Pacific Symposium on Biocomputing 25:226-237(2020)

PhySigs: Phylogenetic Inference of Mutational Signature Dynamics

Sarah Christensen
Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Mark D.M. Leiserson
Dept. of Computer Science, University of Maryland, College Park, MD 20740

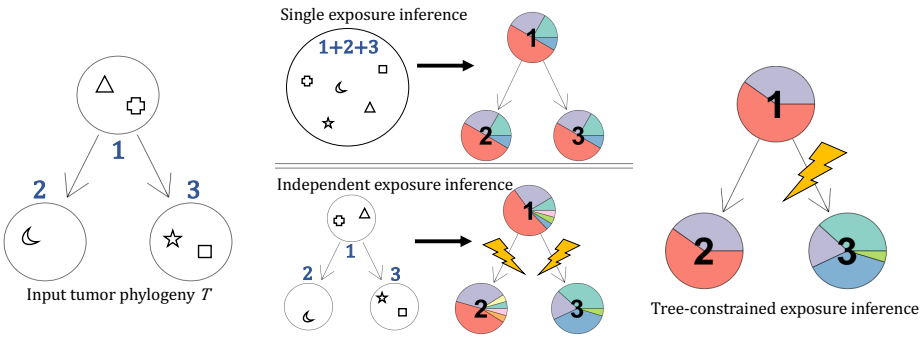
Mohammed El-Kebir†
Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801
†*E-mail: melkebir@illinois.edu*

Distinct mutational processes shape the genomes of the clones comprising a tumor. These processes result in distinct mutational patterns, summarized by a small number of mutational signatures. Current analyses of clone-specific exposures to mutational signatures do not fully incorporate a tumor's evolutionary context, either inferring identical exposures for all tumor clones, or inferring exposures for each clone independently. Here, we introduce the Tree-constrained Exposure problem to infer a small number of exposure shifts along the edges of a given tumor phylogeny. Our algorithm, PhySigs, solves this problem and includes model selection to identify the number of exposure shifts that best explain the data. We validate our approach on simulated data and identify exposure shifts in lung cancer data, including at least one shift with a matching subclonal driver mutation in the mismatch repair pathway. Moreover, we show that our approach enables the prioritization of alternative phylogenies inferred from the same sequencing data. PhySigs is publicly available at `https://github.com/elkebir-group/PhySigs`.

Keywords: Intra-tumor heterogeneity; Convex optimization; Phylogeny; Cancer.

1. Introduction

A tumor results from an evolutionary process, where somatic mutations accumulate in a population of cells.[1] To understand the mechanisms by which mutations accumulate, researchers search large databases of somatic mutations and identify *mutational signatures* i.e. patterns of mutations associated with distinct mutational processes across different types of cancer.[2] In addition to elucidating tumorigenesis, mutational signatures have found clinical applications.[3] One promising application is using a tumor's *exposure* to a signature associated with perturbed DNA damage repair as a biomarker for response to an established therapy, potentially increasing the number of patients who could benefit beyond standard driver-based approaches.[4] Methods for inferring the mutational signatures active in a given tumor are key to realizing this goal. However, initial analyses overlook *intra-tumor heterogeneity*, the pres-



(a) (b) (c)

Fig. 1. **PhySigs unites previous work on inference of clonal mutational signature exposures into one statistical framework by incorporating evolutionary context.** (a) The input is a phylogeny T with nodes representing clones in a patient tumor, and the set of mutations (indicated by shapes) introduced in each clone. (b) Previous work generally falls into two categories and disregards evolutionary structure: While in single exposure inference all mutations are combined into one set for signature exposure inference, signatures for each clone are estimated independently in independent exposure inference. (c) Both previous problems are special cases of the problem solved by PhySigs, which incorporates evolutionary context to return a set of exposure shifts (lighting bolts) as well as the signature exposures for each cluster defined by the shifts.

ence of multiple clones with distinct complements of mutations that may be characterized by distinct mutational signatures. Here, we propose to study the dynamics of exposures to mutational signatures of clones within a tumor, in order to fully understand tumorigenesis and to move towards devising more effective treatment plans.

The evolutionary history of a tumor is described by a phylogeny, whose vertices correspond to clones. Specialized methods exist for tumor phylogeny inference from bulk and single-cell DNA sequencing data.[5] A clone may be distinguished from its parental clone by a unique set of *introduced* mutations that appear in the clone but not in its parent. Introduced mutations provide a record of the mutational signatures acting on the clone at a particular location and time. Previous work to identify exposures to known mutational signatures can be classified in four broad categories. An initial body of work[6–9] aimed to identify a single distribution of mutational signatures for all clones of a tumor, which we refer to as *single exposure inference* (Fig. 1). This was followed by work[10,11] that considered the distribution of exposures for each clone independently, called *independent exposure inference*. In addition to considering clones independently, Jamal-Hanjani et al.[11] clustered mutations into two categories: *clonal* mutations that are present in all clones vs. *subclonal* mutations that are present in only a subset of clones. Finally, Rubanova et al.[12] incorporate even more structure by studying the changes in exposures within a linear ordering of the clones. A similar idea to study the dynamics of APOBEC signature exposure has been explored experimentally in cell lines and patient-derived xenografts.[13]

We build upon this line of work by proposing a model of clonal exposures that explicitly

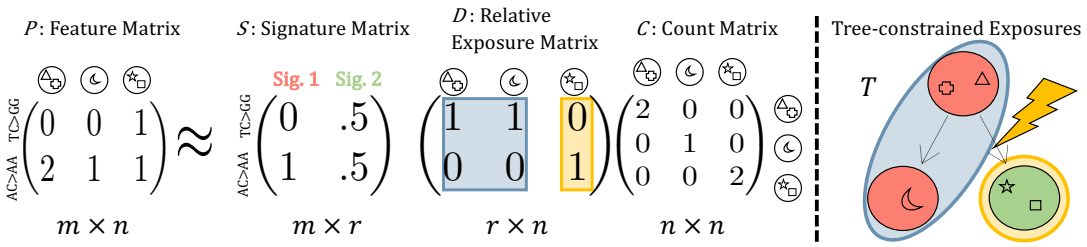


Fig. 2. **PhySigs solves the TE problem for all combinations of exposure shifts.** PhySigs takes as input P, S, C and T and solves the TE problem, identifying for each value of $k \in \{1, \ldots, n\}$ a relative exposure matrix D composed of k identical columns corresponding to clusters of clones with identical exposures (denoted by blue and yellow). Edges between these clusters in T are interpreted to be where exposure shifts occurred (denoted by a lightning bolt). PhySigs uses the Bayesian Information Criterion to select the number k^* of clusters that best explain the data (here $k^* = 2$).

incorporates the tumor phylogeny relating clones. As new mutations interfere with key DNA repair pathways or carcinogenic environmental factors are added or removed, we would expect to see a change in the corresponding exposures along edges of the tumor phylogeny. Such *exposure shifts* induce a partition of the set of clones into disjoint clusters, where within each cluster the clones are ascribed the same set of relative signature exposures. To identify exposure shifts, we formulate the *Tree-constrained Exposure* (TE) problem (Fig. 1). We provide an algorithm, PhySigs, which solves the TE problem and provides a principled way to select the number of exposure shifts such that the mutational patterns observed at each clone are accurately reconstructed without overfitting. PhySigs interpolates between single exposure inference and independent exposure inference, thus uniting previous work under one statistical framework. On simulated data, we demonstrate that PhySigs accurately recovers exposures and shifts. While PhySigs does not detect any exposure shifts in ovarian cancer,[10] it identifies several exposure shifts in non-small-cell lung cancers,[11] including at least one case with strong support from a driver mutation in the corresponding subclone. Moreover, PhySigs enables one to prioritize alternative, equally-plausible phylogenies inferred from the same input DNA sequencing data.

2. Preliminaries and Problem Statement

We consider n samples with single nucleotide variants (SNVs) classified into the $m = 96$ mutation categories most commonly used for mutation signature analysis.[2] We assume each sample's SNVs are the product of r signatures of underlying mutational processes. The $m \times n$ *feature matrix* $P = [p_{ij}]$ indicates the number of mutations of category i in sample j. The $m \times r$ *signature matrix* $S = [s_{i\ell}]$ describes the probability signatures ℓ generate mutations of category i. The $r \times n$ *exposure matrix* $E = [e_{\ell j}]$ contains the number of mutations generated by signatures ℓ in samples j. The three matrices are related as follows:

$$P \approx SE. \tag{1}$$

Beginning with Alexandrov et al.,[2] initial efforts to discover *de novo* mutational signatures shaping cancer genomes used non-negative matrix factorization. These efforts produced a

compendium of 30 validated mutational signatures distributed by the Catalogue of Somatic Mutations in Cancer,[14] and researchers used the signature exposures to reveal signature etiology (e.g. Kim et al.[15]) and other applications (e.g. Trucco et al.[16] and Davies et al.[4]). However, these initial analyses disregard the clonal architecture of individual tumors. To understand the *clonal dynamics of mutational signatures*, we wish to identify signature exposures of the mutations that were introduced in each individual clone.

We start by recognizing that the exposures for each clone are proportional to the number of mutations present in the clone. We formalize this by defining a *relative exposure matrix* $D \in [0,1]^{r \times n}$, a matrix with nonnegative entries between 0 and 1. The relative exposure matrix D corresponding to an absolute exposure matrix E and feature matrix P is obtained by dividing the entries of each column j of E by the total number of mutations in the corresponding sample of P. In other words, we may view exposure matrix E as the product DC of a relative exposure matrix D and a diagonal *count matrix* $C \in \mathbb{N}^{n \times n}$ whose diagonal entries c_{jj} equal the number $\sum_{i=1}^{m} p_{ij}$ of mutations in clone j.

The first problem that we formulate assumes that the mutations introduced in every clone result from the same relative exposures.

Problem 1 (Single Exposure (SE)). *Given feature matrix P, corresponding count matrix C and signature matrix S, find relative exposure matrix D such that $||P - SDC||_F$ is minimum and D is composed of identical columns.*

Current methods[6–9] implicitly solve this problem by estimating signatures of a single sample with mutations pooled across clones, as we will show in Section 3.

By contrast, the second problem assumes that the mutations introduced in each clone result from distinct exposures. In other words, we assume independence between the clones, leading the to the following problem.

Problem 2 (Independent Exposure (IE)). *Given feature matrix P, corresponding count matrix C and signature matrix S, find relative exposure matrix D such that $||P - SDC||_F$ is minimum.*

We note that the above problem is equivalent to the problem solved by current methods for patient-specific exposure inference[6–9] where one replaces patients by clones, as was recently done by Jamal-Hanjani et al.[11]

An *exposure shift* is a significant shift in relative exposures of signatures between two clones. Recognizing that exposure shifts occur on a subset of the edges of a phylogeny T (Fig. 1), we propose the following tree-constrained inference problem, which generalizes both previous problems (Fig. 2).

Problem 3 (Tree-constrained Exposure (TE)). *Given feature matrix P, corresponding count matrix C, signature matrix S, phylogenetic tree T and integer $k \geq 1$, find relative exposure matrix D such that $||P - SDC||_F$ is minimum and D is composed of k sets of identical columns, each corresponding to a connected subtree of T.*

We note that both SE and IE are special cases of TE, where $k = 1$ and $k = n$, respectively (Fig. 1). Moreover, the three problems are identical for a feature matrix P composed of a

single clone ($n = 1$). Finally, for a fixed selection of k subtrees, the TE problem decomposes into k SE instances.

3. Methods

3.1. *Solving the SE problem*

To solve the clone-specific exposure inference problems defined in the previous section, we wish to leverage current methods for patient-specific exposure inference. These current methods[6–9] solve the problem of identifying absolute exposures $\mathbf{e}^* \in \mathbb{R}^r_{\geq 0}$ of a single patient minimizing $||\mathbf{q} - S\mathbf{e}||_F$ given feature vector $\mathbf{q}$ and signature matrix S, as described in Eq. (1).

In the following, we show how to reduce, in polynomial time, any SE instance (P, S) to the patient-specific instance $(\mathbf{q}, S)$. Specifically, we transform feature matrix $P = [p_{ij}]$ composed of $n \geq 1$ clones to a single-clone feature vector $\mathbf{q} = [q_i]$ by setting

$$q_i = \sum_{j=1}^{n} c_j \cdot p_{ij} \qquad \forall i \in [m], \tag{2}$$

where c_j is the number $\sum_{i=1}^{m} p_{ij}$ of mutations introduced in clone j. Let N be the sum of the number of mutations in each sample squared, i.e.

$$N = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} p_{ij} \right)^2 = \sum_{j=1}^{n} c_j^2. \tag{3}$$

We claim that relative exposure matrix D composed of n identical vectors $\mathbf{d}^*$ defined as $\mathbf{d}^* = \mathbf{e}^*/N$ is an optimal solution to SE instance (P, S).

Proposition 1. *Let (P, S) be an instance of* SE*. Let $(\mathbf{q}, S)$ be the corresponding patient-specific instance with optimal solution $\mathbf{e}^*$. Then the relative exposure matrix D composed of n identical vectors $\mathbf{d}^* = \mathbf{e}^*/N$ is an optimal solution to* SE *instance (P, S).*

Proof. We define $\mathbf{c} = [c_j]$ as an n-dimensional row vector, where c_j is the number of mutations introduced in clone j. We begin with (P, S), an arbitrary instance of SE, where we wish to find a vector $\mathbf{d}^*$ that equals

$$\arg\min_{\mathbf{d}} ||P - S\mathbf{d}\mathbf{c}||_F = \arg\min_{\mathbf{d}} \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |p_{ij} - \sum_{\ell=1}^{r} s_{i\ell} \cdot d_\ell \cdot c_j|^2}$$

We now rearrange this equation to reflect the minimization problem for $(\mathbf{q}, S)$, the corresponding patient-specific exposure instance as described above. In doing so, we will show that the set of optimal solutions for $(\mathbf{q}, S)$ has the claimed relationship to the set of optimal solutions for (P, S). We start by squaring and then defining $\hat{p}_{ij} = \sum_{\ell=1}^{r} s_{i\ell} \cdot d_\ell \cdot c_j$ as the reconstructed value for feature i and sample j.

$$\arg\min_{\mathbf{d}} \sum_{i=1}^{m} \sum_{j=1}^{n} |p_{ij} - \sum_{\ell=1}^{r} s_{i\ell} \cdot d_\ell \cdot c_j|^2 = \arg\min_{\mathbf{d}} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(p_{ij}^2 - 2p_{ij}\hat{p}_{ij} + \hat{p}_{ij}^2 \right)$$

We now distribute the inner sum.

$$\arg\min_{\mathbf{d}} \sum_{i=1}^{m} \left[\sum_{j=1}^{n} p_{ij}^2 - 2(\sum_{j=1}^{n} p_{ij}\hat{p}_{ij}) + \sum_{j=1}^{n} \hat{p}_{ij}^2 \right]$$

Next, we remove the first term, which is a constant, followed by substituting $\mathbf{d}$ with $\mathbf{e} = N \cdot \mathbf{d}$.

$$\arg\min_{\mathbf{d}} \sum_{i=1}^{m} \left[-2(\sum_{j=1}^{n} p_{ij}\hat{p}_{ij}) + \sum_{j=1}^{n} \hat{p}_{ij}^2 \right] = \frac{1}{N} \arg\min_{\mathbf{e}} \sum_{i=1}^{m} \left[-2(\sum_{j=1}^{n} p_{ij}\hat{p}_{ij}) + \sum_{j=1}^{n} \hat{p}_{ij}^2 \right]$$

We then update $\hat{p}_{ij}$ terms using $\mathbf{e}$. Let $\hat{q}_i = \sum_{\ell=1}^{r} s_{i\ell} \cdot e_\ell$ be the reconstructed value for mutation category i where $e_\ell = d_\ell \cdot N$. Observe that $\hat{p}_{ij} = \sum_{\ell=1}^{r} s_{i\ell} \cdot d_\ell \cdot c_j = c_j \sum_{\ell=1}^{r} s_{i\ell} \cdot e_\ell / N = c_j \cdot \hat{q}_i / N$.

$$\frac{1}{N} \arg\min_{\mathbf{e}} \sum_{i=1}^{m} \left[-2(\sum_{j=1}^{n} p_{ij} \frac{c_j \cdot \hat{q}_i}{N}) + \sum_{j=1}^{n} \left(\frac{c_j \cdot \hat{q}_i}{N} \right)^2 \right]$$

We now multiply inside the arg min by the positive constant $N > 0$, canceling terms using (3).

$$\frac{1}{N} \arg\min_{\mathbf{e}} \sum_{i=1}^{m} \left[-2(\sum_{j=1}^{n} p_{ij} \cdot c_j \cdot \hat{q}_i) + \hat{q}_i{}^2 \frac{\sum_{j=1}^{n} c_j^2}{N} \right] = \frac{1}{N} \arg\min_{\mathbf{e}} \sum_{i=1}^{m} \left[-2(\sum_{j=1}^{n} p_{ij} \cdot c_j \cdot \hat{q}_i) + \hat{q}_i{}^2 \right]$$

We add back in a constant term.

$$\frac{1}{N} \arg\min_{\mathbf{e}} \sum_{i=1}^{m} \left[(\sum_{j=1}^{n} c_j \cdot p_{ij})^2 - 2(\sum_{j=1}^{n} c_j \cdot p_{ij})\hat{q}_i + \hat{q}_i^2 \right]$$

We now substitute in for $\mathbf{q}$ following (2).

$$\frac{1}{N} \arg\min_{\mathbf{e}} \sum_{i=1}^{m} \left(q_i^2 - 2q_i\hat{q}_i + \hat{q}_i^2 \right) = \frac{1}{N} \arg\min_{\mathbf{e}} \sum_{i=1}^{m} |q_i - \hat{q}_i|^2$$

Finally, we substitute out $\hat{\mathbf{q}}$.

$$\frac{1}{N} \arg\min_{\mathbf{e}} \sum_{i=1}^{m} \left| q_i - \sum_{\ell=1}^{r} s_{i\ell} \cdot e_\ell \right|^2 = \frac{1}{N} \arg\min_{\mathbf{e}} \sqrt{\sum_{i=1}^{m} \left| q_i - \sum_{\ell=1}^{r} s_{i\ell} \cdot e_\ell \right|^2} = \frac{1}{N} \arg\min_{\mathbf{e}} ||\mathbf{q} - S\mathbf{e}||_F = \frac{\mathbf{e}^*}{N}.$$

Observe that this final equation contains the minimization problem for the patient-specific instance $(\mathbf{q}, S)$ as defined in our reduction. Thus, we get that $\mathbf{d}^* = \mathbf{e}^*/N$ as claimed. □

3.2. *Solving the IE and TE problems*

IE problem. In the IE problem, we are given a feature matrix P, a signature matrix S and seek a relative exposure matrix D such that $||P - SDC||_F$ is minimum. We solve this problem by decomposing it into n SE problem instances, each composed of a single clone. For each resulting SE instance, we use the reduction described in Section 3.1 to the patient-specific exposure problem.

TE problem. In the TE problem, we are given a feature matrix P, a signature matrix S, a phylogenetic tree T and an integer $k \geq 1$. The tree T has n nodes and thus $n-1$ edges. To solve this problem, we exhaustively enumerate all $\binom{n-1}{k-1}$ combinations of $k-1$ edge removals that lead to k connected subtrees. Each connected subtree correspond to a single SE instance, which may be solved using the reduction to the patient-specific exposure problem described previously. We select the combination of $k-1$ edges that minimizes the objective function.

3.3. *PhySigs*

Model selection for k. To decide on the number k of subtrees to consider, we use the Bayesian Information Criterion (BIC).[17] That is, we evaluate each optimal solution for each number $k \in \{1, \ldots, n\}$ of subtrees. For a fixed k, the number of observations equals the size of matrix P, i.e. mn, and the number of parameters equals the number of entries in unique columns of D, i.e. kr. Let $L(k) = \min_D ||P - SDC||_F$ be the optimal value for k subtrees, then the corresponding BIC value is

$$\mathrm{BIC}(L(k)) = mn \log(L(k)/(mn)) + kr \log(mn). \tag{4}$$

We select the number k that has the smallest BIC value.

PhySigs. We implemented the above algorithm for the TE problem that includes model selection in R. Our method, PhySigs, uses deconstructSigs[7] as a subroutine for solving the underlying SE problems.[a] PhySigs is available at `https://github.com/elkebir-group/PhySigs`.

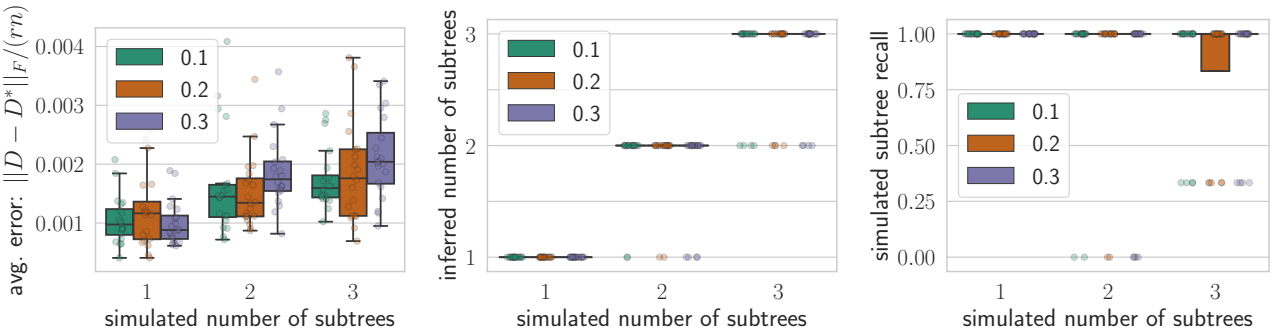
4. Results

We note that the results below were obtained on a laptop with a 2.9 GHz CPU and 16 GB RAM. The majority of input instances completed in minutes, with one notable exception of a large lung cancer instance of 15 clones taking several hours.

4.1. *PhySigs accurately recovers exposures and shifts in simulated data*

We first assess PhySigs's ability to correctly identify model parameters for data generated under the Tree-constrained Exposure model. To do so, we generate simulated data with clone-specific mutations in $m = 96$ categories resulting from exposure to $r = 30$ COSMIC v2 signatures,[14] comprising matrix S. Specifically, we simulate 20 phylogenetic trees T with $n \in \{5, 7\}$ clones, each clone containing between 20 and 200 mutations, as described by the count matrix C. For each phylogenetic tree T, we generate a partition of $k \in \{1, 2, 3\}$ connected subtrees, assigning each subtree a relative exposure vector $\mathbf{d}$ by drawing from a symmetric Dirichlet distribution (with concentration parameter $\alpha = 0.2$). For each combination of T and k, this yields a relative exposure matrix D. Next, we introduce Gaussian noise with mean $\mu = 0$ and standard deviation $\sigma \in \{0.1, 0.2, 0.3\}$, amounting to a $m \times n$ matrix X. Finally, we generate the feature matrix P as $SDC + X$. Thus, we have a total of 180 TE problem instances (T, P).

[a]We note that while deconstructSigs is a heuristic and does not solve the SE problem optimally, it was found by Huang et al.[8] to give comparable results to the optimal solution in most patients.



(a) (b) (c)

Fig. 3. **Simulations show that PhySigs is robust to noise, accurately reconstructing simulated relative exposures as well as the number and location of exposure shifts.** Each boxplot contains results from 20 trees, colors indicate Gaussian noise standard deviations σ. (a) Error between the inferred and simulated relative exposure matrices. (b) The number of inferred vs. simulated subtrees. (c) The fraction of correctly recalled simulated subtrees.

Fig. 3 shows that PhySigs identifies relative exposures D^* that are close to their corresponding simulated exposures D in varying noise regimes and simulated number of exposure shifts (Fig. 3a). Moreover, the number of exposure shifts is correctly identified (Fig. 3b), as well as their exact locations (Fig. 3c). In summary, our simulations demonstrate that PhySigs is robust to noise and is able to accurately reconstruct relative exposures and exposure shifts within this model.

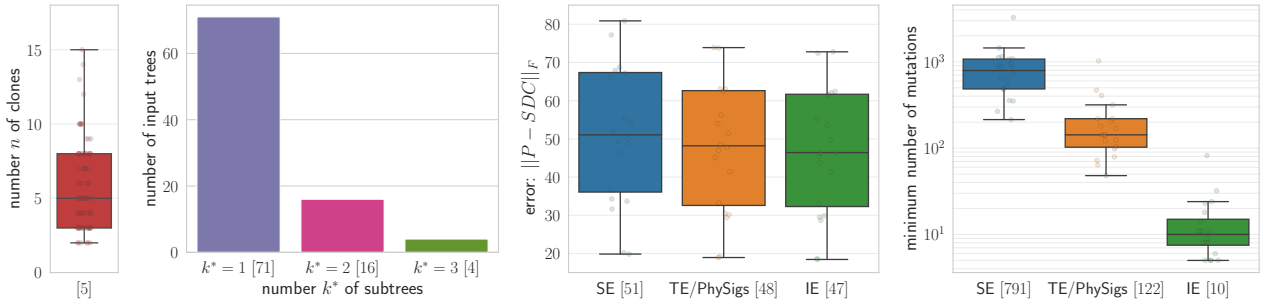
4.2. *PhySigs suggests the absence of exposure shifts in ovarian cancer*

We run PhySigs on a ovarian cancer dataset[10] composed of 7 tumors, containing between 3 to 9 clones, each with a median of 468 mutations. We apply deconstructSigs's[7] trinucleotide normalization to correct feature matrix P by the number of times each trinucleotide is observed in the genome, as this is whole-genome sequencing data. We focus our attention on COSMIC v2 Signatures 1, 3, 5, which have been designated as occurring in ovarian cancer.[b]

We find that PhySigs does not identify any exposure shifts (i.e. $k^* = 1$), assigning identical relative exposures to all clones within each patient (data not shown). This finding is corroborated when comparing PhySigs's inference error to the error obtained when solving the Independent Exposure (IE) problem, showing only a marginal decrease (median error of 149 for IE compared to 150 for PhySigs).

We see similar patterns in exposure shifts and content when additionally including *BRCA* associated signatures 2 and 13, as well as including all breast-cancer associated mutational signatures (1, 2, 3, 5, 6, 8, 10, 13, 17, 18, 20, 26 and 30).[b] It is known that ovarian cancer is predominantly driven by structural variants and copy number aberrations, which has recently motivated the use of copy number signatures rather than SNV signatures to study mutational patterns in ovarian carcinomas.[18] Indeed, examining the exposures, we find that these are

[b] `https://cancer.sanger.ac.uk/signatures_v2/matrix.png`



(a) (b) (c) (d)

Fig. 4. **PhySigs infers accurate exposures without overfitting in a lung cancer cohort of phylogenetic trees.**[11] Median values of each box plot are in square brackets. (a) This cohort contains 91 patients with 2 to 15 clones. (b) PhySigs partitions the trees into $k^* \in \{1, \ldots, 3\}$ subtrees, solving the Tree-constrained Exposure (TE) problem and selecting k^* following the Bayesian Information Criterion. (c) The relative exposure matrix D inferred by PhySigs has smaller error compared to solving the Single Exposure (SE) problem, and comparable error to solving the Independent Exposure (IE) problem. (d) The latter results in overfitting as evidenced by the small number of mutations in the smallest cluster (median of 10 [green] vs. 122 for PhySigs [orange]).

dominated by Signatures 1 and 3—Signature 1 is a clock-like signature[19] and Signature 3 is highly correlated with clock-like Signature 5 (cosine similarity of 0.83). Thus, in the absence of evidence otherwise, PhySigs will not identify exposure shifts.

4.3. *PhySigs identifies exposures shifts in a lung cancer cohort*

Jamal-Hanjani et al.[11] reconstructed phylogenetic trees for 91 lung cancer patients, with 2 to 15 clones per patient (Fig. 4b). Here, we use PhySigs to study the clonal dynamics of mutational signatures in this cohort. Since these data have been obtained using whole exome sequencing, we use deconstructSigs'[7] exome normalization feature to correct feature matrix P. We restrict our attention to COSMIC v2 Signatures 1, 2, 4, 5, 6, 13 and 17, which are associated with non-small-cell lung carcinoma.[b]

PhySigs identifies exposure shifts in 20 out of 91 patients, with a single exposure shift in 16 patients and two exposure shifts in 4 patients (Fig. 4b). To understand why PhySigs identified exposures shifts in these 20 patients, we compare the error $||P - SDC||_F$ of PhySigs's solution to the Tree-constrained Exposure (TE) against the errors of solutions to the Single Exposure (SE) problem and the Independent Exposure (IE) problems. We find that the median error of the SE problem is 51, compared to 48 for TE/PhySigs and 47 for the IE problem (Fig. 4c). The decrease between SE and TE/IE suggests that enforcing a single exposure results is a poor explanation. On the other hand, the marginal decrease between TE and IE (48 vs. 47) suggests that the exposures inferred for each cloned independently by IE likely suffer from overfitting. Indeed, Fig. 4d shows the input to the IE problem instance is composed of clones with a median number of only 10 mutations, resulting in poorly supported clone-specific exposures.

We next sought to validate the exposures inferred by PhySigs by identifying branches of phylogenetic trees with a significant change in exposure that could be explained by other

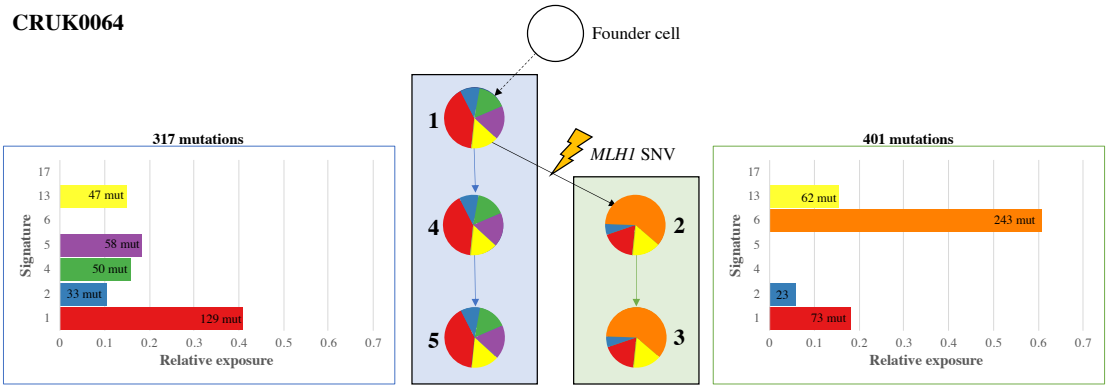


Fig. 5. **PhySigs detects a large increase in DNA mismatch repair-associated Signature 6 (orange) along one branch (clusters 2 and 3; green) of the CRUK 0064 tree.** In support of this finding, the branch includes a subclonal driver mutation to DNA mismatch repair gene *MLH1*.

observations of the tumor. We reasoned that tumors with a subclonal mutation to a gene in the DNA mismatch repair (MMR) pathway could lead to a large increase in Signature 6 (previously associated with DNA mismatch repair[2]) along one branch of the tree. Indeed, we find one such example in the lung cancer dataset, which we illustrate in Fig. 5. PhySigs finds that one subclonal branch of the tree for CRUK0064 has a putative driver mutation to MMR gene *MLH1*[c] and a high percentage of mutations from MMR-associated Signature 6 (60.7% of the 401 mutations). The remaining cancer cells outside this branch have zero Signature 6 exposure, supporting the claim that the mutation in *MLH1* is indeed driving the increase in Signature 6 exposure. We note that using an approach that does not fully incorporate a phylogenetic tree, such as the linearly ordering mutations by cancer cell fractions (CCFs) proposed by Rubanova et al.,[12] may overlook exposure shifts that are only in one branch of the tree as the signal may be drowned out by mutations in other branches with similar CCFs.

Jamal-Hanjani et al.[11] identified multiple trees for 25 out of 91 patients. This is due to the underdetermined nature of the phylogeny inference problem from bulk DNA sequencing samples.[20,21] We show that PhySigs provides an additional criterion for prioritizing alternative phylogenetic trees. Patient CRUK0025 has two alternative trees, T_1 and T_2, each composed of $n = 7$ clones, with uncertainty in the placement of clone 5 (Fig. 6). Examining the error for varying number $k \in \{1, \ldots, n\}$ of subtrees (Fig. 6a), we find that T_1 (Fig. 6b) has smaller error than T_2 (Fig. 6c) for the selected number $k^* = 3$ of subtrees according to the BIC (1,106 for T_1 vs. 1,122 for T_2), with only two exposure shifts. Moreover, to achieve a similar error in tree T_2, three exposure shifts are required (Fig. 6d). Assuming the more parsimonious explanation is more likely for a fixed magnitude of error, PhySigs's optimization criterion enables the prioritization of alternative trees in the solution space, preferring T_1 over T_2 for this patient.

5. Discussion

Based on the idea that exposures may change along edges of a tumor phylogeny, we introduced a model that partitions the tree into disjoint sets of clusters, where the clones within each clus-

[c]According to the driver mutation classifications provided by Jamal-Hanjani et al.[11]

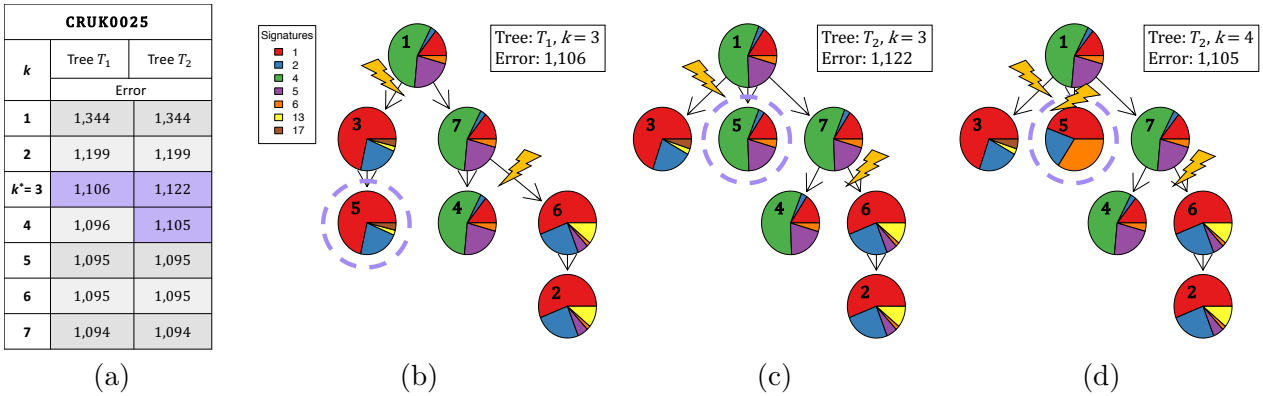


Fig. 6. **PhySigs provides an additional criterion for prioritizing tumor phylogenies when multiple solutions exist.** Jamal-Hanjani et al.[11] identified two potential tumor phylogenies for lung cancer patient CRUK0025 with discrepancies in the placement of clone 5. (a) For each tree, we show the minimum error identified by PhySigs for all k. (b) The optimal exposures inferred by PhySigs for tree T_1 for $k = 3$. Note that this solution was selected by the BIC. (c) The optimal exposures inferred by PhySigs for tree T_2 for $k = 3$. With the same number of exposure shifts, T_2 results in a higher error than T_1. (d) Three exposure shifts ($k = 4$) in T_2 are necessary to achieve the same level of error as two exposure shifts in T_1, suggesting that T_1 is the more accurate tree reconstruction.

ter are ascribed the same set of relative signature exposures. Using this model, we formulated the *Tree-constrained Exposure* (TE) problem and provided an algorithm PhySigs that includes a principled way to select the number of exposure shifts such that the mutational patterns observed at each clone are accurately reconstructed without overfitting. PhySigs unites previous work under one statistical framework, interpolating between single exposure inference[6–9] and independent exposure inference.[10,11] Our simulations demonstrated that PhySigs accurately recovers exposures and shifts. On real data, we found that while PhySigs does not detect any exposure shifts in ovarian cancer,[10] it identified several exposure shifts in non-small-cell lung cancers,[11] at least one of which is strongly supported by an observed subclonal driver mutation in the mismatch repair pathway. In addition, we showed that PhySigs enables the prioritization of alternative, equally-plausible phylogenies inferred from the same input data.

There are several avenues of future work. First, the hardness of the TE problem remains open for the case where $k = O(n)$. Second, PhySigs exhaustively enumerates all 2^n partitions of the n nodes of input tree T. It will be worthwhile to develop efficient heuristics that return solutions with small error. Third, we plan to assess statistical significance of solutions returned by PhySigs using permutation tests or bootstrapping, similarly to Huang et al.[8] Fourth, we plan to release PhySigs as a Bioconductor package. Fifth, building on our results of tree prioritization using PhySigs, we may use our model to resolve additional tree ambiguities such as polytomies (nodes with more than two children), akin to previous work in migration analysis of metastatic cancers.[22] Sixth, we plan to use our model to study population-level trajectories of clonal exposures to mutational signatures. Finally, population-level analysis of clone-specific mutations may lead to better identification of mutational signatures rather than the current tumor-level analysis.[2]

Acknowledgments. S.C. was supported by the National Science Foundation (grant: IIS 15-13629). M.E-K. was supported by the National Science Foundation (grant: CCF 18-50502).

References

1. P. C. Nowell, The clonal evolution of tumor cell populations, *Science* **194**, 23 (October 1976).
2. L. B. Alexandrov *et al.*, Signatures of mutational processes in human cancer, *Nature* **500**, 415 (August 2013).
3. A. V. Hoeck, N. H. Tjoonk, R. v. Boxtel and E. Cuppen, Portrait of a cancer: mutational signature analyses for cancer diagnostics, *BMC Cancer* **19**, p. 457 (2019).
4. H. Davies, D. Glodzik, S. Morganella, L. R. Yates *et al.*, HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures, *Nature Medicine* **23**, 517 (2017).
5. R. Schwartz and A. A. Schäffer, The evolution of tumour phylogenetics: principles and practice, *Nature Reviews Genetics* **18**, p. 213 (2017).
6. L. B. Alexandrov *et al.*, Deciphering Signatures of Mutational Processes Operative in Human Cancer, *Cell reports* **3**, 246 (January 2013).
7. R. Rosenthal, N. McGranahan, J. Herrero, B. S. Taylor and C. Swanton, deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution, *Genome biology* **17**, p. 31 (December 2016).
8. X. Huang, D. Wojtowicz and T. M. Przytycka, Detecting presence of mutational signatures in cancer with confidence, *Bioinformatics* **34**, 330 (September 2017).
9. F. Blokzijl, R. Janssen, R. van Boxtel and E. Cuppen, MutationalPatterns: comprehensive genome-wide analysis of mutational processes, *Genome Medicine* **10**, 1 (December 2018).
10. A. McPherson *et al.*, Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer, *Nature Genetics* (May 2016).
11. M. Jamal-Hanjani *et al.*, Tracking the Evolution of Non-Small-Cell Lung Cancer., *New England Journal of Medicine* **376**, 2109 (June 2017).
12. Y. Rubanova, R. Shi, R. Li, J. Wintersinger, N. Sahin, A. Deshwar, Q. Morris, P. Evolution, H. W. Group and P. network, TrackSig: reconstructing evolutionary trajectories of mutations in cancer, *bioRxiv* , p. 260471 (November 2018).
13. M. Petljak *et al.*, Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis., *Cell* **176**, 1282 (March 2019).
14. J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare *et al.*, COSMIC: the Catalogue Of Somatic Mutations In Cancer, *Nucleic Acids Research* **47**, D941 (October 2018).
15. J. Kim, K. W. Mouw, P. Polak, L. Z. Braunstein *et al.*, Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors, *Nature Genetics* **48**, 600 (2016).
16. L. D. Trucco, P. A. Mundra, K. Hogan, P. Garcia-Martinez, A. Viros *et al.*, Ultraviolet radiationinduced DNA damage is prognostic for outcome in melanoma, *Nature Medicine* , 1 (2018).
17. G. Schwarz, Estimating the Dimension of a Model, *The Annals of Statistics* **6**, 461 (March 1978).
18. G. Macintyre *et al.*, Copy number signatures and mutational processes in ovarian carcinoma, *Nature Genetics* **50**, 1262 (September 2018).
19. L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell *et al.*, Clock-like mutational processes in human somatic cells, *Nature Genetics* **47**, 1402 (2015).
20. M. El-Kebir, G. Satas, L. Oesper and B. J. Raphael, Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures, *Cell Systems* **3**, 43 (July 2016).
21. Y. Qi, D. Pradhan and M. El-Kebir, Implications of non-uniqueness in phylogenetic deconvolution of bulk DNA samples of tumors, *Algorithms for Molecular Biology* **14**, p. 19 (2019).
22. M. El-Kebir, G. Satas and B. J. Raphael, Inferring parsimonious migration histories for metastatic cancers, *Nature Genetics* **50**, 718 (May 2018).