFs only could fit these SVs as SV only nodes. An example of this is shown in Figure 4 where node containing SVs highlighted in yellow is able to be placed more correctly as a child to an existing node in the SNV tree. In Figure 4A, this SV-only node cannot pass constraints as a child to an existing node, so although it contains SVs that arose in Sample d, it is placed as a separate lineage to Sample d. This trend was consistent in our experiments as we increased the number of SVs being assigned to the tree. These results show that there is indeed value in using SVClone to estimate corrected VAFs that account for copy number in samples.

3.2.2 SV assignment in chromothriptic liposarcoma. For this dataset, SV calling was done on all seven samples with TARDIS and potential somatic deletions were identified by comparison with matched control. From this 40 deletions were randomly selected as input into Meltos to investigate the ability to of Meltos to assign SVs to a lineage tree built on relatively lower quality SNVs and a chromothriptic genetic evolutionary history.

As shown in Figure 5B, 19 of the 40 SVs were successfully placed onto the tree by Meltos. These SVs were compared with the GROC-SV calls that were published for this liposarcoma dataset (Spies *et al.*, 2017). Of the assigned calls, 11/19 (58%) overlap with GROC-SV calls while of the unassigned calls, 4/21 (19%) overlap with GROC-SV calls. However, in this case no SVs were able to cluster with a pre-existing node. This speaks to the reliance of Meltos on an accurate tree built on deeply sequenced SNVs. Note that inaccuracy in the SNV VAFs used to build the tree resulted in little space being left within node clusters to fit new additions/SVs based on the perfect phylogeny constraints. While this is believed to contribute to the observed assignments, it is worth noting that this observation is in agreement with observations made by Spies et al. (Spies *et al.*, 2017), where SVs were not found to cluster with SNVs in an evolutionary context for the same dataset.

4 Discussion

In this work, we showed that phylogenetic information from SNV trees increases our ability to identify true somatic SVs among multiple related samples and provide a more comprehensive view of the somatic mutation profile of tumors. We were able identify SVs in a breast cancer patient that appear to follow a similar evolutionary pattern as called SNVs as well as identify subclonal SVs that clustered into SV-only nodes and arose later in the evolution of the

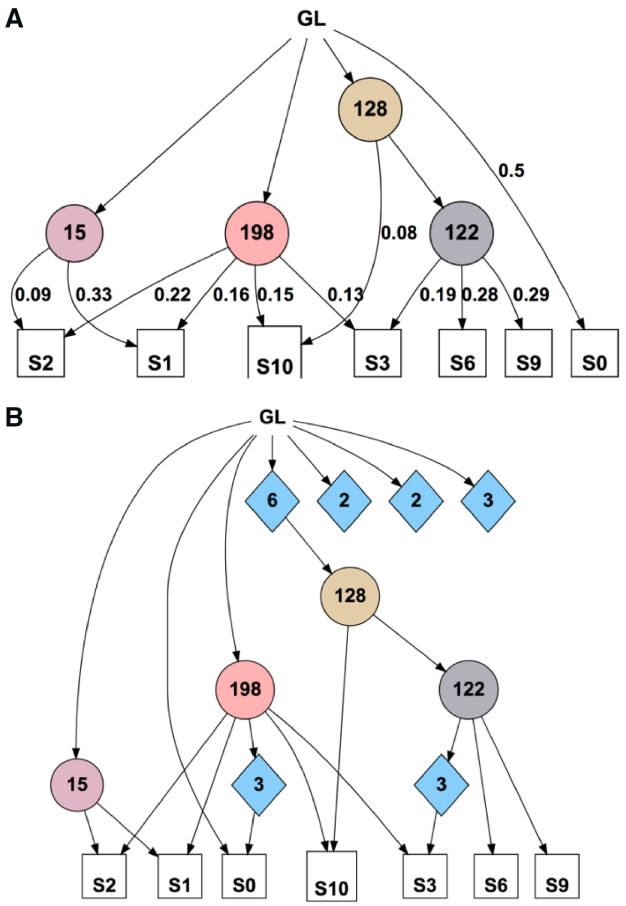

Fig. 5. (**A**) LICHeE lineage tree built from somatic SNVs for sarcoma. (**B**) Meltos assignments of SVs to tree. GL—germline healthy cell, Diamond node—SV only node which was added by Meltos and contains no SNVs

tumor. We also saw the tendency of true calls to get placed into the phylogeny tree over false, unvalidated calls. Signifying the important potential of this approach to building tumor phylogeny trees and filtering SVs.

Our model will indeed benefit from more robust and accurate somatic SV callers. While there is an on-going effort in the community to improve somatic callers both from technology and algorithms perspectives, there is still a lot of room for improvements. For example, integrative methods to handle data from different sequencing platforms simultaneously have the potential to further improve quality of call sets. While we used the GROC-SVs calls from linked-read data to assess tree assignments where applicable, we also identify that incorporating standard short-read and linked-read data together in a future work for reconstruction of lineage trees may also be beneficial and lead to increased accuracy. Furthermore, our model for VAF calculation is simple and a more sophisticated model will help better capture complexity of cancer genomes and determine whole cancer genomes.

For some SVs, Meltos was able to cluster them together in possible SV only clusters but was not able to assign these clusters to the tree based on the constraints. A majority of these were less than 500 bp in size and may be a result of difficulty in accurately estimating the VAF of these small SVs. Further work will aim to rectify these issues by building higher quality datasets through which we can tune the algorithm and utilize experimental validation of assigned SVs. However, we do expect there to be cases where the constraints of perfect phylogeny are violated due to the complexity of cancer genomes. For example, back mutations cannot be accounted for by Meltos currently due to the restrictions of its perfect phylogeny constraints. Cancers can exhibit loss of SNVs due to CNVs so Meltos attempts to correctly place copy number variants on the tree as well as employ other tools

for correction (El-Kebir, 2018). We showed the value of incorporating copy number correction through tools such as SVClone and using that along with Meltos for lineage tree refinement. We also highlighted the potential of this approach to aid in identifying false positive calls by using their fit within the evolutionary structure of the tumor as a filtering mechanism.

Acknowledgements

The authors acknowledge Arend Sidow for providing valuable biological insights. They also thank Nicole Lustgarten for great help in SV analysis and manual curation. I.H. acknowledges a Simons Fellowship in connection to the Algorithmic Challenges in Genomics program that facilitated early discussions on this work.

Funding

This work was supported by start-up funds (Weill Cornell Medicine) to I.H. C.R. and D.S. were also supported by the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant 1T32GM083937). This work was also supported in part by a US National Science Foundation (NSF) Award [IIS-1840275] and a US National Institute of Health (NIH) grant [R01CA183904].

Conflict of Interest: none declared.

References

Alkan,C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.

Bozic,I. and Nowak,M.A. (2014) Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers. *Proc. Natl. Acad. Sci. USA*, **111**, 15964–15968.

Cheng,C. *et al.* (2016) Whole-genome sequencing reveals diverse models of structural variations in esophageal squamous cell carcinoma. *Am. J. Hum. Genet.*, **98**, 256–274.

Chong,Z. *et al.* (2017) novobreak: local assembly for breakpoint detection in cancer genomes. *Nat. Methods*, **14**, 65–67.

Cmero,M. *et al.* (2017) Svclone: inferring structural variant cancer cell fraction. *bioRxiv*, 172486.

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

Diaz,L.A. *et al.* (2012) The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*, **486**, 537–540.

Ding,L. *et al.* (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**, 506.

Donmez,N. *et al.* (2017) Clonality inference from single tumor samples using low-coverage sequence data. *J. Comput. Biol.*, **24**, 515–523.

Easton,J. *et al.* (2017) Genome-wide segregation of single nucleotide and structural variants into single cancer cells. *BMC Genomics*, **18**, 906906.

Eaton,J. *et al.* (2018) Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics (Oxford, England)*, **34**, i357–i365.

El-Kebir,M. (2018) Sphyr: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics (Oxford, England)*, **34**, i671–i679.

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.

Fan,X. *et al.* (2014) Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC Bioinformatics*, **15**, 299.

Funnell,T. *et al.* (2018) Integrated single-nucleotide and structural variation signatures of DNA-repair deficient human cancers. *bioRxiv*, 267500.

Govindan,R. *et al.* (2012) Genomic landscape of non-small cell lung cancer in smokers and never smokers. *Cell*, **150**, 1121–1134.

Ha,G. *et al.* (2014) Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, **24**, 1881–1893.

Hajirasouliha,I. *et al.* (2014) A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, **30**, i78.

Hormozdiari,F. *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.*, **19**, 1270–1278.

Hormozdiari,F. *et al.* (2011) Simultaneous structural variation discovery among multiple paired-end sequenced genomes. *Genome Res.*, **21**, 2203–2212.

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.

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

Kawazu,M. *et al.* (2017) Integrative analysis of genomic alterations in triple-negative breast cancer in association with homologous recombination deficiency. *PLoS Genet.*, **13**, e1006853.

Layer,R.M. *et al.* (2014) Lumpy: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.

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 (sam) format and samtools. *Bioinformatics*, **25**, 2078–2079.

Li,Y. *et al.* (2016) Allele-specific quantification of structural variations in cancer genomes. *Cell Syst.*, **3**, 21–34.

Malikic,S. *et al.* (2015) Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics (Oxford, England)*, **31**, 1349–1356.

Marass,F. *et al.* (2016) A phylogenetic latent feature model for clonal deconvolution. *Ann. Appl. Stat.*, **10**, 2377–2404.

McPherson,A. *et al.* (2017) destruct: accurate rearrangement detection using breakpoint specific realignment. *bioRxiv*, p. 117523.

Morrissy,A.S. *et al.* (2016) Divergent clonal selection dominates medulloblastoma at recurrence. *Nature*, **529**, 351.

Network,T.C.G.A.R. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.

Oesper,L. *et al.* (2014) Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, **30**, 3532.

Popic,V. *et al.* (2015) Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.*, **16**, 91.

Pugh,T.J. *et al.* (2013) The genetic landscape of high-risk neuroblastoma. *Nat. Genet.*, **45**, 279–284.

Ramrez,F. *et al.* (2016) deeptools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.

Rausch,T. *et al.* (2012) Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333.

Roth,A. *et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396.

Satas,G. and Raphael,B.J. (2017) Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics*, **33**, i152–i160.

Sindi,S.S. *et al.* (2012) An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.*, **13**, R22.

Soylev,A. *et al.* (2017) Toolkit for automated and rapid discovery of structural variants. *Methods*, **129**, 3.

Spies,N. *et al.* (2017) Genome-wide reconstruction of complex structural variants using read clouds. *Nat. Methods*, **14**, 915–920.

Viswanathan,S.R. *et al.* (2018) Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing. *Cell*, **174**, 433–447.

Wang,J. *et al.* (2016) Clonal evolution of glioblastoma under therapy. *Nat. Genet.*, **48**, 768.

Yates,L.R. and Campbell,P.J. (2012) Evolution of the cancer genome. *Nat. Rev. Genet.*, **13**, 795–806.

Yates,L.R. *et al.* (2015) Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.*, **21**, 751–759.

Yuan,G.-C. *et al.* (2017) Challenges and emerging directions in single-cell analysis. *Genome Biol.*, **18**, 84.

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

Zare,H. *et al.* (2014) Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.*, **10**, e1003703–e1003715.