PMCE: efficient inference of expressive models of cancer evolution with high prognostic power

Fabrizio Angaroni[1], Kevin Chen [2,3], Chiara Damiani[4,5], Giulio Caravagna[6], Alex Graudenzi [7,8,*], Daniele Ramazotti[2,9,10,*]

[1] Dept. of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Milan, Italy, [2] Dept. of Computer Science, Stanford University, USA, [3] Google LLC, Menlo Park, USA, [4] Dept. of Biotechnology and Biosciences, Univ. of Milan-Bicocca, Milan, Italy, [5] Sysbio Centre for Systems Biology, Milan, Italy, [6] Dept. of Mathematics and Geosciences, Univ. of Trieste, Trieste, Italy, [7] Inst. of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy, [8] Bicocca Bioinformatics, Biostatistics and Bioimaging Centre (B4), Milan, Italy, [9] Dept. of Pathology, Stanford University, USA, [10] Dept. of Medicine and Surgery, Univ. of Milan-Bicocca, Monza, Italy.

1 Abstract

Motivation: Driver (epi)genomic alterations underlie the positive selection of cancer subpopulations, which promotes drug resistance and relapse. Even though substantial heterogeneity is witnessed in most cancer types, mutation accumulation patterns can be regularly found and can be exploited to reconstruct predictive models of cancer evolution. Yet, available methods cannot infer logical formulas connecting events to represent alternative evolutionary routes or convergent evolution.
Results: We introduce PMCE, an expressive framework that leverages mutational profiles from cross-sectional sequencing data to infer probabilistic graphical models of cancer evolution including arbitrary logical formulas, and which outperforms the state-of-the-art in terms of accuracy and robustness to noise, on simulations.
The application of PMCE to 7866 samples from the TCGA database allows us to identify a highly significant correlation between the predicted evolutionary paths and the overall survival in 7 tumor types, proving that our approach can effectively stratify cancer patients in reliable risk groups.
Availability: PMCE is freely available at https://github.com/danro9685/HESBCN. The code to replicate all the analyses is available at: https://github.com/BIMIB-DISCo/PMCE.
Contacts: daniele.ramazzotti@unimib.it, alex.graudenzi@ibfm.cnr.it

2 Introduction

Many natural phenomena are characterized by the presence of ordered sequences of discrete states or events, such as the accumulation of somatic mutations during cancer progression [1]. A particular class of mathematical models used to represent such phenomena is provided by Bayesian Networks (BNs) and related extensions, such as Conjunctive Bayesian Networks (CBNs) [2, 3, 4] and Suppes-Bayes Causal Networks (SBCNs) [5, 6, 7]. Such models capture the temporal ordering and the conditional dependencies among the events, and can be inferred by pooling data of multiple patients.

For instance, cross-sectional mutational profiles of cancer patients generated, e.g., via variant calling from bulk sequencing data, can be used to infer the most likely trends of accumulation of somatic mutations during the development of a certain cancer type. These trends enable *prediction* of the the next step(s) of the disease evolution of any given patient, with evident implications on prognostic strategies and therapeutic

interventions [8]. Yet, the inference problem is complicated by the high levels of heterogeneity typically observed in most tumor types, which are due to the existence of multiple independent evolutionary trajectories, often involving shared subsets of events with complex dependencies [9].

In this regard, most existing computational approaches allow one to model conjunctive processes (i.e., `AND` logical operator), according to which a certain event can occur only if all its direct ancestors (i.e., parent events) have occurred. This strategy allows one to cover a wide range of evolutionary processes observed in real-world scenarios, yet it struggles when more complex dependencies are present, such as *disjuctions* (i.e., `OR`), according to which a certain event can occur if at least one of its parents has occurred, or *mutual exclusivity* relations (i.e., `XOR`), in which a given event can occur if only one of its parents has occurred, e.g., due to synthetic lethality [10].

Therefore, some methods have been recently devised to overcome the limitations of conjunctive models, by allowing one to assess the existence of arbitrary logical formulas connecting events [5, 7]. However, such approaches require that such formulas are identified prior to the inference and provided as input, thus requiring either to possess a deep biological knowledge on the underlying evolutionary process, or to employ ad hoc computational strategies to identify specific patterns among events [11].

In this work we propose PMCE (Predictive Models of Cancer Evolution), a computational framework for the inference of expressive probabilistic graphical models of cancer evolution from cross-sectional mutational profiles of cancer samples, which are named Hidden Extended Suppes-Bayes Causal Networks (ESBCNs). The main novelty of PMCE is the automated identification of logical formulas connecting the events, which is achieved via an efficient Markov Chain Monte Carlo (MCMC) search scheme. In addition, PMCE employs an Expectation-Maximization (EM) strategy on a continuous-time Hidden Markov Model (HMM), in order to assign an error probability to the dataset and an evolutionary time (i.e., the expected waiting time required to observe a given sample) to any node of the output model. On the one hand, this allows one to evaluate the *predictability* of any given tumor, as originally proposed in [4], and which roughly measures the repeatability of the evolutionary patterns observed in a given tumor. On the other hand, by attaching samples to the evolutionary paths of the output model, it is possible to estimate the evolutionary time of any given sample with respect to the expected progression of the corresponding tumor type (notice that time is measured in arbitrary units and the temporal position of the samples is relative to that of the other samples of the dataset).

We assessed the performance of our approach via extensive simulations, in which we tested the capability of PMCE, HCBNs [4] and standard Bayesian Networks [12] in recovering the ground-truth topology, with respect to generative models with distinct sample size, topological complexity and noise. PMCE significantly outperforms competing methods in most settings, proving its superior expressivity and robustness to noise.

Finally, we applied PMCE to 7866 samples from 16 cancer types from The Cancer Genome Atlas (TCGA) database [13]. In addition to the evolutionary models, we computed the predictability of each cancer type, also assessing the possible correlation with the overall mutational burden. Importantly, we executed a combined regularized Cox regression [14, 15] and Kaplan-Meier survival analyses by employing the patient-specific evolutionary paths and evolutionary times inferred from the PMCE models. This allowed us to identify 7 cancer types in which a risk-based stratification of patients defines statistically significant differences in the overall survival. This important result proves that PMCE can be employed to generate experimental hypotheses with translational relevance and high prognostic power and which might, in turn, drive the design of cancer- and patient-specific therapeutic strategies.

3 Methods

Bayesian Networks (BNs) [12] have been often employed to model the likely temporal trends of accumulation of somatic variants in cancer, in many cases by fitting binarized mutational profiles generated from cross-sectional bulk sequencing experiments [16, 17]. The different existing approaches are characterized by distinct levels of expressivity, ranging from tree models [18, 19], to conjunctive Bayesian networks (CBNs) [20, 2] and to more complex representations involving logical formulas among genomic events, including Suppes-Bayes Causal Networks (SBCNs) [5, 1, 21, 22, 23] and Extended Suppes-Bayes Causal Networks (ESBCNs) [7, 24]. In particular, since the search of logical formulas implies an exponential growth in computational complexity, all available algorithms require a set of logical formulas as input, which are then tested in the search step

[7, 24].

In this work, we first introduce the Hidden Extended Suppes-Bayes Causal Network (HESBCN) model, in which a HMM is added to ESBCNs to simulate the stochastic processes related to the accumulation of genetic alterations during cancer development, as well as the measurement error. In particular, we assume that the time between two (independent) events is exponentially distributed with rate λ_i, whereas measurement errors are modeled via a Bernoulli process with error probability ϵ.

We then present a new algorithmic framework named PMCE for the inference of HESBCNs from cross-sectional mutational profiles of cancer samples, which includes a two-step procedure: (*i*) a MCMC search to infer the Maximum a Posteriori (MAP) structure of HESBCNs and the concomitant automated inference of logical formulas, (*ii*) an EM procedure to infer the parameters of the HMM (see Figure 1).

More in detail, a HESBCN is a *probabilistic graphical model* [25] defined by

- the set $\{N\} = \{\psi_1, \ldots, \psi_n\}$ of vertices representing logical formulas involving one or more atomic event(s) of a model. In our case, the atomic events are binary variables modeling the presence/absence of genomic variants (e.g., single-nucleotide or structural variants), whereas logical relations are limited to `AND`, `OR` (soft exclusivity) and `XOR` (hard exclusivity).

- The set $\{\triangleright\} = \{\triangleright_1, \ldots, \triangleright_w\}$ of w edges representing the conditional dependencies among the vertices of the model (e.g., $\psi_i \triangleright \psi_j$) and which are also referred to as *evolutionary steps* in the case studies.

  We note that, by construction, HESBCNs model Direct Acyclic Graphs (DAGs), in which a vertex can have more than one parent and models with multiple roots and/or disconnected components are allowed.

- The set of conditional probabilities associated to the vertices, $\{\theta\} = \{\theta_1, \ldots, \theta_n\}$, where θ_i is the conditional probability that a logical formula $\psi_i \in N$ is true, given that its predecessor formulas are true.

- The error probability ϵ of the dataset, by assuming that the false positives/negatives rates in the input binary data are modeled with a Bernoulli process, as in [3]. In our case, we assume that the rate of false positives and false negatives is identical and $= \epsilon/2$.

- The set $\{\lambda\} = \{\lambda_1, \ldots, \lambda_n\}$ of rates of the Poisson processes of the continuous-time HMM, associated to the vertices of the model, which allow one to estimate the expected waiting time of a node, given that its predecessor has occurred.

3.1 Structure learning of HESBCNs via MCMC

The PMCE framework first aims at inferring the Maximum a Posteriori (MAP) HESBCN structure. Given a cross-sectional mutational profiles dataset D, generated, e.g., from bulk sequencing data, let r be the total number of atomic events, i.e., the genomic variants observed in at least one sample. Let be $d^i \in \{0,1\}^r$ the binarized mutational profile of length r of the ith sample belonging to the input dataset D, such that $d^i_j = 1$ if the jth mutation is detected in sample i^{th}, 0 otherwise (a discussion on the binarization of variant allele frequency profiles from sequencing data can be found in [5, 26]).

Assuming that all the observations are independent, the posterior probability of a HESCBN model ($\{\triangleright\}$, $\{\theta\}$, ϵ), given the data D, is proportional to:

$$
\begin{aligned}
&P(\{\triangleright\}, \{\theta\}, \epsilon | D) \propto \\
&\propto \left[\prod_{d^i \in D} \sum_{\theta_k \in \{\theta\}} P(d^i | \{\triangleright\}, \theta_k, \epsilon) \right] \cdot P(\{\triangleright\}) P(\{\theta\}) P(\epsilon),
\end{aligned}
\tag{1}
$$

where: $P(\{\theta\}) = \prod_{i=1}^{n} P(\theta_i)$ and $P(\theta_i)$ are Beta-priors with both shape parameters equal to 10^{-5}, as in [3]; $P(\epsilon)$ is the prior for the error probability, chosen as uniform in the structure learning step; $P(\{\triangleright\})$ is the prior for the network structure, which here encodes the conditions of Suppes' probabilistic causation [27], as follows:

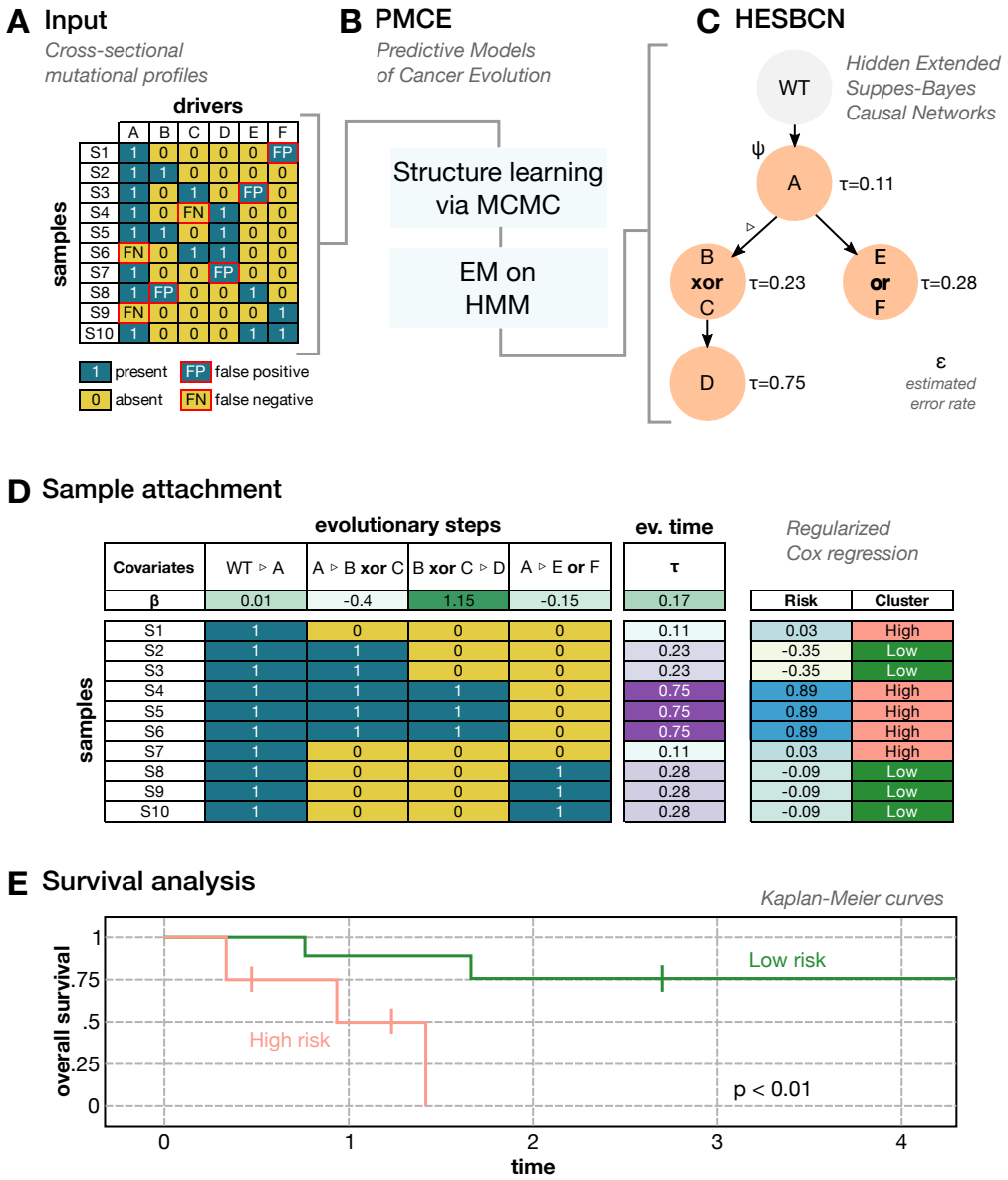

Figure 1: **PMCE framework**. **(A)** PMCE takes as as input binarized cross-sectional mutational profiles of selected genes (e.g., drivers), generated via bulk sequencing experiments. Data can include false positives and false negatives. **(B)** PMCE employs a Monte-Carlo Markov Chain (MCMC) search and an Expectation-Maximization step to infer the Maximum a Posteriori (MAP) structure and the parameters of **(C)** a Hidden Extended Suppes-Bayes Causal Network (HESBCN), which describes the evolutionary steps ▷ from the wild-type (WT). PMCE is capable of inferring logical formuals ψ connecting the events, the evolutionary time τ of each event, and the error rate ϵ of the dataset. **(D)** PMCE attaches the samples to the HESBCN model via maximum likelihood estimation, and employs the evolutionary steps and time in order to perform a Regularized Cox Regression [14, 15] on survival data. This allows one to identify the relevant covariates and the corresponding risk coefficients β. Samples are then stratified in clusters according to the overall risk. **(E)** Risk clusters are compared via Kaplan-Meier survival analysis.

$$P(\{\triangleright\}) = \prod_{i \in \{N\}} \begin{cases} 1 & \text{if} \quad \forall \psi_j \in \{Pa(\psi_i)\} \\ & P(\psi_j) > P(\psi_i) \quad \text{and} \\ & P(\psi_i|\psi_j) > P(\psi_i|\neg\psi_j) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where $\{Pa(\psi_i)\}$ indicates the parent set of the formula ψ_i. Notice that the convergence to the MAP structure in this scenario was demonstrated in [5].

The likelihood term $P(d^i|\{\triangleright\}, \theta_k, \epsilon)$ represents the probability to observe the d^i sample given a HESBCN model. This term is evaluated for each node $k = 1, \ldots, n$ by first solving each formula from the data ($= 1$ if the formula is true, $= 0$ otherwise) and then computing the conditional probability θ_k.

In order to infer the structure of a HESBCN given the data, PMCE employs a Monte Carlo Markov Chain (MCMC) approach with a Gibbs' sampling scheme from the posterior distribution, which is a modification of that proposed in [3]. The MCMC starts from a randomly initialized HESBCN structure and includes 8 different moves:

1. *Modify Theta*: given a randomly chosen node i, θ_i is set to a new value, which is sampled from a uniform distribution: $[0, 1]$.
2. *Create a new parental relation*: a new parental relation is randomly added to the model, by avoiding the creation of cycles.
3. *Delete a parental relation*: a randomly chosen parental relation is deleted from the model.
4. *Add an edge into a transitive closure relation*: we select a node, and considering its transitive closure a new edge is inserted into the model.
5. *Delete an edge from a transitive closure relation*: an edge is deleted from the model, by considering only the edges belonging to the transitive closure of a given node.
6. *Node swap*: two randomly chosen nodes are swapped.
7. *Reincarnation*: a *delete a parental relation* move is followed by a *create a new parental relation* move.
8. *Local restart*: the inference is restarted from a new random configuration, if the previous MCMC has converged to a deadlock.

* During the MCMC, every time that a node (e.g., `A`) happens to have more than one parent (e.g., `B` and `C`) a logical operator is chosen with uniform probability among `AND`, `OR` and `XOR` and associated to the elements of its parent set (e.g., (`B XOR C`) $\triangleright$ `A`).

The acceptance probability ρ of a proposed MCMC sample, is given by:

$$\rho = \min \left\{ 1, \frac{P(\{\triangleright\},\{\theta\},\epsilon|D)' \times \mathrm{MSP}(\{\triangleright\},\{\theta\},\epsilon)' \times \mathrm{TP}(\{\triangleright\},\{\theta\},\epsilon|\{\triangleright\}',\{\theta\}',\epsilon)}{P(\{\triangleright\},\{\theta\},\epsilon|D) \times \mathrm{MSP}(\{\triangleright\},\{\theta\},\epsilon) \times \mathrm{TP}(\{\triangleright\}',\{\theta\}',\epsilon|\{\triangleright\},\{\theta\},\epsilon)} \right\} \tag{3}$$

where $'$ indicates the quantities associated with the structure after the proposed move. MSP stands for move selection probability (see Table 1 for default probabilities). TP is the transition probability from the current structure to the new one, given the selected move. It is important to notice that despite being in the equation, the error rate ϵ is set to 0 during the structure inference.

In order to limit the size of the logical formulas included in the output model, PMCE maximizes a score composed by Equation (1) plus a regularization term, e.g., BIC or AIC. As output, PMCE returns the MAP HESBCN structure and parameters retrieved after a user-selected number of MCMC iterations and restarts.

Table 1: Values of move selection probability (MSP) for the moves described in the Methods section.

Move	MSP default value
Modify theta	0.15
Create a new parental relation	0.4
Delete a parental relation	0.2
Add an edge into a transitive closure relation	0.04
Delete an edge from a transitive closure relation	0.04
Node swap	0.07
Reincarnation	0.1

3.2 Hidden Markov Model

The accumulation of genomic variants during cancer evolution can be modeled as a stochastic time-dependent process, e.g., via a continuous-time Hidden Markov model, as originally proposed in [2]. More in detail, let us suppose that the formula ψ_i of a HESBCN model is satisfied after time t_i. A model composed by n nodes will include n waiting times $t_1, \ldots, t_n$. As commonly done with stochastic branching processes, we assume an exponentially distributed time and, accordingly, we define the waiting time associated to any logical formula ψ_i as:

$$t_i = \max_{\psi_j \in \{Pa(\psi_i)\}} t_j + Z_i, \qquad Z_i \sim exp[\lambda_i], \tag{4}$$

where λ_i is the rate of the exponential distribution of the i-th formula.

Let us then define a *logical formula path* as an ordered sequence of logical formulas:

$$\mathbf{\Psi}_l = (\ldots(\psi_1 \triangleright \psi_2) \triangleright \ldots) \triangleright \psi_l, \tag{5}$$

such that $\psi_{i-1} \in \{Pa(\psi_i)\}, \quad \forall i = 2, \ldots, l$. If we suppose that the probability density of a logical formula path $\mathbf{\Psi}_l$ factorizes according to $P(\mathbf{\Psi}_l) = \prod_{i=1}^{l} P(\psi_i)$, then the following equation holds:

$$P(\mathbf{\Psi}_l) = \prod_{i=1}^{l} \int_0^\infty dt_i \int_0^\infty dt_s \chi(t_i, t_s) f(t_i | \{t_j\}_{\{Pa(\psi_i)\}}) f(t_s), \tag{6}$$

where t_s is time of the diagnosis, $\chi(t_i, t_s) = 1$ if $t_s > t_i$, 0 otherwise, $f(t_s)$ is the probability density of the diagnosis time (which is assumed to be uniform) and $f(t_i | \{t_j\}_{\{Pa(\psi_i)\}})$ is the probability density of t_i, conditioned on its predecessors.

In principle, one could directly maximize Eq. (6) to estimate the set of optimal $\{\lambda_i\}$. However, dealing with real data, it may be sound to account for experimental noise, i.e., the false positives and false negatives possibly included in the input data D.

Thus, let us introduce the set of theoretical genotypes $\{\mathcal{G}_{\mathbf{\Psi}_l}\} = \{g^1, \ldots, g^q\}$ as the set of q possible genotypes subsumed by the HESBCN path Ψ_l and that are represented as ternary vectors: $g^k = \{0, 1, *\}^r$, such that $g^k_j = 1$ if the variant j is present in the theoretical genotype k, $g^k_j = 0$ if it is absent, $g^k_j = *$ if the variant is not included in path Ψ_l and, therefore, it can either be present or absent.

Then, we model the measurement errors via a Bernoulli process with an error probability ϵ. In detail, the probability of observing the genotype $d^i(r)$, given that the theoretical genotype is $g^k(r)$ and the error probability is ϵ, is given by:

$$P(d^i(r) | g^k(r), \epsilon) = \epsilon^{\mathrm{HD}(g^k(r), d^i(r))} (1 - \epsilon)^{r - \mathrm{HD}(g^k(r), d^i(r))}. \tag{7}$$

Here, $\mathrm{HD}(g^k(r), d^i(r))$ denotes the Hamming distance between $d^i(r)$ and $g^k(r)$.

Accordingly, the posterior probability of a HESBCN model becomes:

$$\begin{aligned} &P(\{\triangleright\}, \{\theta\}, \{\lambda\}, \epsilon | D) \equiv P(\{\mathcal{P}\}, \{\lambda\}, \epsilon | D) = \\ &\sum_{\mathbf{\Psi}_l \in \{\mathcal{P}\}} \prod_{d^i(r) \in D} \frac{P(\mathbf{\Psi}_l) P(d^i(r) | \{\mathcal{G}_{\mathbf{\Psi}_l}\}, \epsilon)}{\sum_{\mathbf{\Psi}_l \in \{\mathcal{P}\}} P(\mathbf{\Psi}_l) P(d^i(r) | \{\mathcal{G}_{\mathbf{\Psi}_l}\}, \epsilon)}, \end{aligned} \tag{8}$$

where $\{\mathcal{P}\}$ is set of the logical formula paths included in the HESBCN model, $P(\mathbf{\Psi}_l)$ is the probability density from Eq. (6) and the likelihood is defined by assuming that the probability of a logical formula path is equal to the product of the probability of every related theoretical genotypes, as follows:

$$P(d^i(r)|\{\mathcal{G}_{\mathbf{\Psi}_l}\},\epsilon) = \prod_{g^k(r)\in\{\mathcal{G}_{\mathbf{\Psi}_l}\}} P(d^i(r)|g^k(r),\epsilon), \tag{9}$$

To estimate the sets of $\{\lambda_i\}$ and ϵ while the structure is kept fixed, we employ a EM algorithm to maximize equation (8). Thanks to this procedure, PMCE returns the expectation time λ_i for each node of the model (notice that the choice of not including the search of λ_i and ϵ directly in the MCMC was made to reduce the overall computational complexity and speed the computation up).

PMCE then computes the maximum likelihood (ML) attachment of the samples to the HESBCN model using Eq. (7) with the ϵ estimated in the EM step. This allows one to return: (i) the ML theoretical genotype of each sample and (ii) the ML evolutionary steps, i.e., the set of parental relations among true logical formulas of that sample (see Figure 1).

Finally, since stochastic branching processes are additive, it is possible to compute the evolutionary time of each sample as $\tau = \sum_s \frac{1}{\lambda_s}$, i.e., the expected waiting time to cover all the related evolutionary steps (measured in arbitrary time units, which can be rescaled with respect to the diagnosis time t_s, as in [3]).

3.3 Predictability

After inferring the HESBCN model from the mutational profiles of the samples of a given tumor, it is possible to quantify its *predictability*, i.e., the repeatability of the evolutionary trajectories across all samples [28, 4].

More in detail, if each node of a HESBCN is a state of the evolutionary history of a given tumor type, then it is natural to interpret the set of the logical formula paths $\{\mathcal{P}\}$ as the set of all the possible evolutionary trajectories. Since $1/\lambda_i$ is the expected waiting time of the i-th node, it is intuitive to relate the probability of a single logical formula path of length l to the set of λ associated to the nodes belonging to the path, i.e., [28, 4]:

$$\Pi(\mathbf{\Psi}_l) = \prod_{i=1}^{l} \frac{\lambda_{\psi_i}}{\sum_{\psi_h|\psi_i\in\{Pa(\psi_h)\}} \lambda_{\psi_h}}. \tag{10}$$

It is then possible to define the entropy of the set of logical formula paths of a HESBCN model $\{\mathcal{P}\}$ as [28, 29]:

$$H_{\{\mathcal{P}\}} = -\sum_{\mathbf{\Psi}_l\in\{\mathcal{P}\}} \Pi(\mathbf{\Psi}_l)\log(\Pi(\mathbf{\Psi}_l)). \tag{11}$$

This quantity estimates the amount of uncertainty in selecting a certain path among all possible paths of a given model. As a consequence, a HESBCN model has a maximum entropy if the probability associated to any possible path of the process is the same: $H_{max} = \log(l!)$ [4].
Intuitively, entropy in Eq.(11) could be used to quantify how much of the evolutionary process is "localized" in a few paths. The predictability of a HESBCN model characterized by a set of logical formula paths $\{\mathcal{P}\}$, is then defined as in [4]:

$$\Phi_{\{\mathcal{P}\}} = 1 - \frac{H_{\{\mathcal{P}\}}}{H_{max}}. \tag{12}$$

As a matter of fact, $0 \leq \Phi_{\{\mathcal{P}\}} \leq 1$. If $\Phi_{\{\mathcal{P}\}} = 0$, all possible logical formula paths are equally probable, whereas, if $\Phi_{\{\mathcal{P}\}} = 1$, only one path is possible, which indicates perfect predictability.

3.4 Survival Analysis

The evolutionary path and time of a given sample can be employed as risk estimators, in order to stratify patients into risk groups. To assess the prognostic power of the cancer evolution models returned by PMCE, we implemented a combined regularized Cox regression [14, 15] and Kaplan-Meier survival analysis.

In brief, according to the standard Cox proportional hazards model [30], given a baseline of risk r_0 of a given disease, and a vector of covariates X_i, the risk associated to X_i is given by the hazard function:

$$r(t) = r_0(t)e^{\sum_i \beta_i X_i}, \tag{13}$$

where β_i are coefficients that measure the impact of every covariate (if $\beta_i > 0$ the covariate is associated to a risk increment). In our case, we applied the regularized Cox regression via Coxnet [14, 15], which allows one to select the subset of covariates that minimize the cross validation error, by employing the elastic net regularizer (see Supplementary Figure S1).

In particular, we employed as input covariates for the analysis: (i) all the parental relations between the nodes (i.e., the formulas ψ_i) of the MAP HESBCN model, assessed with respect of the maximum likelihood evolutionary steps associated to each sample (= 1 if both the formulas in that relation are satisfied, = 0 otherwise), and (ii) the evolutionary time τ of each sample (in arbitrary units, see Figure 1).

For each tumor, we selected the set of β_i associated to the minimum cross-validation error and kept as significant only the cancer types displaying at least one covariate with non-zero β_i. We then computed for each sample a risk score as follows:

$$\xi = \sum_i \beta_i X_i. \tag{14}$$

This allowed us to stratify the patients into three different risk clusters: (i) high risk with $\xi > 0$, (ii) medium risk with $\xi = 0$ (baseline, typically including all the samples with impactless covariates), (iii) low risk with $\xi < 0$. The survival curves of the outcoming risk clusters were finally assessed via standard Kaplan-Meier analysis.

4 Results

4.1 Results on simulations

To assess the performance of PMCE with respect to competing approaches, we performed extensive tests on simulations. Synthetic datasets were sampled from randomly generated weekly connected DAGs with $r = 10$ atomic events and density = 0.4. Three levels of topological complexity were implemented and, in particular, 100 DAGs were generated to include only `AND` logical relations (associated to the confluences), 100 DAGs only `OR`'s and 100 DAGs only `XOR`'s, for a total of 300 generative topologies. For each topology, 3 different values of sample size ($m = [50, 100, 200]$) and 5 levels of noise rates ($\epsilon = [0, 0.05, 0.10, 0.15, 0.20]$) were scanned to generate a synthetic mutational profile dataset, for a total of 4500 independent datasets. We compared PMCE with Hidden Conjunctive Bayesian Networks (HCBNs) [2] and with standard Bayesian Networks (BNs) [12] and compared the performance of the methods by assessing the accuracy $\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$ with respect to the ground-truth generative topology. Both PMCE and standard BNs were executed with BIC regularization.

In Figure 2, one can see that PMCE significantly outperforms the competing approaches in all settings, except for the scenarios with high samples size and `AND` generative topologies, in which, as expected, HCBNs display a superior accuracy, likely due to the algorithmic design aimed at specifically inferring conjunctive relations and the smaller search space.

The most difficult setting appear to be that involving `XOR` generative topologies (Figure 2C), in which consistently lower performances are observed for all methods, with respect to synthetic datasets generated with distinct topological complexity. This result is likely related to the presence of spurious dependencies among events due to the properties of the `XOR` logical formula.

Importantly, the overall performance of PMCE is only slightly affected by even significantly high levels of false positives/negatives in the data (in all setting and with all topologies), and this is especially true for configurations with large sample size. This result proves the robustness of our approach in delivering reliable predictions from real-world data.

4.2 Results on real data

We applied PMCE to 7866 samples from 16 cancer types from The Cancer Genome Atlas (TCGA) database [13]. The dataset includes bulk sequencing data, which is often coupled with clinical information on the

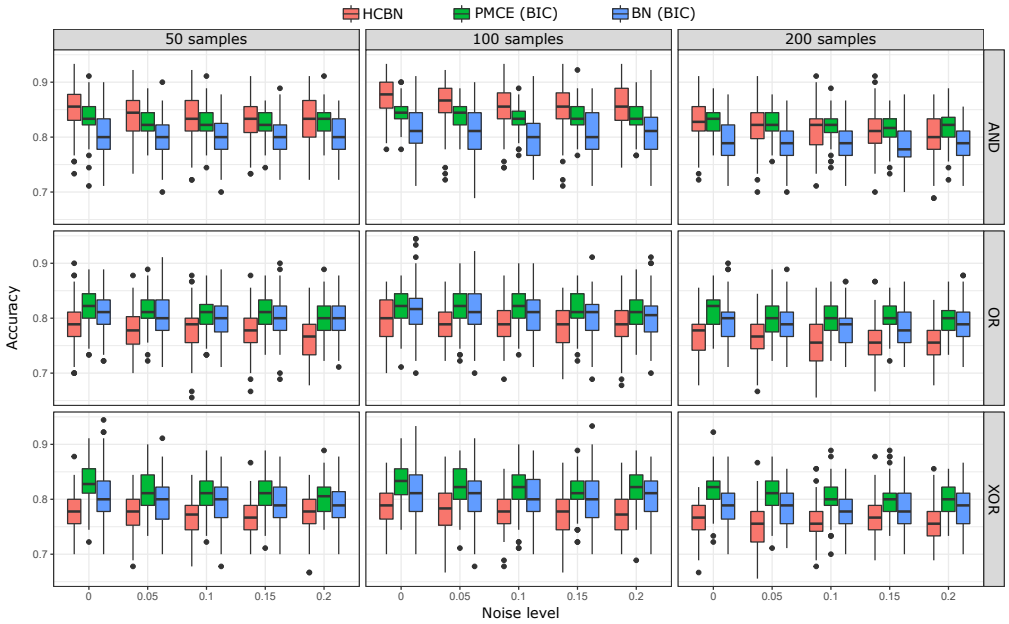

Figure 2: **Performance assessment on simulations.** 4500 independent synthetic datasets were generated from DAG generative topologies with $m = 10$ nodes (representing e.g., genomic variants) and density $= 0.4$, three distinct levels of topological complexity (only `AND` – panels **(A)**, `OR` – panels **(B)** and `XOR` – panels **(C)**, respectively), three sample sizes ($m = [50, 100, 200]$) and 5 noise levels ($\epsilon = [0, 0.05, 0.10, 0.15, 0.20]$). PMCE was compared with HCBNs [2] and standard BNs [12] (with BIC regularization) on overall accuracy with respect to the ground-truth generative topology. Box-plots return the distributions on all simulations.

related patient. In particular, the HESBCN model was inferred separately for every cancer type via PMCE, by considering driver mutations only, as identified in related works [31]. All models are displayed in Supplementary Figures S2-S17 and highlight different degrees of complexity and heterogeneity.

For each model, we evaluated the predictability as per Eq. (12) and the results are shown in Figure 3 **A**-**B**. One can notice that Φ is significantly different across cancer types. For instance, lung squamous cell carcinoma display a value of $\Phi \approx 0.97$, which is likely due to the fact that most patients follow a single path including a driver mutation on TP53. Conversely, both glioblastoma multiforme and colorectal adenocarcinoma show a very low score of $\Phi \approx 0.05$, hinting at the presence of multiple independent evolutionary trajectories in distinct patients.

In addition, we investigated the possible correlation between the predictability and the overall mutational burden, measured via both the median number of total mutations and the mean number of driver mutations (see Figure 3). Notably, the overall correlation appear to be limited in both cases: $R^2 = -0.0836$ and $R^2 = -0.0247$, respectively. Despite a higher number of driver mutations may imply, in principle, a larger number of possible evolutionary trajectories, the overall tumor burden appears not to be a clear indicator of predictability, pointing at the existence of preferred routes of cancer evolution for certain cancer types, as opposed to the limited regularities that characterize the evolution of distinct cancer types [32, 33, 34].

Survival analysis was executed by employing the Regularized Cox Regression via Coxnet [14, 15] with 10000 maximum iterations via elastic net, which allowed us to select the significant subset of covariates for each tumor type. In particular, for every sample, we employed as covariates: (i) each evolutionary step (i.e., parental relation of the HESBCN model) present in the maximum likelihood theoretical genotype and (ii) the evolutionary time τ. In detail, for 7 cancer types we found an non empty set of covariates ($\beta \neq 0$) associated to the minimum cross validation error (See Supplementary Figure S1 and Supplementary File 1). The samples of each of such cancer types, were then divided into three risk clusters, according to Eq. (14).

In Figure 3, we show the Kaplan-Meier plot for 6 cancer types, in addition to the selected covariates and the associated risk coefficient (since in the liver hepatocellular carcinoma one of the two clusters contains only 16 patient, we show the corresponding plots as Supplementary Figure S18). In all cases, the highly significant

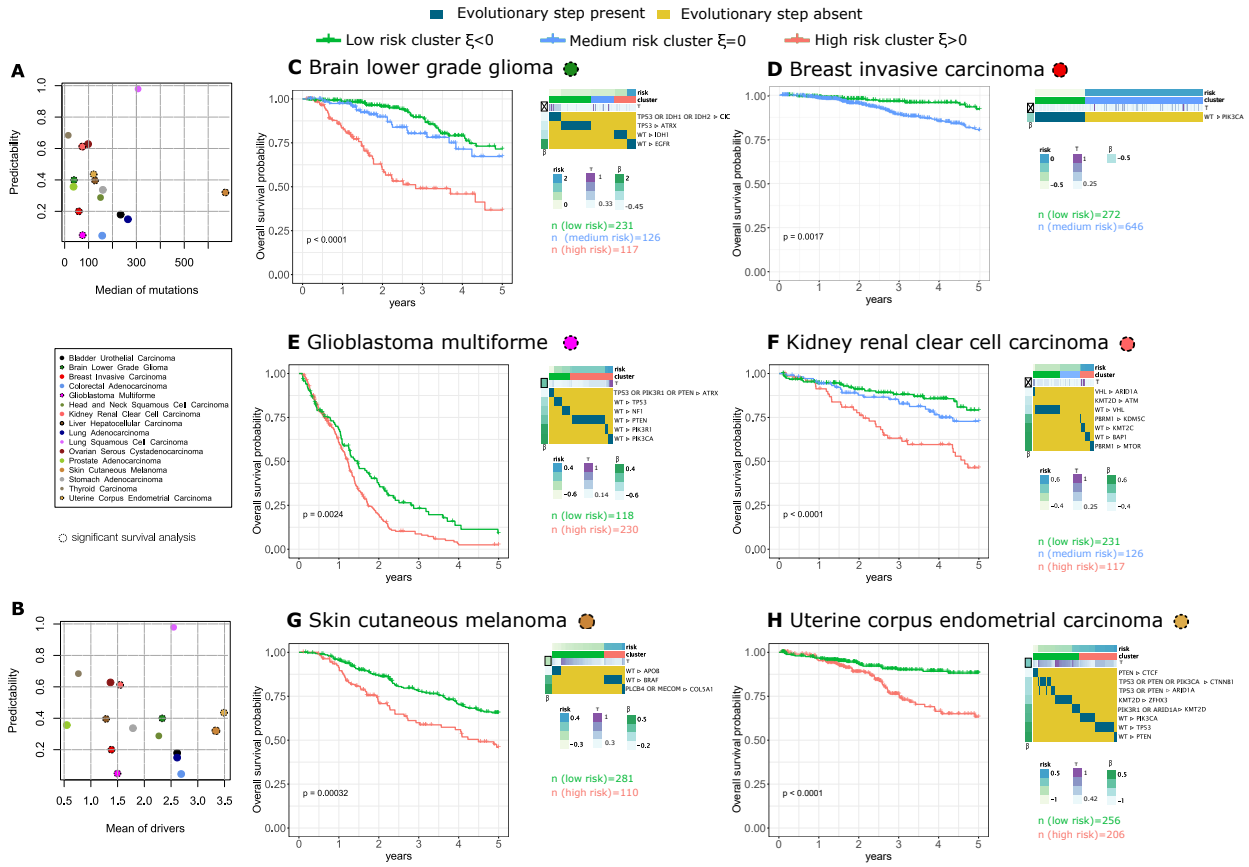

Figure 3: **Application of PMCE to 7866 TCGA samples.** PMCE was applied to 7866 samples of 16 different cancer types from the TCGA database [13], on the binarized profiles of driver mutations from [31]. **(A)** Scatter-plot returning the relation between the predictability of each tumor type and the median number of mutations, and **(B)** the mean number of drivers. The 7 cancer types showing a significant survival analysis p-value (see Methods) are circled with a dashed line. The stratification of patients in risk groups based on evolutionary steps and time obtained via PMCE and selected via Regularized Cox Regression is shown with Kaplan-Meier survival plots for **(C)** brain lower grade glioma, **(D)** breast invasive carcinoma, **(E)** glioblastoma multiforme, **(F)** kidney renal clear cell carcinoma, **(G)** skin cutaneuos melanoma **(H)** uterine corpus endometrial carcinoma. For all tumor types, the selected covariates, the sample size of the clusters, and the Kaplan-Meier p-value are shown. In addition, a heatmap returns the presence/absence of the selected covariates, the value of β_i for each covariate and the values of τ of each sample.

p-values prove that the information retrieved from the evolutionary models is effective in stratifying patients in well-separated risk groups.

More in detail, brain lower grade glioma (Fig. 3C) [35] is characterized by the presence of three risk groups. The low risk cluster is characterized by evolutionary trajectories comprising an `OR` relation between mutations `IDH1` and `IDH2`, which are typically associated with good prognosis [35]. Notably, the high risk cluster is characterized by mutations of `EGFR`, which identifies the "glioblastoma-like" subtype with a known poor prognosis. Consistently, in the glioblastoma multiforme (Fig. 3E) the cluster with better overall survival is associated to the evolutionary steps comprising mutations of `TP53`, which may confirm the role of such gene as a tumor suppressor gene for this cancer type [36]. In the breast invasive carcinoma (Fig. 3D) two clusters are found and the main covariate that seems to affect the overall survival is the presence of the mutation of `PIK3CA`, confirming the hypothesized positive prognostic significance of such mutation [37].

The analysis of the kidney renal cell carcinoma (Fig. 3F) returns a high number of significant covariates, which is consistent with the observed low predictability value for this cancer type. Strikingly, the presence of the mutation of `VHL` is associated to the low risk cluster, which may be in accordance to the fact that the inactivtion of such gene VHL is a common biomarker of bad prognosis [38]. With respect to the skin cutaneous melanoma (Fig. 3G), two different clusters are discovered, with the bad prognosis one associated to mutations on `BRAF` [39]. Finally, the uterine corpus endometrial carcinoma (Fig. 3G) shows an high number of relevant genes, including mutations of known tumors drivers such as `PTEN`, `TP53` and `PIK3CA` [35].

Interestingly, for 3 out of 7 tumors, the evolutionary time τ is selected as a relevant covariate, suggesting that the tumor progression time plays an important role in the clinical outcome.

5 Discussion

The possibility of exploiting sequencing data to reliably predict the likely clinical outcome of a given cancer patient, and possibly intervene to halt or slow down the disease progression, may have an important impact on downstream clinical practices and therapeutic strategies.

In this regard, we have shown that cancer progression models, even when inferred from low-resolution (binarized) mutational profiles of cross-sectional samples, may deliver accurate predictions on patients' survival. Accordingly, targeted sequencing of specific gene panels might be a viable and cost-effective strategy to position a given patient onto the expected cancer progression route at diagnosis time, possibly anticipating the next evolutionary steps.

Clearly, the problem is far from being solved, especially due to the impact of intra-tumor heterogeneity, which is a major cause of therapy failure and relapse [40] and which can undermine the accuracy of any population model. As a consequence, many new methods attempt to deliver cancer evolution models at the resolution of the single tumor [41, 42, 43], taking advantage of the recent advances in single-cell DNA sequencing techniques and, more recently, of the opportunity of calling variants from RNA-sequencing data [44], despite the typically high levels of technical and biological noise [45].

In this respect, it would be interest to investigate how to combine single-tumor models within comprehensive predictive population models as here proposed, for instance by employing transfer learning to find patterns of repeated evolution [26]. In addition, the rise of longitudinal sequencing experiments, e.g., from patient-derived organoids, may allow one to assess the impact of selected therapeutic strategies on the predicted evolution [46], with important translational repercussions.

Acknowledgements

This work was partially supported by a Bicocca 2020 Starting Grant to FA and DR. DR was also supported by a Premio Giovani Talenti dell'Università degli Studi di Milano-Bicocca. We thank Marco Antoniotti, Lucrezia Patruno, Francesco Craighero and Gianluca Ascolani for useful discussions.

References

[1] Caravagna, G. *et al.* Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences* **113**, E4025–E4034 (2016).

[2] Gerstung, M., Baudis, M., Moch, H. & Beerenwinkel, N. Quantifying cancer progression with conjunctive bayesian networks. *Bioinformatics* **25**, 2809–2815 (2009).

[3] Sakoparnig, T. & Beerenwinkel, N. Efficient sampling for bayesian inference of conjunctive bayesian networks. *Bioinformatics* **28**, 2318–2324 (2012).

[4] Hosseini, S.-R., Diaz-Uriarte, R., Markowetz, F. & Beerenwinkel, N. Estimating the predictability of cancer evolution. *Bioinformatics* **35**, i389–i397 (2019).

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

[6] De Sano, L. *et al.* Tronco: an r package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics* **32**, 1911–1913 (2016).

[7] Ramazzotti, D., Graudenzi, A., Caravagna, G. & Antoniotti, M. Modeling cumulative biological phenomena with suppes-bayes causal networks. *Evolutionary Bioinformatics* **14**, 1176934318785167 (2018).

[8] Wang, E. *et al.* Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. In *Seminars in cancer biology*, vol. 30, 4–12 (Elsevier, 2015).

[9] Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics* **20**, 404–416 (2019).

[10] O'Neil, N. J., Bailey, M. L. & Hieter, P. Synthetic lethality and cancer. *Nature Reviews Genetics* **18**, 613–623 (2017).

[11] Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research* **22**, 398–406 (2012).

[12] Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT press, 2009).

[13] Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113–1120 (2013).

[14] Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software* **39**, 1 (2011).

[15] Tibshirani, R. *et al.* Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 245–266 (2012).

[16] Beerenwinkel, N., Schwarz, R. F., Gerstung, M. & Markowetz, F. Cancer evolution: mathematical models and computational inference. *Systematic biology* **64**, e1–e25 (2015).

[17] Schwartz, R. & Schäffer, A. A. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics* (2017).

[18] Desper, R. *et al.* Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of computational biology* **6**, 37–51 (1999).

[19] Loohuis, L. O. *et al.* Inferring tree causal models of cancer progression with probability raising. *PloS one* **9**, e108358 (2014).

[20] Beerenwinkel, N., Eriksson, N. & Sturmfels, B. Conjunctive Bayesian Networks. *Bernoulli* 893–909 (2007).

[21] Bonchi, F., Hajian, S., Mishra, B. & Ramazzotti, D. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics* **3**, 1–21 (2017).

[22] Gao, G., Mishra, B. & Ramazzotti, D. Causal data science for financial stress testing. *Journal of computational science* **26**, 294–304 (2018).

[23] Bonchi, F., Gullo, F., Mishra, B. & Ramazzotti, D. Probabilistic causal analysis of social influence. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1003–1012 (2018).

[24] Ramazzotti, D., Nobile, M. S., Antoniotti, M. & Graudenzi, A. Efficient computational strategies to learn the structure of probabilistic graphical models of cumulative phenomena. *Journal of computational science* **30**, 1–10 (2019).

[25] Pearl, J. *Causality* (Cambridge university press, 2009).

[26] Caravagna, G. *et al.* Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature methods* **15**, 707–714 (2018).

[27] Suppes, P. A probabilistic theory of causality (1973).

[28] Estrada, E. & Hatano, N. Statistical-mechanical approach to subgraph centrality in complex networks. *Chemical Physics Letters* **439**, 247–251 (2007).

[29] Szendro, I. G., Franke, J., de Visser, J. A. G. & Krug, J. Predictability of evolution depends nonmonotonically on population size. *Proceedings of the National Academy of Sciences* **110**, 571–576 (2013).

[30] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202 (1972).

[31] Bailey, M. H. *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018).

[32] Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).

[33] McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell* **27**, 15–26 (2015).

[34] Turajlic, S. *et al.* Tracking cancer evolution reveals constrained routes to metastases: Tracerx renal. *Cell* **173**, 581–594 (2018).

[35] Network, C. G. A. R. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine* **372**, 2481–2498 (2015).

[36] Kraus, J. A. *et al.* Molecular analysis of the pten, tp53 and cdkn2a tumor suppressor genes in long-term survivors of glioblastoma multiforme. *Journal of neuro-oncology* **48**, 89–94 (2000).

[37] Kalinsky, K. *et al.* Pik3ca mutation associates with improved outcome in breast cancer. *Clinical cancer research* **15**, 5049–5059 (2009).

[38] Brauch, H. *et al.* Vhl alterations in human clear cell renal cell carcinoma: association with advanced tumor stage and a novel hot spot mutation. *Cancer Research* **60**, 1942–1948 (2000).

[39] Bhatia, P., Friedlander, P., Zakaria, E. A. & Kandil, E. Impact of braf mutation status in the prognosis of cutaneous melanoma: an area of ongoing research. *Annals of translational medicine* **3** (2015).

[40] Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences* **110**, 4009–4014 (2013).

[41] Jahn, K., Kuipers, J. & Beerenwinkel, N. Tree inference for single-cell data. *Genome biology* **17**, 1 (2016).

[42] Ramazzotti, D., Graudenzi, A., De Sano, L., Antoniotti, M. & Caravagna, G. Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data. *BMC bioinformatics* **20**, 1–13 (2019).

[43] Zafar, H., Navin, N., Chen, K. & Nakhleh, L. Siclonefit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome research* **29**, 1847–1859 (2019).

[44] Moravec, J. C., Lanfear, R., Spector, D., Diermeier, S. & Gavryushkin, A. Cancer phylogenetics using single-cell rna-seq data. *bioRxiv* (2021).

[45] Patruno, L. *et al.* A review of computational strategies for denoising and imputation of single-cell transcriptomic data. *Briefings in Bioinformatics* **bbaa222** (2020).

[46] Ramazzotti, D. *et al.* Longitudinal cancer evolution from single cells. *bioRxiv* **https://doi.org/10.1101/2020.01.14.906453** (2020).

-