BIOINFORMATICS

Vol. 30 ISMB 2014, pages i78–i86
doi:10.1093/bioinformatics/btu284

A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data

Iman Hajirasouliha[1,2,†], Ahmad Mahmoody[1,†] and Benjamin J. Raphael[1,2,*]

[1]Department of Computer Science and [2]Center for Computational Molecular Biology, Brown University, Providence, RI, 02906, USA

ABSTRACT

Motivation: High-throughput sequencing of tumor samples has shown that most tumors exhibit extensive intra-tumor heterogeneity, with multiple subpopulations of tumor cells containing different somatic mutations. Recent studies have quantified this intra-tumor heterogeneity by clustering mutations into subpopulations according to the observed counts of DNA sequencing reads containing the variant allele. However, these clustering approaches do not consider that the population frequencies of different tumor subpopulations are correlated by their shared ancestry in the same population of cells.

Results: We introduce the *binary tree partition* (*BTP*), a novel combinatorial formulation of the problem of constructing the subpopulations of tumor cells from the variant allele frequencies of somatic mutations. We show that finding a BTP is an NP-complete problem; derive an approximation algorithm for an optimization version of the problem; and present a recursive algorithm to find a BTP with errors in the input. We show that the resulting algorithm outperforms existing clustering approaches on simulated and real sequencing data.

Availability and implementation: Python and MATLAB implementations of our method are available at http://compbio.cs.brown.edu/software/

Contact: braphael@cs.brown.edu

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

1 INTRODUCTION

Cancer is a disease driven by somatic mutations that accumulate in the genome during the lifetime of an individual. High-throughput sequencing technologies now provide an unprecedented ability to measure these somatic mutations in tumor samples (Ding *et al*., 2013). Application of these technologies to cohorts of cancer patients has revealed a number of new cancer-causing mutations and cancer genes (Kandoth *et al*., 2013; Lawrence *et al*., 2013; Vogelstein *et al*., 2013). Cancer sequencing studies have also demonstrated that most tumors exhibit extensive *intra-tumor heterogeneity* characterized by individual cells in the same tumor harboring different complements of somatic mutations (Ding *et al*., 2012; Gerlinger *et al*., 2012; Nik-Zainal *et al*., 2012; Schuh *et al*., 2012; Shah *et al*., 2012). Such heterogeneity is a consequence of the fact that cancer is an evolutionary process in a population of cells. The *clonal theory* of cancer evolution (Nowell, 1976) posits that the cells of a tumor descended from a single founder cell. This founder cell contained an advantageous mutation leading to a clonal expansion of a large population of cells descended from the founder. Subsequent clonal expansions occur as additional advantageous mutations accumulate in descendent cells. A sequenced tumor sample thus consists of multiple subpopulations of tumor cells from the most recent clonal expansions (Fig. 1).

Nearly all cancer sequencing efforts thus far sequence DNA from a *single* sample of a tumor at a *single* time. This is because of technical limitations: the most cost-effective DNA sequencing technologies (e.g. Illumina) require input DNA from many tumor cells, and such samples are typically available only when patients undergo surgery (Occasionally, paired samples from two time points, such as a primary tumor and a metastasis, are also sequenced.). While sequencing of multiple samples from the same tumor (Gerlinger *et al*., 2012; Newburger *et al*., 2013; Salari *et al*., 2013) or single-cell sequencing (Hou *et al*., 2012; Navin *et al*., 2011; Xu *et al*., 2012) might eventually provide even better datasets to assess intra-tumor heterogeneity, technical considerations have limited their applicability utility thus far. Thus, there is tremendous interest in methods that infer the relative proportion of different subpopulations of tumor cells in a single sample.

Several recent studies (Ding *et al*., 2012; Nik-Zainal *et al*., 2012; Shah *et al*., 2012) have demonstrated that it is possible to infer the subpopulations of tumor cells by counting the number of DNA sequence reads that contain a somatic mutation. For single-nucleotide mutations, or variants, the *variant allele frequency (VAF)* is defined as the fraction of DNA sequence reads covering the variant position that contains the variant allele rather than the reference/germ line allele. The VAF provides an estimate of the fraction of tumor chromosomes containing the mutation, but with error due to the stochastic nature of the sequencing process (Fig. 1). In addition, technologies currently used in cancer sequencing studies produce short reads that rarely contain more than one somatic mutation. Thus, for any pair of somatic mutations, the only information available to distinguish their subpopulation of origin is the VAF.

To overcome substantial variability in measured VAFs, a common approach is to cluster VAFs and from these clusters infer the number and proportion of various subpopulations of tumor cells in the sample. A number of techniques have been introduced to perform this clustering, with Dirichlet Process Mixture models and related non-parametric models being particularly popular, as they do not fix the number of clusters in advance (Miller *et al*., forthcoming; Nik-Zainal *et al*., 2012; Shah *et al*., 2012). VAF clusters correspond to tumor subpopulations, and the *cellular fraction*, or fraction of tumor cells containing the cluster of somatic mutation, is derived from the VAF of the clusters. In the simplest case, the VAF directly determines the cellular fraction: e.g. a cluster with VAF = 0.5 corresponding to homozygous mutations in 50% of tumor cells, or

*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.

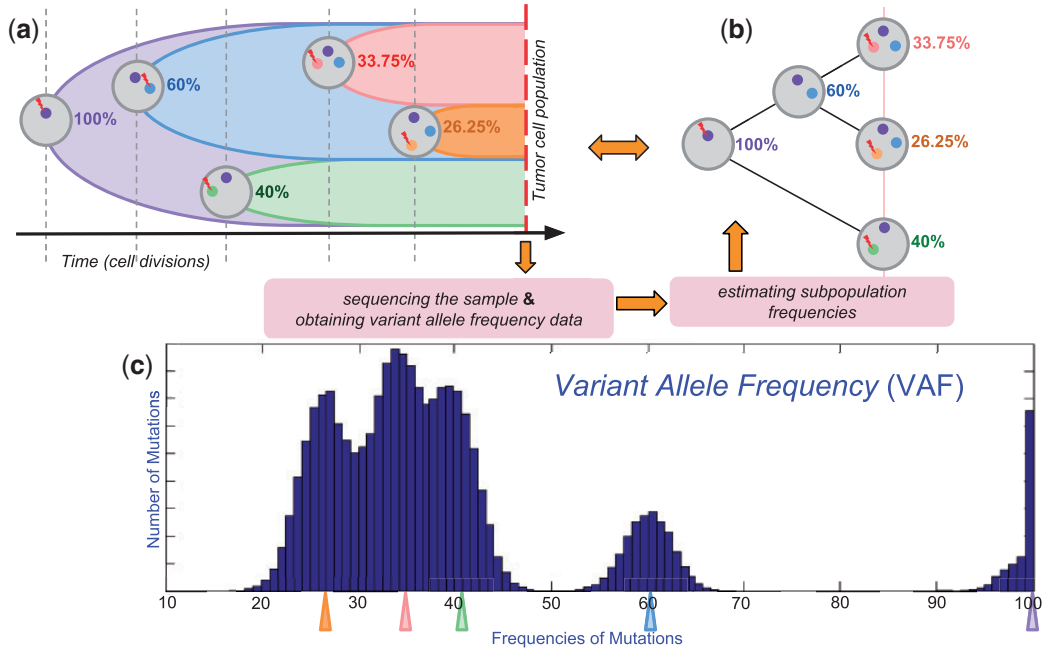

Fig. 1. (**a**) The cells in a tumor descend from a single founder cell via multiple waves of clonal expansion. Each circle represents a population, each dot corresponds to a mutation and the shaded sections indicate the cells descended from each founder cell of the clonal expansion. (**b**) Under mild assumptions, these clonal expansions give rise to a BTP, with nodes representing populations of tumor cells with specific subsets of somatic mutations. (**c**) The variant allele frequencies (VAFs) of somatic mutations are determined from sequencing data and used to infer tumor subpopulations and/or the BTP. Here the clusters correspond to clonal expansions, and center of each cluster estimates the frequency of the newly formed subpopulation (denoted with colored marker). Note that the size of each cluster depends on the number of mutations that accumulate before its expansion

heterozygous mutations in 100% of tumor cells. However, in practice, the inference of cellular fraction is complicated by copy number aberrations and normal admixture (percentage of the tumor sample that is normal cells), and these two factors are themselves correlated (Carter *et al.*, 2012; Oesper *et al.*, 2013; Strino *et al.*, 2013). We will not consider such complications here; rather we restrict attention to heterozygous somatic mutations outside of copy number aberrations; assumptions that were made in sequencing studies including Ding *et al.* (2012).

With two exceptions (Jiao *et al.*, 2014; Strino *et al.*, 2013), current techniques for clustering VAFs treat each subpopulation independently and do not consider that these frequencies are correlated by the fact that they are partitions of the same cellular population. Thus, these approaches do not explicitly construct an evolutionary history of the accumulated somatic mutations in the cells. Strino *et al.* (2013) performs a heuristic search over possible trees, and we discuss Jiao *et al.* (2014) below.

Contributions. In this article, we formulate the problem of inferring subpopulations of tumor cells from VAF data obtained from a single tumor sample as the combinatorial problem of constructing a *Binary Tree Partition (BTP)*. We show that the problem of finding a BTP is NP-complete and present a $\frac{2}{3} - o(1)$-approximation algorithm for a related max -BTP problem. The approximation algorithm is based on *Local Search* and we use the well-known packing bound of Hurkens and Schrijver (1989) for the purpose of analysis. Next, we define ε-BTP, a generalization of the BTP problem that allows for the possibility that VAFs are observed with errors and some VAFs are not observed. We present a straightforward recursive algorithm to find an ε-BTP and show that this algorithm outperforms existing VAF clustering approaches on simulated and real data. This recursive algorithm is fast in practice and runs in less than a minute, on a single CPU, for each of our simulated or real samples.

2 TUMOR SUBPOPULATIONS AND THE BTP

In this section, we formulate the problem of determining tumor subpopulations from VAF clusters. The clonal theory of cancer evolution proposes that the cancerous cells in a tumor are the result of multiple waves of somatic mutation and clonal expansion. Given the relationship between sequence coverage (1000–10 000× for targeted studies) and number of tumor cells that are sequenced in a tumor sample (millions), we first assume that any somatic mutation reported in the data must be present in an appreciable fraction of tumor cells. This implies that the observed somatic mutations were present in at least one clonal expansion. Second, as in other recent studies (Jiao *et al.*, 2014; Salari *et al.*, 2013), we assume that somatic mutations follow the infinite sites assumption such that at most one single mutation occurs at a genomic locus (e.g. single position) during the evolution of the tumor. It follows from this assumption that if a mutation β occurs in a tumor cell subsequent to a mutation α, then the fraction of cells containing mutation α must be at least as large as the fraction of cells containing β. This condition was also recently noted in Jiao *et al.* (2014).

We assume that at any particular time in the cancer progression *at most* one cell in the tumor population acquires a new mutation leading to a clonal expansion. We emphasize that

this assumption restricts only the number of clonal expansions that *begin* at a given time and not the number of clonal expansions that are ongoing at one time. Under this assumption, each clonal expansion splits the present tumor cell population into exactly two subpopulations: the subpopulation P of cells containing the newly acquired somatic mutation and the subpopulation P' of cells without the mutation (and possibly with a different set of new mutations occurring later in time). Thus, the ancestral history of the sequenced tumor cell population is represented by a rooted *binary* tree with nodes corresponding to populations of cells at each clonal expansion, and edges indicating the ancestral relationships between these populations. Consequently, each node v in the tree has a set M_v of somatic mutations that accumulate along the unique edge p_v that connects v to its parent (As the root r does not have a parent, the set M_r represents the somatic mutations that accumulate before the first clonal expansion. Alternatively, we can let p_r be a *hidden* edge that connects the founder cell of the tumor (represented by the root r) to the normal cell from which it is derived.). Each node v also has a frequency a_v representing the proportion of sequenced tumor cells with mutations M_v (Fig. 1). Note that most tumor samples contain admixture by normal cells with no detectable somatic mutations, and thus in general $a_r < 1$. A consequence of these assumptions is that every internal node v in the tree satisfies the *children sum to parents (CSP)* condition: $\sum_{u \text{ child of } v} a_u = a_v$. We make the following definition:

Definition 1.. Given a multiset $\mathcal{L} = \{a_1, \ldots, a_n\}$ with $0 < a_i \leq 1$, a BTP for $\mathcal{L}$ is a complete rooted weighted binary tree $T = (V,E)$ with nodes $V = \{v_1, \ldots, v_n\}$ such that v_i has weight a_i and every internal node satisfies the CSP condition.

Figure 1 shows a simple example of a complete binary tree in which the CSP conditions are satisfied for every internal node (also see Supplementary Appendix D.6 for a more general example). Recall that a complete rooted binary tree is a binary tree wherein there is a unique node of degree 2 (the root), and every node in the tree is either a leaf or has exactly two children. Our goal is to construct such a tree from the measured VAF data. We define the following.

Definition 2 (BTP problem). Given a multiset $\mathcal{L} = \{a_1, \ldots, a_n\}$, find a BTP for $\mathcal{L}$ if one exists.

Note that in some cases, at the split defined by a clonal expansion, cells from P' may survive to the present, without undergoing additional clonal expansions (e.g. such cells may cease dividing, or senesce). In this case, there are no mutations that exclusively occur in P'. We discuss this case in Section 5 below.

3 COMPLEXITY OF THE BTP PROBLEM

In this section, we outline the proof of the following theorem.

Theorem 3. *The BTP problem is NP-complete.*

The proof of this theorem relies on the idea of finding a set of *conflict-free triangles* in the multiset $\mathcal{L}$. This idea is also useful below for deriving an approximation for a related problem of finding a max -BTP, and so we now define the relevant concepts.

Suppose $\mathcal{L} = \{a_1, \ldots, a_n\}$ is a multiset of n elements. For any distinct i, j and k such that $a_i + a_j = a_k$, we define the ordered pair $(k, \{i,j\})$ as a *triangle* in the multiset. See Supplementary Appendix A.1 for an example. We call k as the *peak* of the triangle and i, j as the *tails* of the triangle.

We say that triangles $t = (k, \{i,j\})$ and $t' = (k', \{i',j'\})$ are *in conflict* if $k = k'$ or $\{i,j\} \cap \{i',j'\} \neq \emptyset$. In other words, two triangles are in conflict if and only if they either share a common peak or a common tail. A set Z of triangles is *conflict-free* if no pair of triangles in Z are in conflict. If T is a BTP for $\mathcal{L}$, for each internal node a_k and its children a_i and a_j, we have $a_k = a_i + a_j$, by CSP. Therefore, $(k, \{i,j\})$ is a triangle in $\mathcal{L}$, and thus, T corresponds to a set of conflict-free triangles in $\mathcal{L}$: each internal node and its children form a triangle in $\mathcal{L}$, and no two triangles share a common peak or a common tail.

Because the number of nodes in a complete binary tree is always odd, $|\mathcal{L}|$ being an odd number is a necessary condition for the existence of a BTP for $\mathcal{L}$. For a multiset $\mathcal{L}$, with $|\mathcal{L}| = 2q - 1$, the size of a conflict-free set of triangles is at most $q - 1$. This is because each triangle has exactly two tails, and in a set of conflict-free triangles, all the tails must be distinct elements.

We have the following theorem, whose proof is in Supplementary Appendix A.2.

Theorem 4. *Suppose* $\mathcal{L} = \{a_1, \ldots, a_{2q-1}\}$. *$\mathcal{L}$ has a BTP if and only if there exists a set of $q-1$ conflict-free triangles in $\mathcal{L}$.*

The proof of NP-completeness of the BTP problem (Theorem 3) is by a reduction from the Numerical Matching with Target Sums (NMTS) problem, (Garey and Johnson, 1979). An instance of NMTS is a triple $\mathcal{I} = (X, Y, B)$ where $X, Y, B \subset \mathbb{Z}^+$, and $X = \{x_1, \ldots, x_m\}$, $Y = \{y_1, \ldots, y_m\}$ and $B = \{b_1, \ldots, b_m\}$. The goal is to find two permutations π_X and π_Y on $\{1, \ldots, m\}$ such that $x_{\pi_X(i)} + y_{\pi_Y(i)} = b_i$ for $i = 1 \ldots, m$.

Theorem 5. *Let $\mathcal{I} = (X, Y, B)$ be an instance of NMTS. Then, $\mathcal{I}$ has a solution if and only if a particular multiset $\mathcal{L}_\mathcal{I}$ has a BTP (i.e. a set of $2m-1$ conflict-free triangles). Moreover, each solution of $\mathcal{I}$ can be obtained (in polynomial time) from a BTP of $\mathcal{L}$, and vice versa.*

Proof. For a given instance $\mathcal{I} = (X, Y, B)$, let $\hat{x}_i = 4(m+1)x_i + 1$, $\hat{y}_i = 4(m+1)y_i + 3$ and $\hat{b}_i = 4(m+1)b_i + 4$, $1 \leq i \leq m$. Moreover, let $\beta_j = \sum_{k=1}^{j} \hat{b}_k$, $2 \leq j \leq m$. Now we construct an instance $\mathcal{L}_\mathcal{I}$ of BTP (i.e. a multiset), with $4m-1$ elements as follows: $\mathcal{L}_\mathcal{I} = \{\hat{x}_1, \ldots, \hat{x}_m\} \cup \{\hat{y}_1, \ldots, \hat{y}_m\} \cup \{\hat{b}_1, \ldots, \hat{b}_m\} \cup \{\beta_2, \ldots, \beta_m\}$.

($\Rightarrow$) First assume that we have two permutations π_X, π_Y on $\{1, \ldots, m\}$ such that $x_{\pi_X(i)} + y_{\pi_Y(i)} = b_i$. By definition of $\hat{x}_i$, $\hat{y}_i$ and $\hat{b}_i$ we have also $\hat{x}_{\pi_X(i)} + \hat{y}_{\pi_Y(i)} = \hat{b}_i$. Now we construct a set of conflict-free triangles for $\mathcal{L}_\mathcal{I}$ of size $2m-1$: for each $i \in \{1, \ldots, m\}$, add the triangle $(\hat{b}_i, \{\hat{x}_{\pi_X(i)}, \hat{y}_{\pi_Y(i)}\})$. In addition, for each $i \in \{1, \ldots, m-1\}$, we add all triangles $(\beta_{i+1}, \{\hat{b}_i, \hat{b}_{i+1}\})$. Thus, by Theorem 4 we obtain a BTP from π_X and π_Y in polynomial time.

($\Leftarrow$) Suppose S is a set of $2m-1$ conflict-free triangles for $\mathcal{L}_\mathcal{I}$. We claim that for each $\hat{x}_i$ a triangle $(\hat{b}_{\gamma(i)}, \{\hat{x}_i, \hat{y}_{\alpha(i)}\})$ exists, where α and γ are two permutations on $\{1, \ldots, n\}$. Note that this

completes the proof, by taking $\pi_X = \gamma^{-1}$ and $\pi_Y = (\gamma^{-1} \circ \alpha)$ for the instance $\mathcal{I}$. Node $\hat{x}_i$ cannot be the root, as the largest number in $\mathcal{L}_\mathcal{I}$ is β_m. Therefore, $\hat{x}_i$ has a sibling s and a parent p. For all $i \in \{1, \ldots, m\}$, we have $\hat{x}_i = 1$, $\hat{y}_i = 3$, $\hat{b}_i = 4$ and $\beta_i = 4i$, all in **mod** $4(m+1)$. Thus, if $\hat{x}_i + s = p \in \mathcal{L}_\mathcal{I}$, we have $s = 3 \pmod{4(m+1)}$ and $p = 4 \pmod{4(m+1)}$. This implies $s = \hat{y}_{\alpha(i)}$ and $p = \hat{b}_{\gamma(i)}$ for some $\alpha(i)$ and $\gamma(i)$. Finally, because all the elements of $\mathcal{L}_\mathcal{I}$ are presented uniquely in T, α and γ are two permutations on $\{1, \ldots, m\}$. Note that we construct π_X and π_Y from T in polynomial time, and the proof is complete. □

We note that the proof above shows that the BTP problem in NP-complete in the *strong sense*, i.e. the problem is still NP-complete if the elements of the multiset are polynomially bounded. In addition, the analogous partition problem for non-binary trees is also NP-complete by reduction from the subset sum problem. See Supplementary Appendix A.6.

4 A $\frac{2}{3} - o(1)$ APPROXIMATION FOR MAX-BTP

In the previous section, we showed that for a given multiset $\mathcal{L}$ of $2q-1$ elements, each BTP for $\mathcal{L}$ corresponds to a collection of $q-1$ *conflict-free* triangles. Because $\mathcal{L}$ can have at most $q-1$ conflict-free triangles, we define the max -BTP problem to be the problem of finding the *maximum sized set of conflict-free triangles*. This is a closely related problem to the BTP problem: in the context of VAF data, the maximum sized set of conflict-free triangles denotes partial information about the ancestral relationships among mutations. Moreover, for a multiset of m elements if max -BTP has a solution of size Δ then a BTP with $k = m - 1 - 2\Delta$ additional nodes can be found. See Supplementary Appendix A.7.

We derive a $\frac{2}{3} - o(1)$ approximation algorithm for the max -BTP problem for $\mathcal{L}$. The algorithm is based on Local Search. We start with any collection of conflict-free triangles in $\mathcal{L}$ as a solution and iteratively add another triangle as follows. For a fixed constant $t \geq 1$, we iteratively replace any subcollection of $s \leq t$ triangles in the solution with $s+1$ triangles of $\mathcal{L}$ such that the new collection still contains only conflict-free triangles.

It is easy to see that the above local search terminates in polynomial time. Let OPT be the size of the optimal solution. Because we cannot have more than $q-1$ conflict-free triangles in a solution, $OPT \leq q-1$. After each iteration, the size of the collection increases by 1, and because t is a constant, the search procedure at each iteration is polynomial time. Similar to the technique that was used in Hajirasouliha *et al*. (2007), we use the packing bound of Hurkens and Schrijver (1989) to prove the following theorem (Proof in Supplementary Appendix A.7).

THEOREM 6. *There exists a polynomial time algorithm that gives an approximated solution to the problem of finding maximum set of conflict-free triangles within a factor of* $\frac{2}{3} - \delta$ *for any* $\delta > 0$.

5 THE ε-BTP PROBLEM

Typically on real data, a BTP will not exist—either because the frequencies a_i are determined with some error or the VAF data does not capture the frequency of a subpopulation that does not have mutations that exclusively occur in that subpopulation (VAFs provide information only about the proportion of cells with a mutation, and do not provide information about proportions of cells that have a specific mutation and lack another mutation.). In this section, we introduce the ε-BTP to account for these scenarios. Suppose we have the multiset $\tilde{\mathcal{L}} = \{\tilde{a}_1, \ldots, \tilde{a}_m\}$ of observed frequencies and a corresponding *VAF error vector* $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ for $\tilde{\mathcal{L}}$, where ε_i is the maximum possible error in observing $\tilde{a}_i$ for $1 \leq i \leq m$. To account for subpopulations without distinguishing mutations, we may need to add *auxiliary* frequencies to $\tilde{\mathcal{L}}$ that correspond to the missing subpopulation frequencies. We make the following definitions.

DEFINITION 7 (ε-BTP). Given a multiset $\tilde{\mathcal{L}} = \{\tilde{a}_1, \ldots, \tilde{a}_m\}$ with associated VAF error vector $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$, an ε-BTP with $k \geq 0$ *auxiliary* nodes is a BTP for a multiset $\mathcal{L} = \{a_1, a_2, \ldots a_{m+k}\}$ such that for all $i \leq m : |a_i - \tilde{a}_i| \leq \epsilon_i$. We call the nodes $a_{m+1}, \ldots, a_{m+k}$ the *auxiliary nodes* of the ε-BTP.

DEFINITION 8 (The ε-BTP problem). Given a multiset $\tilde{\mathcal{L}}$ and an associated VAF error vector ε, find an ε-BTP of $\tilde{\mathcal{L}}$ with minimum number of auxiliary nodes such that two auxiliary nodes are not siblings.

The constraint on auxiliary nodes in the definition of ε-BTP problem follows from the assumptions in our model of cancer progression: each branching in the cancer progression happens only when at least one clonal expansion starts. So, the VAF data captures the frequency of the newly formed subpopulation (see Section 2). Thus, at least one of the children of the current subpopulation node is not an auxiliary node.

It is straightforward to show that for any multiset $\mathcal{L}$ of size m, it is always possible to obtain an ε-BTP with $k = m-1$ auxiliary nodes (proof in Supplementary Appendix A.3). Also, when $\varepsilon_i = 0$ for all $1 \leq i \leq m$, a BTP exists for $\mathcal{L}$ if and only if the corresponding ε-BTP has a solution with $k = 0$ auxiliary nodes.

To outline our algorithm, we need the following definitions.

DEFINITION 9 (ε-CSP tree). Given a VAF error vector ε, an ε-CSP tree is a (weighted) binary tree, such that for each internal node $\tilde{a}_i$ we have $\tilde{a}_j + \tilde{a}_k \in [\tilde{a}_i \pm (\epsilon_i + \epsilon_j + \epsilon_k)]$, where $\tilde{a}_j$ and $\tilde{a}_k$ are the children of $\tilde{a}_i$.

We say that an ε-CSP tree $\tilde{T}$ for a multiset $\tilde{\mathcal{L}}$ is *acceptable* if we can obtain a BTP, $\alpha(\tilde{T})$, by replacing each $\tilde{a}_i$ by a value a_i where $|a_i - \tilde{a}_i| \leq \epsilon_i$. Note that $\alpha(\tilde{T})$ is an ε-BTP for $\tilde{\mathcal{L}}$. Also, note that an ε-CSP tree is not necessarily acceptable (See Supplementary Appendix D.8). However, one can easily check whether a given ε-CSP tree $\tilde{T}$ is acceptable by finding a collection of e_i's, where $|e_i| \leq \epsilon_i$, satisfying the following constraints: $(\tilde{a}_i + e_i) = (\tilde{a}_j + e_j) + (\tilde{a}_k + e_k)$, for each internal nodes $\tilde{a}_i$ and its children $\tilde{a}_j, \tilde{a}_k$. This can be easily done via a linear program, which we denote by $LP(\tilde{T})$.

Our Rec-BTP algorithm (Algorithm 1) uses a recursive method that works as follows: at each recursion during the algorithm, we have (i) a partially constructed ε-CSP tree $\hat{T}$, (ii) a multiset of remaining frequencies $\hat{\mathcal{L}}$ and (iii) the number of remaining auxiliary nodes that we are allowed to use. We check if $\hat{T}$ can be extended by attaching two elements of $\hat{\mathcal{L}}$, or one element from $\hat{\mathcal{L}}$ and an auxiliary node, to one of the leaves in $\hat{T}$ (we assign the auxiliary node's weight accordingly). If $\hat{\mathcal{L}}$ is empty, it

means the algorithm has constructed an ε-CSP tree. So we output $\left\{\alpha(\hat{T})\right\}$ if $LP(\hat{T})$ has a feasible solution. Finally, Rec-BTP outputs all the ε-BTPs. Iterating over all values of k from 0 to $m-1$, the algorithm will find the smallest k such that there exists an ε-BTP.

Later in Section 6, for the purpose of benchmarking our results, in case of multiple ε-BTP outputs, we choose only the tree whose list of node frequencies has the minimum root mean square deviation (RMSD) from the original VAFs data (defined below in Section 6).

6 EXPERIMENTAL RESULTS

6.1 Simulated data

We generate simulated mutation data from all complete rooted binary trees with 3, 5, 7 and 9 nodes (Supplementary Appendix D.7). For each tree topology, we generate 1000 random BTPs by assigning a weight a_i to each node i, as follows. For the root r, we set $a_r = 1$, assuming that our tumor sample is pure, i.e. is not contaminated with normal cells. Next, we proceed down the tree: for each parent with weight a_i, we select a pair of real numbers a_j and a_k for the children uniformly at random such that the CSP condition ($a_i = a_j + a_k$) is satisfied. Finally, we generate a set M_i of somatic mutations for each node, with $|M_i|$ selected uniformly from [50 400] independently for each node. We assume all somatic mutations in the set M_i happened independently because the parental cell was created. For each such BTP and set of mutations, we generate a *VAF data* corresponding to each tumor subpopulation. Ideally, the VAF of a tumor subpopulation from node v equals a_v. However, because the observed frequencies are estimated from alignments of sequencing data, the observed frequencies will deviate from the true values. We assume that the observed VAFs for mutations of a subpopulation v are normally distributed with mean a_v and standard deviation σ. Specifically, for each node v, let X_v be a set of $|M_v|$ samples from $\mathcal{N}(a_v, \sigma)$. Here, we present the results when $\sigma = 2$, with results on $\sigma = 1$, 4 in Supplementary Appendix B. The VAF data for the tumor sample is thus $X = \cup_v X_v$. Note that we simulate VAFs directly rather than the number of mapped reads containing a mutation. Because we assume that our sequence coverage is high (>1000×), it follows that the corresponding binomial (or negative binomial) distribution of read counts is well approximated by the normal distribution. For lower coverages, the asymptotic normal approximation may not model the data as accurately; nevertheless, the simulations provide a comparison of the different methods on high-quality data.

For each simulated VAF dataset X, we estimate the number and frequencies of tumor subpopulations using two non-parametric clustering algorithms: (i) Accelerated variational Dirichlet process Gaussian mixture model (AVDPM; Kurihara *et al.*, 2006), as implemented in https://sites.google.com/site/kenichikurihara/academic-software/variational-dirichlet-process-gaussian-mixture-model, which is a general clustering method that we apply directly on VAF data, and (ii) *SciClone* (Miller *et al.*, forthcoming), which is a recent algorithm (with available software but no published paper) that estimates tumor composition from VAF data by clustering the data using a mixture of Gaussian model. Parameter settings for each method are given in Supplementary Appendix C. Also, because SciClone runtimes were extremely long, we down sampled the mutation data by a factor of 20. For each of our synthetic dataset, we implanted the fraction of the mutations (together with their corresponding VAF) on a synthetic chromosome with neutral copy number compatible with the SciClone input format (See Supplementary Appendix C). We ran SciClone with default parameters on each dataset and then extracted the means of reported clusters from SciClone.

Algorithm 1: Rec-BTP($\widehat{\mathcal{L}}, \widehat{T}, \epsilon, k_{\max}$)

Input : Partially constructed tree $\widehat{T}$, remaining frequencies multiset $\widehat{\mathcal{L}}$, ϵ, and an upper bound $k_{\max}$ on the number of remaining auxiliary nodes.

Output: List of all ϵ-BTPs that can been obtained by expanding $\widehat{T}$ using at most $k_{\max}$ auxiliary nodes and the elements of $\widehat{\mathcal{L}}$.

```
begin
    if L̂ = ∅ then
        if LP(T̂) has a feasible solution then
            return {α(T̂)};
        else
            return ∅;
    O ← ∅;
    for each leaf ã_t in T̂ do
        for ∀ã_i ∈ L̂ do
            for each ã_j ∈ L̂ ∩ I_{ε_i+ε_j+ε_t}(ã_t − ã_i) do
                T' ← attach ã_i and ã_j to ã_t;
                L' ← L̂ − {ã_i, ã_j};
                O ← O ∪ Rec-BTP(L', T', ε, k_max)
            if k_max > 0 then
                T' ← attach ã_i and an auxiliary node
                (with weight ã_t − ã_i) to ã_t;
                L' ← L̂ − {ã_i};
                O ← O ∪ Rec-BTP(L', T', ε, k_max − 1);
    return O;
```

From the output of each clustering algorithm, we obtain the input for our Rec-BTP algorithm. We compute the sample mean $\tilde{a}_i$ and standard deviation σ_i for each cluster C_i, and set the VAF error ε_i for the subpopulation frequency of cluster C_i equal to $1.96 \cdot c\frac{\sigma_i}{\sqrt{|C_i|}}$, where c is a constant set to 3. Note that $1.96 \cdot \sigma_i/\sqrt{|C_i|}$ is the radius of the empirical 95% confidence interval in estimating the true subpopulation frequency.

We set $\mathcal{L} = \{\tilde{a}_1, \ldots, \tilde{a}_m\}$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$ as the input to Rec-BTP. We find the minimum k for which there exists a ε-BTP for $\mathcal{L}$ with k auxiliary nodes. In many cases, there are multiple BTPs for $\mathcal{L}$ with exactly k auxiliary nodes. In these cases, we select a single BTP with the *minimum cost* $\text{cost}_X(T) = \sum_{i=1}^{s} \sum_{x \in D_i} |x - a_i|^2$, where $s = n - k$, X is the VAF dataset and $D_i = \{x \in X | \arg\min_j |a_j - x| = i\}$ is the subset of elements of X that are closest to a_i.

We compare the results of our Rec-BTP algorithm (applied to clusters from both AVDPM and SciClone) to the original

clusters output from these algorithms over the 1000 randomly constructed BTPs for each of the seven tree topologies. We use two measures to compare the estimated subpopulation frequencies: (i) the *number* of subpopulations and (ii) the RMSD between the set of estimated subpopulation frequencies and the true subpopulation frequencies.

Number of subpopulations. Figure 2 shows that Rec-BTP outputs the correct number of clusters more frequently than the clustering methods. For trees with 3 and 5 nodes, Rec-BTP does not improve the AVDPM clusters much. However, with larger number of nodes the advantage of Rec-BTP grows. Rec-BTP provides a large improvement over the SciClone clusters.

We further examined the scenarios where each algorithm reported the correct and incorrect number of clusters. Figure 3 compares the fraction of cases where the clustering method and the Rec-BTP assisted method report too few, the correct number or too many clusters. We see that most of the cases where Rec-BTP reports the correct number of cases (blue squares) are those where the clustering algorithm reported too few clusters and the Rec-BTP algorithm created additional clusters. Only in the case of 3 and 5 nodes do AVDPM and SciClone determine the correct number of clusters (green squares) in an appreciable fraction of cases. Overall, we see that the clustering methods tend to underestimate the correct number of clusters (first columns in each table in Fig. 3). In a significant fraction of these cases, Rec-BTP adds auxiliary nodes to obtain the correct number of clusters, although this becomes more difficult with larger trees. We also see that SciClone tends to underestimate the number of subpopulations more frequently than AVDPM.

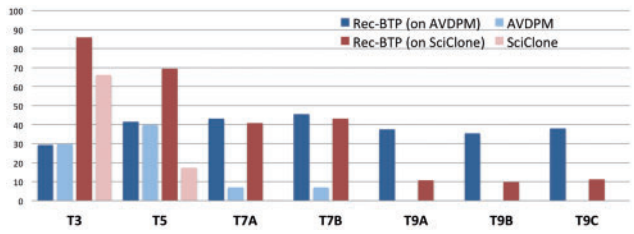

Fig. 2. Percentage of trees where each algorithm finds the correct number of subpopulations. Light blue/red bars are AVDPM and SciClone, respectively. Dark blue/red bars are Rec-BTP results using AVDPM and SciClone clusters, respectively, as input

Accuracy of the subpopulation frequencies. We compare the estimated population frequencies and true population frequencies for each method using the RMSD. Suppose $a_1 \geq \ldots \geq a_n$ are the true subpopulation frequencies, and $\tilde{a}_1 \geq \ldots \geq \tilde{a}_m$ are the estimated subpopulation frequencies. If $m = n$, the RMSD is $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - \tilde{a}_i)^2}$. If $m \neq n$, we add zeros to the shorter sequence so the two sequences have equal length. The zeros reflect the fact that we have not estimated the frequencies of some subpopulations.

Table 1 gives RMSD for AVDPM, SciClone and Rec-BTP built from these clusters. Specifically, we provide the RMSDs when AVDPM or SciClone give (i) the same and (ii) less than the correct number of subpopulations. For some tree topologies, there is no sample in which both Rec-BTP and AVDPM/SciClone give the same number of subpopulations equal to the true number of subpopulations, which we denote by N/A.

In cases where both methods (Rec-BTP and a clustering algorithm) return the same number of subpopulations, they have similar performance in estimating the subpopulation frequencies. Also, as mentioned earlier, these results are for the simulated input data in which the variant allele frequencies deviate from their true value with standard deviation $\sigma = 2$. When $\sigma = 1$ Rec-BTP performs even better.

6.2 Comparison with Phylosub

As noted in the introduction, Jiao *et al.* (2014) is a recent method that clusters VAF frequencies using a tree constraint. In particular, Jiao *et al.* (2014) replace the Dirichlet process mixture for clustering with a Bayesian non-parametric prior over trees satisfying a weak form of the CSP constraint. We compared our Rec-BTP algorithm with *PhyloSub*.

We generated VAF data from a collection of 400 random complete binary trees with three, five, seven and nine nodes with fixed topologies. For each tree, we generated 100 random instances. In contrast to the simulations in the previous section, here we used only one of the two topologies for trees with seven nodes and only one of the three possible topologies for trees with nine nodes. We converted each random simulated VAF data to a

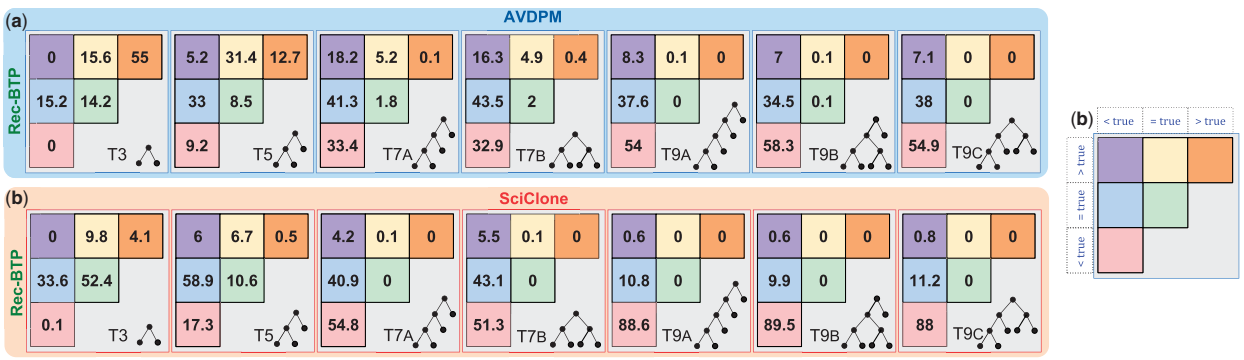

Fig. 3. Estimating the number of subpopulations using different algorithms. (**a**) Each entry in the table represents the fraction of random trees obtained from AVDPM (columns) and Rec-BTP on AVDPM clusters (rows). (**b**) Results for SciClone versus Rec-BTP on SciClone clusters. (**c**) Interpretation of each entry: the reported number of subpopulations by each method compared with the true number of subpopulations

Table 1. Mean and standard deviation of RMSD over 1000 trees for each method

Tree	Rec-BTP = AVDPM		Rec-BTP > AVDPM		Rec-BTP = SciClone		Rec-BTP > SciClone	
	Rec-BTP	AVDPM	Rec-BTP	AVDPM	Rec-BTP	SciClone	Rec-BTP	SciClone
$T3$	0.9 ± 0.8	0.7 ± 1.1	$\mathbf{1.3 \pm 0.6}$	45.1 ± 11.3	$\mathbf{1.2 \pm 0.4}$	1.5 ± 0.4	$\mathbf{2.6 \pm 1.1}$	43.7 ± 11
$T5$	$\mathbf{6.9 \pm 5.6}$	7.5 ± 5.2	$\mathbf{9 \pm 12.8}$	17.4 ± 14.1	$\mathbf{1 \pm 0.5}$	1.2 ± 0.4	$\mathbf{4.2 \pm 6.4}$	15.7 ± 9.2
$T7_A$	8.6 ± 4.8	8.2 ± 4.5	$\mathbf{9.2 \pm 5.9}$	10.1 ± 6.5	N/A	N/A	$\mathbf{6.9 \pm 7.7}$	12.8 ± 7.5
$T7_B$	8.4 ± 2.8	8.1 ± 2.4	$\mathbf{8.8 \pm 6.2}$	11.7 ± 7.1	N/A	N/A	$\mathbf{6.2 \pm 6.9}$	13.2 ± 6.8
$V9_A$	N/A	N/A	8.9 ± 4.7	8 ± 4.6	N/A	N/A	$\mathbf{6.4 \pm 5.4}$	10.8 ± 5.5
$V9_B$	$\mathbf{2 \pm 0}$	2.2 ± 0	$\mathbf{8.1 \pm 4.5}$	8.3 ± 4.4	N/A	N/A	$\mathbf{5.6 \pm 4.9}$	11.4 ± 5.2
$V9_C$	N/A	N/A	$\mathbf{7.5 \pm 4.5}$	9.9 ± 4.4	N/A	N/A	$\mathbf{6.1 \pm 5.1}$	10.1 ± 4.5

Note: Bold face text indicates best performance.

PhyloSub input identical to the procedure performed for the simulations in the *PhyloSub* paper. In creating *PhyloSub* inputs, we assumed the total read counts for every single nucleotide variants (SNV) position is 10 000 (i.e. an ideal uniform coverage of 10 000× for the *PhyloSub* input) and assumed every SNV is heterozygous. On each dataset, we ran the Markov chain Monte Carlo (MCMC) method of *PhyloSub* 100 times, each with 5000 MCMC iterations as per Jiao *et al.* (2014), and used the reported *top trees* (i.e. those trees with best log likelihood) in our comparison.

We found that *PhyloSub* produced trees with many more nodes than the simulated value, and significantly more than Rec-BTP or SciClone (Fig. 4 and Supplementary Appendix D. 9]). Because *PhyloSub* usually tends to report trees with a higher number of nodes, we also considered the size of the smallest tree reported by *PhyloSub* in their provided list of *top trees* for each input. While this value was smaller, it was still much larger than the true value or the values from the other approaches.

The large number of clusters produced by *PhyloSub* might result from the fact that the method does not assume that the trees are binary. The output trees contain many different topologies. Nevertheless, it is surprising that *PhyloSub* does not find binary trees when the data are produced from this topology. As *PhyloSub* reports a higher number of nodes than Rec-BTP, we were unable to directly compare the provided frequencies of the clusters of each method.

6.3 Acute myeloid leukemia sequencing data

We tested our algorithm on VAFs obtained from deep read counts information for SNVs from an acute myeloid leukemia sample (AML1/UPN933124) using data from Ding *et al.* (2012). We used the 386 SNVs reported in the primary AML sample, obtaining the tumor VAF data directly from Supplementary Table S5a in Ding *et al.* (2012). Note that Ding *et al.* (2012) also report data from a relapse sample following chemotherapy. As the relapse-specific mutations form only one cluster, we do not analyze the BTP problem for this sample in our study. Nevertheless, the generalization of the ε-BTP problem for the case where the input data contains two types of VAFs (e.g. both tumor- and relapse-specific mutations) is an interesting open problem. We first ran SciClone on the VAF data, obtaining

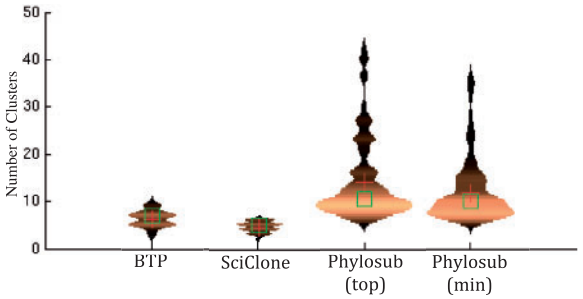

Fig. 4. A violin plot for the number of clusters output by Rec-BTP, SciClone, PhyloSub where the top tree is considered, and PhyloSub where the tree, among the top ones, with the minimum number of nodes is considered. The y-axis shows the number of nodes in each tree, while the histogram in each violin plot corresponds to tree sizes of different experiments

four distinct VAF clusters with means of 47.17, 33.17, 22.42, 3.65%. We then ran Rec-BTP on these clusters, fixing the root of the BTP as an additional node with frequency 100%, reflecting the fact that the tumor sample was pure and started from a single founder clone (Ding *et al.*, 2012).

Figure 5 shows the resulting ε-BTP with corresponding multiset of population frequencies: $\mathcal{L}_{Rec-BTP} = \{100, 53.75, 46.25, 42.86, 32.25, 21.5, 3.39\}$. Ding *et al.* (2012) present a history of clonal expansions that implies an ε-BTP for the multiset $\{100, 53.12, 12.74, 29.04, 5.1\}$ with two auxiliary nodes (Fig. 4b). There are two such ε-BTP, depending on the relative order of two clonal expansions (Fig. 4), one with subpopulation frequencies: $\mathcal{L}_1 = \{100, 53.12, 46.88, 12.74, 34.14, 29.04, 5.1\}$ and another with frequencies $\mathcal{L}_2 = \{100, 34.14, 65.86, 53.12, 12.74, 29.04, 5.10\}$. However, because the frequencies reported in Ding *et al.* (2012) are scaled according to the estimated 93.72% purity of their sample, it is necessary to multiply the frequencies in $\mathcal{L}_1$ and $\mathcal{L}_2$ by 0.9372 before comparing with the clusters obtained from the VAF data.

We compare these different subpopulation frequencies estimates by computing the average ℓ_1 norm between the VAF for each mutation and the closest subpopulation. Table 2 shows this measure for each of the subpopulation multisets. We see that the

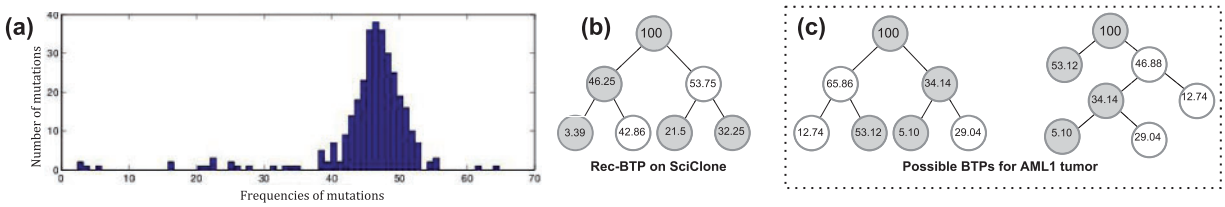

Fig. 5. (**a**) The VAF data for AML1 tumor sample. The VAFs are dense around ~47. (**b**) The ε-BTPs for AML1 from Ding *et al.* (2012) obtained from Rec-BTP using SciClone clusters. (**c**) The two possible ε-BTPs derived from the clonal frequencies and ancestral reconstruction proposed in Ding *et al.* (2012). Auxiliary nodes are shown in white circles

Table 2. Comparison of subpopulation frequencies obtained by Rec-BTP with two possible subpopulation frequency multisets obtained from Ding *et al.* (2012), based on the calculation of $\ell 1$ and $\ell 2$ norms to the original VAF data

Metric	$\mathcal{L}_{Rec-BTP}$	$\mathcal{L}_1$	$\mathcal{L}_2$
Average ℓ_1 norm	1.6367	3.3309	1.8376
Average ℓ_2 norm	0.1161	0.2221	0.1331

Rec-BTP gives a better fit to the VAF data than both estimates $\mathcal{L}_1$ and $\mathcal{L}_2$ given in Ding *et al.* (2012). This shows that using the tree constraint in the BTP provides additional information that is useful for clustering real VAF data.

Note that calculation of ℓ_1 and ℓ_2 norms for the SciClone cluster means would result in 2.4816 and 0.1823, respectively. However, because the number of SciClone clusters is only 4, these numbers cannot be reasonably compared with the ones calculated for the trees in Figure 5.

7 DISCUSSION

In this article, we provide the first rigorous combinatorial formulation of the problem of inferring the composition of tumor subpopulations constrained by a tree in a single tumor sample from the variant allele frequencies (VAFs) of somatic mutations. In our formulation, we introduced a novel definition of the BTP and the ε-BTP. We showed that the problem of finding a BTP (and hence an ε-BTP) is in general NP-complete; however, we derived an approximation algorithm for a related problem, max-BTP. We developed a recursive algorithm for the ε-BTP that works well in practice and showed the advantages on this algorithm on simulated and real sequencing data.

These results show the utility of the BTP, but also suggest additional areas of further investigation. In particular, it would be interesting to combine the clustering of VAFs and the construction of the BTP into a single model. While there has been some work to combine these two steps using a machine learning approach (Jiao *et al.*, 2014), the complexity of the corresponding inference problem is unknown. In our tests, we were unable to obtain satisfactory results using this model, suggesting there is room for additional improvements. One possible direction is to use MCMC or other sampling approaches over the space of BTPs, perhaps combining this inference into a graphical model that better models the features of real sequencing data [e.g. as used in pyClone (Shah *et al.*, 2012; Roth *et al.*, 2014)]. Finally, the extension of the BTP and related approaches to multiple samples from the same tumor (taken at the same or different times) will be increasingly useful as such data become available.

ACKNOWLEDGEMENT

The authors thank Li Ding and Michael McLellan for assistance with the AML data.

Funding: This work was supported by National Science Foundation CAREER Award (CCF-1053753 to B.J.R.) and the National Institutes of Health (R01HG5690 to B.J.R.). B.J.R. is also supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an Alfred P. Sloan Research Fellowship. I.H. is also supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship.

Conflict of Interest: none declared.

REFERENCES

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.

Ding,L. *et al.* (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**, 506–510.

Ding,L. *et al.* (2013) Advances for studying clonal evolution in cancer. *Cancer Lett.*, **340**, 212–219.

Garey,M.R. and Johnson,D.S. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman and Company, New York.

Gerlinger,M. *et al.* (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.*, **366**, 883–892.

Hajirasouliha,I. *et al.* (2007) On completing latin squares. In: *STACS*. pp. 524–535.

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

Hurkens,C.A.J. and Schrijver,A. (1989) On the size of systems of sets every t of which have an SDR, with an application to the worst-case ratio of heuristics for packing problems. *SIAM J. Discret. Math.*, **2**, 68–72.

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

Kandoth,C. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.

Kurihara,K. *et al.* (2006) Accelerated variational dirichlet process mixtures. In: Schölkopf,B. *et al.* (ed.) *NIPS*. MIT Press, pp. 761–768.

Lawrence,M.S. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.

Miller,C.A. *et al.* (forthcoming) Sciclone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol.*

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

Newburger,D.E. *et al.* (2013) Genome evolution during progression to breast cancer. *Genome Res.*, **23**, 1097–1108.
Nik-Zainal,S. *et al.* (2012) The life history of 21 breast cancers. *Cell*, **149**, 994–1007.
Nowell,P.C. (1976) The clonal evolution of tumor cell populations. *Science*, **194**, 23–28.
Oesper,L. *et al.* (2013) Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol.*, **14**, R80.
Roth,A. *et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, **11**, 396–398.
Salari,R. *et al.* (2013) Inference of tumor phylogenies with improved somatic mutation discovery. In: *RECOMB*. pp. 249–263.
Schuh,A. *et al.* (2012) Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, **120**, 4191–1496.
Shah,S.P. *et al.* (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, **486**, 395–399.
Strino,F. *et al.* (2013) TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res.*, **41**, e165.
Vogelstein,B. *et al.* (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.
Xu,X. *et al.* (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, **148**, 886–895.