Bioinformatics, 35, 2019, i379–i388
doi: 10.1093/bioinformatics/btz355
ISMB/ECCB 2019

OXFORD

Collaborative intra-tumor heterogeneity detection

Sahand Khakabimamaghani[1], Salem Malikic[1], Jeffrey Tang[2], Dujian Ding[1], Ryan Morin[2,3], Leonid Chindelevitch[1] and Martin Ester[1,4,*]

[1]School of Computing Science, [2]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, [3]Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4E6 and [4]Vancouver Prostate Centre, Vancouver, BC V6H 3Z6, Canada

*To whom correspondence should be addressed.

Abstract

Motivation: Despite the remarkable advances in sequencing and computational techniques, noise in the data and complexity of the underlying biological mechanisms render deconvolution of the phylogenetic relationships between cancer mutations difficult. Besides that, the majority of the existing datasets consist of bulk sequencing data of single tumor sample of an individual. Accurate inference of the phylogenetic order of mutations is particularly challenging in these cases and the existing methods are faced with several theoretical limitations. To overcome these limitations, new methods are required for integrating and harnessing the full potential of the existing data.

Results: We introduce a method called Hintra for intra-tumor heterogeneity detection. Hintra integrates sequencing data for a cohort of tumors and infers tumor phylogeny for each individual based on the evolutionary information shared between different tumors. Through an iterative process, Hintra learns the repeating evolutionary patterns and uses this information for resolving the phylogenetic ambiguities of individual tumors. The results of synthetic experiments show an improved performance compared to two state-of-the-art methods. The experimental results with a recent Breast Cancer dataset are consistent with the existing knowledge and provide potentially interesting findings.

Availability and implementation: The source code for Hintra is available at https://github.com/sahandk/HINTRA.

Contact: ester@cs.sfu.ca

1 Introduction

Cancer is the result of a gradual accumulation of somatic genetic mutations. While most of the acquired mutations are putatively neutral and have no significant effect on a cell's phenotype, some confer a selective advantage to the host cell; they are known as *driver mutations*. Consequently, individual tumors are heterogeneous and typically consist of multiple populations of cells (subclones), each harboring a distinct set of driver mutations and possessing a distinct phenotype, a phenomenon known as intra-tumor heterogeneity (ITH). Detecting ITH helps identify the key events initiating the development of the disease or leading to metastasis, and allows for the determination of a tumor's subclonal composition.

ITH is usually represented by a phylogenetic tree of tumor evolution, where individual nodes represent subclones. Mutations are placed along the edges of the phylogenetic tree. A mutation is assigned to the unique edge directed into the node (subclone) in which the mutation appears for the first time (see Fig. 3 for an example of a phylogenetic tree). In this work, we assume that the mutations satisfy the infinite sites assumption. This assumption means that each mutation appears exactly once in the phylogenetic tree and is present (conserved) in all the descendants of the subclone in which it first occurs.

Most of the existing methods for studying tumor evolution operate on tumor data from a single cancer patient. The earliest developed methods used bulk sequencing data from a single sample [e.g. rec-BTP (Hajirasouliha *et al.*, 2014), CTPsingle (Donmez *et al.*, 2017)] or multiple samples from the same individual [e.g. PhyloWGS (Deshwar *et al.*, 2015), AncesTree (El-Kebir *et al.*, 2015), LICHeE (Popic *et al.*, 2015), CITUP (Malikic *et al.*, 2015)]. These were followed by the development of several methods that work on single-cell data [e.g. OncoNEM (Ross and Markowetz, 2016), SCITE (Jahn *et al.*, 2016), SiFit (Zafar *et al.*, 2017)]. The most recently introduced methods, B-SCITE (Malikic *et al.*, 2017)

and PhISCS (Malikic *et al.*, 2018), simultaneously utilize the complementary strengths of both single-cell and bulk sequencing data. Indeed, most of the methods above have limitations when faced with the input consisting of a single sample low-to-medium coverage bulk sequencing dataset, which are predominant in existing databases (e.g. TCGA and cBioPortal). Since these type of data contain numerous ambiguous cases (i.e. cases where the input data are consistent with more than one possible phylogenetic tree for the tumor), the existing algorithms for ITH detection based on a single tumor sample [e.g. CTPsingle (Donmez *et al.*, 2017)] will yield several possible solutions for those cases (Malikic *et al.*, 2017).

In addition to ITH, inter-tumor heterogeneity is another phenomenon complicating the understanding and treatment of cancer. Inter-tumor heterogeneity is a direct consequence of the fact that individual tumors are genetically distinct. Despite the inter-tumor heterogeneity, one can still expect partially similar evolutionary trajectories among subsets of tumors (Caravagna *et al.*, 2018; Pathare *et al.*, 2009). Leveraging the phylogenetic similarities among tumors from a cohort of patients in a collaborative fashion can guide the process of exploring the solution space and reduce the above-mentioned ambiguities in inferring tumor phylogenies, especially for cases when the input is low-to-medium coverage bulk sequencing data from a single tumor sample. Most of the existing methods that infer phylogeny at the ensemble level are based on binary mutation data. Some of these methods [e.g. CAPRESE (Loohuis *et al.*, 2014), CAPRI (Ramazzotti *et al.*, 2015) and Beerenwinkel *et al.* (2004)] exploit Suppes' probabilistic causation theory (Suppes, 1970) to determine the pairwise order of mutations. Some other methods [e.g. conjunctive Bayesian networks (Beerenwinkel *et al.*, 2007; Gerstung *et al.*, 2009) and Bayesian mutation landscape (Misra *et al.*, 2014)] model the phylogenetic relationships as a Bayesian network and propose approaches for learning the network structure. RESIC method proposed by Attolini *et al.* (2010) is based on the principles of population genetics and models the tumor samples as individuals in the steady-state of a genetically evolving population and infers the evolutionary history of their genotypes. Since all of these methods use binary mutation data, they do not fully utilize the potential of sequencing data by overlooking the intrinsic information about the timing of evolutionary events. Moreover, these methods gain general knowledge about cancer progression and do not provide personalized evolutionary details.

A recent method, Revolver (Caravagna *et al.*, 2018), uses non-binary sequencing data and exploits the repeating evolutionary patterns for ITH detection in individual tumors by transferring information across all tumors. In this method, the assumption is that a particular mutation usually has the same predictor (preceding mutation) across different tumors in a particular cancer type. Accordingly, the authors consider the frequency of the direct ancestors of a mutation across different tumors and use that information when inferring the phylogeny for a specific tumor. Revolver uses an expectation–maximization (EM) approach for finding the optimum phylogenetic trees. In the first step, an existing method [e.g. ClonEvol (Dang *et al.*, 2017)] is used for deriving a set of high-scoring candidate phylogenetic trees for each tumor, the best of which is chosen as the current tree for each tumor. Then, the frequencies of the direct ancestors of each mutation are learned from the currently selected trees for all tumors. This information, which constitutes the parameters of the distribution over tree topologies, is then used for re-evaluating the tree set for each tumor and selecting the ones with the highest new scores. These two steps of updating the parameters/frequencies (E-step) and updating the current trees based on the new parameters (M-step) continue until convergence or until termination criterion is met.

This approach decreases the uncertainty of phylogenetic structures by incorporating the ancestry information. However, the underlying evolutionary assumption in Revolver, which is the dependency of a mutation only on its direct ancestor (the preceding mutation), is a limitation because earlier mutations inherited by a subclone might also be decisive in the selection of the next mutation during the cancer evolution. Therefore, considering only the direct ancestor as the predictor of a mutation might result in a loss of information. Another issue, which is discussed further in Section 2.3, is that the tree topology distribution used in Revolver is biased toward more branching structures. If not controlled, this bias may produce unrealistic results with too much branching.

In this paper, we discuss the consequences of the above key issues and introduce a collaborative ITH detection method to address them. Our contributions can be summarized as follows:

- We introduce a Probabilistic graphical model (PGM) called Hintra for collaborative ITH detection, as well as a corresponding parameter learning method. The proposed PGM is based on read count data, instead of summary values such as cancer cell fraction (CCF) or variant allele frequency, to account for the uncertainty of the measurements. To reduce the bias of existing methods, we propose a Bayesian EM method that leverages the topology uncertainty when learning the parameters, using a distribution over possible phylogenetic tree topologies instead of a point estimate.
- Hintra includes a novel factorization approach for phylogenetic tree topologies. Addressing the information loss issue mentioned earlier, Hintra considers all the mutations preceding a particular mutation in the phylogenetic tree, instead of only the most recent one. Moreover, the proposed factorization allows for the prediction of the next mutation that might happen in a subclone given its current mutational landscape. This capability, which is lacking in the existing methods, can be used for prognostic clinical applications.

Using both synthetic and real data, we evaluate performance of Hintra and compare it to the state-of-the-art methods including Revolver (as a collaborative ITH detection method) and ClonEvol (as a stand-alone ITH detection method). Our results for synthetic data based on different scenarios indicate that Hintra outperforms the existing methods. Our results for real data were biologically consistent and provided new information of potential clinical interest.

2 Materials and methods

For the sake of simplicity, the methods in Sections 2.1–2.4 are presented assuming that a single sample is available for each tumor. Later, in Section 2.5, we generalize our model to allow multiple samples per tumor.

2.1 Problem definition

We now formally define the collaborative ITH detection problem. We assume that the input consists of read count data across m tumors. For each tumor, we consider read counts for a given set G of n known driver genes. The input data are organized into two matrices, one for the reference read counts denoted by $R = [r_{ij}] \in \mathbb{N}_0^{m \times n}$ and the other for the variant read counts denoted by $V = [v_{ij}] \in \mathbb{N}_0^{m \times n}$, where $\mathbb{N}_0$ denotes the set of non-negative integers. More precisely, r_{ij} and v_{ij} respectively denote the number of reference and variant reads supporting driver gene j in tumor i.

The output is a set of phylogenetic trees $\{T_i\}_{1 \leq i \leq m}$, where T_i is the phylogenetic tree of tumor i indicating the phylogenetic order of mutations in that tumor. A phylogenetic tree is a representation of the evolutionary events that are observed in a tumor. The root of the tree corresponds to the germline cell and the other nodes indicate the subclones of the tumor. Each edge stands for a mutation that occurs in a cell of the subclone corresponding to the edge's tail (the parent) and triggers the growth of the subclone corresponding to the edge's head (the child). Our goal is to infer the tree for each tumor by considering the evolutionary patterns of similar mutations in the other tumors' trees.

As a byproduct, we also learn model parameters that can be used to compute the probability that a particular mutation occurs in a cell having a specific set of mutations. For example, the parameters can contain information on the most frequent mutation occurring in (i.e. providing competitive advantage to) a breast cancer cell already containing the mutations *TP53* and *PIK3CA*. This parameter provides predictive information with prognostic applications.

Although it is theoretically possible to consider the exact position where each of the input mutations occurs within its gene, we chose to analyze the data at the gene level to increase the frequency of each mutation and gain statistical power. For genes affected by multiple mutations in the same tumor, we use the read count data for the most prevalent of such mutations, i.e. the mutation with the largest CCF. The CCF represents the fraction of cells of a tumor that harbor a particular mutation and most of the existing methods preprocess the read counts from sequencing data into CCFs before using them for phylogeny inference. This allows the use of the existing CCF computation tools [e.g. PyClone (Roth *et al.*, 2014)] that can handle complicated cases such as mutations involving copy number variations (CNV). However, it ignores the uncertainty in the computed CCFs, which may lead to incorrect results by assigning high weights to uncertain inputs or vice versa. Incorporating read counts directly into the inference provides a more accurate representation and can help prioritize informative inputs over uncertain ones. Moreover, in cases with CNV, the computed CCFs can be simply translated into read counts based on an appropriate approximation of the locus coverage (e.g. mean sequencing coverage). Therefore, we choose read counts as the format of our input.

2.2 Probabilistic graphical model

The proposed PGM for Hintra is shown in Figure 1 . In this model, each tumor i, for $1 \leq i \leq m$, is associated with a phylogenetic tree T_i, whose structure depends on a parameter β. The tree structure constrains the possible values of the read count data. This is done through a latent variable $\theta_{i.}$, which is a vector of size n of CCFs of driver mutations in tumor i. The dot in the index $i.$ denotes a vector. The CCF of a mutation indicates the proportion of cells in the tumor sample that harbor that mutation. A larger CCF is, in general, evidence of earlier occurrence of the mutation during tumor evolution. Accordingly, $\theta_{i.}$ depends on the tree structure T_i corresponding to tumor i and influences the noisy observed reference and variant read counts for tumor i.

According to the PGM, the joint probability of the model variables is factored as:

$$P(V, R, \theta, T, \beta) = P(V|R, \theta)P(\theta T)P(T|\beta) \tag{1}$$

The first term on the right hand side of Equation (1) is the likelihood term and is defined as below:

$$\begin{aligned} P(V|R, \theta) &= \prod_{i=1}^{m}\prod_{j=1}^{n} P(v_{ij}|r_{ij}, \theta_{ij}) \\ v_{ij}|r_{ij}, \theta_{ij} &\sim \text{Binomial}\,(v_{ij} + r_{ij}, \theta_{ij}/2) \end{aligned} \tag{2}$$

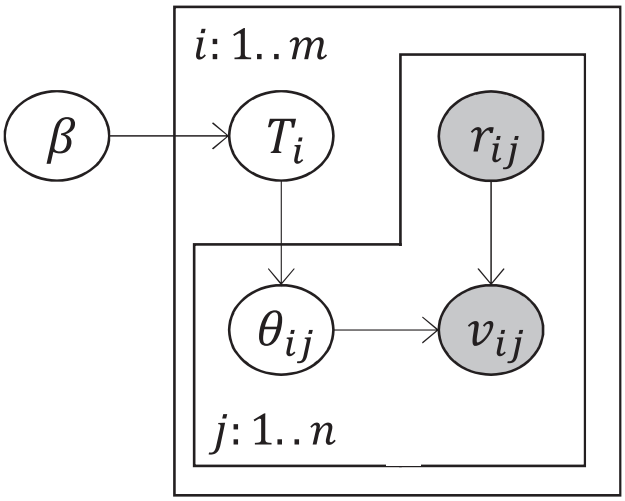


Fig. 1. Probabilistic graphical model of Hintra. Latent and observed variables are indicated by white and shaded circles, respectively

The Binomial distribution parameter is equal to $\theta_{ij}/2$ because CCF is computed as $\theta_{ij} = \frac{2v_{ij}}{v_{ij}+r_{ij}}$ for driver mutation j of tumor i (note the multiplication by 2 in the nominator). The second factor in RHS of Equation (1) is defined as below:

$$\begin{aligned} P(\theta|T) &= \prod_{i=1}^{m} P(\theta_{i.}|T_i) \\ \theta_{i.}|T_i &\sim \text{Uni(possible values)} \end{aligned} \tag{3}$$

The possible values for vector $\theta_{i.}$ are restricted by: (i) the *sum rule* indicating that the CCF for a mutation should not be smaller than the sum of the CCFs of all of its children in the phylogenetic tree T_i (Jiao *et al.*, 2014) and (2) $0 \leq \theta_{ij} \leq 1$ for $1 \leq j \leq m$.

The third factor and its computation are discussed in Section 2.3.

2.3 Prior probability of phylogenetic trees

The underlying assumption of collaborative ITH detection is that some of the evolutionary patterns (i.e. phylogenetic relationships between the evolutionary events in a tumor) are common among different tumors. Accordingly, the goal is to define the entire phylogenetic tree in terms of its substructures representing the evolutionary patterns. One can then investigate the frequency of the patterns to find the more frequent patterns and use them as a reference whenever there is ambiguity for a tumor with respect to the phylogenetic relationships between the events involved in those frequent patterns. Here, ambiguous case refers to the case where multiple phylogenetic trees are consistent with the observed bulk data read counts. For a simple example of an ambiguous case we can consider a tumor with CCF values $[0.2,\ 0.3,\ 1.0]$. In this case, relying solely on CCF values, one can easily observe that both the chain and the branching topology are possible explanations of the observed data. Several more complicated examples for this were recently provided in the analysis of acute lymphoblastic leukemia patients in Malikic *et al.* (2017). For an example of a non-ambiguous case we can consider a tumor with mutations having CCFs $[0.5,\ 0.8,\ 1.0]$. In this case, only the chain topology is consistent with the observed CCFs. Namely, for the branching topology, the frequencies of the two child nodes would add up to 1.3, which is larger than the CCF of their parent, thus violating the sum rule.

To the best of our knowledge, the most recent ITH detection method that is based on the assumption of common evolutionary patterns is Revolver (Caravagna *et al.*, 2018). Revolver assumes independence of the edges and defines the probability of the phylogenetic tree of tumor i as the product of the probabilities of the observed edges (i.e. the probability of attaching a given child node to a particular parent node) as follows:

$$P(T_i|\beta) \propto \prod_{p \to c \in E_i} P(p|c,\beta), \tag{4}$$

where p and c are the parent and child nodes of a given edge $p \to c$ of tree T_i, and E_i is the set of all edges of the tree for tumor i. The parameter β governs the edge probabilities and is shared across all tumors.

In the above approach, each node is assumed to be dependent only on its direct ancestor. However, the selection of the next mutation that brings competitive advantage to a cell does not only depend on the last mutation, but it depends on the entire current mutational burden of the cell. Figure 2A shows a scenario in which the above assumption is violated, leading to a poor performance for Revolver (see Section 3.1). In this scenario, the two trees have different truncal mutations (**a** for topology 1 and **e** for topology 2). Because of this difference, mutation d is attached to different parents in the two trees. However, considering only pairwise relationships, because **d** happens after **b** in 70% of the tumors, the factors of Revolver will assign it under **b** even when inferring trees for a tumor having true topology 2.

Another drawback of the above factorization is that it cannot be translated into a prognostic application of predicting the next driver mutation based on the current mutational landscape of a subclone. The reason is that the conditional probability of the next mutation given all current mutations is not computable as a function of the parameters learned based on the above assumptions. In other words, based on the above assumptions, the next mutation depends only on the most recent ancestor.

To overcome the above limitations, we extend the tree factorization to capture the effect of all existing driver mutations (ancestors) on the next driver mutation (descendant). The occurrence order of the ancestors is not taken into account because the selection of the next mutation depends only on the set of current mutations, but not the order of these mutations. We define the prior tree probability as below:

$$P(T_i|\beta) \propto \prod_{\mathcal{P} \to c \in f(T_i)} \beta_{\mathcal{P},c}, \tag{5}$$

where $\mathcal{P}$ is the set of possible ancestor mutations of c in tree T_i, which we call the *ancestry set* hereafter. An ancestry set consists of all the mutations on the path in T_i from the root to any internal (i.e. non-leaf) node/subclone and captures the mutational landscape of that node/subclone. The function $f(T_i)$ returns the set of edges of T_i consisting of the ancestry sets and their children. The parameter β is a matrix with rows corresponding to all possible ancestry sets for all tumors in the dataset. The columns correspond to the set G of all mutations. The entry β_{kj} of the β matrix indicates the amount of evidence for an edge labeled with the j-th mutation whose tail is a node with the k-th ancestry set. Figure 3 illustrates these concepts.

2.4 Parameter learning

Although the proposed prior probability in Section 2.3 resolves the information loss issue, it inherits the bias toward branching structures. Scenario B in Figure 2 shows a sample situation where this bias can lead to unexpected phylogeny detection. In topology 1 in Scenario B, mutation **b** occurs after **a**, which is not consistent with topology 2, in which **b** occurs after **c**. A similar inconsistency exists between the ancestors of mutation **a** in the two topologies. Accordingly, based on both the Revolver and Hintra factorization approaches, any ambiguous case that suggests a branching topology in which **a** and **b** can occur in parallel has supporting evidence due to the conflicting orders of **a** and **b** in the two topologies, even if it is originally associated with one of the two topologies. However, in case of a slight ambiguity (e.g. a 5% ambiguous cases for each of the two topologies), the evidence for the branching topology is very

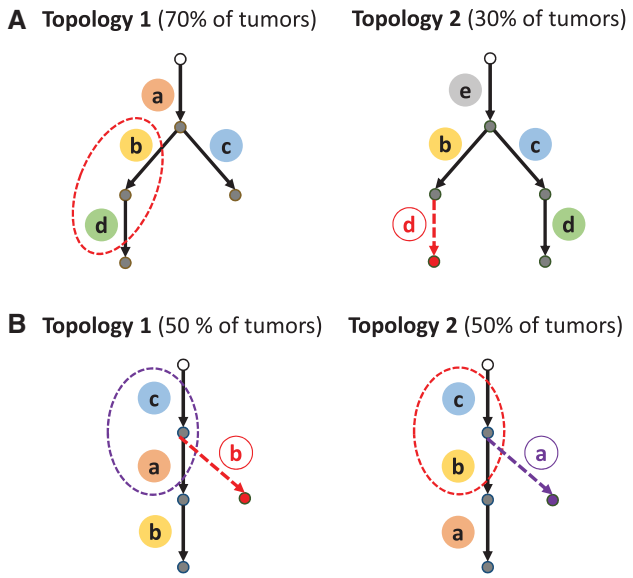
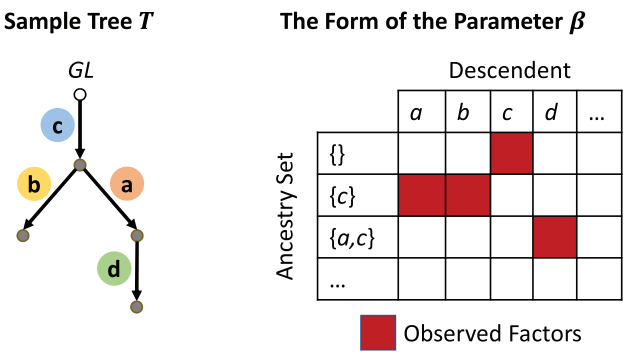


Fig. 2. Two sample scenarios in which tree factorization and parameter learning as in Caravagna *et al.* (2018) results in undesired inference. The small circles denote the tumor subclones and the empty circle is the germline cell. The edges are labeled with the mutations, denoted by letters within larger circles. The true tree topologies are shown with solid edges. Each ambiguous situation is shown in a different color, with dashed ovals indicating the conflicting evidence (source of ambiguity) and the dashed edge indicating the possible mistake due to that evidence



Factorization of Tree T

$$f(T) = \{\{\} \to c, \{c\} \to b, \{c\} \to a, \{a,c\} \to d\}$$

$$P(T|\beta) \propto \beta_{\{\},c}\, \beta_{\{c\},b}\, \beta_{\{c\},a} \beta_{\{a,c\},d}$$

Fig. 3. A sample phylogenetic tree and its factorization

small and the chain topology should be favored (which has support from, for example, 95% of the samples). So, if the inherent bias toward branching topologies in the factorization approaches is not controlled, the methods infer the incorrect branching structure (shown with the dashed edges) for the ambiguous cases. We control this bias by employing a Bayesian EM parameter learning method described next. This method accounts for uncertainty of each of the possible topologies when learning the parameters and, in this scenario, only accepts a branching topology in cases with high certainty (i.e. when the subclones corresponding to **a** and **b** are very small).

We propose a Bayesian EM approach to learn the parameters of the PGM of Hintra. The goal is to optimize the value of β, the topology distribution parameter, by maximizing the marginal likelihood $P(V|R,\beta)$ and utilizing the data's uncertainty. This is performed using an iterative approach with the following steps at each iteration:

1. Compute $\beta^{'}$ using $P(T|\beta,V,R)$ [see Equation (6)].
2. If $P(V|R,\beta) < P(V|R,\beta^{'})$ [see Equation (10)], then set $\beta = \beta^{'}$ and continue; otherwise output $\beta^{'}$ and terminate.

Initially, the tree priors are assumed to be uniform. Then, in the first step, $\beta_{\mathcal{P},c}$ is updated for each ancestry set $\mathcal{P}$ and descendant mutation c using the following equation:

$$\beta^{'}{}_{\mathcal{P},c} = \sum_{i=1}^{m}\sum_{T_i} 1_{f(T_i)}(\mathcal{P}\rightarrow c)\times P(T_i|\beta,v_{i.},r_{i.}) + \epsilon, \tag{6}$$

where $1_{A(x)}$ is the indicator function for $x \in A$ and the value ϵ is the pseudo-count for avoiding zero probabilities. Equation (6) is the sum of evidence for factor $\mathcal{P}\rightarrow c$ over all tumors, where the evidence is weighted by the posterior likelihood of every possible tree topology that contains the factor $\mathcal{P}\rightarrow c$. Accordingly, $\beta^{'}_{\mathcal{P},c}$ indicates the updated evidence for the factor $\mathcal{P}\rightarrow c$. The posterior likelihood for tree topology is computed as:

$$P(T_i\beta,v_{i.},r_{i.}) = \frac{P(v_{i.}|r_{i.},T_i)P(T_i|\beta)}{\sum_X P(v_{i.}|r_{i.},X)P(X|\beta)}. \tag{7}$$

In the above equation, the marginal data likelihood is computed as below:

$$P(v_{i.}r_{i.},T_i) = \int_{\theta_{i.}}\prod_{j=1}^{n} P\left(v_{ij}|r_{ij},\theta_{ij}\right)P\left(\theta_{ij}|T_i\right)\mathrm{d}\theta_{i.}. \tag{8}$$

Because the term containing the integral over the vector $\theta_{i.}$ is not in closed form, we approximate that term using discrete values as below:

$$P\left(v_{i.}|r_{i.},T_i\right) \approx \sum_{\theta_{i.}\in\delta_\Delta(T_i)}\prod_{j=1}^{n} P(v_{ij}|r_{ij},\theta_{ij})P(\theta_{ij}|T_i)\Big), \tag{9}$$

where $\delta_\Delta(T_i)$ is a function that enumerates all discrete values of the vector $\theta_{i.}$ with step-size Δ considering the constraints imposed by topology T_i. In our experiments (Section 3), we use $\Delta = 0.05$.

In the second step, the marginal probability conditional on β (i.e. the maximization objective) is computed as:

$$P(VR,\beta) = \prod_{i=1}^{m}\sum_{T_i} P(v_{i.}|r_{i.},T_i)P(T_i|\beta). \tag{10}$$

Figure 4 illustrates, with an example, the Bayesian EM approach described above as well as the EM approach used in Revolver, which uses MAP point estimate. It explains how using a Bayesian approach that employs the entire spectrum of possible topologies (i.e. data uncertainty) instead of the point estimates as used in Revolver can reduce the bias inherent in both of the tree prior probability definitions used in Revolver and Hintra. The figure shows the first step of different EM approaches on a dataset with three tumors all having two driver mutations, *a* and *b*. Topologies *A*, *B* and *C* constitute all possible trees with the two mutations. Each of the three tumors has a different topology, as shown. The bar charts show the initial posterior probabilities of the topologies [i.e. $P(T|D) \propto P(D|T)P(T|\beta)$] for the tumors computed assuming a uniform initial topology prior $P(T|\beta)$ and hypothetical data likelihoods. Two types of posteriors are computed, one based on the Bayesian estimation (the top row) and one based on the MAP estimation (the bottom row). The evidence β updated based on the two types of posteriors is shown in the middle. At the right, the updated priors based on the updated evidences are presented. Despite the fact that each of the three topologies is observed only once, the bias in the prior definition makes the most branching topology *B* more likely. However, the Bayesian approach considers the entire spectrum of possible topologies (i.e. based on the data likelihood given each possible topology), which reduces the bias. As shown in this figure, ambiguous cases like *B* often have a more uniform distribution over topologies than the other cases, resulting in reduced support for branching. This mitigates the effects of the prior bias during the learning process. Besides this, since we optimize the marginal data likelihood

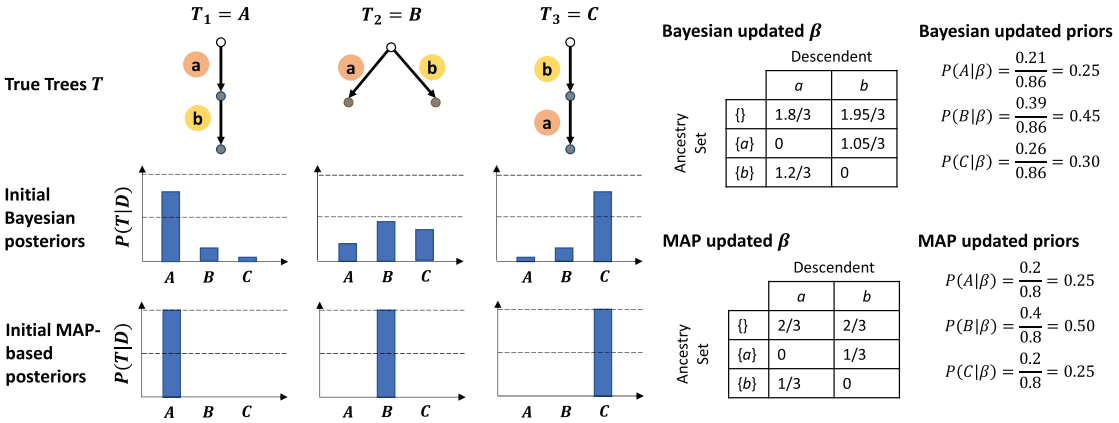


Fig. 4. Bias in the topology prior probabilities and how the Bayesian approach mitigates this bias

[Equation (10)] instead of the maximum likelihood, we adapt the parameter β to the whole information in the observed data as opposed to fitting it to the point estimates. This results in further reduction in bias.

To reconstruct the phylogenetic tree structures, we use MAP estimation after the parameter β has been computed using the above Bayesian EM approach. For each tumor i, we have:

$$T_i = \arg\max_{T_i}\{P(v_{i.}|r_{i.}, T_i)P(T_i|\beta)\} \tag{11}$$

2.5 Generalization to multiple samples per tumor

In the more general case where multiple samples (obtained, for example, by sequencing multiple regions of the tumor) are available for a given tumor, we define the likelihood of tumor data [previously shown in Equation (9) for the single-sample case] as the product of the likelihoods of the individual samples:

$$\begin{aligned} P(v_{i.}|r_{i.}, T_i) &= \prod_{q=1}^{s_i} P(v_{i.}{}^q|r_{i.}{}^q, T_i) \\ &\approx \prod_{q=1}^{s_i} \sum_{\theta_i \in \delta_\Delta(T_i)} \prod_{j=1}^{n} P(v_{ij}{}^q|r_{ij}{}^q, \theta_{ij}) P(\theta_{ij}|T_i), \end{aligned} \tag{12}$$

where s_i is the number of samples for tumor i and $v_{i.}^q$ and $r_{i.}^q$ are the read count data for sample q of tumor i.

2.6 Extracting prognostic information

The likelihood that each mutation c follows an ancestry set $\mathcal{P}$ is computed as:

$$P(c|\mathcal{P}) = \frac{\beta_{\mathcal{P},c}}{\gamma_{\mathcal{P}}}, \tag{13}$$

where $\gamma_{\mathcal{P}}$ is the evidence for $\mathcal{P}$ computed as:

$$\gamma_{\mathcal{P}} = \sum_{i=1}^{m} \sum_{T_i} 1_{g(T_i)}(\mathcal{P}) \times P(T_i|\beta, v_{i.}, r_{i.}) + 2\epsilon, \tag{14}$$

where the function $g(T_i)$ returns all the ancestry sets in tree T_i.

Because $P(c|\mathcal{P})$ is a proportion estimate, the minimum value for the sample size $\gamma_{\mathcal{P}}$ to have a 95% confidence interval of width W can be computed as $4/W^2$ (e.g. $\gamma_{\mathcal{P}} \geq 100$ for $W = 0.2$).

3 Experiments and results

3.1 Experiments with synthetic data

We evaluated the performance of Hintra using synthetic data to have access to the ground-truth phylogenetic trees. The comparison partners included Revolver (Caravagna *et al.*, 2018), as the only method that explores a similar idea of collaborative ITH detection, and ClonEvol (Dang *et al.*, 2017), as the state-of-the-art method for stand-alone ITH detection. We used the same evaluation metric as in Caravagna *et al.* (2018), namely true positive ratio, which is the proportion of predicted edges that exist in the ground-truth tree.

For comparison with Revolver, we conducted three different experiments. The first experiment evaluated the information transfer and de-noising capabilities of Hintra. In this experiment, we followed exactly the same simulation procedure used for evaluating Revolver. The second experiment showcased one of our main contributions, the ability of Hintra to capture more complete evolutionary patterns. This experiment was based on Scenario A in Figure 2. The third experiment examined the ability of Hintra to control the topology distribution bias and it was based on Scenario B shown in Figure 2.

As in Caravagna *et al.* (2018), the sensitivity to CCF noise levels are monitored in the three experiments, where noise follows a Gaussian distribution and was controlled through tweaking the SD (e.g. 0 or 0.05). Moreover, ambiguity was introduced into the ground-truth models as the percentage of tumors with CCFs that had different possible phylogenetic structures (i.e. ambiguous cases). These experiments were conducted assuming a single sample per tumor. To evaluate the effect of the number of samples on the methods' performance, we conducted an additional set of experiments where 2 or 4 samples were generated per tumor, and considering the most difficult simulation configuration, i.e. higher noise and ambiguity. All samples of a tumor were assumed to be ambiguous for ambiguous cases and they were all non-ambiguous otherwise. For each configuration of the parameters, the experiment was repeated 10 times with *de novo* generation of the synthetic data at each repetition. For each single repetition, the average true positive ratio over all tumors was computed and plotted as a point. We simulate a sequencing coverage of 100× in all experiments.

The synthetic data in Caravagna *et al.* (2018) was produced by assuming the same evolutionary tree (a chain structure) consisting of four mutations across all tumors. We repeated the same experiments for a cohort of 50 tumors. Two CCF noise levels of 0 and 0.05 and three different percentages of ambiguous cases (10, 30 and 50%) were simulated as in the original paper. The results are shown in Figure 5. According to Figure 5, unlike ClonEvol, both Hintra and Revolver were able to detect the correct phylogenetic trees for all tumors and for all levels of ambiguity when there was no noise in the CCFs. However, after introducing noise with SD 0.05 to the true CCFs, the level of ambiguity had a stronger effect on performance. Hintra outperformed Revolver in all the datasets with noise, and the gap between the two methods increased with increasing level of ambiguity. This indicates the higher robustness of Hintra to noise and ambiguity. Interestingly, by increasing the number of samples, the

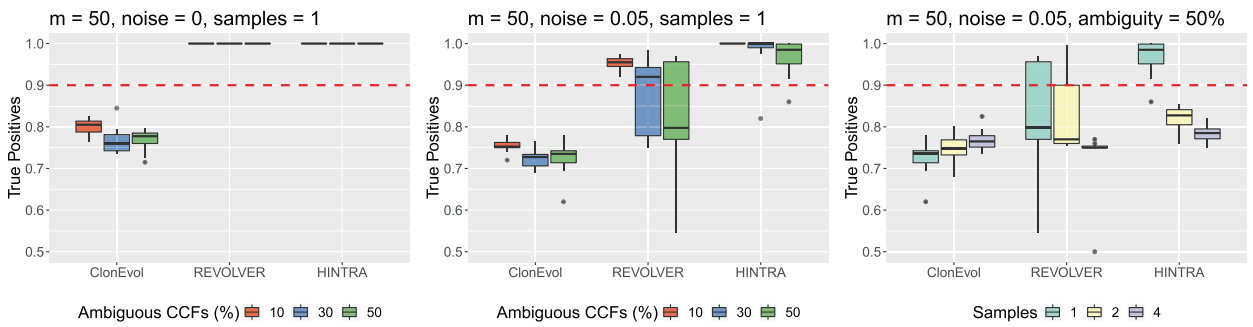


Fig. 5. Results for the synthetic datasets from Caravagna *et al.* (2018)

performance of the stand-alone algorithm improved but the collaborative methods exhibited decreasing accuracy. These results are consistent with the original study (Caravagna *et al.*, 2018). The most likely reason for this is the high level of ambiguity (50%). For ambiguous cases, multiple noisy samples create conflicting phylogenies which leads to a higher probability for branching structures. Thus, transferring information from the ambiguous cases decreases the overall accuracy. The stand-alone method is less sensitive because each tumor is analyzed separately. We note that it is very unlikely that in real data all samples of a tumor with a ground-truth chain structure are ambiguous as we assumed here. Hintra, in general, performs slightly better than the two other methods for larger numbers of samples.

In the second experiment, we simulated Scenario A shown in Figure 2 for a cohort of 50 tumors (35 with tree 1 and 15 with tree 2). Because the ground-truth topologies have branches and so are associated with ambiguous cases, we only investigated the noise level and the number of samples as the variable factor. Two noise levels of 0 and 0.05 were simulated. The results are shown in Figure 6. We observe that there are considerable gaps between the performances of the three methods. As explained earlier, the gap between Revolver and Hintra is due to the differences in the definitions of the tree topology factors, where Hintra looks further back into the evolutionary history of a subclone and provides a more accurate assignment of the mutations based on that richer information. Unlike the previous scenario, having more samples improves the performance of the methods in this scenario. This is due to the fact that the trees are consistently branching topologies in this scenario. These topologies are associated with ambiguous cases and, unlike for chain topologies, having multiple ambiguous noisy samples is not misleading for these cases. Overall, Hintra performs slightly better than the two other methods with a larger number of samples.

For the third experiment, we simulated Scenario B as shown in Figure 2. Two levels of noise (0 and 0.05) were introduced to the CCFs and 6, 10 and 14% were used as the frequency of ambiguous

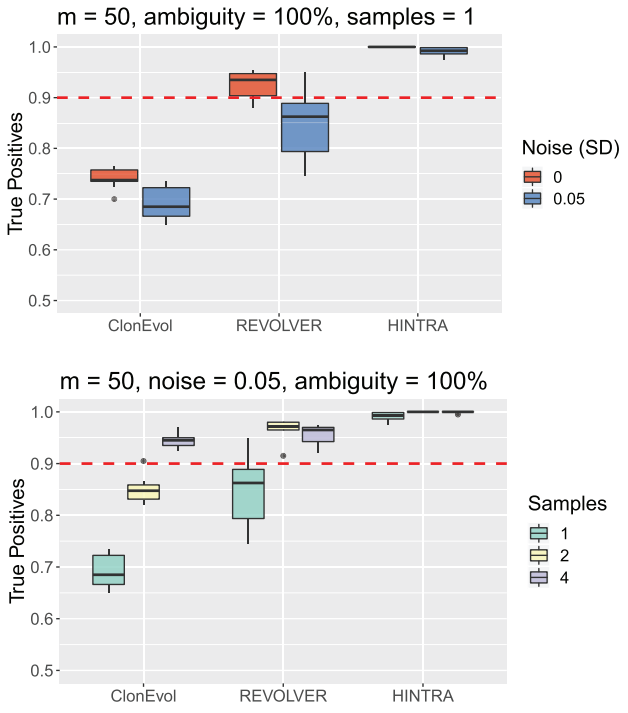


Fig. 6. Results for the synthetic datasets based on Scenario A from Figure 2

cases. The goal of this experiment was to investigate the capability of the methods to control the bias toward branching structures. Accordingly, we set small ambiguity levels to leave enough evidence for the true structures and examined whether the methods could still infer branching structures in absence of direct evidence. The branching structure was still supported indirectly due to the two conflicting structures of trees 1 and 2, but there were not enough ambiguous cases to support that structure. Figure 7 shows the results. Because of the low levels of ambiguity, the true positive rates were in general high for all methods. However, CloneEvol and Hintra performed better than Revolver in this experiment. CloneEvol performed stand-alone phylogeny inference and chose one of the two possible topologies (chain and branching) at random. However, Revolver had a bias toward branching structures and preferred that structure whenever it was possible. Hintra had less bias in the topology distribution due to the Bayesian EM approach and opted for the branching structure only whenever it had high probability. Therefore, Hintra controlled the bias effectively for all levels of ambiguity and noise in our experiments. When more samples were used per tumor with a noise of 0.05 and an ambiguity of 14%, all the methods showed improvement. Due to the small ambiguity, more samples improved the evidence for the correct topology and strengthened the information transferred from those cases, which resulted in a better resolution for the ambiguous cases. Overall, Hintra once again outperformed the other methods when multiple samples were used.

3.2 Experiments with real data

In the absence of ground-truth phylogenetic trees for tumor mutations, performing an objective comparison of the accuracy of Hintra and any other method is difficult. Instead, we evaluated Hintra's performance based on the consistency of the learned parameters with existing biological domain knowledge. We chose Breast Cancer as the subject of study, since it is one of the most studied cancer types for which a rich body of domain knowledge is available. We used a public dataset from Razavi *et al.* (2018), which includes 1756 advanced breast cancer patients. This dataset is the most recently published genomic dataset for Breast Cancer with clinical data.

In the available clinical data, these patients were stratified according to whether they do or do not express the genes for the receptors for the hormones estrogen and progesterone (HR) and HER2, resulting in the HR+/HER2−, HR+/HER2+, HR−/HER2+ and TN (Triple Negative) subtypes. We used this information to separate the patients into four corresponding groups and ran Hintra for each group independently to infer tumor progression and phylogenetic trees. We only included tumors having single nucleotide variations (SNVs) and a normal copy number in the considered loci. We considered mutations in breast cancer genes from COSMIC Cancer Gene Census dataset [cancer.sanger.ac.uk (Harsha *et al.*, 2016)] and augmented the list by the genes mentioned in the original study (Razavi *et al.*, 2018). After limiting to the selected genes and filtering out synonymous mutations, loss of heterozygosity and weak signals (i.e. small read counts and mutations with <1% frequency in each subtype), the number of patients with at least one mutation was reduced to 1348. We also limited the number of mutated genes per tumor to 5 and removed the 47 tumors (3.5%) that did not satisfy this constraint.

A large subset of the cohort was HR+/HER2− cases. In a majority of cases in this subtype, we observed clonal mutation acquisition in signaling cascades (TP53, PIK3CA, AKT, GATA3, PTEN, etc.) as discussed in the original findings (Razavi *et al.*, 2018), suggesting that Hintra is able to reliably detect these early mutational

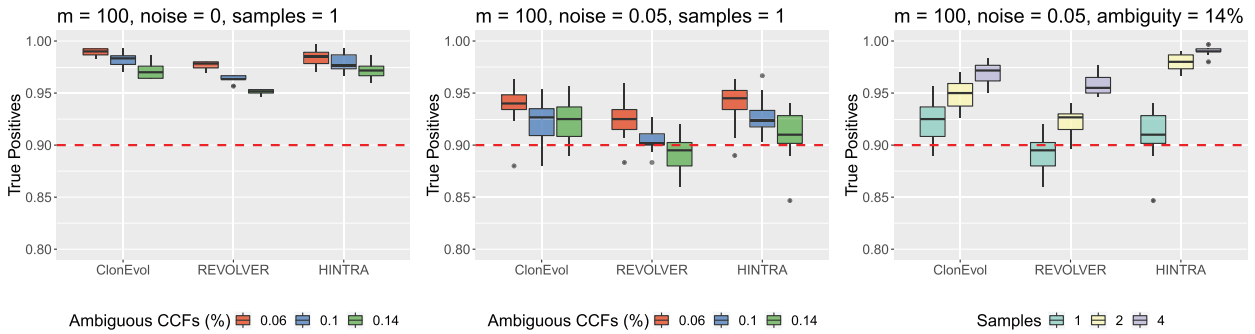


Fig. 7. Results for the synthetic datasets based on Scenario B from Figure 2

associations. Hintra also found that the most likely descendant of TP53 and PIK3CA combinatory events in HR+/HER2− subtype is PTEN, which occurs with probability 0.2. We consulted the literature and found some inconsistencies between studies with regarding the relationship of PTEN to PIK3CA. For example, some studies argue that PIK3CA is mutually exclusive to PTEN (Stemke-Hale *et al.*, 2008), while others state that PIK3CA could be characterized together with PTEN deletions for HR+ subtypes (Mukohara, 2015). In the dataset we used, the mutual exclusivity of PIK3CA and PTEN mutations was observed across the cohort. Our results suggest that TP53 may have some additive effects on PTEN and its association to PIK3CA. This may be a potentially interesting topic since TP53 and PTEN are both tumor suppressors and could provide a tumorigenic advantage to these aggressive subtypes. Consistent with the existing knowledge, Hintra detected TP53 as the most important initiator preceding GATA3, CDH1 and FOXA1, which are commonly associated with invasive lobular carcinoma, a subtype within HR+/HER2−. A high proportion of HR+/HER2− cases acquire a CDH1 mutation, which is a hallmark of lobular carcinoma. Furthermore, it is known that CDH1 loss and PIK3CA gain of function are highly correlated with these outcomes; however, their order is not accounted for in the literature, and when these mutations are mentioned, they are characterized as a group (An *et al.*, 2018). Interestingly, we found that CDH1 is almost three times as likely to be the initiator of this association with PIK3CA, which may provide some insights on the development of lobular carcinoma.

The limited number of samples in the other three subtypes (HR−/HER2+, HR+/HER2+ and TN) resulted in weaker signals. Among the stronger patterns derived from the parameters learned by Hintra, we observed that TP53 is almost twice as likely to be an initiator driver mutation when associated with PIK3CA in HR−/HER2+ and TN tumors. This adds to the results of PCAWG studies such as Gerstung *et al.* (2018) demonstrating that driver mutations in PIK3CA and TP53 are more likely to be clonal.

3.3 Computational resources analysis

The size of the input for collaborative ITH detection can be defined in terms of the CCF discretization hyper-parameter Δ, number of tumors m, the number of mutations per tumor and the number of unique mutation profiles referred to as 'combinations'. Here, 'profile' stands for the set of observed mutations for the corresponding tumor. In addition to these factors, the number of utilized Central Processing Unit (CPU) cores can affect the time and memory resources consumed by the different methods.

To evaluate the effects of these factors, a set of experiments with synthetic data was conducted. The range of values tested and the default values for each of the factors are provided in Table 1. When studying the effect of each factor by changing its value, all other factors were set to their default values. The datasets were generated following the approach in Caravagna *et al.* (2018). Different combinations of mutations were generated based on the assumption that half of the mutations of a new combination should already exist in the previous combinations (each mutation can belong to a different existing combination) and half of them should be new mutations not existing in any of the previous combinations. This mimics the real data distribution in the way that it assigns higher probability of mutation to a few genes and promotes heterogeneity.

Table 1. Problem size factors considered in the running time analysis

Factor	Values
Samples (m)	20, *40*, 60
Mutations per sample	3, *4*, *5*
Combinations	1, *5*, 10
Δ	0.025, *0.050*, 0.100
CPU cores	2, *4*, 6

Note: Default values are italicized.

Both Hintra and Revolver consist of two phases: preprocessing and EM. During Phase I (preprocessing), Revolver uses ClonEvol to construct and score the trees for each tumor and selects the top trees as candidates. Hintra computes the marginal likelihood using Equation (9). The results of this phase could be stored in both algorithms to avoid re-computation costs. During Phase II (EM), both algorithms learn the parameters. The maximum number of EM iterations was set to 100 for both Hintra and Revolver in these experiments. We measured the running times for the two phases separately for better interpretation. The results are shown in Figure 8. According to these results, Revolver performed the first phase more efficiently and was less sensitive to the problem size. This is due to the efficient strategy used in ClonEvol for searching the tree topology space, whereby the search space is pruned based on the consistency of the subtrees with the CCFs, resulting in a considerably smaller search space. On the other hand, the current implementation of Hintra enumerates all possible topologies, whose number is combinatorially related to the number of mutations. Furthermore, unlike Hintra, ClonEvol requires the clonal mutation to be identified in the input and it builds the tree of the rest of the mutations under that clonal mutation. This has the significant effect of reducing the search space by fixing one node.

Another important factor affecting the running time of Hintra is Δ. The running time of Hintra contains a term proportional to $\binom{\Delta^{-1}+x}{x}$, where x is the number of mutations in a sample. Our

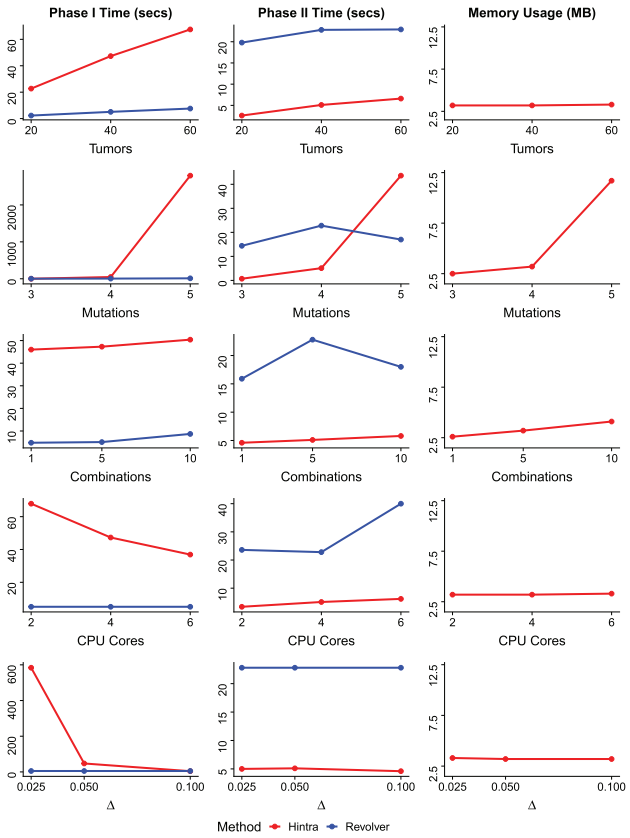


Fig. 8. Results of running time and memory analysis

experiments (results not shown) indicate that there was no noticeable difference between the accuracy of Hintra when $\Delta = 0.05$ or $\Delta = 0.1$ and the latter can be used to improve the speed without sacrificing the accuracy. Yet, using $\Delta = 0.05$, our experiment with breast cancer HR+/HER2− subtype (see Section 3.2), consisting of 1019 samples with up to 5 mutations, took about 50 min. The running time of both methods scales linearly with the number of tumors. In contrast to Revolver, which does not allow parallel processing in Phase I, using more CPU cores improves the running time performance of Hintra (see Fig. 8).

In Phase II, Hintra was in general more efficient than Revolver. However, it was more sensitive to the number of mutations. This can be explained by the fact that Hintra integrates over all tree topologies while Revolver focuses only on a set of top trees selected by ClonEvol, the size of which is bounded independently of the number of mutations.

The running time of Hintra in both Phases I and II can be improved by using alternative approaches. For example, one can use ClonEvol and then integrate only over the selected trees in the probabilistic framework of Hintra. Alternatively, Monte Carlo Markov Chain approaches as in (Ross and Markowetz, 2016) can be used to sample high likelihood trees in constant time. These approaches are expected to result in a small loss of accuracy as the density over tree topologies would be concentrated in a small area of the search space. Another way of improving the running time is limiting the summation in Equation (9) to θ values close to the observed CCFs. Because these values are associated with larger likelihoods, we expect this approximation to be close to the true value.

The memory consumption of Hintra is also shown in Figure 8. Based on these results, the number of mutations per sample is the only important factor for the amount of memory used. This affects the number of possible ancestry sets as well as the total number of mutations n. These two values determine the size of the β parameter. Moreover, the number of mutations per sample determines the number of possible topologies, which indicates the number of marginal likelihoods that need to be computed and stored. While Hintra consumes up to about 12 MB, Revolver uses about 4 GB of memory during its execution (not shown in the figure due to the large magnitude). This may be due to the implementation of Revolver in R, which is very inefficient compared to C++, which we used to implement Hintra.

4 Conclusions

We presented Hintra as a new method for collaborative ITH detection. Hintra is a PGM with a novel tree prior probability that considers all the mutations preceding a particular mutation in the phylogenetic tree, instead of only the most recent one. It uses a Bayesian approach to learn its parameters, which mitigates the bias toward branching topologies found in other tools. We compared Hintra's performance using synthetic and real data against both a stand-alone and a collaborative method. In our experiments on synthetic datasets, we demonstrated the effectiveness of both proposed tree prior probability and Bayesian learning method using different scenarios. Similarly, for synthetic data from the literature, Hintra inferred the true phylogenetic trees with more accuracy compared to the state-of-the-art.

In our experiments on breast cancer data, Hintra's findings were consistent with the existing domain knowledge. Moreover, based on the prognostic parameters learned, Hintra provided new insights of potential interest. We note that the experiments on breast cancer data used only SNVs. CNVs were excluded due to the difficulties that they cause in inferring correct CCFs, which would be later transformed into read counts for phylogeny detection by Hintra. Although there are tools for inferring the CCF values for CNVs [e.g. PyClone (Roth *et al.*, 2014)], their accuracy is limited for low-coverage cross-sectional data. Including CNV data will be considered in future work as both the sequencing technologies and CCF inference tools improve.

In the current implementation of Hintra a limited number of mutations can be considered for each patient. This number depends on the available computational resources. Although in some datasets (e.g. the breast cancer dataset used in Section 3.2) this limitation does not result in a considerable information loss (3.5% of the samples with more than 5 mutations), in general it can limit the findings to only well-known driver genes and a subset of the patients. Part of this problem is resolved by enabling parallel computing. Further improvements in running time can be achieved by using the ideas discussed in Section 3.3 for future implementation.

Although the presented probabilistic framework of Hintra considers a model for read count data, generalization to other increasingly available data types (e.g. binary data from single-cell sequencing) using appropriate distributions (e.g. Bernoulli) is also possible. Investigating the possibility of an intrinsically unbiased prior probability for phylogenetic trees that is suitable for a collaborative framework is another direction for future work.

Acknowledgements

We thank Dr. Giulio Caravagna for answering our questions and providing the source code for Revolver simulations.

Conflict of Interest: none declared.

References

An,Y. *et al.* (2018) Cdh1 and pik3ca mutations cooperate to induce immune-related invasive lobular carcinoma of the breast. *Cell Rep.*, **25**, 702–714.e6.

Attolini,C.S.-O. *et al*. (2010) A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc. Natl. Acad. Sci. USA*, **107**, 17604–17609.

Beerenwinkel,N. *et al*. (2004). Learning multiple evolutionary pathways from cross-sectional data. In: *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology, RECOMB '04*, San Diego, California, USA, pp. 36–44. ACM, New York, NY.

Beerenwinkel,N. *et al*. (2007) Conjunctive Bayesian networks. *Bernoulli*, **13**, 893–909.

Caravagna,G. *et al*. (2018) Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nat. Methods*, **15**, 707–714.

Dang,H.X. *et al*. (2017) ClonEvol: clonal ordering and visualization in cancer sequencing. *Ann. Oncol.*, **28**, 3076–3082.

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

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

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

Gerstung,M. *et al*. (2009) Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, **25**, 2809–2815.

Gerstung,M. *et al*. (2018). The evolutionary history of 2, 658 cancers. *bioRxiv*. Doi: 10.1101/161562.

Hajirasouliha,I. *et al*. (2014) A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, **30**, i78–i86.

Harsha,B. *et al*. (2016) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.

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 Bioinformatics*, **15**, 35.

Loohuis,L.O. *et al*. (2014) Inferring tree causal models of cancer progression with probability raising. *PLoS One*, **9**, e108358.

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*. Doi: 10.1101/234914.

Malikic,S. *et al*. (2018). PhISCS—a combinatorial approach for sub-perfect tumor phylogeny reconstruction via integrative use of single cell and bulk sequencing data. *bioRxiv*. Doi: 10.1101/376996.

Misra,N. *et al*. (2014) Inferring the paths of somatic evolution in cancer. *Bioinformatics*, **30**, 2456–2463.

Mukohara,T. (2015) Pi3k mutations in breast cancer: prognostic and therapeutic implications. *Breast Cancer (Dove Med. Press)*, **7**, 111–123.

Pathare,S. *et al*. (2009) Construction of oncogenetic tree models reveals multiple pathways of oral cancer progression. *Int. J. Cancer*, **124**, 2864–2871.

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

Ramazzotti,D. *et al*. (2015) CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, **31**, 3016–3026.

Razavi,P. *et al*. (2018) The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell*, **34**, 427–438.e6.

Ross,E.M. and Markowetz,F. (2016) OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.*, **17**, 69.

Roth,A. *et al*. (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396–398.

Stemke-Hale,K. *et al*. (2008) An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer. *Cancer Res.*, **68**, 6084–6091.

Suppes,P. (1970). *A Probabilistic Theory of Causality*. North-Holland Publishing Co, Amsterdam, Netherlands.

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