Tumor-specific Causal Inference (TCI): A Bayesian Method for Identifying Causative Genome Alterations within Individual Tumors

Gregory Cooper, Chunhui Cai, Xinghua Lu

Department of Biomedical Informatics, School of Medicine, University of Pittsburgh

Abstract

Precision medicine for cancer involves identifying and targeting the somatic genome alterations (SGAs) that drive the development of an individual tumor. Much of current efforts at finding driver SGAs have involved identifying the genes that are mutated more frequently than expected among a collection of tumors. When these population-derived driver genes are altered (perhaps in particular ways) in a given tumor, they are posited as driver genes for that tumor. In this technical report, we introduce an alternative approach for identifying causative SGAs, also known as "drivers", by inferring causal relationships between SGAs and molecular phenotypes at the individual tumor level. Our tumor-specific causal inference (TCI) algorithm uses a Bayesian method to identify the SGAs in a given tumor that have a high probability of regulating transcriptomic changes observed in that specific tumor. Thus, the method is focused on identifying the tumor specific SGAs that are causing expression changes that are specific to the tumor. Those SGAs that have a high probability of regulating transcriptomic changes related to oncogenic processes are then designated to be the putative drivers of the tumor. In this paper, we describe in detail the TCI algorithm and its implementation.

1. Introduction

Cancer is mainly caused by SGAs, such as somatic mutations (SMs), somatic copy number alterations (SCNAs), chromosome rearrangement and other genomic alterations. A tumor cell commonly hosts hundreds to over a thousand SGAs, among which only a small minority contribute to tumor development by perturbing cellular signaling pathways while most others are passenger SGAs (unrelated to cancers). A foremost task of precision oncology for cancer treatment is to identify and target the driver SGAs of an individual tumor. Current methods of identifying candidate driver SGAs are mostly based on the assumption that, if a gene is mutated at a frequency significantly above the expected rate in a cohort of tumors, the mutation events of the gene are likely positively selected in tumors due to resultant oncogenic advantages. Therefore, such a gene is more likely a cancer driver gene [1-4]. Hereafter, we refer to this family of methods as *frequency-oriented models*. These models do not attempt to explicitly determine the functional role of a driver in cancer development, that is, they cannot provide insight into functional impact of oncogenic processes caused by a driver SGA. In general, frequency-oriented models are constrained by the need to define the baseline mutation rate, and different models for estimating the baseline rate will lead to different results.

It is well accepted that driver genes can contribute to cancer development through various types of genomic alterations, such as chromosome structure variations, non-coding mutations, and epigenetic modifications [3, 5-7]. For example, copy number amplification and promoter mutations of the telomere reverse transcriptase (*TERT*) play important roles in different cancer types [8, 9]. However, to our knowledge, there is no reported principled method to integrate multiple types of SGA events to determine the significance of the corresponding gene in cancer development, nor there is any theoretical method that can systematically infer the functional impact of driver SGAs perturbing a common gene.

Here, we introduce a novel framework that identifies driver SGAs in a tumor-specific and signal-oriented fashion. Our approach is based on the assumption that driver SGAs cause cancer progression by perturbing signaling pathways, and as such their functional impact is reflected by the cellular or molecular phenotypes regulated by these perturbed pathways. Thus, the task is to find the SGAs that causally regulate cancer-related molecular phenotypes, e.g., differential expression of genes involved in oncogenic processes, for each individual tumor. To this end, we designed a tumor-specific causal inference (TCI) algorithm that infers causal relationships between SGAs and differentially expressed genes (DEGs) within a specific tumor.

The Bayesian causal inference framework developed in this study provides a principled approach to not only incorporate biological prior knowledge and theoretical assumptions but also integrate diverse types of genomic and molecular phenotypic data to infer the functional impact of genomic alterations in individual tumors [10, 11]. In these respects, TCI first calculates the prior probability that an SGA is a driver in the tumor of interest. Based on the positive selection assumption underlying the frequency-based methods, we assume that the more often are the SGA events perturbing the corresponding gene in a tumor cohort, the more likely the gene is a driver in the current tumor. As such, the calculation of the prior incorporates the strength of the frequency-oriented methods [1, 3]. In a signal-oriented fashion, TCI further calculates the marginal likelihood that the molecular phenotype change is caused by the SGA. Finally, TCI derives a posterior probability that the SGA is causally responsible for the observed phenotypic change in a tumor. Thus, TCI unifies the frequency-oriented and signal-oriented approaches to determine the functional impact of an SGA event within a specific tumor.

Previously reported approaches, e.g. eQTL, can infer the association between SGAs and gene expression levels across a population of tumors [12, 13]. To our knowledge, however, no previously published method is capable of inferring the causal relationships between SGAs and differentially expressed genes (DEGs) in a tumor-specific manner. In this paper, we introduce such an approach which is tumor specific in two ways. First, for a given DEG *E* in a given tumor *t*, the only SGAs that can possibly cause (drive) *E* are the SGAs in *t*; SGAs that occur in other tumors, but not in *t*, are not candidate drivers of *E* in *t*. Thus, the search space for candidate drivers is focused in a tumor-specific manner. Second, the scoring of a given SGA in *t* being a driver of *E* is scored probabilistically in manner that is tumor-specific, as we explain in Section 2.2.

The change in going from population-based to tumor-specific causal inference is substantial. Since multiple SGAs can perturb a common signaling pathway, we should consider the causal relationships between SGAs perturbing the pathway and a DEG regulated by the pathway as a multiple-to-one relationship. For example, multiple perturbations of the PI3K pathway are known to regulate its downstream gene expression during tumorigenesis [14]. Interestingly, rarely do SGAs perturb a common pathway in an individual tumor, which is a phenomenon referred to as mutual exclusivity [15-19]. Figure 1 illustrates the mutual exclusive tendency among 3 members of the PI3K pathway, *PIK3CA*, *PTEN*, and *PIK3R1* [14]. At the tumor population level, any SGA perturbing the PI3K pathway can cause the expression change of a downstream gene, while in an individual tumor, it is more likely that only one SGA causes the expression change. The multiple-to-one relationship and the mutual exclusivity of the SGAs significantly dampens the strength of statistical association between an individual SGA and the DEG of interest when viewed across all tumors (perhaps of a given type). Therefore, a conventional EQTL analysis may fail to find the causal relationship between a low frequency SGA (e.g., *PIK3R1*) that is perturbing a downstream gene via the PI3K pathway, because it cannot adequately account for the gene expression variance caused by other common SGAs that are perturbing the gene in other tumors (e.g., mutation/amplification of *PIK3CA* and mutation/deletion of *PTEN*). Although at the population level it may not be significantly (statistically) associated with a DEG *E* that is downstream of the PI3K pathway, an alteration of *PIK3R1* should nonetheless be the most probable cause of *E* in an individual tumor when both *PIK3CA* and *PTEN* are normal. The TCI algorithm takes advantage of such tumor-specific relationships between SGAs and DEGs in order to locate the SGAs that are most likely driving the DEGs.

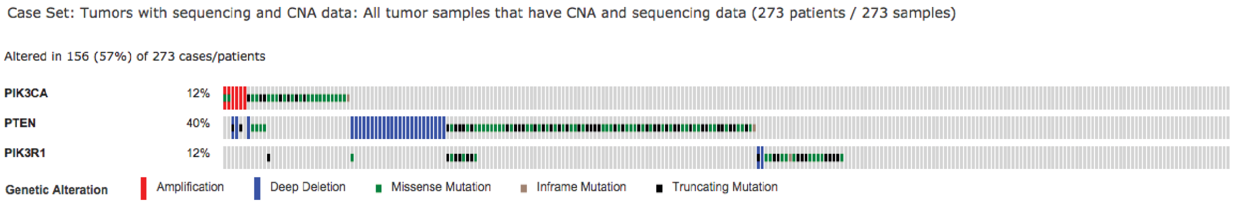

Figure 1. **SGAs mutual exclusivity among *PIK3CA, PTEN,* and *PIK3R1* in the PI3K pathway**. An example of mutual exclusivity among *PIK3CA, PTEN,* and *PIK3R1* affecting the PI3K pathway in 273 Glioblastoma Multiforme tumor samples. Each column represents an individual tumor. The combination of color and bar size denotes genetic alteration types: a long red bar represents a tumor with copy number amplification; a long blue bar represents a tumor with copy number deletion; a short green bar represents a tumor with missense mutation; a short brown bar represents a tumor with inframe mutation; a short black bar represents a tumor with truncating mutation. These three SGAs are altered in 156 (57%) out of 273 brain tumors, and as shown, only one of those SGAs occurs in most of the tumors.

2. Model Specification

Let the genotypes of all genes in a tumor be represented by a set of binary variables, such that the state of a gene is set to 1 if the gene is altered (e.g., mutated), or otherwise it is set to 0; similarly, let the expression states of all genes be represented by a set of binary variables, such that the expression state of a gene is set to 1 if it is differentially expressed, or otherwise it is set to 0. Let **TS** = {T_1, T_2, …, T_t, …, T_N} denote the tumor set which contains a total *N* tumor samples, where *t* indexes over the tumors included in the tumor set. Let **SGA**$_t$ = {A_1, A_2, …, A_h, …, A_m} denote a subset of *m* genes that are altered at the genome level in a tumor *t*, i.e., their genomic states are set to 1, where *h* indexes over the variables in **SGA**$_t$; let **DEG**$_t$ = {E_1, E_2, …, E_i, …, E_n} denote *n* genes that are differentially expressed in the tumor *t*, where *i* indexes over the variables in **DEG**$_t$. Hereafter, we will use **SGA** instead of **SGA**$_t$ and **DEG** instead of **DEG**$_t$ for simplicity of notation. For each tumor, we further include a variable A_0, to collectively represent non-specific factors other than SGAs (e.g., tumor microenvironment) that may affect the gene expression in a tumor. Based on the assumptions that each DEG is likely to be regulated by one aberrant signaling pathway and such a pathway is likely perturbed by only one SGA observed in the tumor (mutual exclusivity), the TCI model further constrains each DEG to be causally regulated by only one SGA (or by A_0) in a given tumor. The TCI model assumes no hidden common causes among the variables in **SGA**∪**DEG**, including the presence of mixture distributions. It is not concerned with modeling the causal relationships among the variables within **DEG** or among the variables within **SGA**. With the above settings, the task is to determine for each variable A_h in **SGA** the probability that it is the cause of one or more variables in **DEG**, which we interpret as the probability that A_h is a driver in tumor t.

For a given tumor, we represent the causal relationships between the variables in **SGA** and those in **DEG** using a bipartite causal Bayesian network (CBN) in which the variables in **SGA** are at level 1 and the variables in **DEG** are at level 2. In such a CBN, arcs always point from **SGA** to **DEG**. A permissible CBN model M has only one arc coming into each variable E_i in **DEG** from one variable A_h in **SGA** or A_0 which means that it is abnormal due to some non-SGA influence. In model M, a given A_h can have zero arcs (a passenger SGA) or one or more arcs emanating from it to the variables in **DEG**; thus, an SGA can causally regulate multiple DEGs. Figure 2 shows an example of a permissible model. Since each tumor generally has a unique **SGA** set and a unique **DEG** set, the model is called tumor-specific.

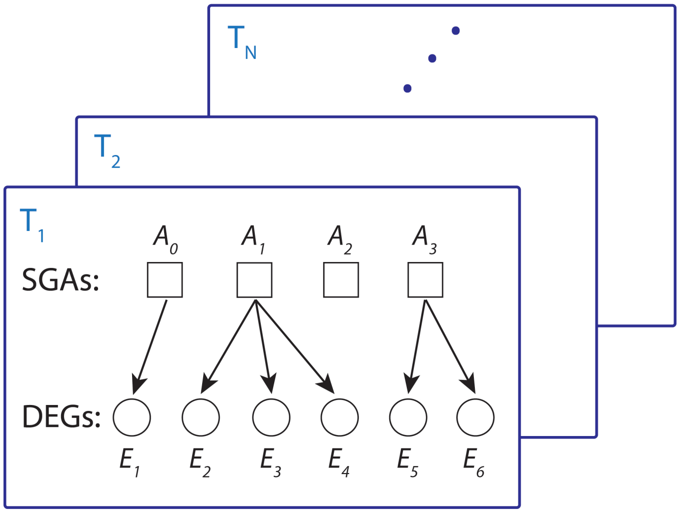

Figure 2. Tumor-specific Bayesian causal inference framework. An example of an admissible CBN for inferring causal relationships between SGAs and DEGs. Here, each plate represents one tumor. For the tumor T_1, set **SGA** has three SGA variables plus the non-specific factor A_0 (m = 4) and set **DEG** has six DEG variables (n = 6). Each E_i must have exactly one arc into it, which represents having one cause among the variables in set **SGA**. In this model, E_1 is caused by A_0, and A_1 and A_3 are drivers of DEGs ({E_2, E_3, E_4} and {E_5, E_6} respectively), while A_2 is not a driver.

2.1. The basic Bayesian framework of the TCI model

Let *M* be a CBN structure and let *D* be an observational training dataset, in which each case denotes a sample that contains a measurement for each of the variables in *M*. We assume that the cases in *D* are i.i.d..

We can derive the posterior probability of a CBN structure *M* as follows:

$$P(M|D) = \frac{P(D,M)}{P(D)} = \frac{P(D,M)}{\sum_{M'} P(D,M')} = \frac{P(D|M)P(M)}{\sum_{M'} P(D \mid M')P(M')}, \tag{1}$$

where the sum is taken over all admissible models *M'*. The term *P*(*M*) denotes prior belief that the data-generating CBN has *M* as its structure.

We call the term *P*(*D, M*) the score of CBN structure *M* in light of data *D*. As shown in Equation 1, it can be expressed as follows:

$$P(D,M) = P(D|M)P(M). \tag{2}$$

We will assume that *P*(*M*) is a modular prior probability that can be expressed as follows:

$$P(M) = \prod_{i=1}^{n} P(A_{g(i)} \to E_i), \tag{3}$$

where $A_{g(i)}$ is a node in **SGA** that is the parent of E_i in *M*, and $P(A_{g(i)}\text{->}E_i)$ is the prior probability that $A_{g(i)}$ is causally influencing E_i in the current tumor. The function *g*(*i*) returns an index of *A*. If *g*(*i*) = 0 then A_0 represents its parent, which means E_i is regulated by a non-SGA factor in the tumor.

The term *P*(*D*|*M*) is the marginal likelihood of *M*, which can be derived by marginalizing out model parameters θ as follows:

$$P(D|M) = \int_{\theta} P(D \mid M, \theta) P(\theta \mid M)\, d\theta, \tag{4}$$

where θ represents the parameters (probabilities) associated with CBN structure *M*.

If we assume parameter independence, parameter modularity, and Dirichlet prior probability distributions, we can solve Equation 4 to derive *P*(*D*|*M*) in closed form[20] as follows:

$$P(D|M) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}, \qquad (5)$$

where:

- *i* indexes over the **DEG** variables included in *M*;
- *n* is the number of DEGs in *M*, i.e., the nodes in the **DEG** set of *M*;
- *j* indexes over the 0/1 values (states) of a gene *A* in **SGA** that is being modeled as the parent of E_i in *M*;
- q_i is the number of values of parent gene *A* of the node E_i, which is 2, because the *A* is modeled as having the values 1 (a somatic genome alteration) and 0 (no alteration);
- *k* indexes over the 0/1 values of the expression states of E_i;
- r_i is the number of values of node E_i, which is 2, because *E* is modeled as having the values 1 (a differential gene expression level) and 0 (a normal level of gene expression);
- N_{ijk} is the number of cases in *D* that node E_i has the value denoted by *k* and its parent has the value denoted by *j*;
- α_{ijk} is a parameter in a Dirichlet distribution that represents prior belief about $P(E_i \mid parent(E_i))$; it can be interpreted as belief equivalent to having previously seen (prior to data *D*) α_{ijk} cases in which node E_i has the value *k* and its parent has the value *j*;
- Γ is the gamma function;
- $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$
- $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$

Substituting Equations 3 and 5 into Equation 2 yields:

$$P(D,M) = \prod_{i=1}^{n} P\left(A_{g(i)} \rightarrow E_i\right) \prod_{j=1}^{q_i} \frac{\Gamma\left(\alpha_{ij}\right)}{\Gamma\left(\alpha_{ij} + N_{ij}\right)} \prod_{k=1}^{r_i} \frac{\Gamma\left(\alpha_{ijk} + N_{ijk}\right)}{\Gamma\left(\alpha_{ijk}\right)}. \tag{6}$$

Let *e*(*g*(*i*), *i*) represent the function that appears inside the outer product of Equation 6. Thus, it is defined as follows:

$$e(g(i),i) = P\left(A_{g(i)} \rightarrow E_i\right) \prod_{j=1}^{q_i} \frac{\Gamma\left(\alpha_{ij}\right)}{\Gamma\left(\alpha_{ij} + N_{ij}\right)} \prod_{k=1}^{r_i} \frac{\Gamma\left(\alpha_{ijk} + N_{ijk}\right)}{\Gamma\left(\alpha_{ijk}\right)}. \tag{7}$$

We can now write Equation 6 as follows:

$$P(D,M) = \prod_{i=1}^{n} e(g(i),i). \tag{8}$$

Equation 7 is the score for a causal arc existing from $A_{g(i)}$ to E_i. However, we wish to have a non-zero score only for a causal relationship that satisfies the following constraint: E_i is more likely to be abnormal (value 1) when $A_{g(i)}$ is abnormal (value 1) than when $A_{g(i)}$ is normal (value 0). Given the Dirichlet distributions we are using, the expectation of these conditional probabilities is as follows:

$$P\left(E_i = k \middle| A_{g(i)} = j\right) = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}. \tag{9}$$

Using conditional probabilities of this form to enforce the constraint mentioned above, leads to the following function:

$$f(g(i),i) = \begin{cases} e(g(i),i), & if\ P\left(E_i = 1 | A_{g(i)} = 1\right) > P\left(E_i = 1 | A_{g(i)} = 0\right) \\ & or\ g(i) = 0; \\ 0, & otherwise \end{cases} \quad . \tag{10}$$

We next use *f* to refine Equation 8 to be the following:

$$P(D,M) = \prod_{i=1}^{n} f(g(i),i). \tag{11}$$

The posterior probability that an SGA A_h causes a DEG E_i in tumor t relative to the **SGA** and **DEG** is calculated as follows,

$$P(A_h \rightarrow E_i | D, \boldsymbol{SGA}_t, \boldsymbol{DEG}_t) = \frac{f(h,i)}{\sum_{h'=0}^{m} f(h',i)}. \tag{12}$$

2.2. Tumor-centric scoring

When assessing the causal relationship between A_h and E_i using Equation 12, we consider the states of A_h and E_i in all tumors in the training set. As mentioned previously, however, E_i could be regulated by multiple distinct SGAs that affect a common signaling pathway. These SGAs tend to be mutually exclusive across all tumors. For example, a gene that is expressed downstream in the PI3K pathway would be differentially expressed in tumors hosting either a *PTEN* deletion/mutation or a *PIK3CA* amplification/mutations, and these two SGA events tend to be mutually exclusive (Figure 1). Under such circumstances, either a *PTEN* alteration or a *PIK3CA* alteration should be sufficient to explain DEG E_{PI3K}.

In this section, we describe a modified Bayesian scoring measure that models SGAs affecting a DEG. Consider the situation in which A^* is the driver of DEG E_i in most tumors. Suppose a tumor t that is currently being analyzed has E_i as a DEG but does not include A^* as an SGA. In this case, we need to locate the SGA that is most likely the driver of E_i in tumor t, in light of most of the tumors in the training set having A^* as the driver of E_i.

Consider the following example. Let E_{PI3K} be a DEG in tumor t. Suppose the expression of E_{PI3K} is regulated by the PI3K pathway. Suppose also that *PIK3CA* is the most commonly perturbed member along that pathway (Figure 1), which leads it to be chosen as A^* according to the methods in Section 2.1. Current tumor t does not contain *PIK3CA* as an SGA, however. Thus, we need a causal explanation for DEG E_{PI3K} in tumor t. Suppose that *PIK3R1* is an SGA in tumor t. The method described below scores *PIK3R1* as a driver of E_{PI3K} for all the tumors in the training set that contain *PIK3R1* as an SGA; the remaining tumors in the training set are scored using *PIK3CA* as their driver; the overall score is a function of these two scores. This method is repeated for other SGAs in tumor *t* as candidate causes of E_{PI3K}. If *PIK3R1* turns out to be the SGA in tumor *t* that results in the highest overall score, then *PIK3R1* is output as the most likely driver of E_{PI3K} in tumor *t*. While this example illustrates the most basic situation in which tumor-specific scoring is called for, the general method can be useful in other circumstances as well.

We now describe the mathematical method we used to implement tumor-specific scoring. In tumor *t*, we want to find the most probable cause of each E_i that has a value of 1 (i.e., is a DEG). Let $A_{g(i)}$ denote a hypothesized gene that is causing E_i to be a DEG in tumor *t*. In order for $A_{g(i)}$ to be a candidate cause, we require that it be altered (i.e., have a value of 1) in tumor *t*. Let $D^1_{g(i)}$ denote the set of tumors in the training set in which variable $A_{g(i)}$ has the value 1, which denotes that these tumors have somatic genome alteration (SGA) in gene $A_{g(i)}$. We can calculate $e(g(i), i, D^1_{g(i)})$ with respect the tumor set $D^1_{g(i)}$ as follows:

$$e\big(g(i), i, D^1_{g(i)}\big) = P\big(A_{g(i)} \rightarrow E_i\big) \prod_{j=1}^{q_i} \frac{\Gamma\big(\alpha_{ij}\big)}{\Gamma\big(\alpha_{ij} + N^1_{ij}\big)} \prod_{k=1}^{r_i} \frac{\Gamma\big(\alpha_{ijk} + N^1_{ijk}\big)}{\Gamma\big(\alpha_{ijk}\big)}, \qquad (13)$$

where N^1_{ijk} is the number of cases in $D^1_{g(i)}$ that node E_i has value *k* and its parent $A_{g(i)}$ has value *j*. Since $D^1_{g(i)}$ represents the tumors for which $A_{g(i)}$ has the value 1, this means that *j* is fixed at the value 1. Thus, we can simplify Equation 13 to be the following:

$$e(g(i), i, D^1_{g(i)}) = P(A_{g(i)} \rightarrow E_i) \frac{\Gamma(\alpha_{i1})}{\Gamma(\alpha_{i1} + N^1_{i1})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{i1k} + N^1_{i1k})}{\Gamma(\alpha_{i1k})}. \qquad (14)$$

Let $D^0_{g(i)}$ denote the set of tumors in the training set in which variable $A_{g(i)}$ has the value 0, which represents that these tumors do not have genome alteration in $A_{g(i)}$. We need to determine the most likely parent of E_i for these tumors. An efficient way to do so, which we use, is to find the most likely gene *A** that causes E_i over all tumors in dataset *D*. Then, we hypothesize that gene as the cause of E_i in $D^0_{g(i)}$. This approach is efficient, because we only have to perform the search once for each E_i prior to seeing the current tumor *t*. More specifically, we use the *f* score from Equation 10 (on the data *D* on all the tumors) to search over all possible genes to find the one that maximizes the score. Let $A_{G(i)}$ denote such a maximal scoring gene. We take $A_{G(i)}$ to be the parent of E_i for all the tumors in $D^0_{g(i)}$. The score for this parent in these tumors in as follows:

$$e(G(i), i, D^0_{g(i)}) = P(G(i) \rightarrow E_i) \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N^0_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N^0_{ijk})}{\Gamma(\alpha_{ijk})}. \qquad (15)$$

We arrive at a tumor-centric score *e(g(i), i)* by taking the product of Equation 14 and Equation 15:

$$e(g(i), i) = e\left(g(i), i, D^1_{g(i)}\right) \cdot e\left(G(i), i, D^0_{g(i)}\right). \qquad (16)$$

In tumor *t*, consider an E_i that is a DEG (i.e., E_i = 1). We search over all the genes that have a value of 1 (i.e., that are SGAs). For each such $A_{g(i)}$, we use Equation 16 to score it. The $A_{g(i)}$ that has the highest score is returned as the most likely cause of E_i in tumor *t*. We perform this procedure for each E_i that is a DEG in tumor *t*. After doing so, we have determined the most likely SGA causing each DEG in tumor *t*.

3. Implementation Details

The following shows the implementation details of our models, i.e., general method and tumor-centric method. Also, in order to apply the methods in the previous section, we need to specify both structure priors and parameter priors.

3.1. Pseudocode

The TCI pseudocode in this section consists of a general method and a tumor-centric method. The general method is used to derive the most probable parent $A_{G(i)}$ for each node E_i across all tumors in training set *D*. Then, the tumor-centric method calculates the Bayesian causal score for each edge $A_h \rightarrow E_i$ in each tumor *t*.

3.1.1 General method

```
var dataset $D$; set SGA', DEG'; array cause; integer $i$, $h$, $m'$, $n'$; real $d$;
SGA' = the set of genes that have aberrant genome alterations in any tumor of $D$ ;
$m'$ = |SGA'|
DEG' = the set of genes that are differentially expressed in any tumor of $D$ ;
$n'$ = |DEG'|
for $i$ = 1 to $n'$ do // search for the global driver $A_{G(i)}$ for each $E_i$ at the population level
    for $h$ = 0 to $m'$ do
        compute $f(h, i)$; //use Equation 10 to compute $f(h, i)$
    $A_{G(i)}$ is identified as the SGA $G(i)$ that has the highest $f(h,i)$ for $E_i$.
```

3.1.2 Tumor-centric causal inference

```
var dataset $D$; set SGA, DEG; array cause; integer $i$, $h$, $m$, $n$; real $d$;
SGA = the set of genes in tumor $t$ that have aberrant genome alterations;
$m$ = |SGA|
DEG = the set of genes in tumor $t$ that are differentially expressed;
$n$ = |DEG|
for $i$ = 1 to $n$ do //populate the values of the cause array
    $d$ = 0,
    for $h$ = 0 to $m$ do
        use the global driver $A_{G(i)}$ for $E_i$ as determined using the general method;
        $d := d + e(h, i)$; //use Equation 16 to compute $e(h, i)$
    for $h$ = 0 to $m$ do
```

cause [*h*, *i*] := *e*(*h*, *i*)/*d*;

3.2. Structure priors

Given the tumor of interest with a unique **SGA** set and a unique **DEG** set, we need to define a tumor specific structure prior *P*(*M*) over permissible CBN structures *M*. Because the **SGA** and **DEG** sets are (with high probability) unique to each tumor, the prior distribution over *M* is also tumor-specific. Assuming the structure prior is modular, we can factorize the *P*(*M*) as a product of prior probabilities for each permissible edge as follows:

$$P(M) = \prod_{i=1}^{n} \prod_{h=0}^{m} P(A_h \rightarrow E_i). \tag{17}$$

$P(M)$ comprises a product of prior probabilities of causal edges of a test. A prior probability of a causal edge from a somatic alteration of gene A_h to a DEG E_i can be stated as $P(A_h \rightarrow E_i)$ (abbreviated as θ_h) and determined according to:

$$\theta_h = (1 - \theta_0) \frac{\mu_h}{\sum_{h'=1}^{m} \mu_{h'}}, \tag{18}$$

where θ_0 is a prior probability that the cause of DEG E_i is not an SGA, and *h′* indexes over the number *m* of genes in tumor *t* that have SGAs.

Additional genomic information can be applied to derive the prior probability of each edge $A_h \rightarrow E_i$ using existing prior knowledge. Consider, for example, the availability of the following information for each gene *h*: (1) the number of unique synonymous mutations observed for *h* among the tumors in *D*, and (2) the number of abnormal somatic copy number alterations (according to a given definition of abnormal) of *h* in a normal population without cancer. Such information can be applied to help account for mutation and copy number alterations that are due to differences in gene lengths and chromosome locations. In particular, using the information in (1) and (2) above, we can calculate μ_h as follows:

$$\mu_h = \sum\nolimits_{t' \in U_h} w_{ht'}, \tag{19}$$

where U_h denotes the tumors in training set *D* that have a somatic alteration in gene A_h, and $w_{ht'}$ denotes a weight proportional to the probability that SGA *h* is a driver in the genome of tumor *t'*. We calculate $w_{ht'}$ as follows:

$$w_{ht'} = \frac{R_{ht'}}{\sum_{h'=1}^{m} R_{h't'}}, \tag{20}$$

where

$$R_{ht'} = \frac{T_{ht'}^{SM}}{R_h^{SM}} + \frac{T_{ht'}^{SCNA}}{R_h^{SCNA}}. \tag{21}$$

In equation 21, $T_{ht'}^{SM}$ denotes whether gene *h* has a non-synonymous somatic mutation (SM) event or not in tumor *t'*, i.e., 1 or 0, respectively; R_h^{SM} denotes the number of unique synonymous mutation events in gene *h* observed in the reference set of tumor genomes, *D*; $T_{ht'}^{SCNA}$ denotes whether gene *h* is affected by an SCNA event or not in tumor *t'*, i.e., 1 or 0, respectively; R_h^{SCNA} denotes the expected number of times gene *h* is affected by copy number alteration among the tumors in *D*, and yet is only a passenger alteration, based on the number of times gene *h* is affected by copy number alteration in a reference set of cases from a normal human population without known cancer.

3.3. Parameter priors

We need parameter priors for when E_i has A_0 as its parent and for when it has an A_h in **SGA** set as its parent. Table 1 addresses the case when E_i has A_0 as its parent. The Dirichlet parameter values in the table represent that every probability of $P(E_i = 1)$ is equally likely a *priori* (i.e., before any data are considered); recall that $E_i = 1$ represents that E_i is abnormal. Table 2 addresses the case in which E_i has some parent A_h. The Dirichlet parameter values in the table make it somewhat more likely that E_i will be normal (abnormal) when its cause A_h is normal (abnormal). To make that pattern stronger, the values of 2.0 could be replaced with larger values, such as 2.5, 3.0, or even higher.

Table 1. Dirichlet parameter values when A_0 is the parent of node E_i.

Prior probability being specified	Dirichlet parameter	Default value
$P(E_i)$	α_{i10}	1.0
	α_{i11}	1.0

Table 2. Dirichlet parameter values when there is one A_h parent of node E_i.

Prior probability being specified	Dirichlet parameter	Default value
$P(E_i \mid A_h = 0)$	α_{i00}	2.0
	α_{i01}	1.0
$P(E_i \mid A_h = 1)$	α_{i10}	1.0
	α_{i11}	2.0

The computations in this paper have used standard arithmetic operations. However, model scores can become extremely small. Therefore, it is generally necessary to use log arithmetic. When doing so, Equation 7, for example, becomes a sum of log terms, rather than a product of terms.

Acknowledgement

Research reported in this publication was supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA *et al*: **Mutational heterogeneity in cancer and the search for new cancer-associated genes**. *Nature* 2013, **499**(7457):214-218.
2. Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N: **Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation**. *Genome Med* 2012, **4**(11):89.
3. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA *et al*: **Mutational landscape and significance across 12 major cancer types**. *Nature* 2013, **502**(7471):333-339.
4. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER *et al*: **MuSiC: identifying mutational significance in cancer genomes**. *Genome Res* 2012, **22**(8):1589-1598.
5. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C: **Emerging landscape of oncogenic signatures across human cancers**. *Nat Genet* 2013, **45**(10):1127-1133.
6. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C: **The landscape of kinase fusions in cancer**. *Nat Commun* 2014, **5**:4846.
7. Dawson MA, Kouzarides T: **Cancer epigenetics: from mechanism to therapy**. *Cell* 2012, **150**(1):12-27.
8. Heidenreich B, Rachakonda PS, Hemminki K, Kumar R: **TERT promoter mutations in cancer development**. *Curr Opin Genet Dev* 2014, **24**:30-37.
9. Xie H, Liu T, Wang N, Bjornhagen V, Hoog A, Larsson C, Lui WO, Xu D: **TERT promoter mutations and gene amplification: promoting TERT expression in Merkel cell carcinoma**. *Oncotarget* 2014, **5**(20):10048-10057.
10. Pearl J: **Causality: Models, Reasoning and Inference**, 2nd edn: Cambridge University Press; 2009.
11. Glymour C, Cooper G: **Computation, Causation, and Discovery**. Cambridge, MA: MIT Press; 1999.
12. Boyle EA, Li YI, Pritchard JK: **An Expanded View of Complex Traits: From Polygenic to Omnigenic**. *Cell* 2017, **169**(7):1177-1186.
13. Nica AC, Dermitzakis ET: **Expression quantitative trait loci: present and future**. *Philos Trans R Soc Lond B Biol Sci* 2013, **368**(1620):20120362.
14. Cully M, You H, Levine AJ, Mak TW: **Beyond PTEN mutations: the PI3K pathway as an integrator of multiple inputs during tumorigenesis**. *Nat Rev Cancer* 2006, **6**(3):184-192.
15. Lu S, Lu KN, Cheng SY, Hu B, Ma X, Nystrom N, Lu X: **Identifying Driver Genomic Alterations in Cancers by Searching Minimum-Weight, Mutually Exclusive Sets**. *PLoS Comput Biol* 2015, **11**(8):e1004257.
16. Cancer Genome Atlas Research N: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways**. *Nature* 2008, **455**(7216):1061-1068.
17. Ciriello G, Cerami E, Sander C, Schultz N: **Mutual exclusivity analysis identifies oncogenic network modules**. *Genome Res* 2012, **22**(2):398-406.

18. Yamamoto H, Shigematsu H, Nomura M, Lockwood WW, Sato M, Okumura N, Soh J, Suzuki M, Wistuba, II, Fong KM *et al*: **PIK3CA mutations and copy number gains in human lung cancers**. *Cancer Res* 2008, **68**(17):6913-6921.

19. Vandin F, Upfal E, Raphael BJ: **De novo discovery of mutated driver pathways in cancer**. *Genome Res* 2012, **22**(2):375-385.

20. Heckerman D, Geiger D, Chickering DM: **Learning bayesian networks: The combination of knowledge and statistical data**. *Machine learning* 1995, **20**(3):197-243.