Oncogenetic Network Estimation with Disjunctive Bayesian Networks: Learning from Unstratified Samples while Preserving Mutual Exclusivity Relations

Phillip B. Nicol [1*], Kevin R. Coombes [2], Courtney Deaver [3], Oksana A. Chkrebtii [4], Subhadeep Paul [4], Amanda E. Toland [5] and Amir Asiaee [6*]

[1]Harvard University, Cambridge, MA 02138, USA, [2]Department of Biomedical Informatics, Ohio State University, Columbus, OH 43210, USA, [3]Natural Sciences Division, Pepperdine University, Malibu, CA 90263, USA, [4]Department of Statistics, Ohio State University, Columbus, OH 43210, USA [5]Departments of Cancer Biology and Genetics and Department of Internal Medicine, Division of Human Genetics, Comprehensive Cancer Center, Ohio State University, Columbus, OH, 43420, USA and [6]Mathematical Biosciences Institute, Ohio State University, Columbus, OH 43210, USA

ABSTRACT

Cancer is the process of accumulating genetic alterations that confer selective advantages to tumor cells. The order in which aberrations occur is not arbitrary, and inferring the order of events is a challenging problem due to the lack of longitudinal samples from tumors. Moreover, a network model of oncogenesis should capture biological facts such as distinct progression trajectories of cancer subtypes and patterns of mutual exclusivity of alterations in the same pathways. In this paper, we present the Disjunctive Bayesian Network (DBN), a novel cancer progression model. Unlike previous models of oncogenesis, DBN naturally captures mutually exclusive alterations. Besides, DBN is flexible enough to represent progression trajectories of cancer subtypes, therefore allowing one to learn the progression network from unstratified data, i.e., mixed samples from multiple subtypes. We provide a scalable genetic algorithm to learn the structure of DBN from cross-sectional cancer data. To test our model, we simulate synthetic data from known progression networks and show that our algorithm infers the ground truth network with high accuracy. Finally, we apply our model to copy number data for colon cancer and mutation data for bladder cancer and observe that the recovered progression network matches known biological facts.

INTRODUCTION

Cancer is an evolutionary process that can be modeled as a sequence of fixation of genetic alterations throughout the tumor cell population (1, 11). Each new driver alteration confers a selective growth advantage to the cell and sweeps through the population, which results in clonal expansion (44). But the alterations and the order in which they accumulate are not arbitrary. Alterations are restricted by tissue and exposure types and their order is determined by the type of conferred advantage. Inferring the order of alterations has been shown to have diagnostic and prognostic importance (1, 11) but is a challenging problem due to the lack of longitudinal samples from tumors. The first model of tumorigenesis by Fearon and Vogelstein (22) was developed for colon cancer and suggested that a *chain* of aberrations is required to transform normal cells into carcinoma. Many cancer types, however, are only diagnosed in the later stages of the disease, meaning that early events driving cancer progression are usually hidden in available data. Our goal is thus to infer the order of alterations from the cross-sectional data.

Recently, it was shown that cancers of the same type in different individuals have very few or no driver mutations in common (44), which suggests that chain models are not enough to capture cancer progression. Desper's *Oncogenetic tree* (20) modeled progression as a rooted directed tree (branching). A mixture of oncogenetic trees (9, 10) was proposed to capture the presence of an aberration in multiple progression paths. *Directed Acyclic Graphs* (DAGs) are the next straightforward generalization of tree-based models, as they allow multiple alterations (parents) to set up the clonal stage for the appearance of a new aberration (the child). Bayesian networks (BN), which are DAGs equipped with a joint probability distribution (4), lend themselves naturally to representing such models. Perhaps the most famous BN model of cancer progression is the Conjunctive Bayesian Network (CBN) (8, 23) which assumes all parent aberrations must be present in order for a child alteration to occur.

From the evolutionary cancer modeling perspective, the assumptions of CBNs are very restrictive because a single advantageous hit is usually enough for clonal expansion and preparation of the tumor for future hits (44). Moreover, it is known that genes of the same pathway are altered mutually exclusively (31) in the population and therefore under the CBN progression assumption those genes cannot share any descendant alterations, Figure 1a. The inability of CBN to capture mutual exclusivity of alterations has motivated a line of work in which the mutual exclusivity restriction and pathway information are introduced artificially to the CBN (17, 24).

*Emails: phillipnicol@college.harvard.edu and asiaeetaheri.1@osu.edu

Furthermore, the CBN progression rule makes the corresponding inference very sensitive to false positives, i.e., passenger alterations. Since any passenger alteration co-occurs with drivers, which are usually more frequent, the CBN assumption requires each passenger alteration to be the child of all drivers and therefore distorts the shape of the graph from the underlying ground truth network, Figure 1b.

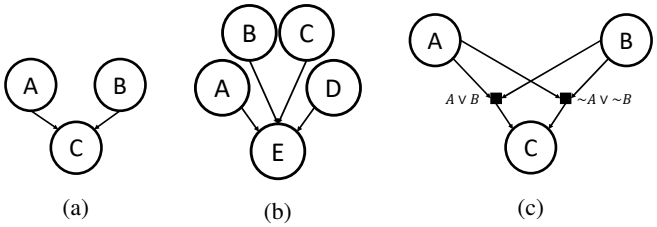


Figure 1. Issues with current state-of-the-art methods of cancer progression inference. (a) In the CBN-based models, mutual exclusive nodes A and B can not share a child C because $A \wedge B$ should be true for progressing to C. **(b)** The CBN-based models are sensitive to passenger alterations, like E. Many edges are added from driver alterations to E to comply with the CBN progression rule, which falsely renders E as an important alteration. **(c)** The probability raising models of progression are unable to capture mutual exclusivity, and therefore extra logical nodes (black squares) are added to the network to enhance the expressiveness of the model.

Related Work

Existing network models of oncogenesis focus on extending the above model to a variety of more complicated settings.

The models of tumorigenesis discussed thus far are discrete-time models. There are continuous-time extensions such as timed oncogenetic trees (20) and continuous-time CBN (6). Progression models alone do not capture the range of observed data, which makes likelihood-based methods assign zero probability to such data sets. One way of addressing this issue is to consider measurement error, i.e., false positive and negative observations, as the source of non-compliance (23, 50). Another approach is to relax the original model and accommodate some deviations. For example, the mixture of oncogenetic trees model (9, 10) captures the independently arising alterations in a separate star-shaped individual tree, which confers the flexibility to all alterations to happen without any parent. Existing approaches considering pathways and their effects in cancer progression either assume that the pathways are an input of the progression inference algorithm (17, 24) or learn them along with the progression network based on the principle of mutual exclusivity of mutations belonging to the same pathway (18, 43). Finally, population genetic models such as Wright-Fisher (7) and Moran (2) processes have been used to model the evolution of cancer as a large absorbing Markov chain whose states are cell population with specific fixated alterations. The absorbing states represent a diagnosed tumor or fully developed tumor and state transition is determined by the fitness values of alterations. These population genetic models are related to the progression network viewpoint of the same process, but the investigation of their connections is beyond the scope of this paper. Thinking about cancer progression in a population genetic framework allows more refined modeling of progression by considering aspects such as the number of cells, alteration rate, and fitness of each alteration (2).

There have been several recent attempts to model the accumulation of alterations by Suppes' probability raising causal framework (15, 19, 34, 41, 42). Intuitively, these methods for two alterations A and B test the following two inequalities using their frequency counts in the given data set to determine if A is a parent of B: $\mathbb{P}(A) > \mathbb{P}(B)$ and $\mathbb{P}(A|B) \geq \mathbb{P}(A|\neg B)$. However, the causality definition of Suppe's has been proven to be insufficient for modeling cause and effect mainly because it is symmetric: A raises B's probability if and only if B does so for A (37, 38). Even in Desper's original paper, the authors address the impossibility of reconstruction of skewed oncotrees, i.e., trees with spurious topological edges (20, 34). Therefore the condition of $\mathbb{P}(A) > \mathbb{P}(B)$ is the only factor that determines the order of two alterations. But this condition just assumes that the more frequent alteration should have happened earlier, which is the core heuristic of all non-causal progression inference method.

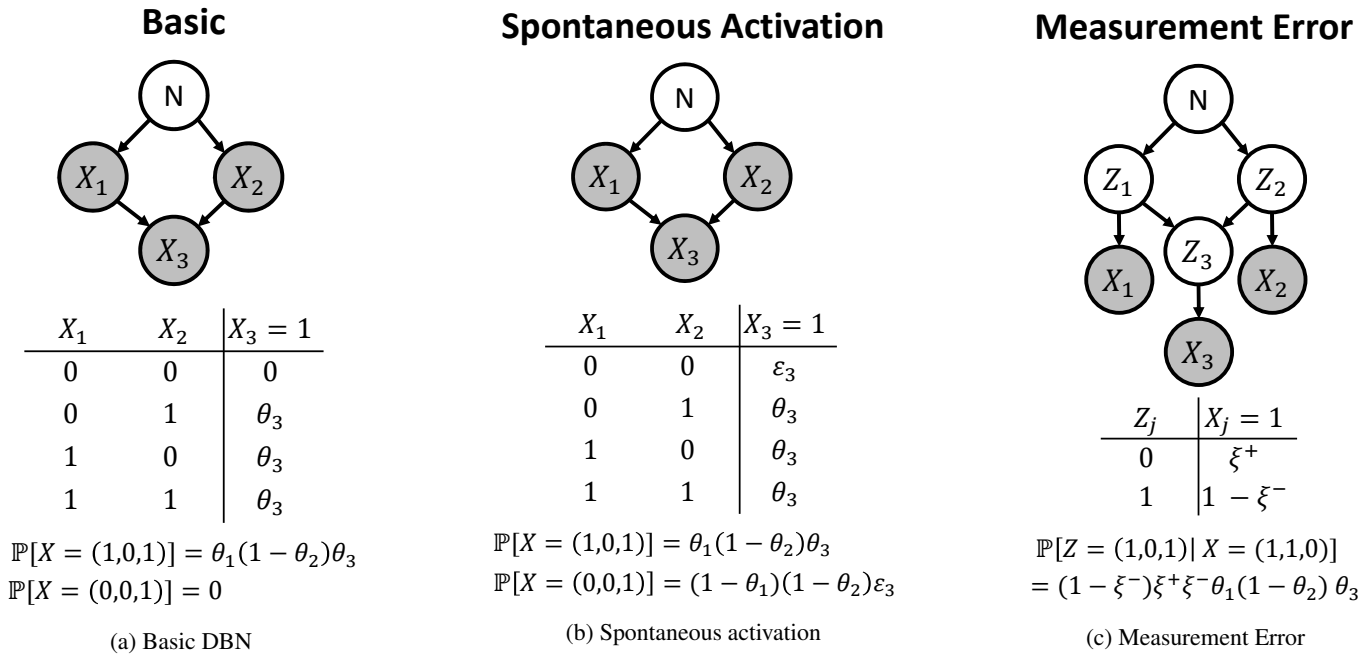
Additionally, mutual exclusivity of causes cannot be modeled directly in the standard Suppes' framework (41). Therefore, it has been suggested (15, 41) to augment the progression network with artificial nodes required for modeling mutual exclusivity. For example, if C has A and B as mutually exclusive parents, $A \vee B$ and $\neg A \vee \neg B$ should be added to the network, Figure 1c. Learning progression networks, under the assumption of mutual exclusivity, using causal discovery methods (32, 48, 49) is an open question.

Lastly, since each cancer subtype has distinct molecular characteristics and (semi-)disjoint progression path, one must first stratify samples to disjoint subtypes and then learn the progression network of each subtype separately. Note that this extra step is required for all of the above models mainly because they cannot capture mutual exclusivity of subtypes naturally. PICNIC (15) is the state-of-the-art pipeline that clusters samples to subtypes, detects driver events, checks for statistically significant mutual exclusivity hypotheses or takes pathway information as an input, and infers the progression network from one of the several available models.

Our Contribution

Due to the intrinsic shortcomings of the state-of-the-art methods, in this paper, we propose the Disjunctive Bayesian Network (DBN), which relaxes the CBN progression assumption. The *DBN progression rule* assumes that each alteration can occur if at least one of its parents has occurred first, Figure 2a. Our results show that DBN can naturally accommodate distinct progression paths for subtypes and capture mutual exclusivity of alterations present in the data. Therefore, one can skip two preprocessing steps that are necessary in state of the art models: stratifying samples by subtype and mutual exclusivity detection.

In DBN the probability of an event does *not* increase with the number of its parents that have occurred. This assumption makes biological sense, because usually an aberration hitting any genes of a pathway is enough to perturb the pathway's function, and give the cell a selective advantage and prepare it for the next alteration. The DBN progression rule is therefore in contrast with the well-known CBN model where the assumption is that *all* parent events should occur to make the child event possible (8).



(a) Basic DBN (b) Spontaneous activation (c) Measurement Error

Figure 2. Bayesian networks of the three cancer progression models investigated in this paper. Node N represents normal cell state, and each random variable X_j is an observed alteration, and the corresponding progression probability parameter is θ_j. In all models, the conditional probability table of X_3 is shown, and probabilities of instance observations are computed. **(a)** Basic DBN model where further progression is impossible if none of the parent alterations have occurred. **(b)** Spontaneous activation model where there is a non-zero chance of a child occurring even if none of its parents are active. **(c)** Measurement error model where each actual unobserved alteration Z_j generates an observation X_j according to universal false positive and negative probabilities.

We consider two extensions of our proposed model. The first extension relaxes the strict disjunction assumption and allows spontaneous alteration. This means that an aberration may occur in spite of unaltered parents, Figure 2b. The second extension directly models false positive and negative errors in measuring alterations, and therefore allows observations to deviate from the standard DBN model, Figure 2c.

Although we are not directly modeling pathways, each set of parent alterations can be thought of as a pathway whose hit of its single element is sufficient for progression. In this way, each alteration can belong to more than one pathways and pathways can have non-linear interaction, which is more general than the state-of-the-art pathway linear progression (43) and pathTiMEx (18) models.

We present a *genetic algorithm* for learning the structure of DBN from cross-sectional tumor data. We characterize a likelihood-equivalence relation over DAGs representing the DBN and use it to speed up the algorithm by only searching through the representative DAGs of each equivalence class. We show that the ability of the proposed algorithm in reconstructing ground truth progression networks from simulated data sets and inferring biologically interpretable progression networks for colon and bladder cancer.

In summary, our scalable ALgorithm for Oncogenesis Network Estimation, ALONE, based on the biologically-derived progression rule of DBN, captures patterns of mutual exclusivity, pathway perturbations, and disjoint progression of cancer subtypes *alone*, without contrived modeling of each one of these issues separately.

Notation. We denote sets by capital script $\mathcal{V}$, matrices by bold capital $\mathbf{V}$, random variables and vectors by capital V, and vectors of their realized values by small bold $\mathbf{v}=(v_1,\ldots,v_p)$ letters. To select a specific index set $\mathcal{S}$ of a vector we use the notation $\mathbf{v}(\mathcal{S})$.

METHODS

We model the observation of genomic events as a binary random vector $(X_1,\ldots,X_p)$, where $X_j=1$ if the j-th event (e.g., mutation or loss and gain of chromosome arms) is detected in the sample. We represent a realization (sample) of the event vector with $\mathbf{x}=(x_1,\ldots,x_p)$. Moreover, we assume that a Bayesian Network (**BN**) governs the order in which the events can occur. The BN consists of a Directed Acyclic graph (**DAG**) G and local Conditional Probability Distributions (**CPD**) $\mathbb{P}(x_j|\mathbf{x}(\mathcal{P}_j);\boldsymbol{\theta})$ where $\mathcal{P}_j$ is the set of parents of event j in G and θ parameterizes the distribution. Local conditional probabilities form the joint distribution of all events as follows:

$$\mathbb{P}(\mathbf{x};G,\boldsymbol{\theta})=\prod_{j=1}^{p}\mathbb{P}(x_j|\mathbf{x}(\mathcal{P}_j);\boldsymbol{\theta}). \quad (1)$$

In this section, we first present the Disjunctive Bayesian Network (**DBN**) progression rule and describe how it determines the form of the CPDs. Next, we improve our initial progression model by presenting two more realistic variants of DBN. For each variant, we derive maximum likelihood estimators for the network parameters $\boldsymbol{\theta}$. Finally, we provide a Genetic Algorithm (**GA**) to search the space of DAGs for the optimal network structure. To simplify this search, we define an equivalence relation between networks and ensure that only one network per equivalence class is searched.

DBN Progression Rule

Basic DBN. The DBN progression rule allows an event to occur if and only if at least one of its parents have occurred. Given that at least one parent of event j occured, event j occurs with probability θ_j. Therefore, CPDs of the basic model, Figure (2a), take the following form:

$$\mathbb{P}(x_j=1|\mathbf{x}(\mathcal{P}_j);\boldsymbol{\theta})=\begin{cases}0, & \mathbf{x}(\mathcal{P}_j)=\mathbf{0}\\ \theta_j, & \text{otherwise}\end{cases}. \tag{2}$$

Although it is known that there exists an order for genetic events, we do not expect the proposed model to perfectly comply with the observed data. For this reason, we investigate two variants of DBN which capture deviations from the basic model (**2**).

Spontaneous Activation Model. One can assume that observations that deviate from the proposed progression model are the results of *spontaneous activation* caused by unknown sources. Therefore, we consider a non-zero spontaneous activation probability $\varepsilon_i>0$ for each node (Figure 2b) such that

$$\mathbb{P}(x_j=1|\mathbf{x}(\mathcal{P}_j);\boldsymbol{\theta})=\begin{cases}\varepsilon_j, & \mathbf{x}(\mathcal{P}_j)=\mathbf{0}\\ \theta_j, & \text{otherwise}\end{cases}. \tag{3}$$

Measurement Error Model. One can attribute deviation from the progression network to measurement error, i.e., presence of false positive and negative observations (50). False positives and negatives can arise from errors in measurement technology. A false negative (failing to observe an event) can also arise from having a single sample from a spatially heterogeneous tumor.

We assume that there are unique (across all events) false positive ξ^+ and negative ξ^- probabilities that generates the observed event $\mathbf{x}$ from the underlying latent event $\mathbf{z}$ as follows:

$$\mathbb{P}(x_j=1|z_j=0)=\xi^+,\quad \mathbb{P}(x_j=0|z_j=1)=\xi^-. \tag{4}$$

The corresponding graphical model is depicted in Figure 2c.

Parameter Estimation

Given n cross-sectional samples and the progression network G, we wish to find $\hat{\boldsymbol{\theta}}_G^{\text{MLE}}$, the maximum likelihood estimator (MLE) for $\boldsymbol{\theta}$ in each variant of the DBN.

MLE for the Basic DBN First, we need to compactly write the joint distribution of events using matrix $\mathbf{A}$, the adjacency matrix of G.

PROPOSITION 1. *The likelihood can be written as*

$$\mathbb{P}(\mathbf{x};\boldsymbol{\theta},G)=\prod_{j=1}^{p}[\theta_j{}^{x_j}(1-\theta_j)^{1-x_j}]^{\mathbb{1}(\mathbf{x}(\mathcal{P}_j)\neq\mathbf{0})}(1-x_j)^{\mathbb{1}(\mathbf{x}(\mathcal{P}_j)=\mathbf{0})} \tag{5}$$

where $\mathbb{1}$ is the indicator function and $\mathbb{1}(\mathbf{x}(\mathcal{P}_j)\neq\mathbf{0})$ checks if any of j's parents has occurred. Note that $\mathbf{x}(\mathcal{P}_j)$ can be computed easily as $(\mathbf{a}_j\odot\mathbf{x})_{\mathcal{P}_j}$ where $\mathbf{a}_j$ is the jth column of $\mathbf{A}$, $\odot$ is the Hadamard product and the subscript $\mathcal{P}_j$ selects parents of j from the vector.

Using the compact representation (**5**) of the likelihood, one can compute the MLE of $\boldsymbol{\theta}$ for the basic DBN model (**2**).

PROPOSITION 2. *Given n independent samples $\{\mathbf{x}_i\}_{i=1}^n$ from the same population defined by G and θ where $\mathbf{x}_i\in\{0,1\}^p$, MLE for θ_j is*

$$\hat{\theta}_j^{MLE}=\frac{\sum_{i=1}^n\mathbb{1}(x_{ij}=1,\mathbf{x}_i(\mathcal{P}_j)\neq\mathbf{0})}{\sum_{i=1}^n\mathbb{1}(\mathbf{x}_i(\mathcal{P}_j)\neq\mathbf{0})}. \tag{6}$$

where x_{ij} is the realization of the jth event in the ith sample.

Intuitively, (**6**) is just a sample proportion. The denominator counts the number of samples in which at least one of the parents of j occurred while the numerator counts those where j occurred along with at least one of its parents.

MLE for the Spontaneous Activation Model. The likelihood of the spontaneous activation model (**3**) is as follows:

$$\mathbb{P}(\mathbf{x};\boldsymbol{\theta},G)=\prod_{j=1}^{p}[\theta_j{}^{x_j}(1-\theta_j)^{1-x_j}]^{\mathbb{1}(\mathbf{x}(\mathcal{P}_j)\neq\mathbf{0})}\varepsilon_j^{\mathbb{1}(\mathbf{x}(\mathcal{P}_j)=\mathbf{0})}. \tag{7}$$

Similarities between likelihoods (**5**) and (**7**) suggest that the MLE for $\boldsymbol{\theta}$ of the spontaneous activation model should be the same as for the $\boldsymbol{\theta}$ of the basic DBN presented in (**6**). Thus we only need to compute the MLE for ε.

PROPOSITION 3. *Given n independent samples from the same population defined by G and θ, the MLE of θ for the spontaneous activation model* (**3**) *is as* (**6**) *and the MLE of ε_j can be computed as:*

$$\hat{\varepsilon}_j^{MLE}=\frac{\sum_{i=1}^n\mathbb{1}(x_{ij}=1,\mathbf{x}_i(\mathcal{P}_j)=\mathbf{0})}{\sum_{i=1}^n\mathbb{1}(\mathbf{x}_i(\mathcal{P}_j)=\mathbf{0})}. \tag{8}$$

where x_{ij} is the realization of the jth event in the ith sample.

The ratio in (**8**) counts the percentage of samples in which j has occurred without any parent and is thus an intuitively reasonable estimator of the spontaneous activation rate.

EM for the Measurement Error Model Assuming ξ^+ and ξ^- are fixed and known, the MLE of $\boldsymbol{\theta}$ can be approximated using the Expectation Maximization (EM) algorithm. Given the tth EM iteration estimate $\boldsymbol{\theta}^{(t)}$ for $\boldsymbol{\theta}$, we set

$$\boldsymbol{\theta}^{(t+1)}=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\sum_{i=1}^{n}\sum_{\mathbf{z}_i}\mathbb{P}(\mathbf{z}_i|\mathbf{x}_i;\boldsymbol{\theta}^{(t)},\xi^+,\xi^-)\ell(\boldsymbol{\theta};\mathbf{x}_i,\mathbf{z}_i), \tag{9}$$

where $\ell(\boldsymbol{\theta};\mathbf{x}_i,\mathbf{z}_i)=\log\mathbb{P}(\mathbf{x}_i,\mathbf{z}_i;\boldsymbol{\theta})$ is the joint log-likelihood of sample i. The update for $\boldsymbol{\theta}^{(t)}$ can be found explicitly as follows.

THEOREM 1 (Closed form EM Update). *Given n independent samples $\{\mathbf{x}_i\}_{i=1}^n$ from the same population defined by G and θ where $\mathbf{x}_i \in \{0,1\}^p$, the EM update* (**9**) *for $\theta_j^{(t+1)}$ has the following closed form:*

$$\theta_j^{(t+1)} = \frac{\sum_{i=1}^n \sum_{\mathbf{z}_i} \mathbb{1}(z_{ij}=1, \mathbf{z}_i(\mathcal{P}_j) \neq \mathbf{0}) \mathbb{P}(\mathbf{z}_i|\mathbf{x}_i)}{\sum_{i=1}^n \sum_{\mathbf{z}_i} \mathbb{1}(\mathbf{z}_i(\mathcal{P}_j) \neq \mathbf{0}) \mathbb{P}(\mathbf{z}_i|\mathbf{x}_i)}. \quad (\mathbf{10})$$

Note that if there is no measurement error, i.e., $\xi^+ = \xi^- = 0$, then (**10**) reduces to (**6**). In practice, computing (**10**) incurs exponential time because the inner sum goes through all 2^p possible realizations of the latent vector $\mathbf{z}_i$. Consequently, this model is only practically useful when $p \leq 12$.

Structure Learning

Given a fixed network G, we have shown that the parameters can be inferred in any of the models under consideration. Learning G presents some challenges. Since the number of possible DAGs is super-exponential in n an exhaustive search is infeasible for even modest values of n. In this section, we present a Genetic Algorithm (**GA**) to approximate the global maximum to the log likelihood function ℓ. The pseudocode of our ALgorithm for Oncogenesis Network Estimation, ALONE, is summarized in Algorithm 1.

Genetic Algorithm Genetic algorithms searches for a global optimum using a "survival of the fittest" strategy. We begin with a population of $2C$ candidate solutions known in the genetic algorithm literature as *chromosomes* and evolve them for T *generations*. Each chromosome is assigned a *fitness value* v which determines its quality. Then, S chromosome pairs are selected preferentially according to their fitness for reproduction. The next generation forms by performing a *crossover operation* on chromosome pairs. In each generation, there is a chance that a *mutation operation* changes each individual chromosome. Mutations help to maintain the genetic diversity of chromosomes, thus avoid local optima by exploring a broader range of potential solutions.

In the setting of our model, chromosomes at generation t are $2C$ DAGs, $\{G_i^t\}_{i=1}^{2C}$ and the fitness of each DAG is its maximum likelihood value. Algorithm 1 summarizes the GA for cancer progression inference. Note that we keep track of the best network (i.e., highest score value) in over all generations and return it as the output of the GA.

In the rest of this section, we first show how to represent DAGs for simplifying application of evolutionary operators. Then, we elaborate on crossover and mutation operations for the GA.

DAG Representation. We need to represent DAGs in such a way that genetic operators of the GA can be easily applied. The most natural way to encode a DAG G is by using its adjacency matrix $\mathbf{A}$. However, perturbing the entries in $\mathbf{A}$ may unintentionally introduce directed cycles into the resulting graph. To avoid this problem, we follow the approach used by Carvalho (16). Any DAG G admits a *topological ordering*, i.e., its vertices can be relabeled so that all edges point from a lower index to a higher index. The adjacency matrix for a topologically ordered DAG is thus *strictly upper triangular*. Consequently, G can be represented as a pair $(\mathbf{O}, \boldsymbol{\pi})$, where

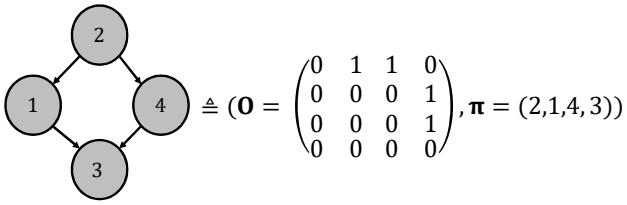


Figure 3. DAG representation for the Genetic Algorithm. The DAG (left) can be decomposed into an upper triangular matrix $\mathbf{O}$ along with a permutation $\boldsymbol{\pi}$. π_i gives the relabeling for node i.

$\mathbf{O}$ is the adjacency matrix for the topological ordering of G, and $\boldsymbol{\pi}$ is a permutation vector describing how the vertices of $\mathbf{O}$ should be relabeled to generate $\mathbf{A}$, (see Figure 3).We consider the ordering $\mathbf{O}$ and permutation $\boldsymbol{\pi}$ as separate chromosomes and evolve each of them individually. We can avoid introducing directed cycles by ensuring that our genetic operators always return an upper triangular matrix. **Crossover.** Each crossover operation is defined to take in two DAGs and produce two offspring so that the number of individuals per generation remains constant. For the two selected DAGs their orderings and permutations are crossed over as follows.

- *Ordering Crossover*: With probability c_o, the two upper triangular matrix chromosomes are recombined by interchanging their rows.
- *Permutation Crossover*: With probability c_π, the permutation chromosomes are recombined using the cycle crossover algorithm which is a standard crossover technique for permutations (35).

If no crossover occurs, the two selected chromosomes are passed down to the next generation unchanged.

Algorithm 1 ALONE: ALgorithm for Oncogenesis Network Estimation

input: Data set $\mathcal{D}$, parameters C, T, and $r \geq 0$.
output: Inferred graph $\hat{G}$
Generate population of random trees: $\mathcal{S}_0 = \{G_i^0\}_{i=1}^{2C}$.
for $t = 1$ to T **do**
  Compute fitness score of each DAG as: $v_i^t = \ell(G_i^t; \hat{\boldsymbol{\theta}}_{G_i^t}^{\text{MLE}}, \mathcal{D})$
  if $r = 0$ **then** ▷ MDL penalty
    $v_i^t = v_i^t + \log n \log p \sum_{j \in G_i^t} |\mathcal{P}_j|$
  end if
  $\mathbf{v}^t = \frac{(v_1^t, v_2^t, \ldots, v_{2C}^t)}{\sum_j^{2C} v_j^t}$ ▷ Selection probabilities
  for $i = 1$ to S **do**
    $(G_i^t, G_{i+1}^t) \leftarrow$ Selection$(\mathbf{v}^t, 2)$ ▷ Select DAGs
    $(G_i^{t+1}, G_{i+1}^{t+1}) \leftarrow$ Crossover(G_i^t, G_{i+1}^t)
    $G_i^{t+1} \leftarrow$ Mutate(G_i^{t+1}, r)
    $G_{i+1}^{t+1} \leftarrow$ Mutate(G_{i+1}^{t+1}, r)
    $G_i^{t+1} \leftarrow \Pi_\sim(G_i^{t+1})$; $G_{i+1}^{t+1} \leftarrow \Pi_\sim(G_{i+1}^{t+1})$
  end for
end for
Return the $\hat{G}$ corresponding to $v_{\max} = \max_{t \in [T], j \in [2C]} v_j^t$

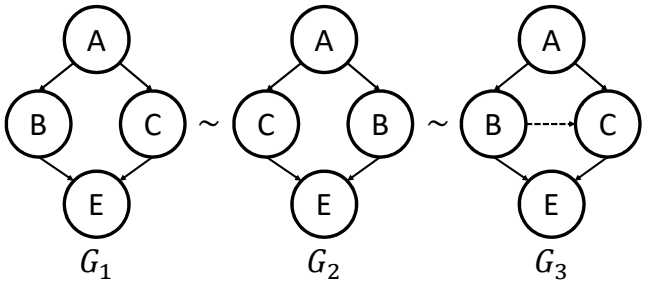


Figure 4. Examples of DAGs from the same equivalence class and their canonical form. For all $\boldsymbol{\theta}$ and $\mathbf{x}$, $\mathbb{P}(\mathbf{x};\boldsymbol{\theta})$ is the same for all of the three network structures shown above. B and C are similar vertices in G_1 and G_2. Edge $B \to C$ is redundant in G_3. By uniquely labeling similar vertices and removing redundant edges we reach G_1 as the canonical form of the other two DAGs.

Mutation. To maintain diversity in the population, we also define several mutation operators.

- *Edge Mutation*: With probability m_e, either an edge is added or an existing edge is removed which is achieved by randomly flipping elements on the upper triangle of $\mathbf{O}$.
- *Branch Mutation*: A branch is defined as a vertex along with all of its descendant vertices. With probability m_b, a branch is randomly selected and is moved to another location. This operation is done by cutting offs all parents of the root and assigning a random parent to it. Also, all parents of the other nodes in the branch that are not member of the branch themselves are cut off.
- *Permutation Mutation*: With probability m_π, two elements in the permutation chromosome $\boldsymbol{\pi}$ are swapped.

Speeding up the GA with DAG Equivalence Classes Since mutation i activates with probability θ_i *irrespective* of which parent mutations are active, many different network structures induce the same probability distribution over $\{0,1\}^p$. We say that $G \sim G'$ if, for every $\boldsymbol{\theta}$ and $\mathbf{x}$, $\mathbb{P}(\mathbf{x};G,\boldsymbol{\theta}) = \mathbb{P}(\mathbf{x};G',\boldsymbol{\theta})$. It is clear that $\sim$ defines an equivalence relation over DAGs.

To make the GA more efficient, we search only one DAG per equivalence class by defining a *canonical form* for each graph.

- An edge e in G is **redundant** if the graph G' obtained by removing e is equivalent to G.
- Vertices A and B are **similar** if swapping their labels yields an equivalent network.
- A DAG $G = (\mathbf{O}, \boldsymbol{\pi})$ is in **canonical form** if it contains no redundant edges and every set of similar vertices are ordered from least to greatest in $\boldsymbol{\pi}$.

Figure 4 shows a canonical form and corresponding DAGs with similar vertices and redundant edge. We show (Supporting Material C) that $G \sim G'$ if and only G and G' have the same canonical form and therefore, the canonical form represents the equivalence class. To ensure that canonical form is preserved through generations, we modify our crossover and mutation operators to only return DAGs in the desired form. The following propositions answers the practical questions of how to determine redundant edges and similar vertices.

PROPOSITION 4. *Under the Basic DBN (Figure 2a) and the measurement error (Figure 2c) models, an edge $A \to B$ is redundant if every path from the root (N) to A contains another parent of B.*

The definition of redundant edges in the spontaneous activation is more complicated and is explained in Section C of the Supporting Material.

PROPOSITION 5. *Vertices A and B are similar if they have the same set of parents and the same set of children.*

Algorithmically, we project back new solution graphs to the state space of canonical forms by removing redundant edges and uniquely labeling similar vertices in function $\Pi_\sim(\cdot)$ (line 12 of Algorithm 1.)

Controlling Complexity To prevent overfitting, we consider two types of penalty to control the complexity of the learned BN. First, if $r = 0$ in Algorithm 1, we perform regularized MLE by using the Minimum Description Length (**MDL**) penalty introduced in (30). MDL penalizes both the number of parameters of CPDs of a BN and the number of parents of each node. Since the number of parameters for all of the CPDs in DBN is one (single parameter θ_j for each node j), MDL penalty for DBNs simplifies to $\log n \log p \sum_{j=1}^{p} |\mathcal{P}_j|$ which penalizes the sum of the number of parents, i.e., number of edges in the BN.

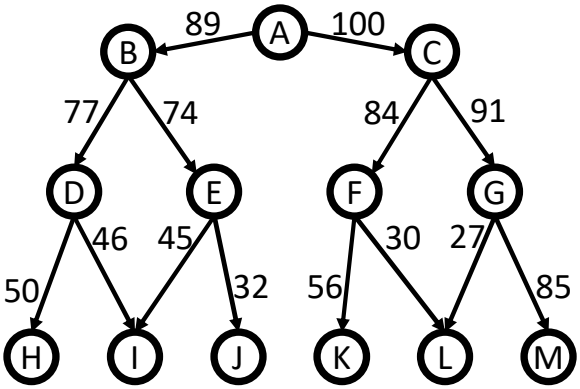
In another approach represented by $r > 0$ in Algorithm 1, we limit the number of parents of each node to a given constant r, i.e., $\max_j |\mathcal{P}_j| \leq r$. This hard penalty is induced first by initializing the first generation solutions as trees where the number of parents is one. Note that mutation is the only operation that can change the number of parents of nodes therefore r is passed to the Mutate function (line 11 of Algorithm 1) to cap the possible increase in the number of parents.

RESULTS

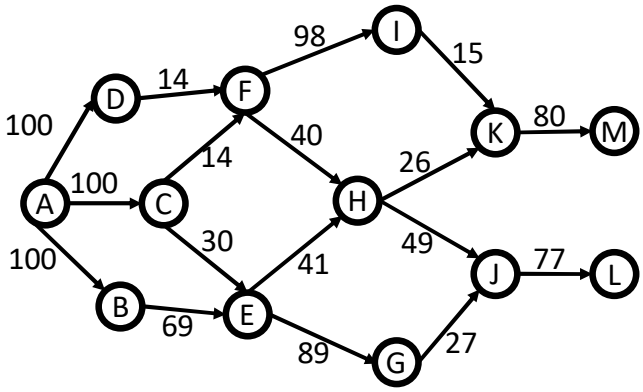
ALONE is implemented in R and the source code is available at https://github.com/phillipnicol/ALONE. The number of solutions per generation of the GA is set to $2C = 100$ and the evolution continues for $T = 300$ generations in all experiments. To quantify uncertainty in the estimated graph G, we report the "mean graph" resulted from running Algorithm 1 on 100 data sets. In the case of simulated data, since we have the underlying ground truth progression network, we obtain 100 data sets by sampling from the probability distribution represented by the DAG. For the real experiment, we obtain 100 bootstrap data sets by uniform sampling with replacement from the original given cancer data. Therefore, in all experiments, we have 100 data sets $\{\mathcal{D}_i\}_{i=1}^{100}$ and overlay the outcomes of Algorithm 1, i.e., $\{\hat{G}_i\}_{i=1}^{100}$ to obtain a weighted graph whose weights represent the "mean presence" of an edge, i.e., average number of times each edge was picked by ALONE.

Comparing DBN-based Models

The genetic algorithm outlined in the previous section can approximate the maximum likelihood estimated network



(a) sDAG: Simple DAG, $\boldsymbol{\theta} = (1,.7,.8,.7,.6,.5,.6,.3,.9,.2,.3,.6,.5)$.



(b) cDAG: Complex DAG, $\boldsymbol{\theta} = (1,.6,.7,.6,.5,.6,.4,.5,.7,.3,.2,.8,.5)$.

Figure 5. True DAGs of the synthetic experiment. $\boldsymbol{\theta}$ is sorted alphabetically and weights w_{ij} of edges are the percentage presence, which shows the number of times that an edge was present in the estimated networks from 100 simulated data sets. We observe that as the distance from the root increases, the recovered edges become less confident (e.g., $A \to C \to F \to K$ in the sDAG). Also, as the number of parents increases, the percentage presence of their edges to their shared child drops. For example, incoming edges to H in the cDAG have smaller weights compared to those for I and G, which are at the same depth as H. Finally, smaller progression probability θ_i of a node makes the inference of the incoming edges harder, which can be due to the fact that i is not frequently observed. For example, in the sDAG $\theta_J < \theta_K$ and $w_{EJ} = 32 < w_{FK} = 56$ and similarly in the cDAG $\theta_L < \theta_M$ and $w_{JL} = 77 < w_{KM} = 80$.

in both the measurement error model (Fig. 2c) and the spontaneous activation model (Fig. 2b). The spontaneous activation model is implemented in `R` while the measurement error model is partially coded in `C++` and is integrated into the `R` code using the `Rcpp` package. Our preliminary experiments show that the measurement error model has a significantly longer run-time as a consequence of its exponential time EM algorithm and cannot handle more than $p = 12$ alterations. In practice, the spontaneous activation model is both efficient and robust and therefore all of the reported results assume universal ε, i.e., $\forall i, \varepsilon_i = \varepsilon$.

Inferring Simulated Ground Truths

To test the performance of ALONE in recovering the ground truth, we applied it to the data generated from four synthetic networks with various levels of structural complexity. We considered a simple DAG (**sDAG**) and a complex DAG (**cDAG**) as ground truth progression networks. For each node we choose $\theta_i \in (0,1)$ uniformly at random once for all simulated data sets. For simplicity, we fix $\varepsilon_j = 0.05$ and avoid estimating it. We then generate $n = 500$ samples from the spontaneous activation model, Figure 2b. We run ALONE with three as the hard limit on number of parents, i.e., $\max_j |\mathcal{P}_j| \leq r = 3$.

Figure 5 shows our result for the sDAG and cDAG where the edge weights are percentage presence (mean presence of the edge times 100) computed from running the algorithm on simulated data sets as explained earlier. For both DAGs, there were only 3 false positive edges that appeared in more than 25 estimated graphs. Therefore, we chose .25 as the threshold for mean presence in the follow-up experiments with real data.

Finally, Figure 6 is the overlay of the trajectory of the log-likelihood of the best solution in each generation of the GA for 100 data sets during the cDAG inference. The black dots are the mean (over simulated data sets) of the fitness score of the best solutions in each generation and the gray shadow encompasses the interquartile range. Figure 6 shows that the log-likelihood of the best solution improves very quickly at first and stabilizes after 150 generations. Therefore, we set maximum number of generations $T = 300$ for the following analyses on real data.

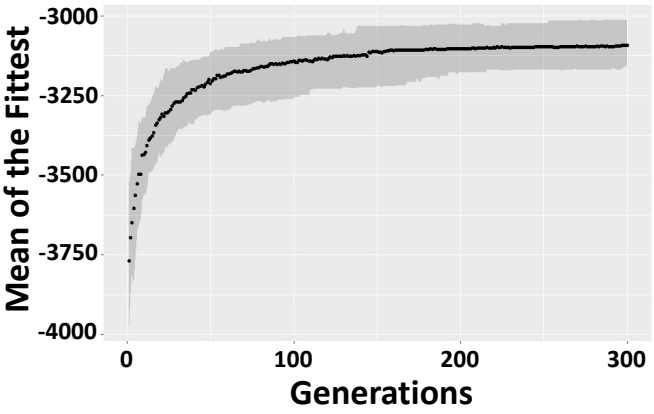


Figure 6. Genetic algorithm progression over generations. The graph is made of the overlay of 100 bootstrapped progression trajectories of the penalized log-likelihood of the best solution, $\text{argmax}_{i \in [S]} v_i^t$, in each generation t of the GA during the cDAG inference. The black dots are the mean over bootstrap trajectories and gray shadow represents the corresponding 90% confidence interval.

Inferring Progression of Copy Number Alterations in Colon Cancer

Our model can work with various types of aberrations and even hybrid data sets consisting of various types of aberrations. In this section, we investigate progression of copy number aberrations (**CNAs**) in colon cancer. We use publicly available CNA data detected by comparative genomic hybridization (CGH) from the Progenetix database (5, 12) which consists of $n = 570$ samples of gains

and losses of chromosomal arms. The data was pre-processed in a recent paper (46) and is available online at https://github.com/RudiSchill/MHN/tree/master/data.

We obtain 100 bootstrap estimates of the graph structure, and report our the mean graph in Figure 7. For clarity, we drop the mean presence of edges and represent them visually where thick, normal, and dashed edges have presence in ranges $[0.75,1]$, $[0.5,0.75)$, and $[0.25,.5)$ respectively.

Our inference based on the DBN assumption recovers four initiating CNA events for colon cancer, i.e., -1p, -18q, +20q, and +Xq. Some of the follow-up events are shared between these events such as -17p, +13q, -4q. Edge thickness of Figure 7 illustrates the mean presence of each edge in the output of 100 bootstrapped experiments and shows that we are most confident about the found roots and -18q being the parent event of -8p and -15q.

Inferring Progression of Mutations in Bladder Cancer

Finally, we use our method to recover the order of *driver* mutations in bladder cancer. We use the Bladder Cancer (BLCA) data from The Cancer Genome Atlas (TCGA) program (14). BLCA has the highest driver mutation rate per sample in the TCGA data set (3) and therefore is suitable to check the scalability of ALONE. Many methods have been developed to distinguish driver mutations from passengers. We use the result of a recent study where 26 computational methods have been applied to the TCGA data to detect driver mutations (3). As a result, we consider 45 mutations and $n=414$ samples. We remove mutations with less than 5% frequency in samples, which leaves us with $p=31$ nodes (30 driver mutations and the normal node N).

We run ALONE with $2S=100$ solutions per generation for $T=300$ generations on 100 bootstrap data sets. The mean progression network is illustrated in Figure 8, where the edge thickness represents the mean presence of the edge according to the same rule described for the colon cancer results. Note that out of $p=31$ nodes, only 18 are inferred in the mean progression network because the remaining 13 are not connected with enough confident to the rest.

We recover three root mutations with a high mean presence for the progression of bladder cancer, i.e., *TP53*, *KDM6A*, and *KMT2D*. From the several children of these roots, three have a mean presence greater than 50%: *RB1*, *STAG2*, and *KMT2C*. Finally, roots with meager mean presence (*ELF3*, *ATM*, and *CREBBP*) and childless *PIK3CA* are mutations for which ALONE can not find enough supporting evidence to place them in the main progression graph. Note that these placements are possible because of the flexibility given to our model based on the spontaneous activation assumption.

DISCUSSION

Simulation Study

Investigating the synthetic experiment helps us understand strengths and weaknesses of ALONE. As we move down the sDAG, Figure 5a, reconstruction becomes more difficult. On the other hand, recovering incoming edges of nodes with more parents are more challenging, e.g., parents of I and L. The cDAG reconstruction result agrees with the hypothesis that having fewer parents simplifies the recovery. For example, for the leaf nodes at the end of the progression network, M and L, reconstruction is highly probable but inferring parents of nodes closer to the root with multiple parents, e.g., F, and K, is harder. Finally, it appears that a smaller progression probability θ_i makes recovery of edges to node i more difficult. This observation could be due to the fact that smaller θ_i results in lower frequency of alteration i in the samples, which makes inferring its parents harder in general.

Colon Cancer

Examining the reconstructed CNA progression network reveals that it is in line with known biological facts. First of all, 20q amplification, which is placed as one of the initiating roots of progression, is known to happen early in many cancers including colon and is suggested to causally drive tumorigenesis (51). Chromosome arm 20q harbors multiple potential oncogenes such as *AURKA* and *SRC* and its gain or amplification have been linked to longer overall survival (40).

Another recovered root is the deletion of 18q which is also known to have a central role in colon cancer and has been observed in 70% of colorectal tumors (39). Among genes on 18q, *DCC*, *SMAD2*, *SMAD4*, and *CABLES* are thought to have a driving role in colon cancer (29, 39). Although -18q has been observed in advanced stages of colorectal tumors where mutations such as *APC* and *KRAS* have already occurred, it seems to be one of the first driving CNAs that happens in colon cancer (22). We should note that the order in which chromosomal instability (such as copy number changes) and mutations occur in colon cancer is not clear (33, 39) and since we are not analyzing such hybrid data our findings are neutral in that regard.

Interestingly, ALONE places the -1p aberration as one of the initiating events. This is in contrast with CBN which places -1p event in the third (last) level of its inferred DAG (23). Loss of 1p has been associated with many colon carcinogenesis pathways and is one of the hot spot defects in the non-neoplastic mucosa associated with the possible initiation of colon carcinogenesis (36).

Finally, there have been arguments for the presence of two mutually exclusive pathways dominated by gains and losses in colon cancer (26) which is confirmed by our results. In the reconstructed progression network of Figure 7, with the exception of +13q and -4q aberrations in the second layer and +8q in the third layer (which all have both gain and loss parents) losses and gains appear to progress exclusively in the colon cancer.

Bladder Cancer

The recovered progression network for bladder cancer reflects existing biological research. First, bladder cancer is known to have two histologically different subtypes known as papillary and non-papillary (27). Papillary tumors are finger-like, which start in the lining and grow toward the center of the bladder. Non-papillary tumors also initiate in the lining but are flat in shape. Both types can be muscle-invasive, which means the tumor has grown outward, escaped the lining, and infiltrated bladder muscles, or non-muscle invasive (27). All of the bladder cases in TCGA are muscle-invasive, but papillary and non-papillary cases are not known.

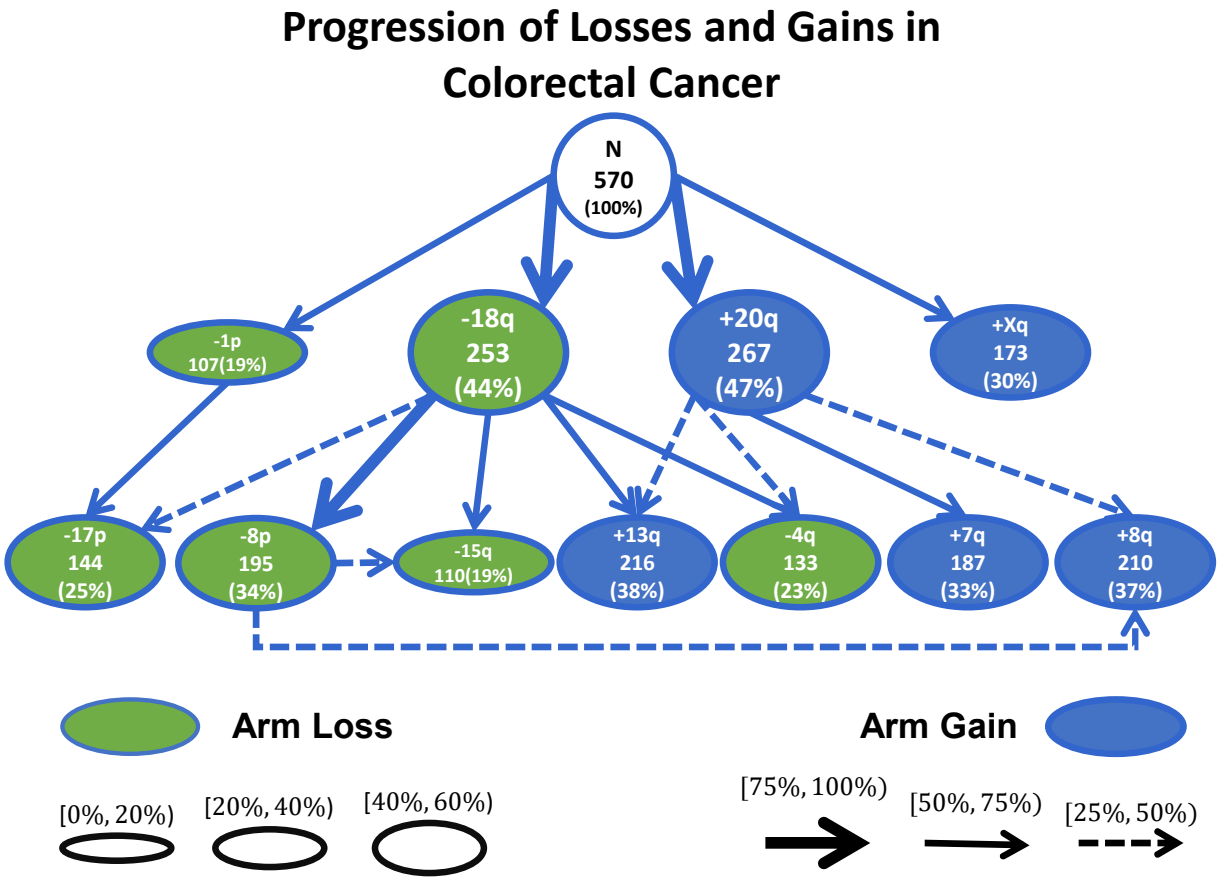


Figure 7. CNA progression network of colon cancer inferred by ALONE. The frequency of each aberration and its percentage are provided inside each node. Also, the size of each node is approximately proportional to its frequency. The mean presence of thick, normal, and dashed edges are in ranges $[0.75, 1]$, $[0.5, 0.75)$, and $[0.25, .5)$ respectively for 100 bootstrapped network estimation. Events -18q and +20q are known important initiating and driving alterations in colorectal cancer, and -1p has been argued to play a role in colon carcinogenesis. Besides, losses and gains are reported to occur mutually exclusively in colon cancer. All of these facts have been captured by ALONE.

There are known molecular signature for papillary and non-papillary bladder cancers. Mutations in *TP53*, *RB1*, and *KMT2D* (green nodes in Figure 8) are very frequent in non-papillary subtype while *KDM6A*, *STAG2*, and *FGFR3* (blue nodes in Figure 8) are hallmarks of papillary tumors (13, 21, 25, 47). Focusing on the high confident recovered roots (*TP53*, *KDM6A*, and *KMT2D*) and their descendants, our inferred progression network of Figure 8 shows separate progression paths for papillary and non-papillary subtypes. The middle sub-graph rooted at *KDM6A* contains *KDM6A*, *STAG2*, and *FGFR3* mutations and is mostly separated from the rest of the network. Therefore we can match it to the progression of the papillary subtype. Sub-graphs on the right and left of the figure (rooted at *TP53* and *KMT2D*) are enriched with molecular hallmarks of non-papillary subtype. Our result shows the ability of ALONE to infer the cancer progression network while maintaining subtype-specific biology. This unsupervised learning of subtype information is in sharp contrast with the existing state-of-the-art methods that need subtype information as input and infer progression for each subtype separately (15, 19, 41).

Another common pre-processing step in cancer progression inference is to detect sets of fitness-equivalent groups of mutually exclusive events by statistical tests or biological priors such as pathways information (15) and then use that knowledge in the inference engine. It has been reported that in bladder cancer (*KDM6A*, *KMT2D*) and (*TP53*, *CDKN2A*) are mutually exclusive pairs (28, 45). Figure 8 shows that without manually detecting mutual exclusiveness relations and taking them as the input for progression inference, ALONE automatically places the mutually exclusive genes in separate branches of the inferred progression network.

In addition to detecting mutual exclusivity in data, one can expect such patterns in biological pathway information as well. We know that usually single perturbation of a pathway is enough for the manifestation of a cancer hallmark, and therefore another mutated gene in the same pathway does not confer a selective advantage. Thus, patterns of mutual exclusivity of cancer events arise among genes in the same pathways. In bladder cancer, high rate of alteration of p53/Rb, RTK/Ras/PI(3)K, and histone modification pathways are observed (13). Figure 8 highlights the corresponding pathways of genes with different outline color for each pathway. It confirms that the two subtypes (papillary and non-papillary) both have perturbation in p53, RTK/Ras/PI(3)K, methylation, and acetylation pathways. The only mutation that is shared between the two subtypes is *EP300*, which corresponds to acetylation.

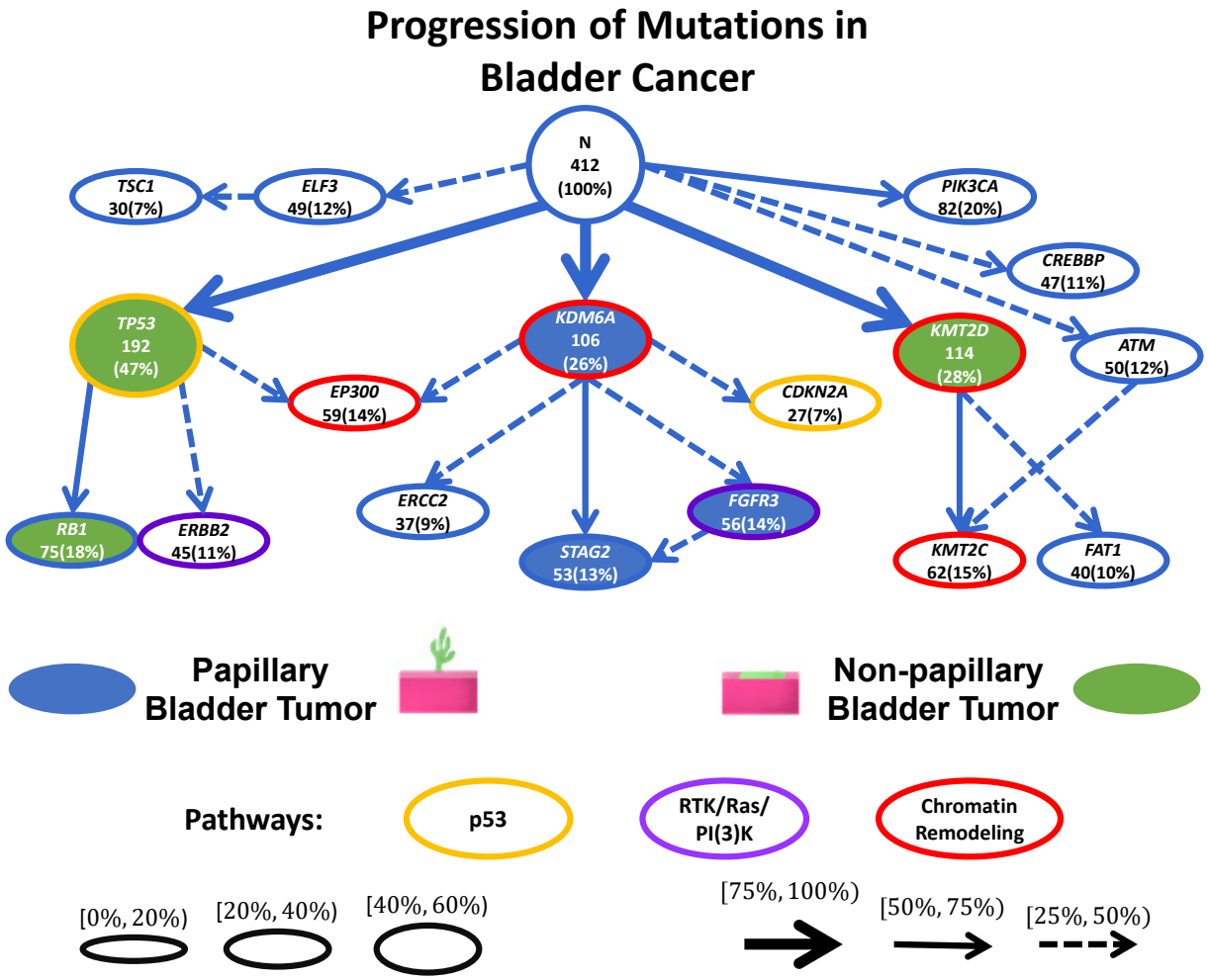


Figure 8. Mutation progression network of bladder cancer inferred by ALONE. Focusing on the three high confident roots (*TP53*, *KDM6A*, and *KMT2D*), the two subtypes of bladder cancer are clearly separated. The middle subgraph (rooted in *KDM6A*) is enriched for hallmark aberrations of the papillary subtype (blue nodes), and the other two subgraphs correspond to flat tumors (green nodes). Known mutual exclusive alteration pairs such as (*KDM6A*, *KMT2D*) and (*TP53*, *CDKN2A*) are occurring in different subgraphs. Four established highly perturbed pathways of bladder cancer are represented with varying outline colors. Each subtype has at least one mutated gene from these pathways is its subgraphs, therefore in both subtypes, all of the four pathways are perturbed.

ACKNOWLEDGEMENTS

Authors would like to thank the Mathematical Biosciences Institute (MBI) at Ohio State University, for partially supporting this research through National Science Foundation grants DMS 1440386 and DMS 1757423.

REFERENCES

1. Philipp M Altrock, Lin L Liu, and Franziska Michor. The mathematics of cancer: integrating quantitative models. *Nature reviews. Cancer*, 15(12):730–745, December 2015.
2. Camille Stephan-Otto Attolini, Yu-Kang Cheng, Rameen Beroukhim, Gad Getz, Omar Abdel-Wahab, Ross L Levine, Ingo K Mellinghoff, and Franziska Michor. A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41):17604–17609, October 2010.
3. Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, Patrick Kwok-Shing Ng, Kang Jin Jeong, Song Cao, Zixing Wang, Jianjiong Gao, Qingsong Gao, Fang Wang, Eric Minwei Liu, Loris Mularoni, Carlota Rubio-Perez, Niranjan Nagarajan, Isidro Cortés-Ciriano, Daniel Cui Zhou, Wen-Wei Liang, Julian M Hess, Venkata D Yellapantula, David Tamborero, Abel Gonzalez-Perez, Chayaporn Suphavilai, Jia Yu Ko, Ekta Khurana, Peter J Park, Eliezer M Van Allen, Han Liang, MC3 Working Group, Cancer Genome Atlas Research Network, Michael S Lawrence, Adam Godzik, Nuria Lopez-Bigas, Josh Stuart, David Wheeler, Gad Getz, Ken Chen, Alexander J Lazar, Gordon B Mills, Rachel Karchin, and Li Ding. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385.e18, April 2018.
4. David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.
5. M Baudis and M L Cleary. Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics*, 17(12):1228–1229, December 2001.
6. N Beerenwinkel and S Sullivant. Markov models for accumulating mutations. *Biometrika*, 96(3):645–661, September 2009.
7. Niko Beerenwinkel, Tibor Antal, David Dingli, Arne Traulsen, Kenneth W Kinzler, Victor E Velculescu, Bert Vogelstein, and Martin A Nowak. Genetic progression and the waiting time to cancer. *PLoS computational biology*, 3(11):e225, November 2007.
8. Niko Beerenwinkel, Nicholas Eriksson, and Bernd Sturmfels. Conjunctive bayesian networks. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 13(4):893–909, November 2007.

9. Niko Beerenwinkel, Jörg Rahnenführer, Martin Däumer, Daniel Hoffmann, Rolf Kaiser, Joachim Selbig, and Thomas Lengauer. Learning multiple evolutionary pathways from cross-sectional data. *Journal of computational biology: a journal of computational molecular cell biology*, 12(6):584–598, July 2005.
10. Niko Beerenwinkel, Jörg Rahnenführer, Rolf Kaiser, Daniel Hoffmann, Joachim Selbig, and Thomas Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–2107, May 2005.
11. Niko Beerenwinkel, Roland F Schwarz, Moritz Gerstung, and Florian Markowetz. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–25, January 2015.
12. Haoyang Cai, Nitin Kumar, Ni Ai, Saumya Gupta, Prisni Rath, and Michael Baudis. Progenetix: 12 years of oncogenomic data curation. *Nucleic acids research*, 42(Database issue):D1055–62, January 2014.
13. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315–322, March 2014.
14. Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nature genetics*, 45(10):1113–1120, October 2013.
15. Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Rebeca Sanz-Pamplona, Luca De Sano, Giancarlo Mauri, Victor Moreno, Marco Antoniotti, and Bud Mishra. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences of the United States of America*, 113(28):E4025–34, July 2016.
16. Arthur Carvalho. A cooperative coevolutionary genetic algorithm for learning bayesian network structures. May 2013.
17. Yu-Kang Cheng, Rameen Beroukhim, Ross L Levine, Ingo K Mellinghoff, Eric C Holland, and Franziska Michor. A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. *PLoS computational biology*, 8(1):e1002337, January 2012.
18. Simona Cristea, Jack Kuipers, and Niko Beerenwinkel. pathTiMEx: Joint inference of mutually exclusive cancer pathways and their progression dynamics. *Journal of computational biology: a journal of computational molecular cell biology*, 24(6):603–615, June 2017.
19. Luca De Sano, Giulio Caravagna, Daniele Ramazzotti, Alex Graudenzi, Giancarlo Mauri, Bud Mishra, and Marco Antoniotti. TRONCO: an R package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics*, 32(12):1911–1913, June 2016.
20. 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: a journal of computational molecular cell biology*, 6(1):37–51, 1999.
21. Colin P N Dinney, David J McConkey, Randall E Millikan, Xifeng Wu, Menashe Bar-Eli, Liana Adam, Ashish M Kamat, Arlene O Siefker-Radtke, Tomasz Tuziak, Anita L Sabichi, H Barton Grossman, William F Benedict, and Bogdan Czerniak. Focus on bladder cancer. *Cancer cell*, 6(2):111–116, August 2004.
22. E R Fearon and B Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767, June 1990.
23. Moritz Gerstung, Michael Baudis, Holger Moch, and Niko Beerenwinkel. Quantifying cancer progression with conjunctive bayesian networks. *Bioinformatics*, 25(21):2809–2815, November 2009.
24. Moritz Gerstung, Nicholas Eriksson, Jimmy Lin, Bert Vogelstein, and Niko Beerenwinkel. The temporal order of genetic and pathway alterations in tumorigenesis. *PloS one*, 6(11):e27136, November 2011.
25. Yaoting Gui, Guangwu Guo, Yi Huang, Xueda Hu, Aifa Tang, Shengjie Gao, Renhua Wu, Chao Chen, Xianxin Li, Liang Zhou, Minghui He, Zesong Li, Xiaojuan Sun, Wenlong Jia, Jinnong Chen, Shangming Yang, Fangjian Zhou, Xiaokun Zhao, Shengqing Wan, Rui Ye, Chaozhao Liang, Zhisheng Liu, Peide Huang, Chunxiao Liu, Hui Jiang, Yong Wang, Hancheng Zheng, Liang Sun, Xingwang Liu, Zhimao Jiang, Dafei Feng, Jing Chen, Song Wu, Jing Zou, Zhongfu Zhang, Ruilin Yang, Jun Zhao, Congjie Xu, Weihua Yin, Zhichen Guan, Jiongxian Ye, Hong Zhang, Jingxiang Li, Karsten Kristiansen, Michael L Nickerson, Dan Theodorescu, Yingrui Li, Xiuqing Zhang, Songgang Li, Jian Wang, Huanming Yang, Jun Wang, and Zhiming Cai. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nature genetics*, 43(9):875–878, August 2011.
26. Mattias Höglund, David Gisselsson, Gunnar B Hansen, Torbjörn Säll, Felix Mitelman, and Mef Nilbert. Dissecting karyotypic patterns in colorectal tumors: two distinct but overlapping pathways in the adenoma-carcinoma transition. *Cancer research*, 62(20):5939–5946, October 2002.
27. Ashish M Kamat, Noah M Hahn, Jason A Efstathiou, Seth P Lerner, Per-Uno Malmström, Woonyoung Choi, Charles C Guo, Yair Lotan, and Wassim Kassouf. Bladder cancer. *The Lancet*, 388(10061):2796–2810, December 2016.
28. Jaegil Kim, Rehan Akbani, Chad J Creighton, Seth P Lerner, John N Weinstein, Gad Getz, and David J Kwiatkowski. Invasive bladder cancer: Genomic insights and therapeutic promise. *Clinical cancer research: an official journal of the American Association for Cancer Research*, 21(20):4514–4524, October 2015.
29. Sandra D Kirley, Massimo D'Apuzzo, Gregory Y Lauwers, Fiona Graeme-Cook, Daniel C Chung, and Lawrence R Zukerberg. The cables gene on chromosome 18Q regulates colon cancer progression in vivo. *Cancer biology & therapy*, 4(8):861–863, August 2005.
30. Wai Lam and Fahiem Bacchus. LEARNING BAYESIAN BELIEF NETWORKS: AN APPROACH BASED ON THE MDL PRINCIPLE. *Computational Intelligence. An International Journal*, 10(3):269–293, August 1994.
31. Mark D M Leiserson, Hsin-Ta Wu, Fabio Vandin, and Benjamin J Raphael. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome biology*, 16:160, August 2015.
32. Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *Annals of statistics*, 37(6A):3133–3164, December 2009.
33. Franziska Michor, Yoh Iwasa, Christoph Lengauer, and Martin A Nowak. Dynamics of colorectal cancer. *Seminars in cancer biology*, 15(6):484–493, December 2005.
34. Loes Olde Loohuis, Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. Inferring tree causal models of cancer progression with probability raising. *PloS one*, 9(10):e108358, October 2014.
35. I M Oliver, D J Smith, and J R C Holland. A study of permutation crossover operators on the traveling salesman problem. In *Proceedings of the Second International Conference on Genetic Algorithms on Genetic Algorithms and Their Application*, pages 224–230, Hillsdale, NJ, USA, 1987. L. Erlbaum Associates Inc.
36. Claire M Payne, Cheray Crowley-Skillicorn, Carol Bernstein, Hana Holubec, and Harris Bernstein. Molecular and cellular pathways associated with chromosome 1p deletions during colon carcinogenesis. *Clinical and experimental gastroenterology*, 4:75–119, May 2011.
37. Judea Pearl. Bayesianism and causality, or, why I am only a Half-Bayesian. In David Corfield and Jon Williamson, editors, *Foundations of Bayesianism*, pages 19–36. Springer Netherlands, Dordrecht, 2001.
38. Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.
39. Maria S Pino and Daniel C Chung. The chromosomal instability pathway in colon cancer. *Gastroenterology*, 138(6):2059–2072, June 2010.
40. Ryan N Ptashkin, Carlos Pagan, Rona Yaeger, Sumit Middha, Jinru Shia, Kevin P O'Rourke, Michael F Berger, Lu Wang, Robert Cimera, Jiajing Wang, David S Klimstra, Leonard Saltz, Marc Ladanyi, Ahmet Zehir, and Jaclyn F Hechtman. Chromosome 20q amplification defines a subtype of microsatellite stable, Left-Sided colon cancers with wild-type RAS/RAF and better overall survival. *Molecular cancer research: MCR*, 15(6):708–713, June 2017.
41. Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, Alex Graudenzi, Ilya Korsunsky, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026, September 2015.
42. Daniele Ramazzotti, Alex Graudenzi, Giulio Caravagna, and Marco Antoniotti. Modeling cumulative biological phenomena with Suppes-Bayes causal networks. *Evolutionary bioinformatics online*, 14:1176934318785167, July 2018.
43. Benjamin J Raphael and Fabio Vandin. Simultaneous inference of cancer pathways and tumor progression from cross-sectional mutation data. *Journal of computational biology: a journal of computational molecular cell biology*, 22(6):510–527, June 2015.

44. Johannes G Reiter, Marina Baretti, Jeffrey M Gerold, Alvin P Makohon-Moore, Adil Daud, Christine A Iacobuzio-Donahue, Nilofer S Azad, Kenneth W Kinzler, Martin A Nowak, and Bert Vogelstein. An analysis of genetic heterogeneity in untreated cancers. *Nature reviews. Cancer*, August 2019.
45. A Gordon Robertson, Jaegil Kim, Hikmat Al-Ahmadie, Joaquim Bellmunt, Guangwu Guo, Andrew D Cherniack, Toshinori Hinoue, Peter W Laird, Katherine A Hoadley, Rehan Akbani, Mauro A A Castro, Ewan A Gibb, Rupa S Kanchi, Dmitry A Gordenin, Sachet A Shukla, Francisco Sanchez-Vega, Donna E Hansel, Bogdan A Czerniak, Victor E Reuter, Xiaoping Su, Benilton de Sa Carvalho, Vinicius S Chagas, Karen L Mungall, Sara Sadeghi, Chandra Sekhar Pedamallu, Yiling Lu, Leszek J Klimczak, Jiexin Zhang, Caleb Choo, Akinyemi I Ojesina, Susan Bullman, Kristen M Leraas, Tara M Lichtenberg, Catherine J Wu, Nicholaus Schultz, Gad Getz, Matthew Meyerson, Gordon B Mills, David J McConkey, TCGA Research Network, John N Weinstein, David J Kwiatkowski, and Seth P Lerner. Comprehensive molecular characterization of Muscle-Invasive bladder cancer. *Cell*, 171(3):540–556.e25, October 2017.
46. Rudolf Schill, Stefan Solbrig, Tilo Wettig, and Rainer Spang. Modelling cancer progression using mutual hazard networks. *Bioinformatics*, June 2019.
47. David A Solomon, Jung-Sik Kim, Jolanta Bondaruk, Shahrokh F Shariat, Zeng-Feng Wang, Abdel G Elkahloun, Tomoko Ozawa, Julia Gerard, Dazhong Zhuang, Shizhen Zhang, Neema Navai, Arlene Siefker-Radtke, Joanna J Phillips, Brian D Robinson, Mark A Rubin, Björn Volkmer, Richard Hautmann, Rainer Küfer, Pancras C W Hogendoorn, George Netto, Dan Theodorescu, C David James, Bogdan Czerniak, Markku Miettinen, and Todd Waldman. Frequent truncating mutations of STAG2 in bladder cancer. *Nature genetics*, 45(12):1428–1430, December 2013.
48. Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.
49. Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 499–506, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
50. Aniko Szabo and Kenneth Boucher. Estimating an oncogenetic tree when false negatives and positives are present. *Mathematical biosciences*, 176(2):219–236, April 2002.
51. Yuval Tabach, Ira Kogan-Sakin, Yosef Buganim, Hilla Solomon, Naomi Goldfinger, Randi Hovland, Xi-Song Ke, Anne M Oyan, Karl-H Kalland, Varda Rotter, and Eytan Domany. Amplification of the 20q chromosomal arm occurs early in tumorigenic transformation and may initiate cancer. *PloS one*, 6(1):e14632, January 2011.