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

Vyacheslau Tsivina[1], Alex Zelikovsky[1], and Pavel Skums[1,2]

[1]Department of Computer Science, Georgia State University, Atlanta, GA, USA
[2]*Email: pskums@gsu.edu*

Abstract

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, which provides an unprecedented insight into intra-tumor heterogeneity allowing to study and quantify selective advantages of individual clones. Here we present SCIFIL, a computational tool for inference of fitness landscapes of heterogeneous cancer clone populations from single cell sequencing 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 and utility of our approach on simulated and experimental data. SCIFIL can be used to provide new insight into the evolutionary dynamics of cancer. Its source code is available at https://github.com/compbel/SCIFIL

1 Introduction

Cancer is responsible for $\sim 25\%$ of deaths in the USA annually [17]. 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 complex heterogeneous populations, which include multiple clonal subpopulations constantly evolving to compete for resources, metastasize, escape immune system and therapy [9, 18, 12, 36]. Clonal heterogeneity plays key role in tumor progression [25, 23], and has important implications for diagnostics and therapy, since rare drug resistant variants could become dominant and lead to relapse in the patient [9, 6, 20]. Therefore cancer is now viewed as a dynamic evolutionary processes defined by complex interactions between clonal variants, which include both competition and cooperation [12, 36, 2].

Recent advances in sequencing technologies promise to have a profound effect on oncological research. Next-generation sequencing (NGS) already allowed for creation of large publicly available databases of cancer genomic data such as The Cancer Genome Atlas [33]. Study of genomic data for different tumors led to progress in understanding evolutionary mechanisms of cancer [36, 12, 18]. Most of cancer data have been obtained using bulk sequencing, which produces admixed populations of cells. However, the most promising recent 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 [4, 31, 35, 27], 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 members of heterogeneous populations (competition or cooperation) and the role of epistasis (non-linear interaction of different SNVs or genes). These questions could be addressed by estimation of evolutionary parameters for cancer lineages from NGS data [35, 31].

One of the most important evolutionary parameters is the collection of replicative fitnesses of individual genomic variants, commonly termed *fitness landscape* in evolutionary biology [11]. Several computational methods have been proposed for *in vitro* estimation of fitness landscapes [29, 22, 14, 10]. 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 [30]. 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 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 [35, 31]. In particular, recent findings on structures of fitness landscapes of cancer from bulk sequencing data [34] initiated a lively scientific discussion published in several papers [31, 27, 35]. It can be anticipated that single cell sequencing 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 [30]. 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 [30, 5, 35].

Contributions. We propose a computational method SCIFIL (Single Cell Inference of FItness Landscape) for *in vivo* inference of clonal selection and estimate of fitness landscapes of heterogeneous cancer clone populations from single cell sequencing data. Instead of assuming that sampled populations are in the equilibrium state, our method is based on a more relevant 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 Methods

We propose a maximum likelihood approach, which estimates fitnesses of individual haplotypes 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 [16], 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 haplotype. For simplicity of further calculations we assume that there is a leaf attached to every internal node, with some leafs having an abundance 0 (or rather some small number $\delta << 1$). Unlike [16], generally we do not need to employ the infinite site assumption (or that the mutation tree represent a *perfect phylogeny*), i.e. repeats of mutations are allowed provided that mutation profiles of all haplotypes in a tree are unique. It

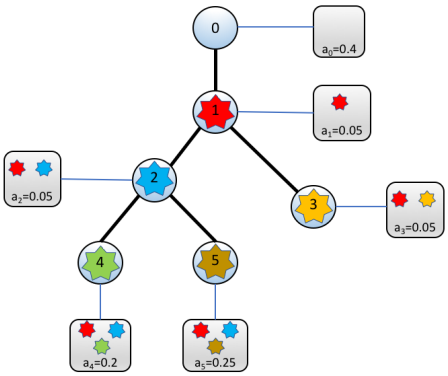

Figure 1: Mutation tree

agrees with recent findings [19]. A mutation tree can be constructed using currently available tools, such as SCITE (without [16] or with [19] infinite site assumption) or SiFit [37].

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

- mutation tree T with $n+1$ leafs corresponding to haplotypes. We assume that internal nodes of T are labeled $0, 1, ..., n$ and the ith haplotype 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 $\mathcal{A} = (a_0, ..., a_n)$ of haplotypes.
- Cancer cells mutation rate θ. It is well-studied evolutionary parameter with estimations provided by various prior studies [32, 13].

The goal is to find haplotype fitnesses $\mathcal{F} = (f_0, ..., f_n)$ that maximize the likelihood

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

The current section is organized as follows. First we introduce our evolutionary model of choice and the general 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 proposed method of estimation of fitnesses $\mathcal{F}$ maximizing (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 node set, internal node set and an 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, which occur at rate θ. The mutation event corresponding to a node i happens at time t_i; at the event the haplotype p_i, gives birth to a haplotype i. The dynamics of the cancer clone population is described by the piecewise continuous function $x = (x_0, ..., x_n)$, where $x_i = x_i(t)$ is the relative abundance of the ith haplotype. The discontinuity points of x correspond to mutation events. Let r, i, j be 3 consecutive mutation events with times $t_r < t_i < t_j$. Between mutation events i and j at interval $[t_i, t_j]$ haplotype frequencies $x_k = x_k^{(i)}$ follows the system of ODEs [28]:

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

with initial conditions

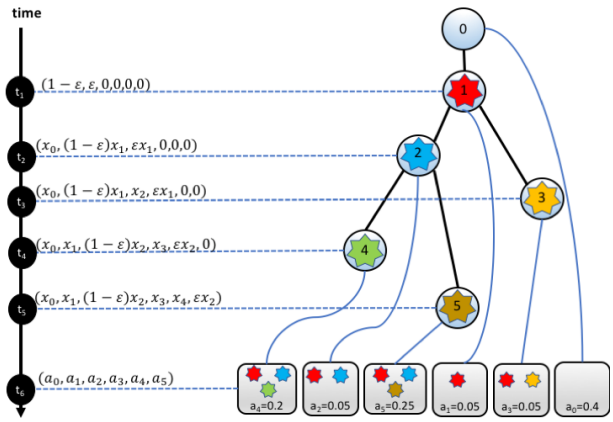

Figure 2: Illustration of the evolutionary model. Each node of the tree represent a mutation event whose times are marked on the time axis. Leafs represent the sampling event. Next to each time point the distribution of haplotype abundances after the corresponding event is shown.

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

Here subtraction of the term $x_k^{(i)} \sum_{l=1}^{n} f_l x_l^{(i)}$ in (2) ensures that the total relative abundance of haplotypes sum up to 1. Initial conditions (3) link haplotype abundances before and after the mutation event i and indicate that at time t_i the haplotype i is generated by the haplotype p_i. The parameter ε is a small number (by default $\varepsilon = 10^{-5}$). At time 0, the root haplotype (healty tissue) gives birth to the first mutation, with the corresponding haplotypes having relative abundances $1-\varepsilon$ and ε.

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

Consider the mutation event in the vertex σ_j. The event occurs if 2 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 haplotype p_{σ_j} rather than in other haplotypes 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 follow a Poisson distribution with the mean $\frac{1}{\theta}$. Assuming that mutations are random, the probability of (b) is equal to the frequency $x_{p_{\sigma_j}}(t_{\sigma_j})$ of the haplotype 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 $\mathcal{M}(a_0, ..., a_n | x_0(t_{n+1}), ..., x_n(t_{n+1}))$. After putting all probabilities together, we have

$$p(\mathcal{A}|T, \mathcal{F}, \Theta, \theta) = \prod_{j=2}^{n+1} Pois(t_{\sigma_j} - t_{\sigma_{j-1}}, \frac{1}{\theta}) \cdot \prod_{j=1}^{n} x_{p_j}(t_j) \cdot \mathcal{M}(a_0, ..., a_n | x_0(t_{n+1}), ..., x_n(t_{n+1})) \tag{4}$$

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

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

The probabilities $\mathcal{M}(a_0, ..., a_n|x_0(t_{n+1}), ..., x_n(t_{n+1}))$, $\prod_{j=2}^{n+1} Pois(t_{\sigma_j} - t_{\sigma_{j-1}}, \frac{1}{\theta})$ and $\prod_{j=1}^{n} x_{p_{\sigma_j}}(t_{\sigma_j})$ are further referred as *abundance likelihood*, *time likelihood* and *mutation likelihood*, respectively. For the tree shown on Figures 1 and 2 it is equally feasible that the mutation 2 appeared before the mutation 3 or vise versa. However, haplotype 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 and abundance likelihoods allow to choose between these two alternatives.

Reduction to discrete optimization. The standard way to solve the maximum likelihood problem (5) is to optimize $\mathcal{F}$ and Θ jointly using Markov Chain Monte Carlo (MCMC) sampling. However, our experiments suggest that the function (1) has too many local optimums which makes MCMC search over the continuous space of possible solutions inefficient. Therefore we suggest a different approach, which transforms the problem (5) into a discrete optimization problem akin to a scheduling problem. Although the modified problem is not exactly equivalent to (5), validation of our approach suggests that it is efficient.

Firstly, we assume that all fitnesses are relative with respect to a fitness of a haplotype 0 which is set to be $f_0 = 1$. By default, this haplotype corresponds to the normal tissue. For the problem of inference of clonal selection such assumption does not restrict the predictive power. Next, note that any assignment of event times Θ defines the order of appearance μ_i for each node $i \in V(T)$ (e.g. on Figure 2 $\mu_i = i$ for $i = 1, ..., 5$). This order agrees with the natural vertex order induced by T, i.e. $\mu_i < \mu_j$ whenever i is an ancestor of j. In turned out that conversely any order μ defines times Θ^μ and fitnesses $\mathcal{F}^\mu$ which maximize the partial likelihood

$$\prod_{j=2}^{n+1} Pois(t_{\sigma_j} - t_{\sigma_{j-1}}, \frac{1}{\theta}) \cdot \mathcal{M}(a_0, ..., a_n|x_0(t_{n+1}), ..., x_n(t_{n+1})) \tag{6}$$

More precisely, the following proposition holds.

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

$$t_i = \frac{\mu_i - 1}{\theta}, i = 1, ..., n, \quad t_{n+1} = \frac{n}{\theta} \tag{7}$$

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

Here A_i is the set of ancestors of a node i.

Proof. Poisson probability achieve maximums at its mean, i.e. the time likelihood is maximal, when for consecutive events i, j we have $t_j - t_i = \frac{1}{\theta}$. This yields the solution (7).

The multinomial probability $\mathcal{M}(a_0, ..., a_n|x_0(t_{n+1}), ..., x_n(t_{n+1}))$ is maximal when $x_i(t_{n+1}) = a_i$ for all $0 = 1, ..., n$. This can be rewritten as

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

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 $x = (x_i(t))_{i=0}^{n+1}$ [1]. We use the observation that the system (2) is invariant with respect to the transition to relative abundances of any pair of haplotypes. Namely, for each haplotype pair $i, j = 0, ..., n$ 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

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

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

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

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

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

Finding optimal ordering. The problem (13) could be considered as a variant of scheduling problem with precedent constraints [7] and with non-linear cumulative cost function. Such problems are usually NP-hard. We solve it by a branch-and-bound search in the space of feasible orderings. Although in the worst case this algorithm is exponential, real mutation trees tend to be path-like [16, 21], which dramatically decrease the search space and allows for more efficient pruning.

The general scheme of our method is described by Algorithm 1. It starts with the pre-order μ' consisting only of the root of T and recursively extends it via branching. The algorithm additionally stores

- a queue Q containing possible extensions of the current pre-order μ', which is initialized by the children of the root.
- a collection $\mathcal{X}$ of haplotype abundance distributions $x^j = (x_i(t_j))_{i=0}^n$ at times t_j and a collection of fitnesses $\mathcal{F}'$ for all nodes j from the current pre-order μ'.

The branching is carried out by recursively extracting the nodes in Q. At the moment when a node $i \in Q$ is extracted, all its ancestors should have been considered at the previous iterations of the algorithm. Thus, we can calculate the time t_i and the fitness f_i using (7),(8),

[1] It is possible to find an exact solution for (9) when discontinuity of x is taken into account. However, the derivation of the corresponding formulas is long, therefore we omit it due to the space restrictions. Furthermore, approximate solution allows to achieve almost the same accuracy as an exact solution.

the distribution $x^i = (x_0(t_i), ..., x_n(t_i))$ from the system (2)-(3) and the log-likelihood of the partial solution $\mu' \cup \{i\}$: $L_{\mu' \cup \{i\}} = \sum_{j \in \mu'} \log(x_{p_j}(t_j)) + \log(x_{p_i}(t_i))$. If the fitness f_i is outside of predefined boundaries $[f_{min}, f_{max}]$ or the partial log - likelihood $L_{\mu' \cup \{i\}}$ is smaller than the current optimum, then i is pruned. Otherwise, i is added to the current pre-order and is replaced in Q by its non-leaf children $N^-(i)$. Once the queue Q is empty and the complete order μ is constructed, its likelihood (4) is compared with the current optimum and the latter is updated, if necessary.

Algorithm 1 Branch-and-Bound algorithm for node ordering

```
opt = −∞; μ_opt = ∅ μ' = {0}, f_0 = 1, F' = {f_0}, Q = N⁻(0), X = ∅.
while Q ≠ ∅ do
    choose a branching node i ∈ Q and extract it from Q;
    calculate t_i, f_i and x^i using (2),(3),(7),(8).
    Calculate partial solution log-likelihood L_{μ'∪{i}} = Σ_{j∈μ'} log(x_{p_j}(t_j)) + log(x_{p_i}(t_i)).
    if f_i < f_min OR f_i > f_max OR L_{μ'∪{i}} ≤ opt then
        prune the branch i
    else
        μ' = μ' ∪ {i}; F' = F' ∪ {f_i}
        if μ' is a complete order and L_{μ'} > opt then
            opt = L_{μ'}; μ_opt = μ.
        else
            add N⁻(i) to Q and recursively process μ', Q and X.
        end if
    end if
end while
```

3 Results

3.1 Simulated data

We simulated 60 test examples with $m = 30, 40, 50$ mutations, which correspond to numbers of mutations for real single cell sequencing data analyzed in previous studies [16, 18, 21]. For each test example, clonal evolution was simulated as follows: (a) at each moment of time a mutation event happens with the probability $\theta = 0.05$; at the event a random haplotype i selected with the probability equal to its current relative abundance gives birth to a new haplotype j with the random fitness $f_j \in [1, f_{max}]$ by acquiring a random mutation m_j. (b) When there is no mutation event, abundances of existing haplotypes are updated according to the model (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 haplotype abundances were used as an input for SCIFIL. The tool was run with the timeout of 1hr.

We quantified the performance of SCIFIL using two performance 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 vectors of true and inferred fitnesses.

MRA and SC highlight different aspects of the problem. MRA measure the accuracy of fitness value estimation, while in our case SC measures how well we are able to qualitatively detect selective advantage of particular clones over other clones. In particular, when the population contains several clones with close fitnesses, even small error in fitness estimation may cause low SC. Correct fitness ranking can be used in evolutionary studies even when actual fitness values are missing or inaccurate [3].

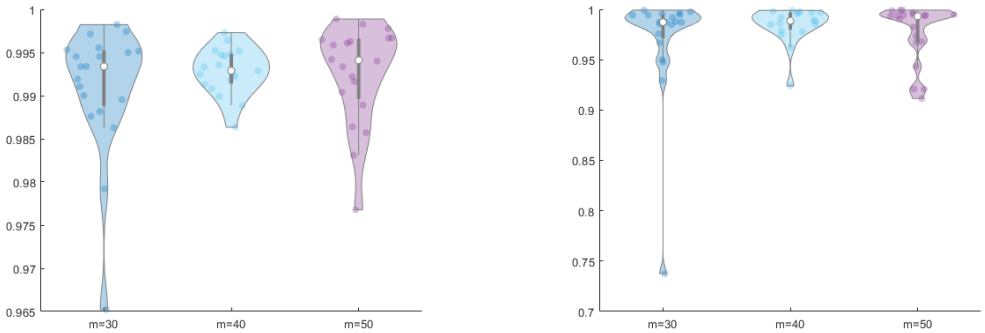

Figure 3: Performance of SCIFIL on simulated data with $m = 30, 40, 50$ mutations. Left: violin plots of the mean relative accuracy of fitness estimation. Right: violin plots of Spearman correlation between true and inferred fitness vectors

The results of SCIFIL evaluation on simulated data are shown on Fig. 3. The algorithm demonstrated high accuracy in a majority of cases with average and median MRA being 0.9921 and 0.9934 (standard deviation 0.0057), and average and median SC being 0.9768 and 0.9887 (standard deviation 0.0388). In general, for the generated test cases the number of mutations does not have a great impact on the algorithm's accuracy which could be explained by the fact that SCIFIL implements an exact algorithm for discrete optimization problem rather than a statistical sampling.

We compared our approach with the previously published tool QuasiFit [30], that also reconstructs replicative fitnesses of members of heterogeneous genomic populations. Although originally designed for viruses, QuasiFit is based on quasispecies model, which is applicable to both intra-host viral populations and cancer clone populations [8] and is essentially a fully continuous version of the model used by SCIFIL. 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. Furthermore, SCIFIL is a discrete optimization approach, while QuasiFit implements Markov Chain Monte Carlo sampling. QuasiFit was run with the per-cell mutation rate $\mu = \varepsilon\theta$ (which is a fully continuous analogue of the parameters used by SCIFIL) and inferred fitnesses were considered after a burn-in of 100000 iterations. As QuasiFit uses a different fitness vector normalization, following [30] we used only the parameter SC for the comparison. The results are shown on Fig. 4. SCIFIL significantly outperforms QuasiFit indicating that the proposed model is more adequate for the inference of clonal selection in finite-time settings than the equilibrium state assumption.

The algorithm's running time significantly depends on topology of the input mutation tree, which defines the size of the search space for the branch-and-bound, and can vary from less than a second to 1 hour. On our test cases, running times for $30, 40, 50$ mutations on a simple desktop computer were in average 5.75, 8.35 and 12.94 minutes. For two experimental datasets discussed in the next section the running times were 0.0039 and 0.3937 minutes, respectively. It should be noted that the current version of the algorithm is not parallelized. It is expected that parallelization will improve the running time.

3.2 Experimental data

We used SCIFIL to infer fitness landscapes for several recently published experimental cancer datasets. Due to the page restrictions, here we present the results for 2 datasets which are particularly illustrative. The first dataset is single-cell sequencing data from a JAK2-negative myeloproliferative neoplasm (essential thrombocythemia) [15, 16], the second one represents metastatic colon cancer [21]. The latter dataset includes SNVs sampled from the main tumor and two metas-

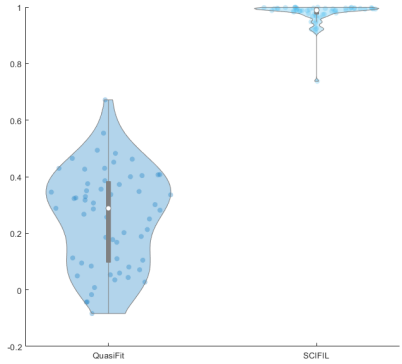

Figure 4: Spearman correlation between true and inferred fitness vectors for QuasiFit and SCIFIL

tases. 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 [16], and fitnesses and 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. First, we calculated a distance between each pair of haplotypes defined as the sum of their hamming distance and the absolute difference of their orders of appearance. Obtained distances were used to map haplotypes to the plain $\mathbb{R}^2$ using multidimensional scaling. Fitness values of the points corresponding to haplotypes were interpolated using biharmonic splines, and the resulting surface was visualized as a contour plot with the tree placed on top of the plot. The obtained plots are shown on Fig. 5, where colors represent fitness values, and distance from each tree node to the root reflects its appearance time.

For myeloproliferative neoplasm (Fig. 5, 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 [16], we may predict that recent subclones will replace ancestral clones. However, based on the available information it is hard to decide whether one of subclone lineages will out-compete another one, or they will continue to coexist.

Evolution of the colon tumor (Fig. 5, right) follows different scenario, with 3 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 on Fig. 5).

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 and inference of fitness landscapes for individual tumors may provide valuable information for understanding mechanisms of disease progression and for design of personalized cancer treatment. Single cell sequencing provides an unprecedented insight into intra-tumor heterogeneity promising to allow to study fitness landscapes at finest possible resolution and quantify selective advantages on the level of individual clones.

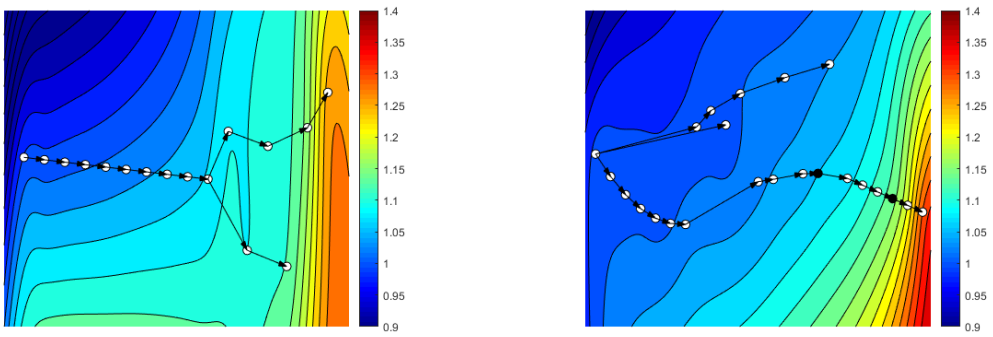

Figure 5: Fitness landscape and mutation tree for JAK2-negative myeloproliferative neoplasm [16, 15] (left) and colorectal cancer (right) [21] inferred by SCIFIL. Colors represent fitness values, and distance from each tree node to the root is approximately proportional to its time of appearance.

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 single cell sequencing. 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. In addition it differs from previous approaches by employing more general and realistic evolutionary model rather than assuming 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 and/or mutation patterns 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 taken into account and/or 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 single cell sequencing data. There exist plethora of tools which can estimate clone abundances from composite bulk and single cell sequencing data (see, e.g. [1, 26]). In addition, such composite data can be employed to increase an accuracy of mutation trees reconstruction [24]. We expect SCIFIL reliability to increase when it will be combined with these tools.

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 [7]. Such problems are generally NP-hard, although the complexity of our problem is unknown. We solve it using branch-and-bound search, which is efficient by still can be exponential in the worse case. Real mutation trees tend to be path-like [16, 21], so for such data and for the numbers of mutations in the present datasets our approach is scalable enough. However, in the future the scalability question will become essential. 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 [7]. 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.

References

[1] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

[2] Rudy Bonavia, Webster K Cavenee, Frank B Furnari, et al. Heterogeneity maintenance in glioblastoma: a social network. *Cancer research*, 71(12):4055–4060, 2011.

[3] Kristina Crona, Alex Gavryushkin, Devin Greene, and Niko Beerenwinkel. Inferring genetic interactions from comparative fitness data. *eLife*, 6:e28629, 2017.

[4] Alexander Davis, Ruli Gao, and Nicholas Navin. Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1867(2):151–161, 2017.

[5] Koen Deforche, Ricardo Camacho, Kristel Van Laethem, Philippe Lemey, Andrew Rambaut, Yves Moreau, and A-M Vandamme. 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(1):34–41, 2007.

[6] Li Ding, Timothy J Ley, David E Larson, Christopher A Miller, Daniel C Koboldt, John S Welch, Julie K Ritchey, Margaret A Young, Tamara Lamprecht, Michael D McLellan, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, 2012.

[7] Alexandre Dolgui, Valery Gordon, and Vitaly Strusevich. Single machine scheduling with precedence constraints and positionally dependent processing times. *Computers & Operations Research*, 39(6):1218–1224, 2012.

[8] Wodarz Dominik and Komarova Natalia. *Computational biology of cancer: lecture notes and mathematical modeling*. World Scientific, 2005.

[9] Maria A Doyle, Jason Li, Ken Doig, Andrew Fellowes, and Stephen Q Wong. Studying cancer genomics through next-generation dna sequencing and bioinformatics. *Clinical Bioinformatics*, pages 83–98, 2014.

[10] Andrew L Ferguson, Jaclyn K Mann, Saleha Omarjee, Thumbi NdungâĂŹu, Bruce D Walker, and Arup K Chakraborty. Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, 38(3):606–617, 2013.

[11] Sergey Gavrilets. *Fitness landscapes and the origin of species (MPB-41)*, volume 41. Princeton University Press, 2004.

[12] Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306, 2012.

[13] Dapeng Hao, Li Wang, and Li-jun Di. Distinct mutation accumulation rates among tissues determine the variation in cancer risk. *Scientific Reports*, 6:19458, 2016.

[14] Trevor Hinkley, João Martins, Colombe Chappey, Mojgan Haddad, Eric Stawiski, Jeannette M Whitcomb, Christos J Petropoulos, and Sebastian Bonhoeffer. A systems analysis of mutational effects in hiv-1 protease and reverse transcriptase. *Nature genetics*, 43(5):487, 2011.

[15] Yong Hou, Luting Song, Ping Zhu, Bo Zhang, Ye Tao, Xun Xu, Fuqiang Li, Kui Wu, Jie Liang, Di Shao, et al. Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. *Cell*, 148(5):873–885, 2012.

[16] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):86, 2016.

[17] Ahmedin Jemal, Rebecca Siegel, Elizabeth Ward, Yongping Hao, Jiaquan Xu, Taylor Murray, and Michael J Thun. Cancer statistics, 2008. *CA: a cancer journal for clinicians*, 58(2):71–96, 2008.

[18] Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1867(2):127–138, 2017.

[19] Jack Kuipers, Katharina Jahn, Benjamin J Raphael, and Niko Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome research*, 2017.

[20] Dan A Landau, Scott L Carter, Petar Stojanov, Aaron McKenna, Kristen Stevenson, Michael S Lawrence, Carrie Sougnez, Chip Stewart, Andrey Sivachenko, Lili Wang, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, 152(4):714–726, 2013.

[21] Marco L Leung, Alexander Davis, Ruli Gao, Anna Casasent, Yong Wang, Emi Sei, Eduardo Sanchez, Dipen Maru, Scott Kopetz, and Nicholas E Navin. Single cell dna sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome research*, pages gr–209973, 2017.

[22] Jingming Ma, Carrie Dykes, Tao Wu, Yangxin Huang, Lisa Demeter, and Hulin Wu. vfitness: a web-based computing tool for improving estimation of in vitro hiv-1 fitness experiments. *BMC bioinformatics*, 11(1):261, 2010.

[23] Carlo C Maley, Patricia C Galipeau, Jennifer C Finley, V Jon Wongsurawat, Xiaohong Li, Carissa A Sanchez, Thomas G Paulson, Patricia L Blount, Rosa-Ana Risques, Peter S Rabinovitch, et al. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature genetics*, 38(4):468, 2006.

[24] Salem Malikic, Katharina Jahn, Jack Kuipers, Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *bioRxiv*, page 234914, 2017.

[25] Lauren MF Merlo, Najaf A Shah, Xiaohong Li, Patricia L Blount, Thomas L Vaughan, Brian J Reid, and Carlo C Maley. A comprehensive survey of clonal diversity measures in barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer prevention research*, 3(11):1388–1397, 2010.

[26] Sumit Mukherjee, Yue Zhang, Joshua Fan, Georg Seelig, and Sreeram Kannan. Scalable preprocessing for sparse scrna-seq data exploiting prior knowledge. *Bioinformatics*, 34(13):i124–i132, 2018.

[27] Javad Noorbakhsh and Jeffrey H Chuang. Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nature genetics*, 49(9):1288, 2017.

[28] Martin A Nowak and Robert M May. Virus dynamics, 2000.

[29] Mark R Segal, Jason D Barbour, and Robert M Grant. Relating hiv-1 sequence variation to replication capacity via trees and forests. *Statistical applications in genetics and molecular biology*, 3(1):1–18, 2004.

[30] David Seifert, Francesca Di Giallonardo, Karin J Metzner, Huldrych F Günthard, and Niko Beerenwinkel. A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory. *Genetics*, pages genetics–114, 2014.

[31] Maxime Tarabichi, Inigo Martincorena, Moritz Gerstung, Armand M Leroi, Florian Markowetz, Paul T Spellman, Quaid D Morris, Ole Christian Lingjaerde, David C Wedge, and Peter Van Loo. Neutral tumor evolution? *Nature Genetics*, page 1, 2018.

[32] Ian PM Tomlinson, MR Novelli, and WF Bodmer. The mutation rate and cancer. *Proceedings of the National Academy of Sciences*, 93(25):14800–14803, 1996.

[33] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[34] Marc J Williams, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature genetics*, 48(3):238, 2016.

[35] Marc J Williams, Benjamin Werner, Timon Heide, Christina Curtis, Chris P Barnes, Andrea Sottoriva, and Trevor A Graham. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature genetics*, page 1, 2018.

[36] Lucy R Yates and Peter J Campbell. Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11):795, 2012.

[37] Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. Sifit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome biology*, 18(1):178, 2017.