Bioinformatics, 35, 2019, i398–i407
doi: 10.1093/bioinformatics/btz392
ISMB/ECCB 2019

Inference of clonal selection in cancer populations using single-cell sequencing data

Pavel Skums[1,*], Viachaslau Tsyvina[1] and Alex Zelikovsky[1,2]

[1]Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA and [2]The Laboratory of Bioinformatics, I.M. Sechenov First Moscow State Medical University, Moscow 119991, Russia

*To whom correspondence should be addressed.

Abstract

Summary: Intra-tumor heterogeneity is one of the major factors influencing cancer progression and treatment outcome. However, evolutionary dynamics of cancer clone populations remain poorly understood. Quantification of clonal selection and inference of fitness landscapes of tumors is a key step to understanding evolutionary mechanisms driving cancer. These problems could be addressed using single-cell sequencing (scSeq), which provides an unprecedented insight into intra-tumor heterogeneity allowing to study and quantify selective advantages of individual clones. Here, we present Single Cell Inference of FItness Landscape (SCIFIL), a computational tool for inference of fitness landscapes of heterogeneous cancer clone populations from scSeq data. SCIFIL allows to estimate maximum likelihood fitnesses of clone variants, measure their selective advantages and order of appearance by fitting an evolutionary model into the tumor phylogeny. We demonstrate the accuracy our approach, and show how it could be applied to experimental tumor data to study clonal selection and infer evolutionary history. SCIFIL can be used to provide new insight into the evolutionary dynamics of cancer.

Availability and implementation: Its source code is available at https://github.com/compbel/SCIFIL.

Contact: pskums@gsu.edu

1 Introduction

Cancer is responsible for more than 600 000 deaths in the USA annually (Siegel *et al.*, 2018). It is a disease driven by the uncontrolled growth of cancer cells having series of somatic mutations acquired during the tumor evolution. Cancer clones form heterogeneous populations, which include multiple subpopulations constantly evolving to compete for resources, metastasize, escape immune system and therapy (Doyle *et al.*, 2014; Greaves and Maley, 2012; Kuipers *et al.*, 2017a; Yates and Campbell, 2012). Clonal heterogeneity plays key role in tumor progression (Merlo *et al.*, 2010), and has important implications for diagnostics and therapy, since rare drug resistant variants could become dominant and lead to relapse in the patient (Doyle *et al.*, 2014; Landau *et al.*, 2013). Therefore, cancer is now viewed as a dynamic evolutionary process defined by complex interactions between clonal variants, which include both competition and cooperation (Bonavia *et al.*, 2011; Greaves and Maley, 2012; Yates and Campbell, 2012).

Recent advances in sequencing technologies promise to have a profound effect on oncological research. Study of genomic data for different tumors produced by next-generation sequencing (NGS) led to progress in understanding evolutionary mechanisms of cancer (Greaves and Maley, 2012; Kuipers *et al.*, 2017a; Yates and Campbell, 2012). Most of cancer data have been obtained using bulk sequencing, which produces admixed populations of cells. Recently, the most promising technological breakthrough was the advent of *single-cell sequencing* (scSeq), which allows to access cancer clone populations at the finest possible resolution. scSeq protocols combined with NGS allow to analyze genomes of individual cells, thus providing deeper insight into biological mechanisms of tumor progression.

The cornerstone of such analysis is an estimation of parameters defining the evolution of heterogeneous clonal populations. Currently, there is no scientific consensus about the rules guiding the evolution of cancer cells (Davis *et al.*, 2017; Noorbakhsh and Chuang, 2017; Tarabichi *et al.*, 2018; Williams *et al.*, 2018), with multiple competing theories being advanced by different researchers. The open questions include the rules of evolution (neutral, linear, branching or punctuated), ways of interaction between clonal variants (competition or cooperation) and the role of epistasis (non-linear interaction of single nucleotide variant (SNVs) or genes). These questions could be addressed by estimation of evolutionary parameters for cancer lineages from NGS data (Tarabichi *et al.*, 2018; Williams *et al.*, 2018).

One of the most important evolutionary parameters is the collection of replicative fitnesses of individual genomic variants, commonly termed *fitness landscape* in evolutionary biology (Gavrilets, 2004). Several computational tools have been proposed for *in vitro* estimation of fitness landscapes (Ferguson *et al.*, 2013; Hinkley *et al.*, 2011; Ma *et al.*, 2010; Segal *et al.*, 2004). However, *in vitro* studies are cost- and labor-intensive, consider organisms removed from their natural environments and does not allow to capture all population genetic diversity (Seifert *et al.*, 2014). One of the possible ways to infer fitness landscape *in vivo* is to analyze follow-up samples taken from a patient at multiple time points and compute fitnesses directly by measuring changes of frequencies of genomic variants over time. However, follow-up samples are very scarce, and the overwhelming majority of data represent individual samples.

Quantification of clonal selection from individual samples is computationally challenging, but extremely important for understanding mechanisms of cancer progression (Tarabichi *et al.*, 2018; Williams *et al.*, 2018). In particular, recent findings on structures of fitness landscapes of cancer from bulk sequencing data (Williams *et al.*, 2016) initiated a lively scientific discussion published in several papers (Noorbakhsh and Chuang, 2017; Tarabichi *et al.*, 2018; Williams *et al.*, 2018). It can be anticipated that scSeq data will be able to shed light into this important problem. It is known that relative abundances of genomic variants alone are not indicative of variant fitnesses (Seifert *et al.*, 2014). Existing methods for inference of fitnesses from single samples utilize more sophisticated approaches, but have various limitations including reliance on the assumption that the population is in equilibrium state, or disregard of population heterogeneity and variability of fitness landscapes, or customization to bulk sequencing data (Deforche *et al.*, 2008; Seifert *et al.*, 2014; Williams *et al.*, 2018).

1.1 Contributions

We propose a computational method Single Cell Inference of FItness Landscape (SCIFIL) for *in vivo* inference of clonal selection and estimate of fitness landscapes of heterogeneous cancer clone populations from scSeq data. SCIFIL estimates fitnesses of clonal variants rather than alleles, and does not assume allele independence which allows to take into account the effects of epistasis. Instead of assuming that sampled populations are in the equilibrium state, our method estimates fitnesses of individual clone types using a maximum likelihood approach. We demonstrate that the proposed method allows for accurate inference of fitness landscapes and quantification of clonal selection. We conclude by applying SCIFIL to real tumor data.

2 Materials and Methods

We propose a maximum likelihood approach, which estimates fitnesses of individual clonal variants by fitting into the tumor phylogeny an evolutionary model with the parameters explaining the observed data with the highest probability. We first establish the ordinary differential equations (ODE) model for the tumor evolutionary dynamics, and define the likelihood of the observed data given the model parameters. We conclude with finding fitnesses maximizing the likelihood by reducing the problem to finding the most likely mutation order and applying branch-and-bound search to solve that problem.

Traditionally, evolutionary histories are represented using binary phylogenetic trees. Following Jahn *et al.* (2016), we use an alternative representation of an evolutionary history of a tumor using a *mutation tree*. The internal nodes of a mutation tree represent mutations, leafs represent single cells, internal nodes are connected according to their order of appearance during the tumor evolution and the mutation profile of each cell equals the set of mutations on its path to the root (Fig. 1). In addition, we accumulate all leafs attached to the same internal node into a single leaf with an abundance representing a particular clone. For simplicity, we assume that there is a leaf attached to every internal node, with some leafs having an abundance 0 (or rather a small number $\delta \ll 1$). Generally, we do not need to employ the infinite site assumption, i.e. repeats of mutations are allowed provided that mutation profiles of all clones in a tree are unique. It agrees with recent findings (Kuipers *et al.*, 2017b). A mutation tree can be constructed using currently available tools, such as SCITE (Jahn *et al.*, 2016), infSCITE (Kuipers *et al.*, 2017b) or SiFit (Zafar *et al.*, 2017).

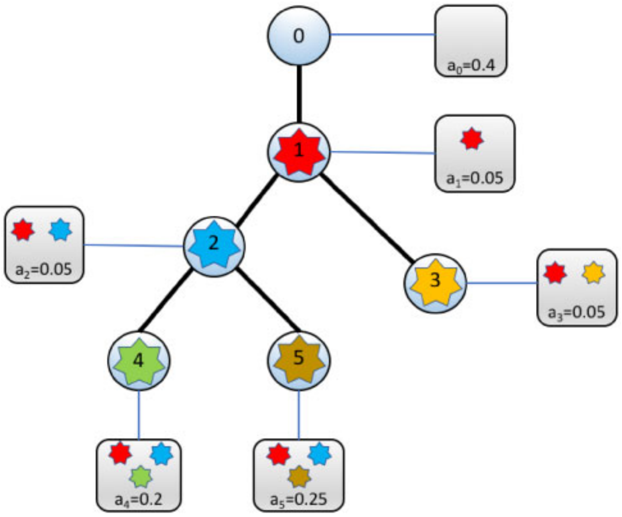

Fig. 1. Mutation tree

Formally, we consider the following algorithmic problem. Given are:

- mutation tree T with $n+1$ leafs corresponding to clonal variants. We assume that internal nodes of T are labeled 0, 1,. . ., n and the ith clone is attached to the node i. The root of T correspond to the mutation 0, which represent absence of somatic mutations or healthy tissue.
- observed relative abundances $\mathcal{A} = (a_0, \dots, a_n)$ of clones.
- Mean cancer cells mutation rate θ. This is a well-studied parameter with estimations provided by prior studies (Hao *et al.*, 2016).

The goal is to find fitnesses $\mathcal{F} = (f_0, \dots, f_n)$ maximizing the likelihood

$$p(\mathcal{A}|T, \mathcal{F}, \theta) \qquad (1)$$

This section is organized as follows. First, we introduce our evolutionary model of choice and the definition of the probability (1). Next, we describe how the likelihood is modified to transform the maximum likelihood problem (1) into a discrete optimization problem. Finally, we describe the method of estimation of fitnesses $\mathcal{F}$ maximizing (1).

2.1 Evolutionary model

We consider tumor evolution as a branching process described by the mutation tree T. Let $V(T)$, $V_I(T)$ and $E(T)$ be the node set, the internal node set and an the arc set of T, respectively. Let also p_i denote the parent of a node $i \in V_I(T)$. We assume that nodes $V_I(T)$ represent mutation events, with jth event occurring at rate θ_j. The mutation event corresponding to a node i happens at time t_i; at the event the clonal variant corresponding to the parent node p_i gives birth to a variant i. The dynamics of the cancer clone population is described by the *piecewise continuous* function $x = (x_0, \dots, x_n)$, where $x_i = x_i(t)$ is the relative abundance of the ith clonal variant. The discontinuity points of x correspond to mutation events. Let r, i, j be three consecutive mutation events with times $t_r < t_i < t_j$, and $x_k^{(i)}$ be the restriction of x_i to the interval $[t_i, t_j]$. Between mutation events i and j clonal frequencies $x_k^{(i)}$ follow the system of ODEs (Nowak and May, 2000):

$$\frac{d}{dt}x_k^{(i)} = f_k x_k^{(i)} - x_k^{(i)} \sum_{l=0}^{n} f_l x_l^{(i)}, \quad k = 0, \dots, n \tag{2}$$

with initial conditions

$$x_k^{(i)}(t_i) = \begin{cases} \varepsilon x_{p_i}^{(r)}(t_i), & \text{if } k = i \\ (1-\varepsilon)x_k^{(r)}(t_i), & \text{if } k = p_i \\ x_k^{(r)}(t_i), & \text{otherwise.} \end{cases} \tag{3}$$

Subtraction of the term $x_k^{(i)} \sum_{l=1}^{n} f_l x_l^{(i)}$ ensures that relative abundances of variants sum up to 1. Initial conditions (3) link clone abundances before and after the mutation event i and indicate that at time t_i the clone i is generated by the clone p_i. The parameter $\varepsilon \ll 1$ is a small number. At time 0, the root clonal variant (healthy tissue) gives birth to the first mutation, with the corresponding clones having relative abundances $1 - \varepsilon$ and ε. The model (2) is a branching-type variant of the quasispecies model, which is applicable to cancer evolution (Wodarz and Komarova, 2005) and agrees or extends several classical population genetics concepts (Wilke, 2005), including those describing genetic systems governed by mutation and selection (Kimura and Maruyama, 1966; Moran, 1976). It does not include specific assumptions about clonal competition or cooperation.

2.2 Likelihood definition

In addition to n mutation events, we consider the $(n+1)$th event representing cell sampling. Suppose that times of mutation events $\Omega = (t_i)_{i=1}^{n+1}$ and mutation rates between events $\Theta = (\theta_i)_{i=1}^{n}$ are given. Let $\sigma = (\sigma_1, \dots, \sigma_{n+1})$ be the permutation of events in order of their appearance, i.e. $0 = t_{\sigma_1} < t_{\sigma_2} < \dots < t_{\sigma_n} < t_{\sigma_{n+1}}$. The probability of observing abundances $\mathcal{A}$ given T,$\mathcal{F}$, Ω, Θ and θ is defined as the product of probabilities of mutation events and probabilities of observed clone abundances.

The mutation event in the vertex σ_j occurs if two conditions are met:

a. no mutation events have been observed over the time interval $(t_{\sigma_{j-1}}, t_{\sigma_j})$;
b. at time t_{σ_j} the mutation happened in the clone p_{σ_j} rather than in other clones which exist at that time.

Appearance of mutation is a classical rare event, and therefore we assume that the time intervals between consecutive mutation events i and j follow a Poisson distribution with the mean $\frac{1}{\theta_i}$. Mutation rates are distributed normally with the mean θ and the standard deviation ν. Assuming that mutations are random, the probability of (b) is equal to the frequency $x_{p_{\sigma_j}}(t_{\sigma_j})$ of the clone p_{σ_j} at time t_{σ_j} according to the system (2). Finally, we assume that the probability of seeing observed frequencies given model-based frequencies at the sampling time follows a multinomial distribution $\mathcal{M}(a_0, \dots, a_n | x_0(t_{n+1}), \dots, x_n(t_{n+1}))$. After putting all probabilities together, we have

$$\begin{aligned} p(\mathcal{A}|T, \mathcal{F}, \Omega, \theta) = \prod_{j=2}^{n+1} Pois\left(t_{\sigma_j} - t_{\sigma_{j-1}}, \frac{1}{\theta_{j-1}}\right) \cdot \prod_{j=1}^{n} \mathcal{N}(\theta_j, \theta, \nu) \cdot \\ \prod_{j=1}^{n} x_{p_j}(t_j) \times \mathcal{M}(a_0, \dots, a_n | x_0(t_{n+1}), \dots, x_n(t_{n+1})) \end{aligned} \tag{4}$$

Our goal is to find best fitting fitnesses $\mathcal{F}_{ML}$, rates Θ_{ML} and times Ω_{ML} by solving the following maximum likelihood problem:

$$(\mathcal{F}_{ML}, \Theta_{ML}, \Omega_{ML}) = \arg\max_{\mathcal{F}, \Theta, \Omega} p(\mathcal{A}|T, \mathcal{F}, \Omega, \theta) \tag{5}$$

The probabilities $\prod_{j=2}^{n+1} Pois(t_{\sigma_j} - t_{\sigma_{j-1}}, \frac{1}{\theta_{j-1}})$, $\prod_{j=1}^{n} \mathcal{N}(\theta_j, \theta, \nu)$, $\prod_{j=1}^{n} x_{p_{\sigma_j}}(t_{\sigma_j})$ and $\mathcal{M}(a_0, \dots, a_n | x_0(t_{n+1}), \dots, x_n(t_{n+1}))$ are further referred to as *time likelihood*, *rate likelihood*, *mutation likelihood* and *abundance likelihood*, respectively. For the tree shown in Figure 2, it is equally feasible that the mutation 2 appeared before the mutation 3 or vice versa. However, clone 2 later produces mutations 4 and 5, and therefore the mutation likelihood suggests that at that mutation events it had high abundance. This situation is probable if either 2nd mutation appeared earlier or it appeared later but has a high fitness. Time, rate and abundance likelihoods allow to choose between these two alternatives.

2.3 Reduction to discrete optimization

The standard way to solve the maximum likelihood problem (5) is to optimize $\mathcal{F}$, Θ and Ω jointly using Markov Chain Monte Carlo (MCMC) sampling. However, our experiments have shown that the function (1) has too many local optima which makes MCMC search over the continuous space of possible solutions inefficient. Therefore, we suggest an alternative heuristic approach, which transforms the problem (5) into a discrete optimization problem akin to a scheduling problem. This problem is then solved using a specifically designed combinatorial heuristic search.

First, we assume that all fitnesses are relative with respect to a fitness of a clone 0 which is set to be $f_0 = 1$. By default, this clone corresponds to the normal tissue. For the problem of inference of clonal selection such assumption does not restrict the predictive power. Next, we observe that any assignment of event times Ω defines the order of appearance μ_i for each node $i \in V(T)$ (e.g. in

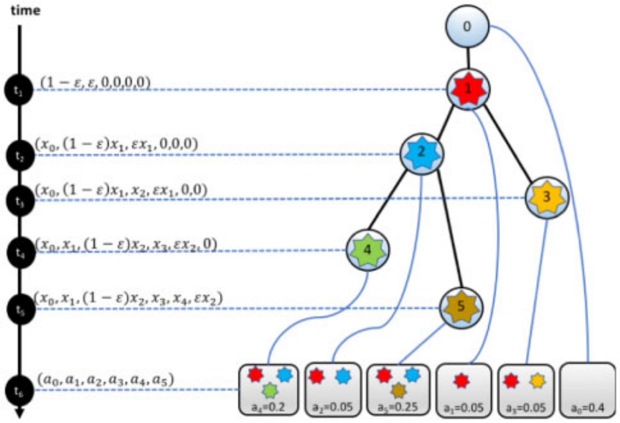
time
0
1
2
3
4
5
t1
t2
t3
t4
t5
t6
$(1-\varepsilon, \varepsilon, 0,0,0,0)$
$(x_0, (1-\varepsilon)x_1, \varepsilon x_1, 0,0,0)$
$(x_0, (1-\varepsilon)x_1, x_2, \varepsilon x_1, 0,0)$
$(x_0, x_1, (1-\varepsilon)x_2, x_3, \varepsilon x_2, 0)$
$(x_0, x_1, (1-\varepsilon)x_2, x_3, x_4, \varepsilon x_2)$
$(a_0, a_1, a_2, a_3, a_4, a_5)$
a4=0.2
a2=0.05
a5=0.25
a1=0.05
a3=0.05
a0=0.4

Fig. 2. Depiction of the evolutionary model. Tree nodes represent mutation events whose times are marked on the time axis. Leafs represent the sampling event. For each node, the distribution of clone abundances after the corresponding event is shown

Fig. 2 $\mu_i = i$ for $i = 1, \dots, 5$). This order agrees with the natural vertex order induced by T, i.e. $\mu_i < \mu_j$ whenever i is an ancestor of j. It turned out that conversely any order μ defines times Ω^μ, rates Θ^μ and fitnesses $\mathcal{F}^\mu$ which maximize the partial likelihood

$$\prod_{j=2}^{n+1} Pois\left(t_{\sigma_j} - t_{\sigma_{j-1}}, \frac{1}{\theta_{j-1}}\right) \cdot \prod_{j=1}^{n} \mathcal{N}(\theta_j, \theta, \nu) \cdot \\ \times \mathcal{M}(a_0, \dots, a_n | x_0(t_{n+1}), \dots, x_n(t_{n+1})) \tag{6}$$

More precisely, the following proposition holds.

Proposition 1. *For a given order vector μ, times Ω^μ, rates Θ^μ and fitnesses $\mathcal{F}^\mu$ maximizing (6) can be estimated as follows:*

$$\theta_i = \theta, \;\; t_i = \frac{\mu_i - 1}{\theta}, i = 1, \dots, n, \;\; t_{n+1} = \frac{n}{\theta} \tag{7}$$

$$f_i = 1 - \theta \sum_{j \in A_i \setminus \{0\}} \frac{1}{n - \mu_j + 1} \log\left(\frac{\varepsilon}{1-\varepsilon} \frac{a_{p_j}}{a_j}\right), i = 1, \dots, n. \tag{8}$$

Here, A_i is the set of ancestors of a node i (including itself).

Proof. Poisson and Gaussian probabilities achieve maximums at their means, i.e. the rate and time likelihoods are maximal, when for consecutive events i, j we have $\theta_i = \theta$, $t_j - t_i = \frac{1}{\theta}$. This yields the solution (7). The multinomial probability $\mathcal{M}(a_0, \dots, a_n | x_0(t_{n+1}), \dots, x_n(t_{n+1}))$ is maximal when $x_i(t_{n+1}) = a_i$ for all $i \in [n]$. This can be rewritten as

$$\frac{x_i(t_{n+1})}{x_i(t_{n+1}) + x_{p_i}(t_{n+1})} = \frac{a_i}{a_i + a_{p_i}} \quad \text{for all } i = 1, \dots, n. \tag{9}$$

Our goal is to find fitnesses $\mathcal{F}$ such that (9) holds. We find an approximate solution to this problem by disregarding the discontinuity of the abundances $x = (x_i(t))_{i=0}^{n+1}$. We use the observation that the system (2) is invariant with respect to the transition to relative abundances of any pair of clones. Namely, for each clone pair $i, j = 0, \dots, n$ dynamics of their relative abundances with respect to each other $y_i = \frac{x_i}{x_i + x_j}$ and $y_j = \frac{x_j}{x_i + x_j}$ is described by the system of ODEs of the same form as (2):

$$\begin{aligned} \dot{y}_i &= f_i y_i - y_i(f_i y_i + f_j y_j), \\ \dot{y}_j &= f_j y_j - y_j(f_i y_i + f_j y_j), \end{aligned} \tag{10}$$

On the interval $[t_i, t_{n+1}]$ relative abundance $y_i = \frac{x_i}{x_i + x_{p_i}}$ satisfy the system (10) with the initial condition $y_i(t_i) = \varepsilon$. After shifting time interval to $[0, t_{n+1} - t_i]$, this system can be linearized and solved in closed form, producing a solution

$$y_i(t) = \frac{\varepsilon e^{f_i t}}{(1-\varepsilon) e^{f_{p_i} t} + \varepsilon e^{f_i t}} \tag{11}$$

After putting the expressions (11) into the equation (9) with $t = t_{n+1} - t_i$, we get the following system of equations to find fitnesses $\mathcal{F}$:

$$f_{p_i} - f_i = \frac{1}{t_{n+1} - t_i} \log\left(\frac{\varepsilon}{1-\varepsilon} \frac{a_{p_i}}{a_i}\right), i = 1, \dots, n; \;\; f_0 = 1 \tag{12}$$

Solving it with t_i described by (7) yields the solution (8). □

Algorithm 1. Algorithm for node ordering

```
Let $U$ be the list of nodes of $T$ sorted in inverse order of their discovery by Breadth First Search from the root; $T' = T$;
for $u \in U$ do
  while $u$ has more than 1 child do
    Choose sibling paths $P_1$ and $P_2$ with the start node $u$
    Join $P_1$ and $P_2$ into a single path $P$ using Algorithm 2
    Modify $T'$ by replacing $P_1$ and $P_2$ by $P$
  end while
end for
```

Algorithm 2. Algorithm for path joining

```
Input Sibling paths $P_1$ and $P_2$
Output is calculated by calling MergePaths($\phi$,1)
  MergePaths($Y$, $i$)
    ▷ $Y$ is the current $k$-subset, $i$ is the next element to be added to it
     ▷ $\mu_{opt}$ and $opt$ are the current optimal order and its likelihood
if $|Y| = k$ or $i > k + l$ then
  return
end if
$Y_{new} = Y \cup \{i\}$, $\mu' = \mu_{Y_{new}}$
while $\mu'$ is not a total order do
  $w_1 = P_1^Y(1)$, $w_2 = P_2^Y(1)$, $j = |\mu'| + 1$
  $t = \frac{j-1}{\theta}$, $f_{w_1} = f_{p_{w_1}} + \frac{1}{t_{n+1} - t} \log\left(\frac{\varepsilon}{1-\varepsilon} \frac{a_{p_{w_1}}}{a_{w_1}}\right)$,
  $f_{w_2} = f_{p_{w_2}} + \frac{1}{t_{n+1} - t} \log\left(\frac{\varepsilon}{1-\varepsilon} \frac{a_{p_{w_2}}}{a_{w_2}}\right)$
  if $f_{w_1} \leq f_{w_2}$ then
    $\mu' = \mu' \cup \{w_1\}$, $P_1^Y = P_1^Y \setminus \{w_1\}$
  else
    $\mu' = \mu' \cup \{w_2\}$, $P_2^Y = P_2^Y \setminus \{w_2\}$
  end if
end while
MergePaths ($Y$, $i + 1$)
if $L_{\mu_Y} > opt$ then
  $opt = L_{\mu'}$, $\mu_{opt} = \mu'$
  MergePaths ($Y_{new}$, $i + 1$)
end if
```

Using Proposition 1, we replace the maximum likelihood problem (5) with the following discrete problem: find the ordering μ maximizing the mutation log-likelihood

$$L_\mu = \log(p(\mu)) = \sum_{j=1}^{n} \log(x_{p_j}(t_j)) \tag{13}$$

with times Ω^μ and fitnesses $\mathcal{F}^\mu$ described by (7), (8) subject to the constraint that μ agrees with the ancestral-descendant order of T.

2.4 Finding optimal ordering

The problem (13) could be considered as a variant of scheduling problem with precedent constraints and with non-linear cumulative cost function (Dolgui *et al.*, 2012). Here, mutations play roles of jobs, ordering of mutations corresponds to scheduling of jobs on a single processor, mutation tree represent job precedence constraints and the objective (13) indicates that the cost of job processing depends on the previously processed jobs. Such problems are usually NP-hard (Dolgui *et al.*, 2012). For small number of mutations, it can be solved by a branch-and-bound search in the space of feasible orderings via backtracking over the mutation tree. In general, we solve it by a heuristic approach combined with the search in the space of feasible sub-orderings of nodes of the mutation tree T. The proposed scheme is described by Algorithm 1. The algorithm starts with the initial tree $T' = T$ and iteratively transforms it into a total order as follows. We call two simple paths of T' *sibling paths*, if they share the starting vertex. We traverse the nodes of T' in a bottom-up direction and merge sibling paths into one path representing optimal sub-order of their nodes with respect to the objective (13). The algorithm stops when all nodes form a single path.

Merging of sibling paths P_1 and P_2 is performed by Algorithm 2. We note that feasible orders of paths' nodes bijectively correspond to k-subsets of the set $[k+l]$: for a given k-subset X, a feasible order μ_X is obtained by placing nodes from $P_1 \setminus \{u\}$ (resp., $P_2 \setminus \{u\}$) at positions from X (resp., $[k+l] \setminus X$) in order of their appearance in P_1 (resp., P_2); inverse is also true. Algorithm 2 recursively generates k-subsets via branching and prune branches, if the corresponding orders are likely to be sub-optimal.

The k-subsets are generated recursively (Nijenhuis and Wilf, 2014) using the property that every k-subset X of $[k+l]$ is either k-subset of the set $[2:k+l]$ or has the form $X = \{1\} \cup Y$, where Y is a $k-1$-subset of $[2:k+l]$. Suppose that at a given iteration a partial k'-subset Y, $k' \leq k$, and the corresponding pre-order μ_Y has been constructed. For all nodes v covered by μ_Y, we calculate their appearance times t_v and fitnesses f_v using (7), (8), and abundance distributions $x^v = (x_0(t_v), \ldots, x_n(t_v))$ from the system (2)–(3) (in fact, it is not necessary to recalculate all values since some of them has been already calculated at previous iterations). Next, we heuristically extend μ_Y to a total order as described below. If the likelihood of the constructed solution is below the current optimum, then the recursion tree branch of the partial solution Y is pruned. Otherwise, the current optimum is updated and the recursion continues.

Finally, we describe how an order μ_Y is extended (lines 5–14 of Algorithm 2). We consider the subpaths P_1^Y and P_2^Y formed by the nodes of P^1 and P^2 that are not covered by μ_Y. For the first nodes of these subpaths, we calculate their provisional fitnesses under the assumption that each node is added to μ_Y as the next element. The node with the smaller provisional fitness is added to μ_Y. This procedure is repeated until μ_Y covers all nodes. The logic behind this approach is based on the observation that according to (2) the frequency of a clone grows while its fitness is larger than the average fitness of the population, and declines otherwise. For a given iteration, adding clone with a smaller fitness slows down the average fitness growth. As a result, for preceding clones probabilities of appearances of their children in the future may become higher.

3 Results

3.1 Simulated data

We simulated 100 test examples with the numbers of mutations ranging from $m=30$ to $m=120$, which correspond to numbers of mutations for real scSeq data analyzed in previous studies (Jahn *et al.*, 2016; Kuipers *et al.*, 2017a; Leung *et al.*, 2017). For each test example, clonal evolution was simulated as follows. (a) Mutations 1,..., m are generated randomly. For the time interval between mutation events i and $i+1$ the current mutation rate θ_i is sampled from the normal distribution with the mean $\theta=0.01$ and standard deviation $\sigma \in \{0.1 \cdot \theta, 0.5 \cdot \theta, 0.9 \cdot \theta\}$. At each moment of time of that interval a mutation event happens with the probability θ_i; at the event a random clone p selected with the probability equal to its current relative abundance gives birth to a new clone j with the random fitness f_j by acquiring a random mutation $i+1$. In our primary fitness sampling scheme, new fitness is sampled uniformly from the interval $[\phi, f_{max}]$, where ϕ is an average fitness of the population at the time of mutation event. This scheme accounts for the fact that according to the evolutionary model (2) the clone with the fitness below ϕ is not viable and will not be observed at sampling time. In additional set of experiments, the secondary sampling scheme has been employed, when new fitness is sampled uniformly from the interval $[f_{min}, f_{max}]$ (by default $f_{min} = 1$, $f_{max} = 1.2$). When there is no mutation event, abundances of existing clones are updated according to (2). After the end of the simulation, final abundances were randomly perturbed by 10% to incorporate the possible noise in their estimation. The simulated mutation tree and clone abundances were used as an input for SCIFIL.

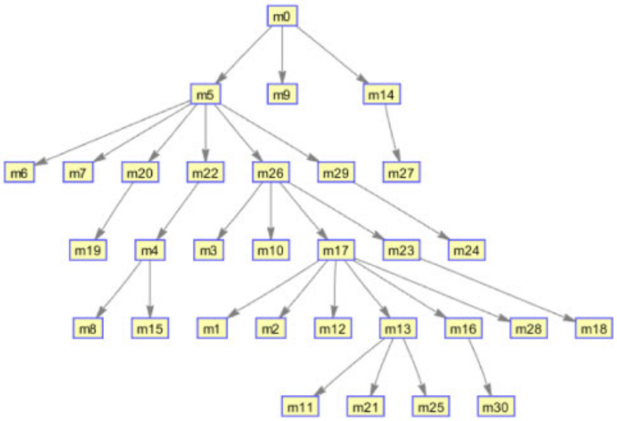

Fig. 3. Example of simulated mutation tree

It should be noted that the construction of the proposed algorithm implies that its performance would be higher on mutation trees with monoclonal structure, both in terms of speed and accuracy. However, our simulation scheme predominantly produces trees with polyclonal structures (see Fig. 3), thus providing no a priori advantage to SCIFIL.

We quantified the performance of SCIFIL using two measures:

1. Mean relative accuracy $MRA = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{|f_i^*-f_i|}{f_i^*}$, where f_i^* and f_i are true and inferred fitnesses, respectively.
2. Spearman correlation SC between true and inferred fitnesses.

MRA and *SC* highlight different aspects of the problem. MRA measure the accuracy of fitness value estimation, while SC measures how well we are able to qualitatively detect selective advantage of particular clones over other clones. Fitness ranking can be used in evolutionary studies even when actual fitness values are missing or inaccurate (Crona *et al.*, 2017).

The results of SCIFIL evaluation on simulated data are shown in Figures 4 and 5. The algorithm demonstrated high accuracy as measured by both parameters. The number of mutations (Fig. 4) does not have a great impact on the Spearman correlation, which averages 97.35% (standard deviation 1.2%) over all analyzed test cases. MRA decreases when the number of mutations grows, but remains above 88% for all datasets. Increase in variation of mutation rate (Fig. 5) does not significantly affect SC, and results in slight decrease

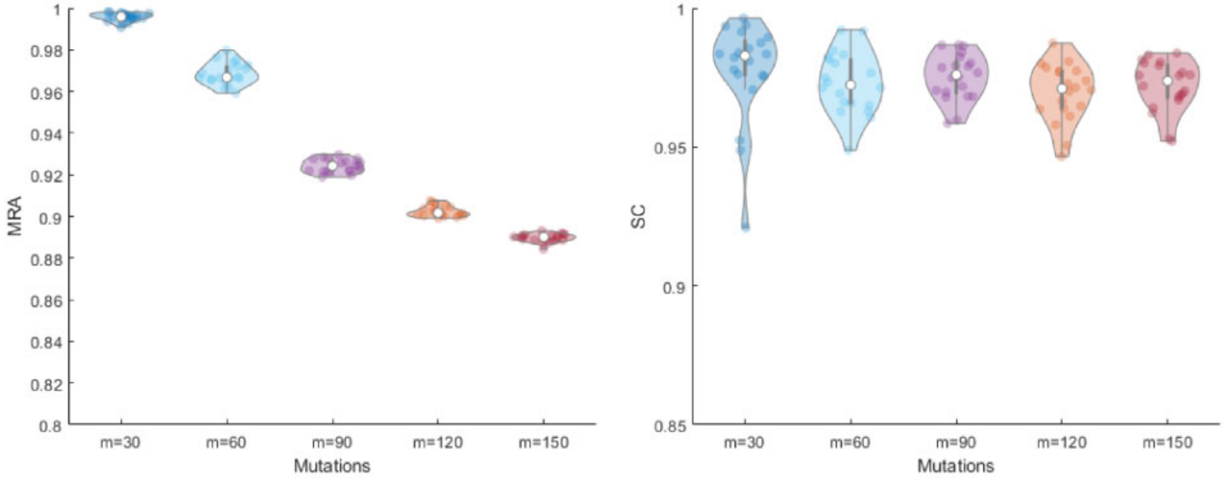

Fig. 4. Performance of SCIFIL on simulated data with *m* mutations and fixed standard deviation of mutation rate. (**Left**): Mean relative accuracy of fitness estimation. (**Right**): Spearman correlation between true and inferred fitness vectors

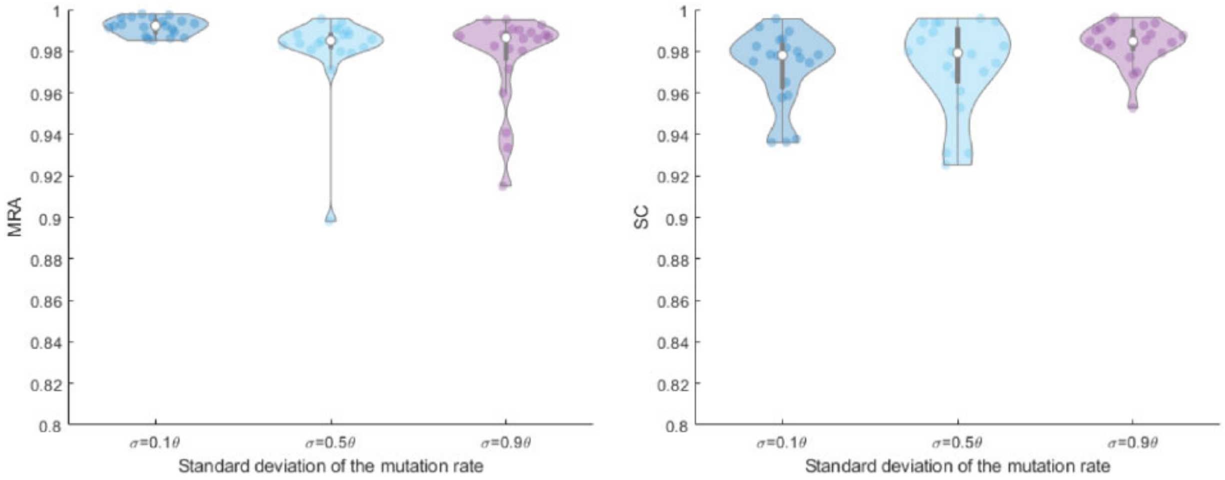

Fig. 5. Performance of SCIFIL on simulated data with $m = 50$ mutations and different standard deviations of mutations rates. (**Left**) Mean relative accuracy of fitness estimation. (**Right**) Spearman correlation between true and inferred fitness vectors

of average MRA. Relative robustness of SCIFIL to the variation of mutation rates (which also introduce variation in mutation times) indirectly suggests, that the proposed algorithm is able to well approximate the original maximum likelihood problem (4). In the case of near-neutral selection ($f_{max} = 1.01$), MRA does not significantly change and SC declines to 87.54%.

Additionally, we have compared SCIFIL output with the topology of input mutation trees to evaluate the contribution of the tree-based prior information to the algorithm's accuracy. Specifically, the clones have been ranked by their estimated fitnesses and by their tree heights, and Spearman correlation SC^T between fitnesses and tree ranks have been calculated (combined with the permutation test to account for the presence of clones of the same rank). The experiment has been repeated two times using the primary and secondary fitness sampling schemes, with the latter being a completely random uniform sampling from the constant interval. For the first sampling scheme, the average correlation between fitness and tree ranks was $SC_1^T = 0.698$ (with $SC = SC_1 = 0.969$). For the second sampling scheme, SC^T drops to $SC_2^T = 0.314$, while the correlation between real and estimated fitnesses decreases to $SC_2 = 0.871$. The value $\tau = 100 \cdot \frac{SC_1 - SC_2}{SC_1^T - SC_2^T}$ (decrease in accuracy per one percent decrease in tree/fitness correlation) may serve as a measure of contribution of a tree topology to the SCIFIL quality. In our case, this value is equal to 25.7%. Transition to near-neutral selection ($f_{max} = 1.01$) has the similar effect, with the correlations being $SC = 0.875$ and $SC^T = 0.379$.

ScSeq data are prone to errors. To evaluate SCIFIL's robustness to trees inferred from noisy data, random errors were introduced to clone mutation profiles at false negative rates $\alpha \in \{0.1, 0.2\}$ and the false positive rate $\beta = 10^{-5}$, and mutation trees were reconstructed from these profiles using the state-of-the-art tool SCITE (Jahn *et al.*, 2016). The simulated/reconstructed mutation trees were used as an input for SCIFIL. It turned out that in ~8% of cases SCIFIL was not able to produce a feasible solution. This issue could be resolved by performing several additional steps of the local search with the same tree modification operations as SCITE and with the objective (4). With this modification, SCIFIL reconstructs fitnesses accurately, although, as expected, the accuracy decreases with the error rate's growth (Fig. 6). To check the influence of undersampling, we assumed that $\gamma = 10\%$ of clones with lowest frequencies were not observed at the sampling time. For such clones, the auxiliary frequency $\delta \ll \varepsilon$ has been assigned before running SCITE. For $m = 50$,

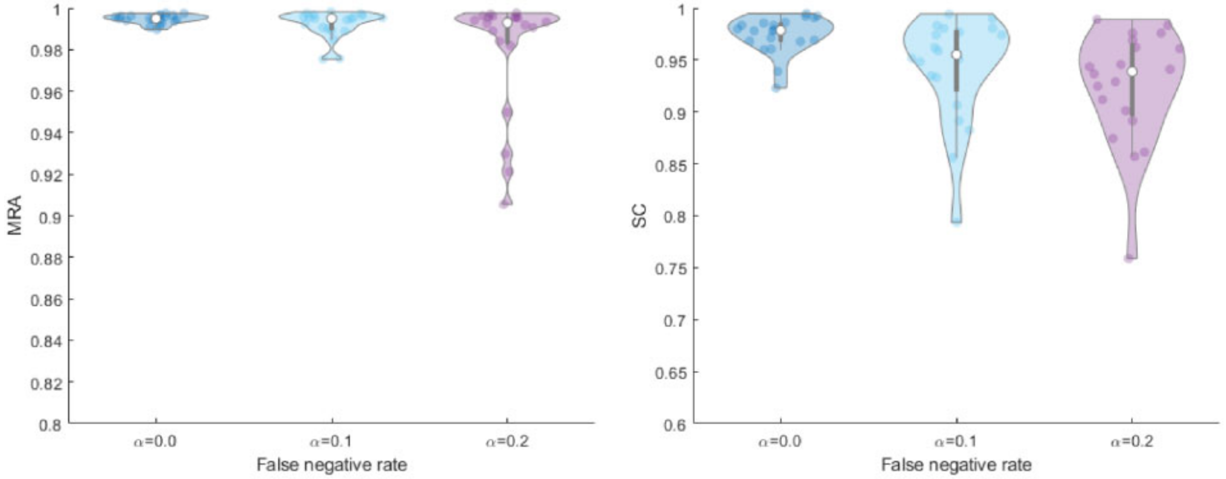

Fig. 6. Performance of SCIFIL on simulated data with different false negative error rates α and with mutation trees reconstructed by SCITE (Jahn *et al.*, 2016). (**Left**) Mean relative accuracy of fitness estimation. (**Right**) Spearman correlation between true and inferred fitness vectors

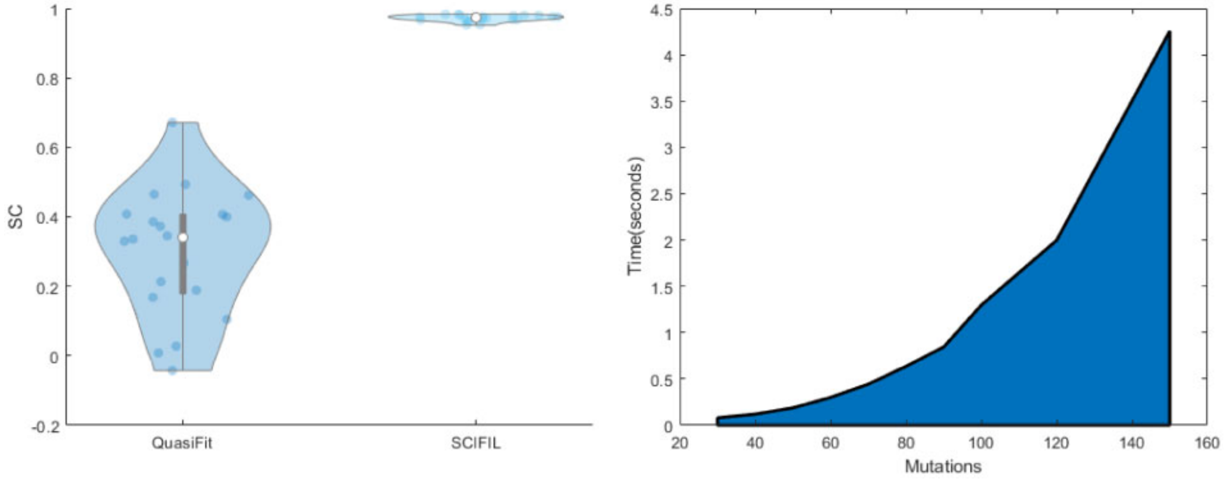

Fig. 7. (**Left**) Spearman correlation between true and inferred fitness vectors for QuasiFit and SCIFIL. (**Right**) Running time of SCIFIL

the average MRA decreased from 0.99 to 0.96 in comparison to the complete data, but SC remained stable (0.972 and 0.968, respectively).

Finally, we compared our approach with the previously published tool QuasiFit (Seifert *et al.*, 2014). Although originally designed for viruses, QuasiFit is based on quasispecies model, which is applicable to both intra-host viral populations and cancer clone populations (Wodarz and Komarova, 2005) and is essentially a fully continuous version of the model used by SCIFIL. Both QuasiFit and SCIFIL reconstruct replicative fitnesses of individual clones (rather than alleles). In addition to genomic data, both algorithms utilize other information: SCIFIL uses a mutation tree, while QuasiFit assumes that the population is in equilibrium state of the quasispecies model. Thus, SCIFIL has access to information about partial clones order encoded by the mutation tree, while equilibrium site assumption allows QuasiFit to eliminate from consideration the temporal component. Furthermore, SCIFIL is a discrete optimization approach, while QuasiFit implements MCMC sampling.

QuasiFit was run with the per-cell mutation rate $\mu = \varepsilon\theta$ (which is a fully continuous analog of the parameters used by SCIFIL) and fitnesses were estimated after a burn-in of 10^5 iterations. As QuasiFit uses a different fitness vector normalization, following Seifert *et al.* (2014), we used only the parameter *SC* for the comparison. The results are shown in Fig. 7(left). On our simulated data, SCIFIL outperforms QuasiFit indicating that in certain settings the proposed model could be more accurate for the inference of clonal selection than the equilibrium state assumption.

Computational experiments suggest that the algorithm's running time scales quadratically with the number of mutations (Fig. 7, correlation = 0.981). It allows SCIFIL to finish in a few seconds for all analyzed datasets when run on a simple desktop computer.

3.2 Experimental data

3.2.1 Fitness landscapes

We used SCIFIL to infer fitness landscapes for two recently published experimental cancer datasets. The first dataset is scSeq data from a JAK2-negative myeloproliferative neoplasm (essential thrombocythemia) (Hou *et al.*, 2012), the second one represents metastatic colon cancer (Leung *et al.*, 2017). The latter dataset includes SNVs sampled from the main tumor and two metastases. We confined our analysis only to the primary tumor, since it is biologically meaningful to compare fitnesses of clones sampled from the same environment. For both datasets, their mutation trees were reconstructed using SCITE (Jahn *et al.*, 2016), and fitnesses and

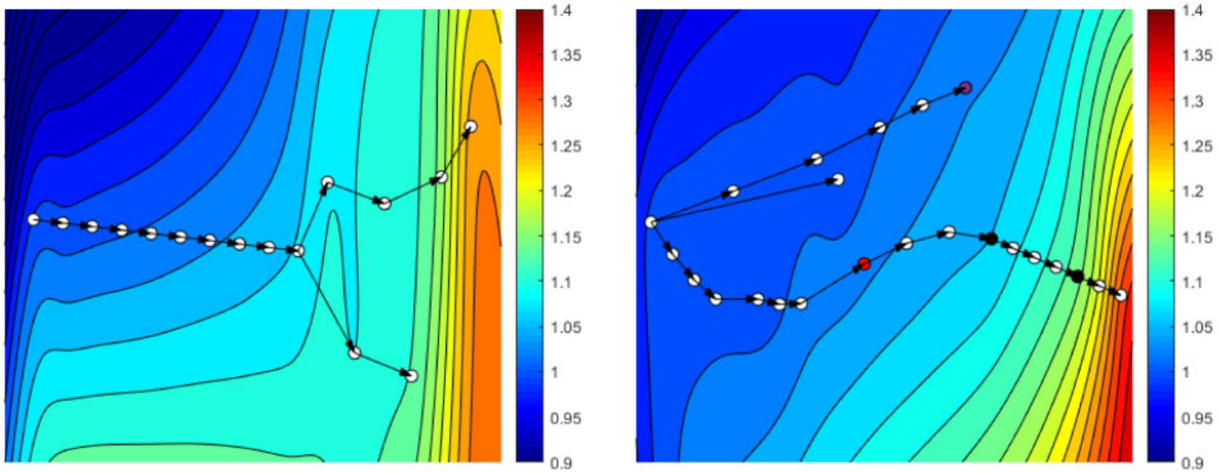

Fig. 8. Fitness landscape and mutation tree for JAK2-negative myeloproliferative neoplasm (Hou *et al.*, 2012) (**left**) and colorectal cancer (**right**) (Leung *et al.*, 2017) inferred by SCIFIL. Colors represent fitness values and distance from each tree node to the root is approximately proportional to its time of appearance

mutation appearance times were inferred by SCIFIL with the cell-wise mutation rate 10^{-6}. It is important to note that varying SCIFIL parameters may change absolute values of inferred fitnesses, but preserve relations between them. The relations are the most informative factors for evolutionary analysis.

We visualized inferred fitness landscapes as follows. We calculated pairwise distances between clones defined as the sum of their hamming distance and the absolute difference of their orders of appearance. The distances were used to map clones to the plain $\mathbb{R}^2$ using multidimensional scaling. Fitness values of the points corresponding to clones were interpolated using biharmonic splines, and the resulting surface was visualized as a contour plot (Fig. 8), where colors represent fitness values, and distance from each tree node to the root reflects its appearance time.

For myeloproliferative neoplasm (Fig. 8, left) we observe linear accumulation of mutations with slight selective advantages at the beginning of tumor evolution, followed by the subclone expansion of two lineages with significantly faster fitness growth. The rate of fitness growth after the branching event is ~3 times higher than before it. Thus, answering the question posed in Jahn *et al.* (2016), we may predict that recent subclones will replace ancestral clones. However, based on the available information it is hard to decide whether one of the subclone lineages will out-compete the other one, or they will continue to coexist.

Evolution of the colon tumor (Fig. 8, right) follows different scenarios, with three independent lineages co-existing at the beginning without a clear selective advantage enjoyed by any of them. This stage is followed by the fast expansion of one of the lineages, which climbs a fitness peak and acquires selective advantage over other lineages. Exactly, at this stage, the advantageous lineage seeded the metastatic tumor at two seeding events (highlighted in black in Fig. 8).

Experimental data also allow to emphasize how SCIFIL estimations extend predictions implied by the underlying evolutionary model. Although the model suggests positive selection with fitness growth along each path of the mutation tree as the most probable scenario, it does not imply any restrictions on the comparative fitnesses of different lineages. In particular, fitness advantages of clones are not defined only by their distances from the root, as emphasized by the fitness landscape of the colon tumor, where, for instance, the node highlighted in purple has higher fitness than the node highlighted in red. The reason is that clone abundances contribute to the estimation of fitness values as much as the evolutionary model and the topology of mutation tree.

3.1.2 Recurrent mutations

Until recently, most studies of tumor evolution utilized *infinite sites assumption*, which states that every genomic position mutates at most once over the evolutionary history. However, recently it has been demonstrated using ScSeq data, that the infinite site assumption could be violated, with the same genomic positions mutationally affected multiple times over the tumor evolution (Kuipers *et al.*, 2017b). Without infinite site assumption, the number of possible alternative evolutionary histories accurately explaining the observed ScSeq data increases, and it becomes challenging to choose the most appropriate one.

We utilized SCIFIL for the analysis possible evolutionary histories with recurrent mutations for a JAK2-negative myeloproliferative neoplasm (Hou *et al.*, 2012). We used infSCITE (Kuipers *et al.*, 2017b) to generate the perfect phylogeny and 18 mutation trees T_{m_i} under the assumption that one of 18 mutations m_i has a recurrence (*recurrence trees*). Just as reported in Kuipers *et al.* (2017b), the results strongly support recurrent mutations: the average log-likelihood for recurrence trees produced by infSCITE in our experiments was −313.45 (standard deviation 1.065), while the log-likelihood of the perfect phylogeny was equal to −319.08 (Fig. 9). However, differences between log-likelihoods of recurrence trees were small in comparison to their difference with the one of the perfect phylogeny, thus impeding the reliable selection of the single most likely recurrence tree. To choose such tree, we utilized evolutionary likelihood estimated by SCIFIL. Among 18 trees, only two have evolutionary likelihoods higher than for the perfect phylogeny (Fig. 9). Notably, the log-likelihood of the tree T_{ASNS} is significantly higher than for other recurrence trees (−518.62 versus −674.696 in average (standard deviation 25.62)), thus providing the strong support for that particular evolutionary history with respect to other possible histories. These results indicate that SCIFIL's can be efficiently used in conjunction with infSCITE or other similar tool for detection of the most probable evolutionary scenarios.

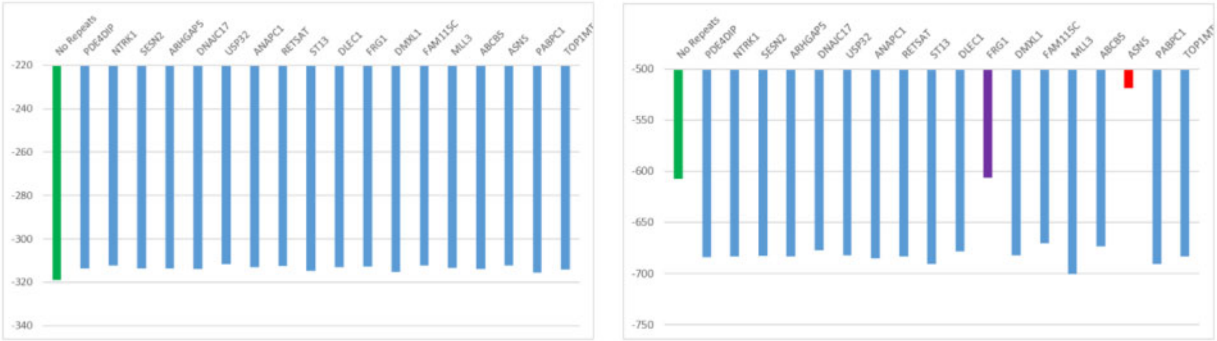

Fig. 9. Log-likelihoods of trees with and without recurrent mutations. (**Left**) Log-likelihoods produced by infSCITE. (**Right**) Evolutionary likelihoods produced by SCIFIL. Likelihoods of perfect phylogeny are shown in green. Purple and red: trees with the evolutionary likelihoods higher than for the perfect phylogeny

4 Discussion

Intra-tumor heterogeneity is one of the major factors influencing cancer progression and treatment outcome. Cancer clones form complex populations of genomic variants constantly evolving to compete for resources, proliferate, metastasize and escape immune system and therapy. Quantification of clonal selection for tumors may provide valuable information for understanding mechanisms of disease progression and for design of personalized treatment. scSeq provides an unprecedented insight into intra-tumor heterogeneity allowing to study fitness landscapes at finest possible resolution and quantify selective advantages on the level of individual clones.

In this paper, we presented SCIFIL, a likelihood-based method for inference of fitnesses of clonal variants. Unlike other available methods for related problems, SCIFIL takes full advantage of the information about structure and evolutionary history of clonal population provided by scSeq. It uses individual cells as evolutionary units, in contrast to the tools based on bulk sequencing which perform their analysis on the level of subpopulations or lineages. Furthermore, SCIFIL can also handle bulk sequencing data as long as clones are reconstructed and mutation tree is constructed using available tools such as AncesTree (El-Kebir *et al.*, 2015), PhyloSub (Jiao *et al.*, 2014) and CITUP (Malikic *et al.*, 2015).

In contrast to previous approaches, SCIFIL employs dynamic evolutionary model rather than assumption that the population achieved the equilibrium state. We have demonstrated that our approach allows for accurate inference of fitness landscapes and can be used for analysis of evolutionary history and clonal selection for real tumors. We envision that SCIFIL can be also used to infer epistasic interactions and to identify combinations of mutations driving the tumor growth. In addition, it can be applied to other highly mutable heterogeneous populations, such as viral quasispecies or bacterial communities.

The proposed approach has limitations which should be addressed in the future work. Fitness is not defined by the genetic composition alone and depends on the environment. Thus SCIFIL quantitative predictions are more reliable when the analyzed clones are sampled from the same tumor. Fitness inference relies on the observed clone abundances, and therefore significant inaccuracies in abundance estimation may affect accuracy of fitness reconstruction. For single-cell data, it is particularly important owing to its susceptibility to allelic dropouts and PCR bias. However, this problem can be addressed by using a combination of bulk and scSeq data. There exist a plethora of tools which can estimate clone abundances from composite bulk and scSeq data (see, e.g. Baron *et al.*, 2016; Mukherjee *et al.*, 2018). In addition, such composite data can be employed to increase an accuracy of mutation trees reconstruction (Malikic *et al.*, 2017). We expect SCIFIL reliability to increase when it will be combined with these tools.

Another set of limitations arise from the selected evolutionary model (2). It was selected due to its generality (Wodarz and Komarova, 2005) and suitability for fitness landscape inference (Nowak, 2006). However, it has certain underlying assumptions: the mutation rates are supposed to be normally distributed, while the dynamical system (2) implies positive selection with the gradual growth of average population fitness. It should be noted that in many cases such assumptions are sufficiently realistic, and have been used in several studies to obtain valuable insights into the dynamics of tumor evolution (Bozic *et al.*, 2010; Jones *et al.*, 2008). In particular, other studies demonstrated that even a normal mutation rate is sufficient to produce significant intra-tumor heterogeneity and emphasized the relative importance of selection over both the size of the cell population and the mutation rate (Beerenwinkel *et al.*, 2007). Although equations (12) suggest that in most cases fitness growths along each path of the mutation tree, the model does not imply any restrictions on the comparative fitnesses of different lineages. Furthermore, observed relative abundances of clones are independent of the model, and their contribution to the estimated fitness values is paramount. Nevertheless, we expect that our approach can be extended by incorporating other models capturing different evolutionary scenarios, such as gradual mutation rate growth over the course of tumor evolution, and clonal competition/cooperation, as well as spatial tumor heterogeneity. It should be noted, though, that currently there is no universal evolutionary model for tumor progression. Alternative models will inevitably introduce other limitations and can be less practical for fitness estimation.

On algorithmic side, the optimization problem behind our approach can be viewed as the type of scheduling problem with precedent constraints and with non-linear objective (Dolgui *et al.*, 2012). Such problems are generally NP-hard, although the complexity of our problem is unknown. It is known that for certain simple objectives and well-structured precedence constraints (e.g. defined by series-parallel graphs) the corresponding scheduling problems are polynomially solvable (Dolgui *et al.*, 2012). For our problem precedence, constraints have the form of a tree. It gives a certain hope of existence of exact polynomial or a good approximation algorithm, although the complex objective function may keep our problem NP-hard. This question requires additional study.

Funding

This work was supported by NIH grant 5R01EB025022.

References

Baron,M. *et al.* (2016) A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Syst.*, **3**, 346–360.

Beerenwinkel,N. *et al.* (2007) Genetic progression and the waiting time to cancer. *PLoS Comput. Biol.*, **3**, e225.

Bonavia,R. *et al.* (2011) Heterogeneity maintenance in glioblastoma: a social network. *Cancer Res.*, **71**, 4055–4060.

Bozic,I. *et al.* (2010) Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. USA*, **107**, 18545–18550.

Crona,K. *et al.* (2017) Inferring genetic interactions from comparative fitness data. *eLife*, **6**, e28629.

Davis,A. *et al.* (2017) Tumor evolution: linear, branching, neutral or punctuated? *Biochim. Biophys. Acta*, **1867**, 151–161.

Deforche,K. *et al.* (2008) Estimation of an in vivo fitness landscape experienced by HIV-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment. *Bioinformatics*, **24**, 34–41.

Dolgui,A. *et al.* (2012) Single machine scheduling with precedence constraints and positionally dependent processing times. *Comput. Oper. Res.*, **39**, 1218–1224.

Doyle,M.A. *et al.* (2014) Studying cancer genomics through next-generation DNA sequencing and bioinformatics. *Clin. Bioinform.*, **1168**, 83–98.

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

Ferguson,A.L. *et al.* (2013) Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, **38**, 606–617.

Gavrilets,S. (2004) *Fitness Landscapes and the Origin of Species (MPB-41)*, Vol. **41**. Princeton University Press, Princeton, NY.

Greaves,M. and Maley,C.C. (2012) Clonal evolution in cancer. *Nature*, **481**, 306.

Hao,D. *et al.* (2016) Distinct mutation accumulation rates among tissues determine the variation in cancer risk. *Sci. Rep.*, **6**, 19458.

Hinkley,T. *et al.* (2011) A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat. Genet.*, **43**, 487.

Hou,Y. *et al.* (2012) Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. *Cell*, **148**, 873–885.

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

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

Jones,S. *et al.* (2008) Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci. USA*, **105**, 4283–4288.

Kimura,M. and Maruyama,T. (1966) The mutational load with epistatic gene interactions in fitness. *Genetics*, **54**, 1337.

Kuipers,J. *et al.* (2017a) Advances in understanding tumour evolution through single-cell sequencing. *Biochim. Biophys. Acta*, **1867**, 127–138.

Kuipers,J. *et al.* (2017b) Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.*, **27**, 1885–1894.

Landau,D.A. *et al.* (2013) Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, **152**, 714–726.

Leung,M.L. *et al.* (2017) Single cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Res.*, **27**, 1287–1299.

Ma,J. *et al.* (2010) vfitness: a web-based computing tool for improving estimation of in vitro HIV-1 fitness experiments. *BMC Bioinform.*, **11**, 261.

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

Malikic,S. *et al.* (2017) Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *bioRxiv*, 234914.

Merlo,L.M. *et al.* (2010) A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer Prev. Res.*, **3**, 1388–1397.

Moran,P.A. (1976) Global stability of genetic systems governed by mutation and selection. *Math. Proc. Cambridge Philos. Soc.*, **80**, 331–336.

Mukherjee,S. *et al.* (2018) Scalable preprocessing for sparse scrna-seq data exploiting prior knowledge. *Bioinformatics*, **34**, i124–i132.

Nijenhuis,A. and Wilf,H.S. (2014) *Combinatorial Algorithms: For Computers and Calculators*. Academic Press, Inc, New York, USA.

Noorbakhsh,J. and Chuang,J.H. (2017) Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nat. Genet.*, **49**, 1288.

Nowak,M.A. (2006) *Evolutionary Dynamics*. Harvard University Press, Cambridge, MA.

Nowak,M.A. and May,R.M. (2000) *Virus Dynamics*. Oxford University Press, Oxford, UK.

Segal,M.R. *et al.* (2004) Relating HIV-1 sequence variation to replication capacity via trees and forests. *Stat. Appl. Genet. Mol. Biol.*, **3**, 1–18.

Seifert,D. *et al.* (2014) A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory. *Genetics*, **199**, 191–203.

Siegel,R.L. *et al.* (2018) Cancer statistics, 2018. *Cancer J. Clin.*, **68**, 7–30.

Tarabichi,M. *et al.* (2018) Neutral tumor evolution? *Nat. Genet.*, **50**, 1630–1633.

Wilke,C.O. (2005) Quasispecies theory in the context of population genetics. *BMC Evol. Biol.*, **5**, 1.

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

Williams,M.J. *et al.* (2018) Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.*, **50**, 895–903.

Wodarz,D. and Komarova,N. (2005) *Computational Biology of Cancer: Lecture Notes and Mathematical Modeling*. World Scientific, Singapore.

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

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