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

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

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

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/BIMIB-DISCo/PMCE, in addition to the code to replicate all the analyses presented in the manuscript.
Contacts: daniele.ramazzotti@unimib.it, alex.graudenzi@ibfm.cnr.it.

1 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 [12, 19]. 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) [23, 25, 45] and Suppes-Bayes Causal Networks (SBCNs) [17, 41, 42]. 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 the *prediction* of the the next step(s) of the disease evolution of any given patient, with evident implications on prognostic strategies and therapeutic interventions [56]. 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 [53].

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 [36].

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 [41, 42]. 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 [14].

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 (HESBCNs). 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 initially proposed in [25], and which roughly measures the repeatability of the evolutionary patterns observed in a given tumor by quantifying the entropy of the different trajectories of the model. 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 [25] and standard Bayesian Networks [28] 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 [57]. 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 [47, 52] and Kaplan-Meier survival analyses by employing the patient-specific evolutionary paths and evolutionary times inferred from the HESBCN 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.

2 Methods

Bayesian Networks (BNs) [28] 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 [5, 19, 46]. The different existing approaches are characterized by distinct levels

of expressivity, ranging from tree models [18, 32], to conjunctive Bayesian networks (CBNs) [4, 23] and to more complex representations involving logical formulas among genomic events, including Suppes-Bayes Causal Networks (SBCNs) [7, 8, 12, 22, 41] and Extended Suppes-Bayes Causal Networks (ESBCNs) [42, 44]. 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 [42, 44].

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* [39] 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 [45]. 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.

2.1 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 i^{th} sample belonging to the input dataset D, such that $d^i_j = 1$ if the j^{th} 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 [11, 41]).

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

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

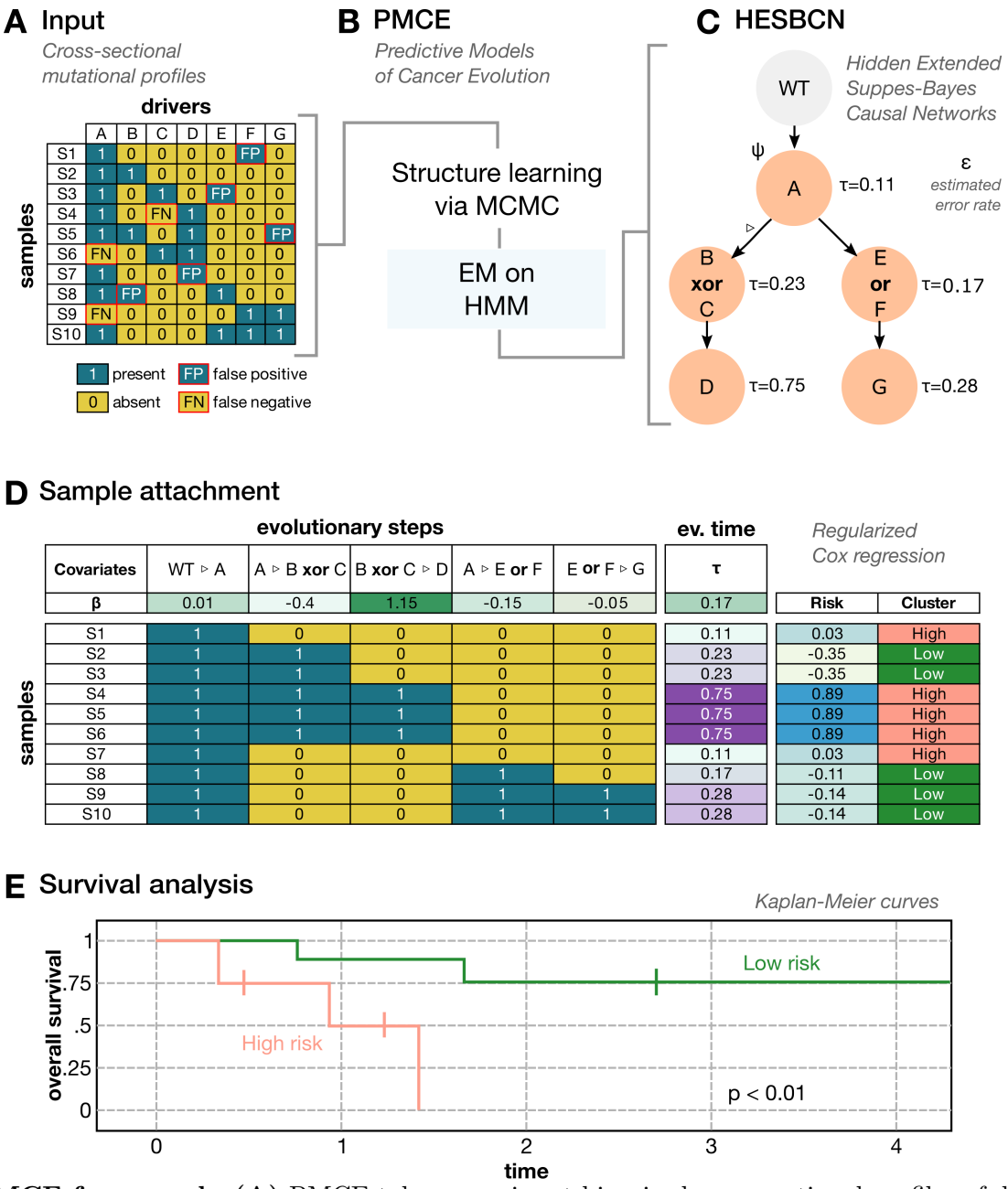

Fig 1. PMCE framework. **(A)** PMCE takes as as input binarized cross-sectional profiles of driver mutations, generated via bulk sequencing experiments. Data can include false positives and false negatives. **(B)** PMCE employs a Monte-Carlo Markov Chain (MCMC) search and an Expectation-Maximization 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 [47, 52] 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.

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 [45]; $P(\epsilon)$ 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 [49], as follows:

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

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 [41].

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 [45]. 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. $(B \quad \texttt{XOR} \quad C) \triangleright A$).

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.

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

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.

2.2 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 [23]. 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 $t_1, \dots, t_n$. 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:

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

such that $\psi_{i-1} \in \{Pa(\psi_i)\}, \quad \forall i = 2, \dots, l$. 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 $\{\mathcal{G}_{\mathbf{\Psi}_l}\} = \{g^1, \dots, g^q\}$ 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:

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

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 [45]).

2.3 Predictability

After inferring the HESBCN model from the mutational profiles of the samples of a given tumor, it is possible to quantify its *predictability* [20, 25], that is a quantity directly related to the entropy computed considering the probability of all the possible evolutionary paths of the model.

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. [20, 25]:

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

It is then possible to define the entropy of the set of logical formula paths of a HESBCN model $\{\mathcal{P}\}$ as [20, 50]:

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

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: $H_{max} = \log(l!)$ [25].
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 [25]:

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

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.

2.4 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 [47, 52] and Kaplan-Meier survival analysis.

In brief, according to the standard Cox proportional hazards model [15], 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 [47, 52], 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.

3 Results

3.1 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) [23] and with standard Bayesian Networks (BNs) [28] 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 2**C**), 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.

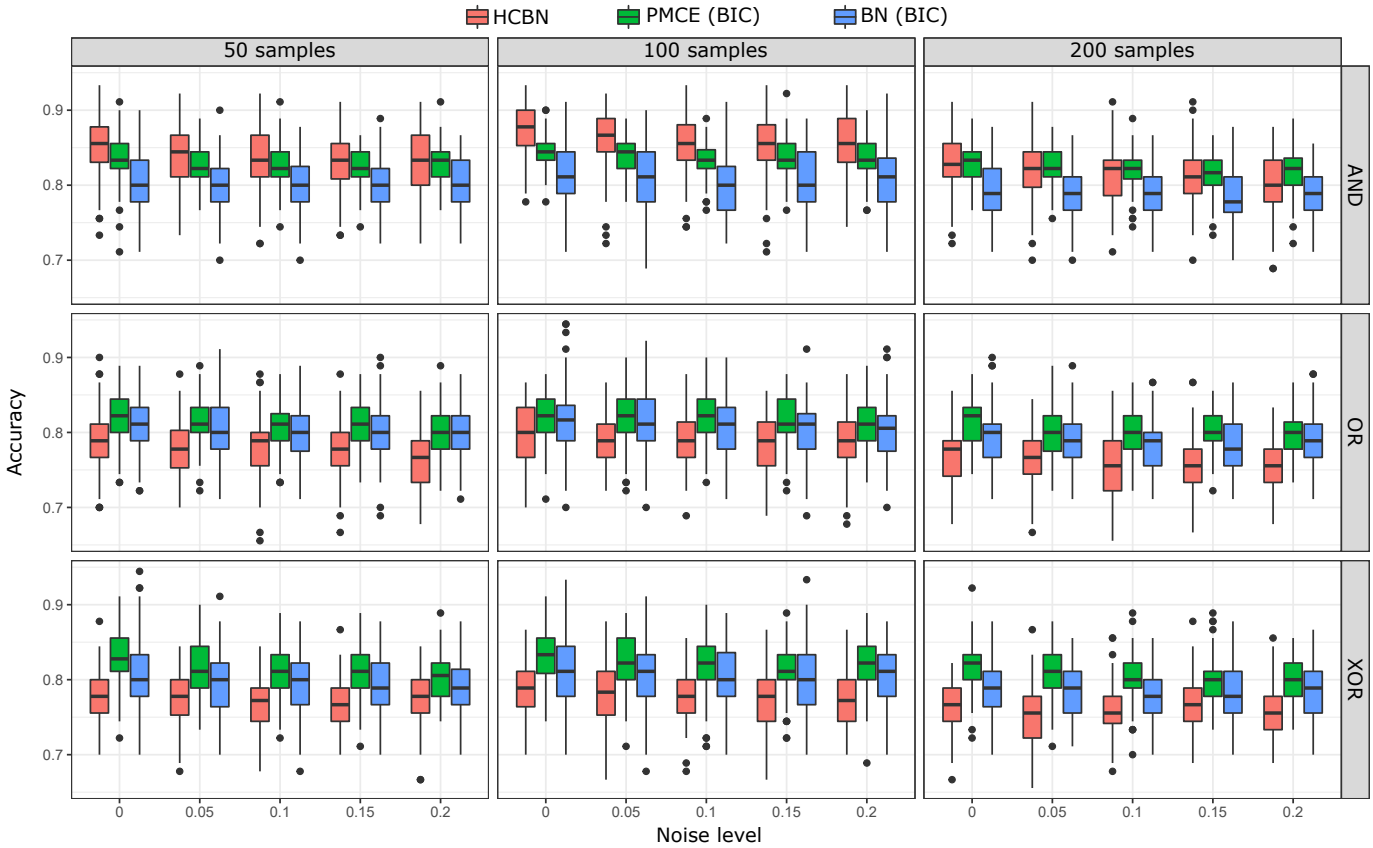

Fig 2. Performance assessment on simulations. 4500 independent synthetic datasets were generated from DAG generative topologies with $m = 10$ nodes (representing e.g. genomic variants) and density $= 0.4$, three distinct levels of topological complexity (only `AND` – panels **(A)**, `OR` – panels **(B)** 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 [23] and standard BNs [28] (with BIC regularization) on overall accuracy with respect to the ground-truth generative topology. Box-plots return the distributions on all simulations.

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.

3.2 Results on real data

We applied PMCE to 7866 samples from 16 cancer types from The Cancer Genome Atlas (TCGA) database [57]. The dataset includes bulk sequencing data, which is often coupled with clinical information on the related patient. In particular, the HESBCN model was inferred separately for every cancer type via PMCE, by considering putative driver mutations only, as identified in related works [3]. All models are displayed in Supplementary Figures S2-S17 and highlight different degrees of complexity and heterogeneity.

Predictability. For each model, we evaluated the predictability as per Eq. (12) and the results are shown in Figure 3**A-C**. 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.

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**A-B**). Notably, the overall correlation appear to be limited in both cases: $R^2 = -0.087$ and $R^2 = -0.025$, 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 [10, 33, 54].

Conversely, the number of formulas included in the evolution models shows a moderate anti-correlation with the predictability ($R^2 = -0.407$, Figure 3**C**). This result would suggest that, as intuitively expected, tumors with a larger number of possible trajectories are indeed the less predictable.

We also assessed the contribution of each genomic alteration included in the models to the overall predictability of the respective tumors. In particular, for each tumor type and each genomic event, we computed the predictability value of the subgraph defined by considering only the paths from the root to that event. With this analysis we obtained a predictability score for each genomic alteration with respect to every cancer type, which is displayed as a heatmap in Supplementary Figure S20.

Such analysis highlighted that driver mutations involved in many evolutionary paths, i.e. representing an early or necessary mutational event for a given tumour type, typically provide limited information in terms of predictability for that cancer. The most evident example is represented by the mutation of *TP53*, that is a pivotal driver in most cancer types, which shows always a very limited predictability score (see Supplementary Figure S20). The same pattern was observed for tissue-specific drivers, see, for instance, the mutations hitting *APC* in colorectal cancers, *BRAF* in thyroid cancers, *IDH1/2* in brain lower grade gliomas or *VHL* and *PBRM1* in kidney carcinomas. Interestingly, we also found driver mutations showing a mixed behaviour, being pivotal and scarcely contributing to predictability in some cancer types, but defining high predictable subtypes in others. A very interesting example in this regard is the mutation of *PIK3CA*, that is a major cancer driver of ER+ breast carcinomas, which displays a very low predictability in breast cancer, but a very high predictability score, e.g. in brain lower grade gliomas and stomach cancers.

Survival analysis. Survival analysis was executed by employing the Regularized Cox Regression via Coxnet [47, 52] with 10000 maximum iterations via elastic net, which allowed us to identify 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

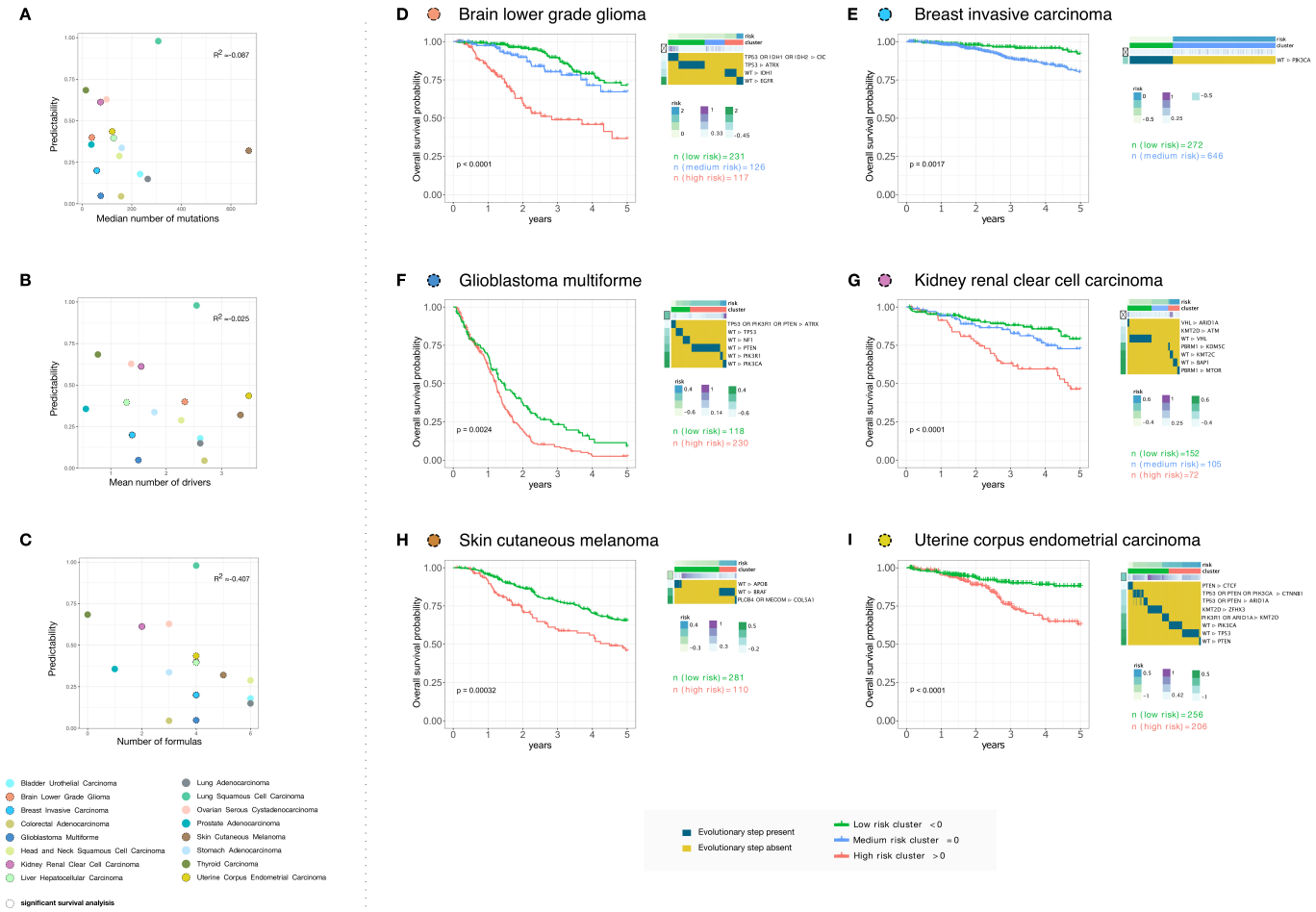

Fig 3. Application of PMCE to 7866 TCGA samples. PMCE was applied to 7866 samples of 16 different cancer types from the TCGA database [57], on the binarized profiles of driver mutations from [3]. **(A)** Scatter-plot returning the relation between the predictability of each tumor type and the median number of mutations, **(B)** the mean number of drivers, **(C)** the number of logical formulas present in the HESBCN model. 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 **(D)** brain lower grade glioma, **(E)** breast invasive carcinoma, **(F)** glioblastoma multiforme, **(G)** kidney renal clear cell carcinoma, **(H)** skin cutaneos melanoma **(I)** 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.

set of covariates (i.e. $\beta \neq 0$) 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).

We first estimated the stability of such results, by performing cross-validation as follows: (*i*) we randomly selected 80% of the samples of each cancer type, (*ii*) we performed the regularized Cox regression analysis for each sampled dataset and estimated the β coefficients, (*iii*) we independently repeated the sampling procedure 1000 times; (*iv*) we finally computed the Pearson correlation coefficient among the β coefficients obtained from the sampled dataset and those from the full dataset, for each of the 1000 independent samplings. As a result, the median Pearson correlation coefficients for the 7 considered cancer types are all larger than 0.90 (i.e. high correlation), proving the significant stability of the analysis (full results are provided in Supplementary Figure S19).

In Figure 3**D-I**, we show the Kaplan-Meier plots 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 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 gliomat (Fig. 3**D**)
[51] is characterized by the presence of three risk groups. The low risk cluster is characterized by evolutionary trajectories comprising an `OR` relation between mutations of *IDH1* and *IDH2*, which are typically associated with good prognosis [51]. 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. 3**F**) 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 [29]. In the breast invasive carcinoma (Fig. 3**E**) 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 [27].

The analysis of the kidney renal cell carcinoma (Fig. 3**G**) 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 [9]. With respect to the skin cutaneous melanoma (Fig. 3**H**), two different clusters are discovered, with the bad prognosis one associated to mutations on *BRAF* [6]. Finally, the uterine corpus endometrial carcinoma (Fig. 3**I**) shows an high number of relevant drivers, including mutations of *PTEN*, *TP53* and *PIK3CA* [51].

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. All in all, even though high-level clinical descriptors such as the overall survival might be affected by the presence of unconsidered variables and clinical covariates, we have here shown that a relatively limited number of features of the progression models inferred by PMCE are sufficient to stratify patients in well-divided risk groups, which in turn points at key molecular differences that might be further investigated.

Analysis of disjunctive relations in glioma models. Disjunctive relations may help grouping driver mutations with similar influence on phenotype, since a tumor might, for instance, exploit independent trajectories hitting the same pathway or cellular function [2]. As a proof of principle, we here focused our analysis on the disjunctive relations of the HESBCN models of gliomas that are found to be significant covariates in the Regularized Cox Regression analysis.

In the case of brain lower grade glioma (Fig. 3**D**), the HESBCN model includes a significant disjunctive relation, namely mutations of: *TP53* `OR` *IDH1* `OR` *IDH2* $\triangleright$ *CIC*. Genes *IDH1* and *IDH2* encode for the isocitrate dehydrogenase enzyme in different cell compartments, namely in cytosol and peroxysomes the former, in mithocondria the latter. This enzyme catalyzes the conversion of isocitrate to alpha-ketoglutarate, while reducing NADP to NADPH. There is evidence that variations in the utilization of this reaction are associated with increased glutamine reductive carboxylation and affect redox balance, glycolysis and oxidative

phosphorylation [16, 35]. We can speculate that mutations in these genes cause similar rearrangements in cell metabolism. The third gene involved in the disjunctive rule (i.e. *TP53*) encodes a tumor suppressor transcription factor instead. Although recent studies have shown that *TP53* has a role in the regulation of both glycolysis and oxidative phosphorylation [31], the encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair. Hence, speculating that mutations in this gene may phenocopy mutations in *IDH1* and *IDH2* is a tempting, but more hazardous hypothesis.

Along similar lines, the glioblastoma multiforme model (Fig. 3**F**) highlights the presence of the disjunction involving mutations of: *TP53* `OR` *PIK3R1* `OR` *PTEN* ▷ *ATRX*. *PIK3R1* is a member of the PI3K/AKT signal transduction pathway [37] and plays an important role in the metabolic action of insulin and, hence, in the regulation of glycolysis [1]. *PTEN* is known to be a major antagonist of PI3K activity in the PI3K-AKT pathway [55] and it is also supposed to be involved in the regulation of energy metabolism in the mitochondria [30]. *PTEN* also regulates P53 protein levels and activity via distinct mechanisms [21], whereas the role of mutant P53 in signaling pathways associated with glioblastoma multiforme is more elusive. These considerations support the hypothesis of a complex interplay involving such genes in this tumor type, which is consistent with the inferred disjunctive relation included in the HESBCN model.

Overall, these results show that the logical formulas inferred by PMCE include mutations of genes possibly involved in key molecular pathways and higher-level cellular functions (e.g. energy metabolism), and which the tumor might independently exploit to further progress.

Application of PMCE to tumors including distinct molecular subtypes. When reconstructing population-level models of tumor evolution from cross-sectional binarized mutational profiles of cancer patients, it is good practice to first stratify the samples into distinct molecular subtypes (as suggested, e.g. in [12]), in order to reduce the possible confounding effects deriving from processing highly heterogeneous mixtures of samples.

However, this task may present pitfalls and an effective stratification of samples into distinct subtypes might not always be possible. Hence, in order to assess the capability of PMCE in dissecting tumor types characterized by high heterogeneity, we performed two additional case studies.

We first applied PMCE to a pan-glioma dataset obtained by merging all lower-grade glioma (510 samples) and glioblastoma (338 samples) tumours (Supplementary Figure S21). We then mapped the samples of the distinct subtypes onto the evolution model inferred with our approach. Importantly, the HESBCN model highlights the presence of subtype-specific non-overlapping evolutionary trajectories. More in detail, consistently with [51], *IHD1*-mutant cancers are found to be enriched for lower-grade gliomas and to be involved in two major trajectories: the first one presenting additional mutations in *TP53* and *ATRX* genes (CIMP subtypes [13]) and the second one involving mutations in *CIC* and *NOTCH1* (codel subtype [13]). *IDH1* wild-type cancers [13] are instead mostly glioblastomas and are characterized by molecular evolutionary trajectories involing mutations in *EGFR*, *NF1*, *PTEN* or *RB1* genes.

We also analyzed the HESBCN model inferred from colorectal adenocarcinoma samples, by focusing on the two known molecular subtypes, i.e. microsatellite stable (MSS, 396 samples) and microsatellite instable (MSI, 62 samples). The HESBCN model associates canonical colorectal cancer drivers such as mutations of *APC*, *TP53* and *KRAS* to MSS cancers, while MSI tumours show a broad range of driver mutations comprising molecular trajectories involving *DMD*, *SPTA1*, *FBXW7* and *KMT2D* (see the Supplementary Figure S22). Collectively, these results demonstrate the effectiveness of PMCE in representing different molecular subtypes within a unique evolution model.

4 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. From the computational perspective, one limitation is that structural learning of Bayesian networks is an extremely hard task and there is no guarantee of converging or reaching global optima via MCMC search schemes [42, 44].

We also note that important results on the inference of the temporal ordering of genetic lesions in single tumours were achieved by estimating cancer timelines from Variant Allele Frequency profiles of single patients, via league model analysis and subclonal deconvolution [24]. However, as PMCE pools together data from multiple patients and drops Variant Allele Frequency information, it cannot – by construction – shed any light on the temporal evolution sequence that holds for a specific patient. Instead, PMCE allows one to infer a statistically robust population-level estimator of cancer evolution. Ideas from these two complementary approaches might lead to a more comprehensive characterization of tumours timelines, which we leave for future works.

A further pitfall of approaches processing low-resolution (binarized) data from bulk samples is due to the impact of intra-tumor heterogeneity, which is a major cause of therapy failure and relapse [48] 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 [26, 43, 58], 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 [34], despite the typically high levels of technical and biological noise [38].

In this respect, it would be interesting 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 [11]. 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 [40], with important translational repercussions.

Acknowledgements

We thank Marco Antoniotti, Lucrezia Patruno, Francesco Craighero, Davide Maspero and Gianluca Ascolani for useful discussions. We also thank Pablo Herrera Nieto and Ramon Diaz-Uriarte for their valuable comments on the first version of the manuscript.

Funding

This work was partially supported by a Bicocca 2020 Starting Grant to FA and DR. DR was also supported by a Premio Giovani Talenti of the University of Milan-Bicocca. This work was also supported by the CRUK/AIRC Accelerator Award #22790 "Single-cell Cancer Evolution in the Clinic". Financial support from the Italian Ministry of University and Research (MIUR) through the grant 'Dipartimenti di Eccellenza 2017' to the Department of Biotechnology and Bioscences of University of Milan-Bicocca is also acknowledged.

References

[1] T. Asano, M. Fujishiro, A. Kushiyama, Y. Nakatsu, M. Yoneda, H. Kamata, and H. Sakoda. Role of phosphatidylinositol 3-kinase activation on insulin action and its alteration in diabetic conditions. *Biological and Pharmaceutical Bulletin*, 30(9):1610–1616, 2007.

[2] Ö. Babur, M. Gönen, B. A. Aksoy, N. Schultz, G. Ciriello, C. Sander, and E. Demir. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome biology*, 16(1):1–10, 2015.

[3] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.

[4] N. Beerenwinkel, N. Eriksson, and B. Sturmfels. Conjunctive Bayesian Networks. *Bernoulli*, pages 893–909, 2007.

[5] N. Beerenwinkel, R. F. Schwarz, M. Gerstung, and F. Markowetz. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–e25, 2015.

[6] P. Bhatia, P. Friedlander, E. A. Zakaria, and E. Kandil. Impact of braf mutation status in the prognosis of cutaneous melanoma: an area of ongoing research. *Annals of translational medicine*, 3(2), 2015.

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

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

[9] H. Brauch, G. Weirich, J. Brieger, D. Glavač, H. Rödl, M. Eichinger, M. Feurer, E. Weidt, C. Puranakanitstha, C. Neuhaus, 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(7):1942–1948, 2000.

[10] R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.

[11] G. Caravagna, Y. Giarratano, D. Ramazzotti, I. Tomlinson, T. A. Graham, G. Sanguinetti, and A. Sottoriva. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature methods*, 15(9):707–714, 2018.

[12] G. Caravagna, A. Graudenzi, D. Ramazzotti, R. Sanz-Pamplona, L. De Sano, G. Mauri, V. Moreno, M. Antoniotti, and B. Mishra. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences*, 113(28):E4025–E4034, 2016.

[13] M. Ceccarelli and et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, 164(3):550–563, 2016.

[14] G. Ciriello, E. Cerami, C. Sander, and N. Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, 2012.

[15] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[16] C. Damiani, R. Colombo, D. Gaglio, F. Mastroianni, D. Pescini, H. V. Westerhoff, G. Mauri, M. Vanoni, and L. Alberghina. A metabolic core model elucidates how enhanced utilization of glucose and glutamine, with enhanced glutamine-dependent lactate production, promotes cancer cell growth: The warburq effect. *PLoS computational biology*, 13(9):e1005758, 2017.

[17] L. De Sano, G. Caravagna, D. Ramazzotti, A. Graudenzi, G. Mauri, B. Mishra, and M. Antoniotti. Tronco: an r package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics*, 32(12):1911–1913, 2016.

[18] R. Desper, F. Jiang, O.-P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of computational biology*, 6(1):37–51, 1999.

[19] R. Diaz-Uriarte and C. Vasallo. Every which way? on predicting tumor evolution using cancer progression models. *PLOS Computational Biology*, 15(8):1–29, 08 2019.

[20] E. Estrada and N. Hatano. Statistical-mechanical approach to subgraph centrality in complex networks. *Chemical Physics Letters*, 439(1-3):247–251, 2007.

[21] D. J. Freeman, A. G. Li, G. Wei, H.-H. Li, N. Kertesz, R. Lesche, A. D. Whale, H. Martinez-Diaz, N. Rozengurt, R. D. Cardiff, et al. Pten tumor suppressor regulates p53 protein levels and activity through phosphatase-dependent and-independent mechanisms. *Cancer cell*, 3(2):117–130, 2003.

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

[23] M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel. Quantifying cancer progression with conjunctive bayesian networks. *Bioinformatics*, 25(21):2809–2815, 2009.

[24] M. Gerstung, C. Jolly, I. Leshchiner, S. C. Dentro, S. Gonzalez, D. Rosebrock, T. J. Mitchell, Y. Rubanova, P. Anur, K. Yu, et al. The evolutionary history of 2,658 cancers. *Nature*, 578(7793):122–128, 2020.

[25] S.-R. Hosseini, R. Diaz-Uriarte, F. Markowetz, and N. Beerenwinkel. Estimating the predictability of cancer evolution. *Bioinformatics*, 35(14):i389–i397, 2019.

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

[27] K. Kalinsky, L. M. Jacks, A. Heguy, S. Patil, M. Drobnjak, U. K. Bhanot, C. V. Hedvat, T. A. Traina, D. Solit, W. Gerald, et al. Pik3ca mutation associates with improved outcome in breast cancer. *Clinical cancer research*, 15(16):5049–5059, 2009.

[28] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[29] J. A. Kraus, N. Glesmann, M. Beck, D. Krex, T. Klockgether, G. Schackert, and U. Schlegel. Molecular analysis of the pten, tp53 and cdkn2a tumor suppressor genes in long-term survivors of glioblastoma multiforme. *Journal of neuro-oncology*, 48(2):89–94, 2000.

[30] J. Liu and Z. Feng. Pten, energy metabolism and tumor suppression. *Acta Biochim Biophys Sin*, 44(8):629–631, 2012.

[31] J. Liu, C. Zhang, W. Hu, and Z. Feng. Tumor suppressor p53 and metabolism. *Journal of molecular cell biology*, 11(4):284–292, 2019.

[32] L. O. Loohuis, G. Caravagna, A. Graudenzi, D. Ramazzotti, G. Mauri, M. Antoniotti, and B. Mishra. Inferring tree causal models of cancer progression with probability raising. *PloS one*, 9(10):e108358, 2014.

[33] N. McGranahan and C. Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell*, 27(1):15–26, 2015.

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

[35] A. R. Mullen, W. W. Wheaton, E. S. Jin, P.-H. Chen, L. B. Sullivan, T. Cheng, Y. Yang, W. M. Linehan, N. S. Chandel, and R. J. DeBerardinis. Reductive carboxylation supports growth in tumour cells with defective mitochondria. *Nature*, 481(7381):385–388, 2012.

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

[37] B. Oskouian and J. D. Saba. Cancer treatment strategies targeting sphingolipid metabolism. *Sphingolipids as Signaling and Regulatory Molecules*, pages 185–205, 2010.

[38] L. Patruno, D. Maspero, F. Craighero, F. Angaroni, M. Antoniotti, and A. Graudenzi. A review of computational strategies for denoising and imputation of single-cell transcriptomic data. *Briefings in Bioinformatics*, bbaa222, 2020.

[39] J. Pearl. *Causality*. Cambridge university press, 2009.

[40] D. Ramazzotti, F. Angaroni, D. Maspero, G. Ascolani, I. Castiglioni, R. Piazza, M. Antoniotti, and A. Graudenzi. Longitudinal cancer evolution from single cells. *bioRxiv*, https://doi.org/10.1101/2020.01.14.906453, 2020.

[41] D. Ramazzotti, G. Caravagna, L. Olde Loohuis, A. Graudenzi, I. Korsunsky, G. Mauri, M. Antoniotti, and B. Mishra. Capri: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026, 2015.

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

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

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

[45] T. Sakoparnig and N. Beerenwinkel. Efficient sampling for bayesian inference of conjunctive bayesian networks. *Bioinformatics*, 28(18):2318–2324, 2012.

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

[47] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.

[48] A. Sottoriva, I. Spiteri, S. G. Piccirillo, A. Touloumis, V. P. Collins, J. C. Marioni, C. Curtis, C. Watts, and S. Tavaré. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, 110(10):4009–4014, 2013.

[49] P. Suppes. A probabilistic theory of causality. 1973.

[50] I. G. Szendro, J. Franke, J. A. G. de Visser, and J. Krug. Predictability of evolution depends nonmonotonically on population size. *Proceedings of the National Academy of Sciences*, 110(2):571–576, 2013.

[51] Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015.

[52] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

[53] S. Turajlic, A. Sottoriva, T. Graham, and C. Swanton. Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics*, 20(7):404–416, 2019.

[54] S. Turajlic, H. Xu, K. Litchfield, A. Rowan, T. Chambers, J. I. Lopez, D. Nicol, T. O'Brien, J. Larkin, S. Horswell, et al. Tracking cancer evolution reveals constrained routes to metastases: Tracerx renal. *Cell*, 173(3):581–594, 2018.

[55] F. Vazquez and W. R. Sellers. The pten tumor suppressor protein: an antagonist of phosphoinositide 3-kinase signaling. *Biochimica et biophysica acta*, 1470(1):M21–35, 2000.

[56] E. Wang, N. Zaman, S. Mcgee, J.-S. Milanese, A. Masoudi-Nejad, and M. O'Connor-McCourt. Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. In *Seminars in cancer biology*, volume 30, pages 4–12. Elsevier, 2015.

[57] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[58] H. Zafar, N. Navin, K. Chen, and L. Nakhleh. Siclonefit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome research*, 29(11):1847–1859, 2019.

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

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

[1] Dept. of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Milan, Italy,
[2] Dept. of Computer Science, Stanford University, 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.
*To whom correspondence should be addressed.

References

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

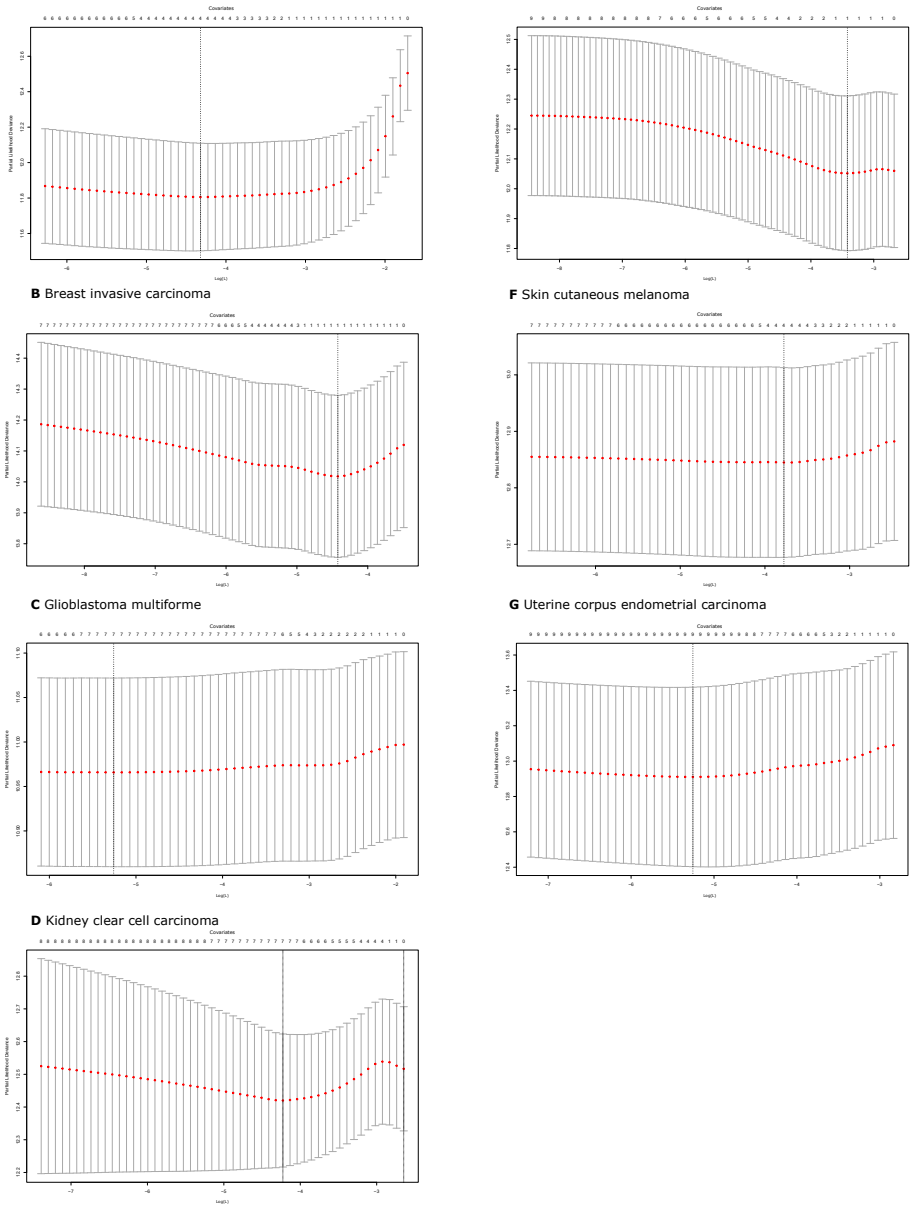

Fig. 1. Cross-validated error rates obtained with the method by Simon et al. (2011). The X axis is associated to the values of L (i.e. the coefficient of the elastic net penalty), with error bars providing a confidence interval for the cross-validated error rate. The vertical bars indicate the minimum error. The X axis gives the optimal number of covariates for a model. In this figure, we present only the cancer types that show the size of the model different from 0.

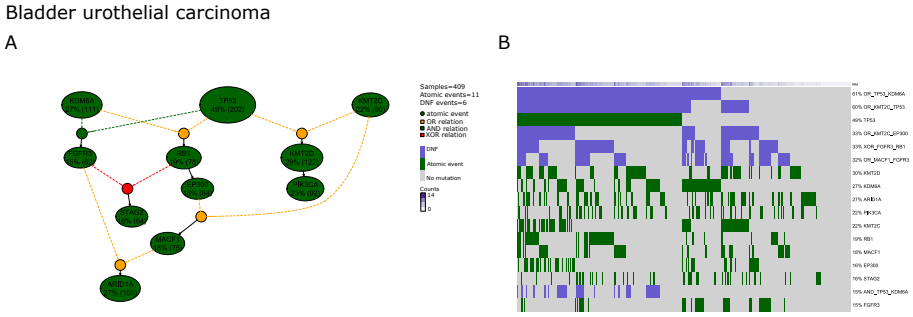

Fig. 2. A) PMCE model for bladder urothelial carcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

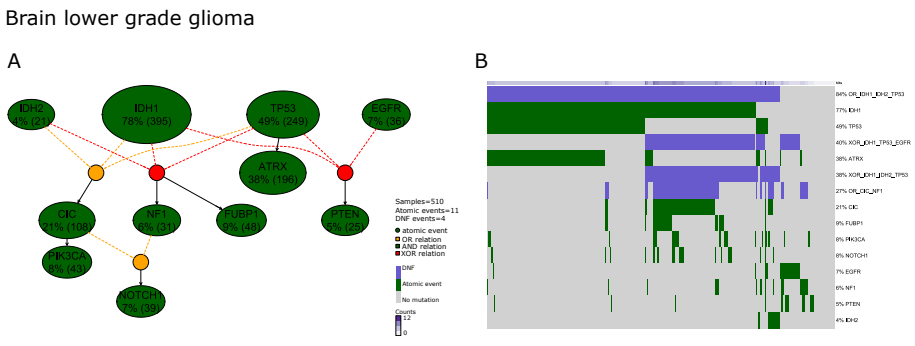

Fig. 3. A) PMCE model for brain lower grade glioma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

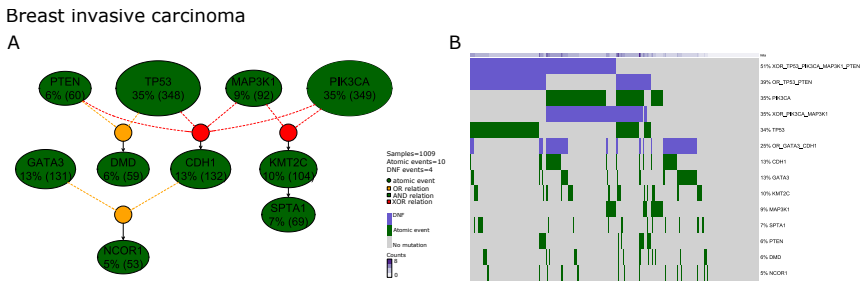

Fig. 4. A) PMCE model for breast invasive carcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

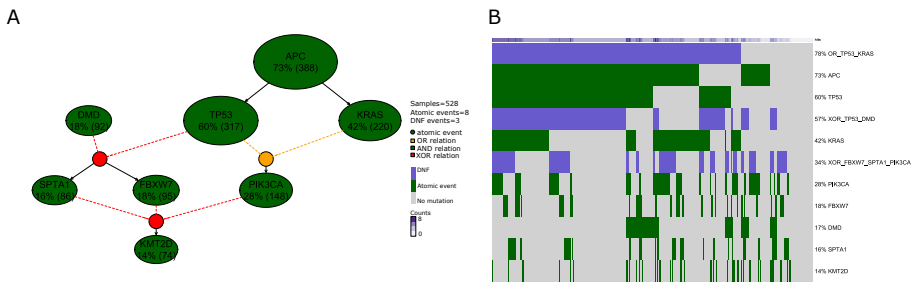

Fig. 5. A) PMCE model for colorectal adenocarcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

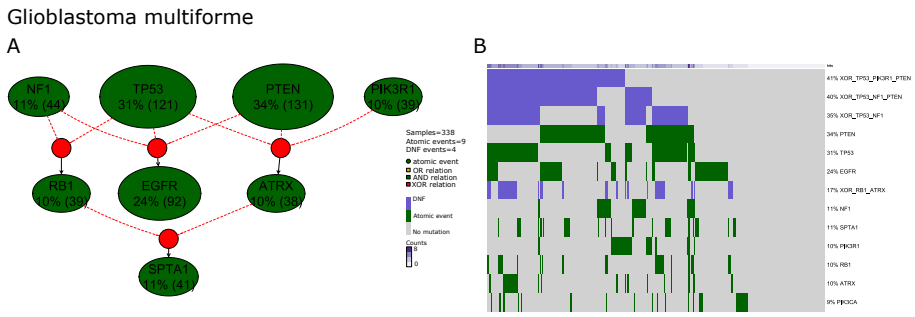

Fig. 6. A) PMCE model for glioblastoma multiforme; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

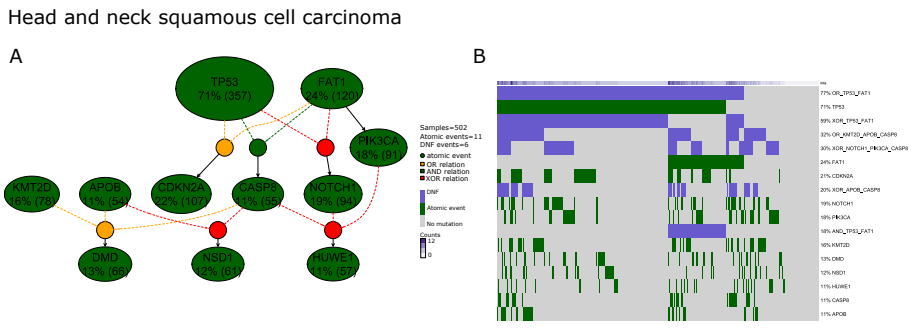

Fig. 7. A) PMCE model for head and neck squamous cell carcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

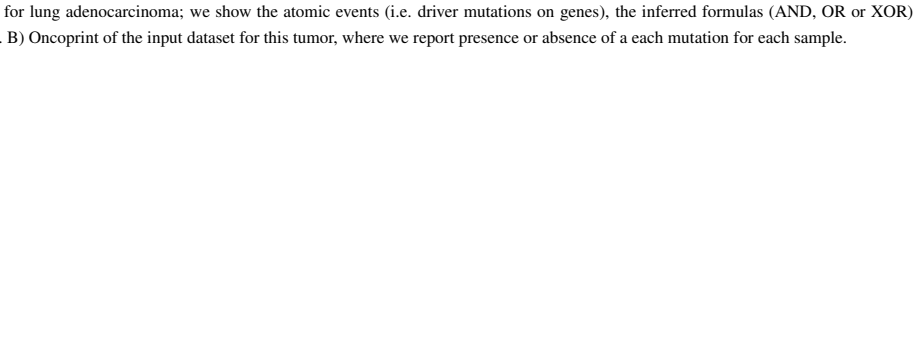

Fig. 8. A) PMCE model for kidney renal clear cell carcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

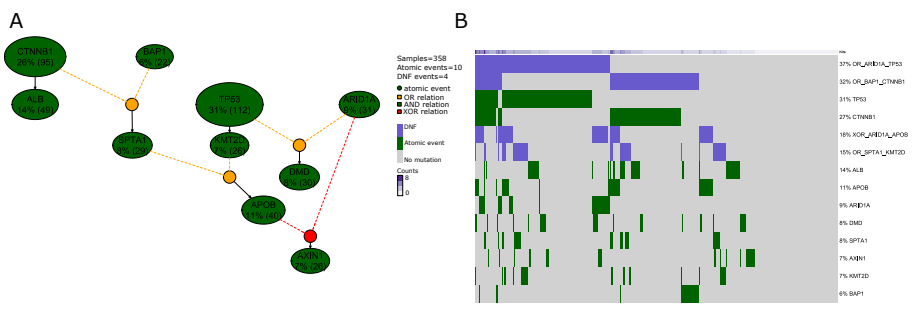

Fig. 9. A) PMCE model for liver hepatocellular carcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

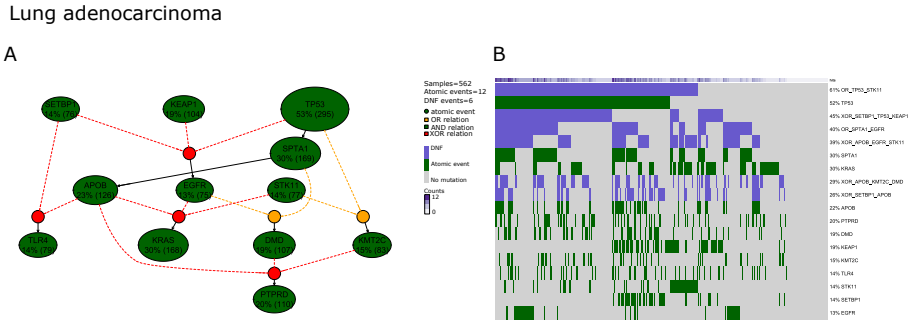

Fig. 10. A) PMCE model for lung adenocarcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

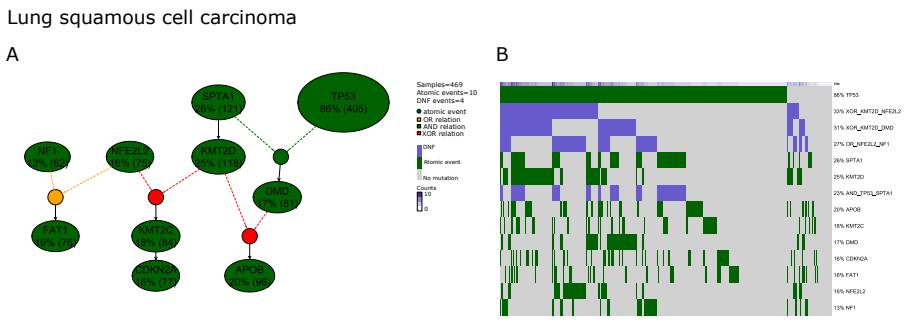

Fig. 11. A) PMCE model for lung squamous cell carcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

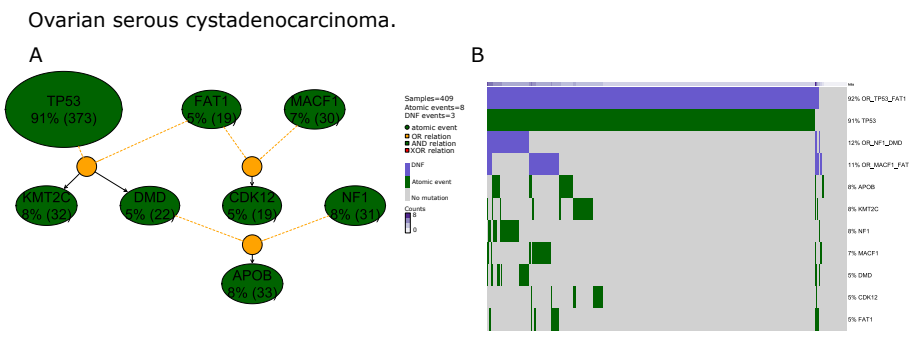

Fig. 12. A) PMCE model for ovarian serous cystadenocarcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

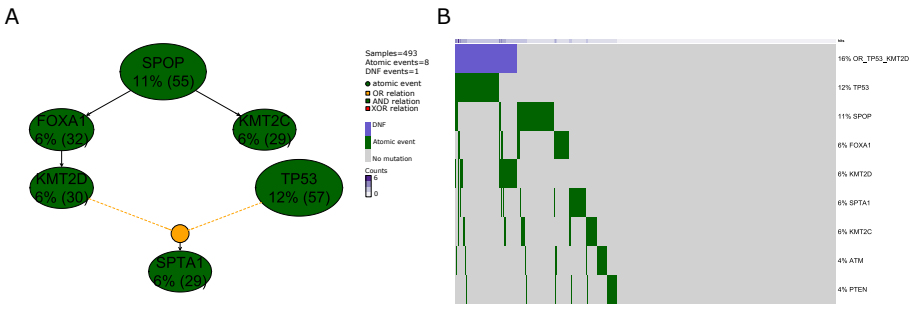

Fig. 13. A) PMCE model for prostate adenocarcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

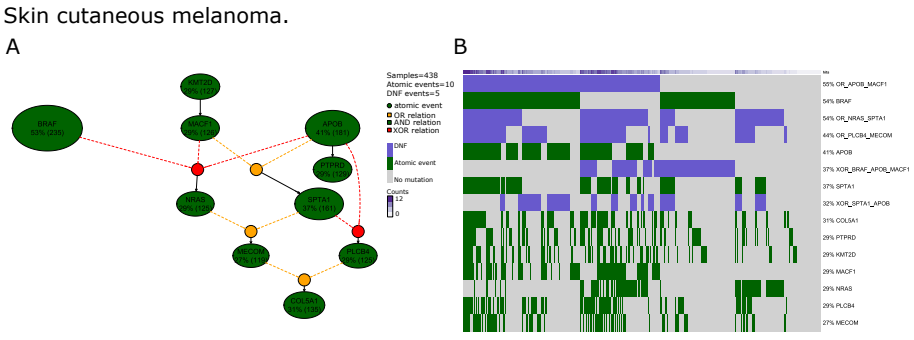

Fig. 14. A) PMCE model for skin cutaneous melanoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

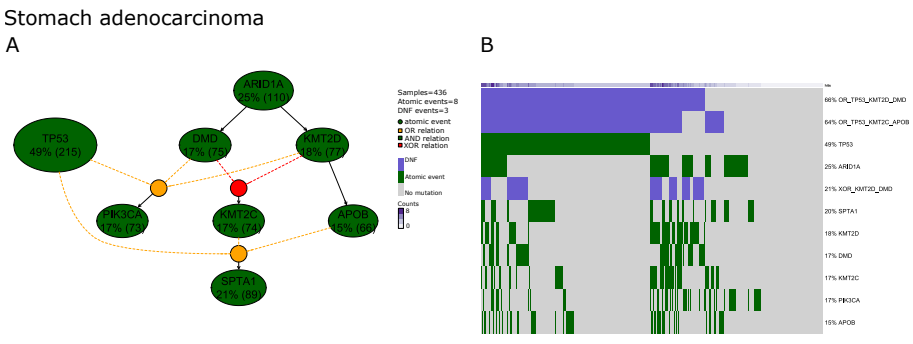

Fig. 15. A) PMCE model for stomach adenocarcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

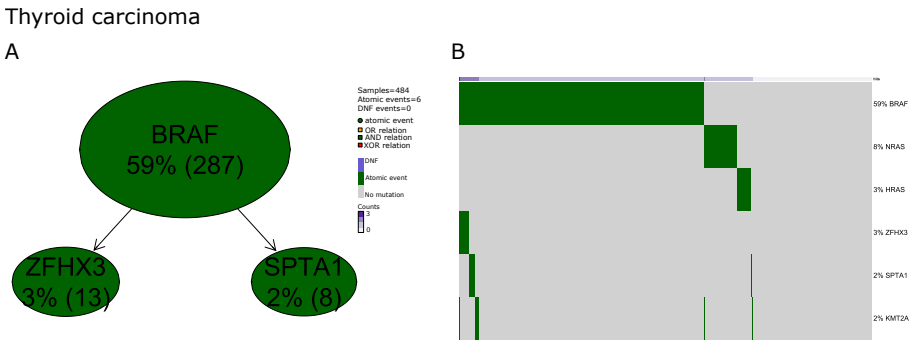

Fig. 16. A) PMCE model for thyroid carcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

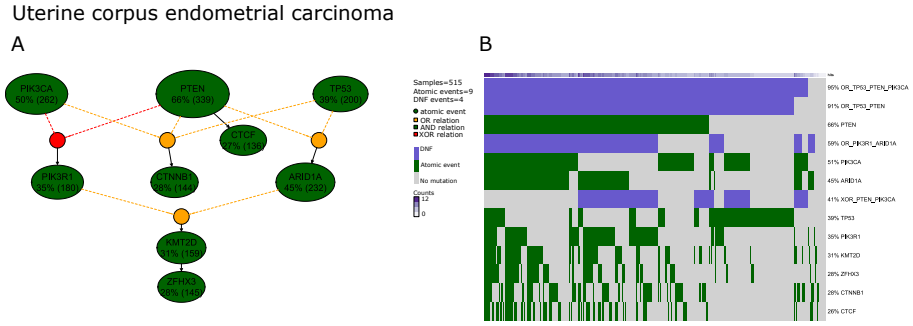

Fig. 17. A) PMCE model for uterine corpus endometrial carcinoma; we show the atomic events (i.e. driver mutations on genes), the inferred formulas (AND, OR or XOR) and their prevalence in the considered cancer samples. B) Oncoprint of the input dataset for this tumor, where we report presence or absence of a each mutation for each sample.

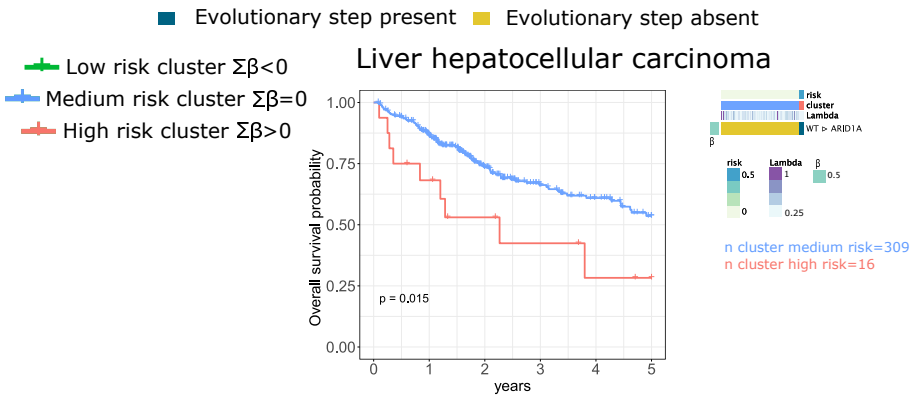

Fig. 18. Risk groups for Liver hepatocellular carcinoma.

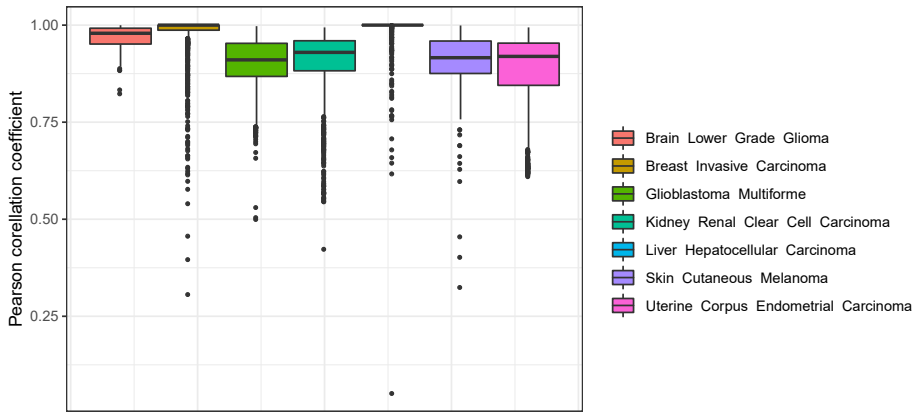

Fig. 19. Results of the cross-validation performed to estimate the stability of the survival analysis discussed in the main text. For each of the 7 cancer types that display a non-empty set of covariates associated to the minimum cross-validation error of the Regularized Cox Regression via Coxnet: (i) 80% of the samples are selected, (ii) the regularized Cox regression analysis is performed on the sampled dataset, (iii) the sampling procedure is repeated 1000 times; (iv) the Pearson correlation coefficient among the β coefficients obtained from the sampled dataset and those from the full dataset is computed. The plot returns the distribution of the Pearson correlation coefficients for the 1000 independent sampling, with respect to each cancer type.

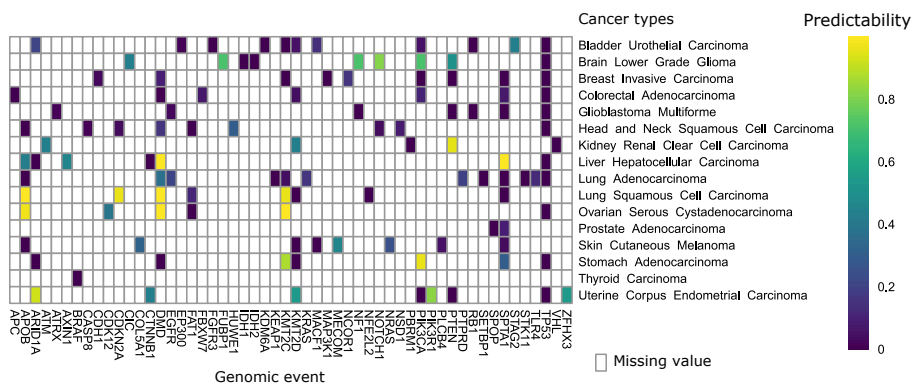

Fig. 20. Heatmap reporting the predictability score (range $[0, 1]$) for each genomic event included in the HESBCN models of the 16 considered cancer types. For each cancer type, for each genomic event, the predictability value of the subgraph defined by considering only the paths from the root to that event is computed as per Eq. (12) of the main text and reported. White cells in the heatmap indicates missing information, i.e. genomic events that are not present in the HESBCN of a given cancer type.

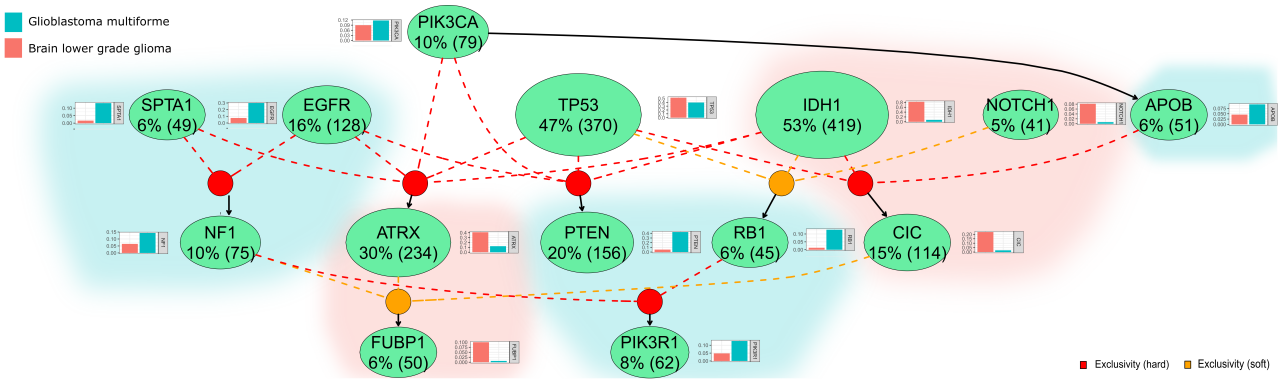

Fig. 21. HESBCN model for a pan-gliomas dataset including 510 lower grade gliomas (red) and 338 glioblastomas (green). Next to each genomic event, we show the barplot reporting the proportion of samples of the two distinct tumor types displaying that event; the molecular trajectories in which the prevalence of a given tumor is dominant are highlighted with a colored shades.

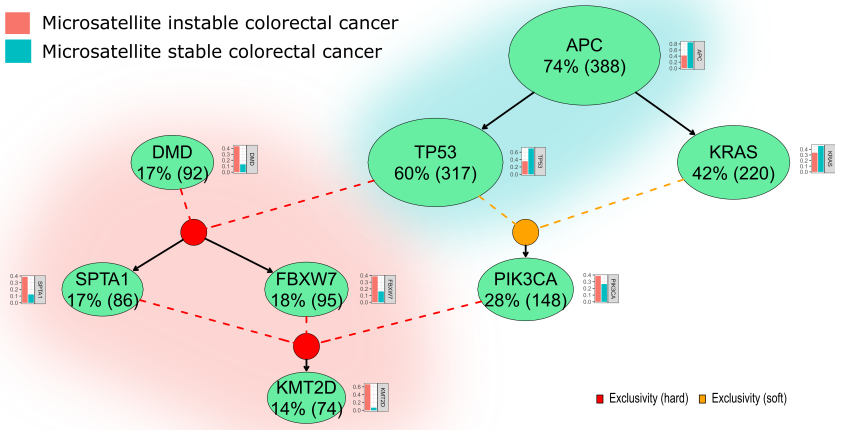

Fig. 22. HESBCN model for a dataset including 458 colorectal tumours comprising 396 microsatellite stable (MSS, green) and 62 microsatellite instable (MSI, red) tumours. Next to each genomic event, we show the barplot reporting the proportion of samples of the two distinct tumor types displaying that event; the molecular trajectories in which the prevalence of a given tumor is dominant are highlighted with colored shades.