Influence network model uncovers new relations between biological processes and mutational signatures

Bayarbaatar Amgalan, Damian Wojtowicz, Yoo-Ah Kim*, Teresa M. Przytycka*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Abstract. There is a growing appreciation that mutagenic processes can be studied through the lenses of mutational signatures – characteristic mutation patterns attributed to individual mutagens. However, the link between mutagens and observed mutation patterns is not always obvious. While some mutation signatures have been connected to specific causes such as UV light exposure, smoking, or other biochemical processes, elucidating the causes of many signatures, especially those associated with endogenous processes, remains challenging. In addition, endogenous mutational processes interact with each other, as well as with other cellular processes, in ways that are not fully understood.

To gain insights into the relations between mutational signatures and cellular processes, we developed a network-based approach termed GeneSigNet. The main idea behind the approach is to utilize gene expression and signature activities for the construction of a directed network containing two types of nodes corresponding to genes and signatures respectively. The construction utilizes a sparse partial correlation technique complemented with a higher moment-based approach assigning edge directionality when possible.

Application of the GeneSigNet approach to breast and lung cancer data sets allowed us to capture a multitude of important relations between mutation signatures and cellular processes. In particular, the model suggests a causative influence of the homologous recombination deficiency signature (SBS3) on a clustered APOBEC mutation signature and linked SBS8 with the NER pathway. Interestingly, our model also uncovered a relation between APOBEC hypermutation and activation of regulatory T Cells (Tregs) known to be relevant for immunotherapy, and a relation between the APOBEC enzyme activity (SBS2) and DNA conformation changes. GeneSigNet is freely available at `https://github.com/ncbi/GeneSigNet`.

1 Introduction

Traditionally, research in cancer genomics has been focused on the identification of cancer driving mutations, which confer a growth advantage to the cancer cells. However, since cancer cells emerge as a result of various mutagenic processes such as UV light or a faulty DNA repair mechanism, they also accumulate numerous mutations with seemingly no direct role in carcinogenesis. Importantly, these so-called passenger mutations can provide valuable information about mutagenic processes that cells have undergone. The key property facilitating the identification of these mutagenic processes is that different mutagenic processes leave characteristic mutation imprints on the cancer genome. Starting from the pioneering work of Alexandrov et al. [1], several computational methods have been developed allowing to decompose a cancer genome's mutation catalogs into characteristic mutation patterns termed mutational signatures. Some of these mutational signatures have already been linked to specific mutagenic processes [2,3]. Mutational processes can be caused by both intrinsic (e.g., DNA repair deficiency) and extrinsic (e.g., UV radiation, tobacco smoking) factors. However, the etiology of many of these signatures still remains unknown or not fully understood. In addition, little is known about the interactions between mutagenic processes and other cellular processes. Yet, such interactions are known to exist. As a case in point, tobacco smoking is not only mutagenic itself but also believed to activate an immune response [4]. Conversely, a perturbation of some cellular processes, such as DNA replication or repair pathways, can be also mutagenic. Furthermore, mutagenic processes themselves have been known to interact between each other: homologous recombination deficiency (HRD) for instance is typically accompanied by a mutational signature related to APOBEC activity [5,6]. Elucidating the interactions between mutagenic processes as well as the interactions of mutagenic processes with cellular pathways is of fundamental importance for a better understanding of carcinogenesis and the design of novel cancer treatments. Deficiencies in the activities of several genes have been linked to specific signatures, including MUTYH [7], ERCC2 [8], MSH6 [9], and FHIT [10]. In addition, a correlation between the expression of the APOBEC family of genes with the strength of signatures SBS2 and SBS13 (the so-called APOBEC mutational signatures) has frequently been observed [2,11]. Indeed, interrogating the correlation between gene expression and the strength of a mutational signature can provide important clues on the etiology of the signatures. For example, Kim et al. identified coherently expressed groups of genes associated with specific combinations of mutational signatures [6]. While their analysis provided important insights into the etiology of some signatures, it also pointed to the limitations of such cluster-based analysis. For example, while the analysis captured the dependency between some signatures (e.g. clustered APOBEC mutations and HRD), the directionality of the relationship could not be untangled.

To fill this gap, we introduce a network-based method, named GeneSigNet (Gene and Signature Influence Network Model), for inferring relationships between gene expression and mutagenic processes. To this end, GeneSigNet constructs a *Gene-Signature Network* (GSN) defined as a sparse, weighted, and directed network consisting of two types of nodes – genes and SigStates (Signature States). SigState nodes are in one-to-one correspondence with signatures and each SigState represents a general cell state associated with the emergence of the corresponding mutational signature (see Section 2.1 for a detailed description). Importantly, both genes and SigStates have their own associated activities. Activities of genes are defined by their expression profiles across samples while activities of SigStates are measured by the strengths of the corresponding signatures across samples. Node activities will provide the basis for inferring edges of the GSN. Mathematically, both gene and SigState nodes are treated identically, however the interpretation of the uncovered relations can be different.

Inferring causal relationships between biological entities is challenging yet crucial in various biological applications. Approaches to infer directions are often based on perturbation data or prior information such as transcription factors binding, protein-protein interaction networks, among others [12,13,14,15]. Unlike gene regulatory networks where causal relations originate from transcription factors to target genes, no perturbation data or prior knowledge is available in our setting. Hence, the directionality of the relationship between SigStates as well as between other pairs of nodes in the GSN is undetermined and GeneSigNet relies on inferring dependencies based solely on the activities of nodes. Aiming at achieving deeper mechanistic insights, our approach goes beyond simple bivariate correlations and strives to identify dominating influences and causal relationships. To identify such dominant relations, GeneSigNet leverages a sparse partial correlation technique (SPCS). Sparse estimations of partial correlations based on a penalized regression [16,17] have been previously introduced to construct biological networks, although neither of them determines the

causality directions. Under the conditional dependency assumption, GeneSigNet selects, for each node, the optimal combination of a sparse set of explanatory factors (genes or SigStates), resulting in a weighted directed network. However, some edges inferred by SPCS can have similar weights in both directions, in which case a higher moment-based strategy is used to resolve directionality. We note that other methods that infer gene regulatory networks from gene expression data [18,19,20,21] may be applicable in our setting even though their development was meant for a different purpose. In particular, GENIE3, considered to be the best method in this class [20,22,23], infers a weighted complete directed graph. Thus we tested if edge weights computed by GENIE3 can be utilized for directionality assignment. However, we found that GeneSigNet outperforms such approach.

Our results demonstrate that network sparsification combined with directionality information regarding the influence between the nodes provides advanced insight going beyond general GO enrichment analysis and suggests more mechanistic explanations. The relations inferred by the GeneSigNet model are consistent with current knowledge and include several new and interesting findings. In particular, the model suggests a causative relation from the homologous recombination deficiency signature (SBS3) to a clustered APOBEC mutation signature and linked Signature 8 (SBS8) to the NER pathway. The last connection is consistent with the recent findings based on an experimental study in mouse [24]. In addition, GeneSigNet identified a relation between APOBEC hypermutation and activation of regulatory T Cells which presents an important implication in immunotherapy and captured a relation of APOBEC signature (SBS2) with DNA conformation changes among other findings.

In what follows, we provide a general description of the method, its evaluation, and the key biological results. For completeness, a formal description, mathematical details, and additional information about biological results are provided in Supplementary Information.

2 Results

2.1 Overview of GeneSigNet

The main idea of the approach is to construct a *gene-signature network* (GSN), consisting of nodes corresponding to genes and signatures (SigStates), and use this graph to establish relations between mutagenic processes and other cellular processes (Fig. 1A). Thus, the GSN extends the concept of a gene network to include, in addition to genes, meta-nodes corresponding to signatures. Both types of nodes have associated node activities: gene expression for the gene nodes and strength of mutational signatures for the corresponding SigState nodes (Fig. 1C).

Importantly, we introduce a concept of SigStates, defined as the abstract representation of cellular states associated with a specific mutational signature. The activity of a SigState is based on the number of mutations attributed to the corresponding signature. Note that while gene expression can have a causative effect on generating mutational signatures, signatures themselves are simply mutation patterns created by mutagenic processes and cannot directly influence gene expression. Instead, the activities of cellular processes related to signatures might influence gene expression, which motivated us to introduce SigStates. A mutagenic process can be directly related to a perturbation of a specific biological pathway or gene, or be triggered by an environmental factor such as smoking, which in turn might additionally affect other cellular processes. Given that the outcome of mutatgenic processes accumulate over time, it is challenging to fully ensure that a SigState represents the processes directly causing the signature. Therefore, unless additional information is given, genes linked to a SigState in either direction are assumed to be associated with the respective mutagenic process rather than causing or being caused by it (Fig. 1B).

Having defined the nodes of the GSN we now turn to inferring the relations between them. Mathematically, our network inference does not distinguish between the two types of nodes. The approach consists of two main steps. First, a sparse partial correlation technique is used to construct a sparse directed network that captures dominant relations between the nodes. Next, a higher moment-based directionality determining technique is used to reduce the number of edges with unresolved directions (bidirectional edges). The workflow of the GeneSigNet method is presented in Figure 2.

Let X be a combined input matrix consisting of gene expressions and exposures of mutational signatures (SigStates) across cancer samples (Fig. 2A). The rows of X denote the samples, whereas the columns represent genes and SigStates. In other words, for i-th sample, x_{ij} denotes expression of a gene (for $1 < j < m$) and

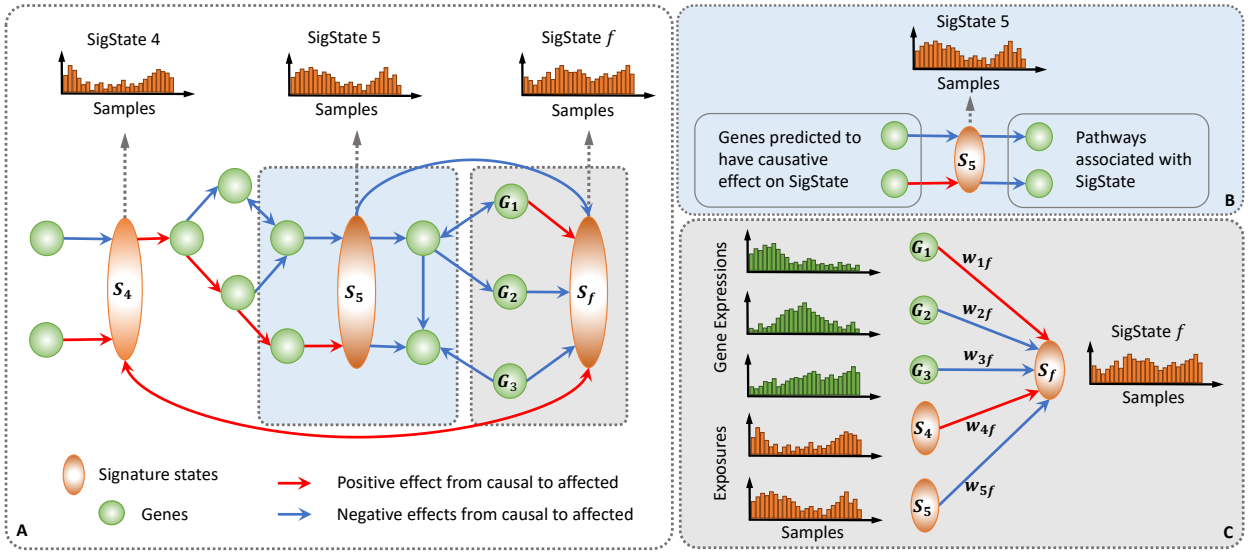

Fig. 1. Gene-Signature Network (GSN). (A) GSN is a directed network consisting of two types of nodes: genes (green circles) and SigStates (orange ovals). SigStates are in one-to-one correspondence with signatures. The activity of a gene is defined by its patient-specific expression level, whereas the activity of a SigState is represented by the strength of the corresponding mutational signature across samples. Edges of a GSN represent inferred dependencies and might be either positive (red) or negative (blue). The edges are unidirectional when the direction of the influence is resolved or bidirectional otherwise. **(B)** Interpretation of gene–SigState edges. A directed edge from a gene to a SigState represents a putative causal relation from the gene to the SigState and thus to the corresponding signature. The genes that are targets of a SigState are interpreted as *associated* with the mutational signature. These downstream genes can be perturbed DNA repair pathways and thus be involved in producing the signature imprints in a genome or can be other dysregulated pathways (e.g. immune response) caused by SigState. **(C)** For each node (gene or SigState), GeneSigNet identifies potential regulators (genes or SigStates) as a sparse, optimized set of nodes whose activities are partially correlated (correlated after accounting for confounding correlations) with the activity of a given node. Formally, GeneSigNet infers, for every node k, the weight w_{kf} of the influence of the putative regulator k on the activity of node f (see also the method workflow in Fig. 2).

exposure of a mutational signature (for $m+1 < j < n$). GeneSigNet decomposes the problem of inferring the network of n nodes into n different variable selection subproblems (Fig. 2B). For each subproblem, a single node is considered a target and sparse partial correlation selection (SPCS) is used to find the weights of incoming effects from the other $n-1$ nodes. The l_1 norm constraint is combined with the least square minimization to avoid over-fitting. For each node, the sparsity constraint allows selecting only dominant incoming effects as potential causal factors (see Sections S1.1 and S1.2 in Supplementary Information for a detailed mathematical description of the SPCS model).

Next, we use a higher moment-based strategy to decide directionality for each bidirectional edge (j, f) inferred by SPCS (Fig. 2C). Let r_j and r_f be the column vectors denoting the residual activities over the p samples, corresponding to the nodes j and f, respectively. These vectors are calculated by removing the confounding effects from the activities of the nodes j and f due to the presence of the remaining $n-2$ nodes (for more details on the confounding effect removal, see Equation S4 and Section S1.3 in Supplementary Information). Hence, r_j and r_f can be assumed to have a bivariate normal distribution since the incoming confounder effects from the other $n-2$ factors were removed from the activities of the nodes j and f. The direction of causal effects between the pair of nodes is determined based on higher moment statistics, skewness and kurtosis of r_j and r_f [25]. If both moments support the same direction with the sparse partial correlation, the edge corresponding to the opposite direction is removed, otherwise, both edges remain in the network (for details, see Section S1.3 in Supplementary Information). Finally, a matrix normalization algorithm, alternate scaling [26], is used to bring the total incoming and outgoing effects of each node to the same range (for details, see Section S1.4 in Supplementary Information), and a weighted-directed network is constructed to represent the regulatory flows over all nodes. A more detailed description of GeneSigNet is provided in Supplementary Information (Section S1).

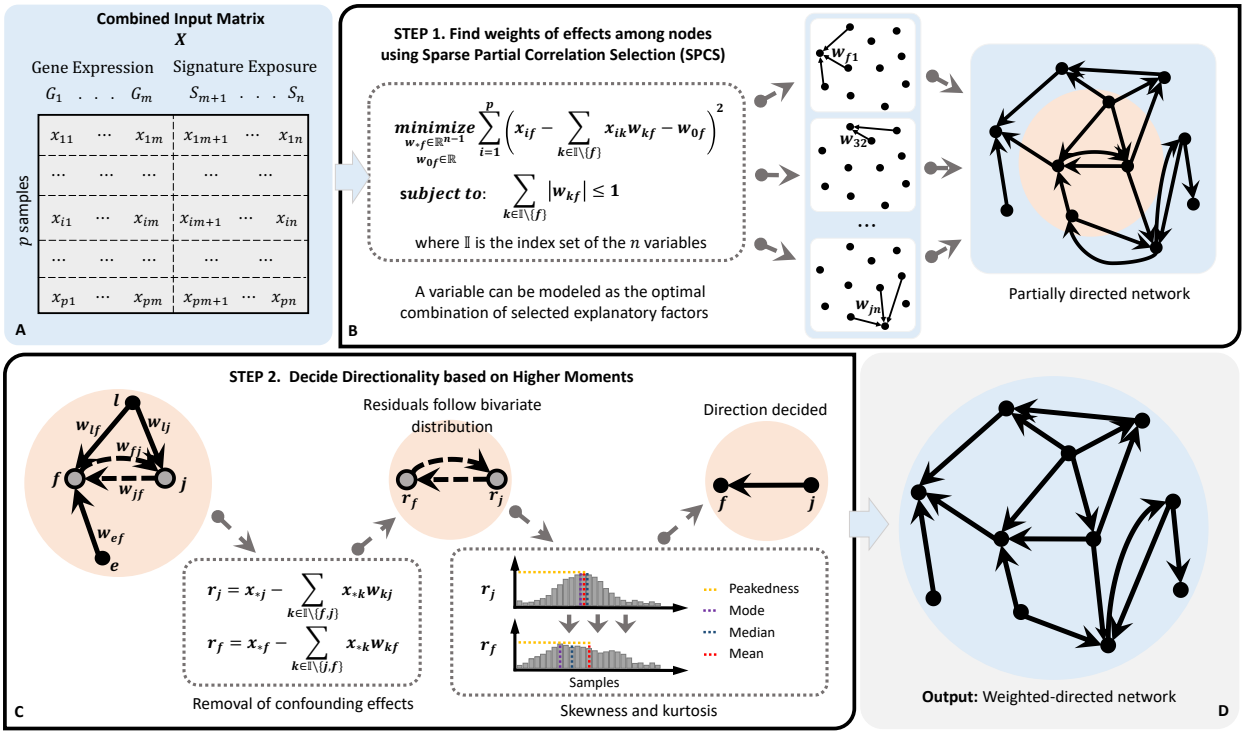

Fig. 2. Workflow of GeneSigNet. (A) A combined input matrix X is given by concatenating gene expression data and exposures of mutation signatures (SigStates) across p samples (patients). The i-th row represents the data of the i-th sample with the first m values corresponding to expression data of m genes followed by $n-m$ values of exposures for $n-m$ mutational signatures. **(B)** Given the input matrix X, we infer a network of n nodes. For each node, Sparse Partial Correlation Selection (SPCS) is used to simultaneously estimate the weights of incoming effects from the other $n-1$ nodes. w_{kf} is the edge weight denoting the strength of the effect on node f coming from node k. An l_1 norm constraint is used to avoid over-fitting. This sparsity constraint allows selecting only dominant effects as causal nodes by filtering out insignificant confounding effects due to noise. **(C)** For each bidirectional edge (j, f), the residual vectors r_j and r_f corresponding to nodes j and f are obtained by removing effects of the $n-2$ nodes other than the two nodes of the considered edge. The direction of causal effects between the pair of nodes is determined based on higher moment statistics, skewness and kurtosis of r_j and r_f. If both moments support the same direction with the partial correlation (see Equation S5 in Supplementary Information), the edge corresponding to the opposite direction is removed, otherwise, both edges remain in the network. **(D)** Finally, the total incoming and outgoing effects of each node are normalized respectively in the same range using the alternate scaling algorithm, and a weighted-directed network is constructed to represent the dependency flows over all nodes as output.

Evaluation of the directionality inference We benchmarked the performance of GeneSigNet in correctly inferring gene interaction directionality. As mentioned previously, a number of adjacent methods inferring gene regulatory networks are also based on node activities to determine weighted directed graphs. The best method in this class is considered to be GENIE3 [20,22,23]. For every node (a gene) in the network, GENIE3 assigns weights (influence scores) from all other genes to this gene and, as a result, constructs a fully connected, weighted, and directed graph. Given these properties, we tested whether utilizing edge weights inferred by GENIE3 and selecting the direction corresponding to the heavier of the two opposing edges could outperform the strategy employed by GeneSigNet. We compared the performance of the methods on breast cancer (BRCA) and lung cancer (LUAD) data sets (see Materials and Methods for details). We used mutational signatures' exposure data and gene expression data to construct a GSN for each cancer type and each method separately. For evaluating the inferred gene interactions, protein-DNA interactions from the ChEA database [27] were used while excluding all bidirectional edges and self-interactions and retained the remaining directed edges as our gold standard. The enrichment of correctly inferred directed interactions was computed by comparing the number of correctly assigned directions to the number of inconsistent directions; the significance of the enrichment was computed using a binomial test.

	Methods	GeneSigNet SPCS	GeneSigNet	GENIE3 naïve	GENIE3 modified
BRCA	Consistent fraction	0.554	0.614	0.455	0.583
	p value	1.1e-06	3.1e-07	1	3.4e-03
	Consistent directions	1,084	316	9,240	162
	Inconsistent directions	874	202	11,052	116
LUAD	Consistent fraction	0.544	0.571	0.505	0.557
	p value	1.0e-05	5.4e-04	0.0731	1.8e-03
	Consistent directions	1,285	310	11,049	370
	Inconsistent directions	1,077	233	10,833	294

Table 1. The evaluation of directionality inference. Using ChEA database as a gold standard, a binomial test was performed to evaluate the directionality inference by comparing the number of correctly assigned directions to the number of inconsistent directions. 'Consistent directions' is the number of inferred directions consistent with the ChEA database, whereas 'Inconsistent directions' denote the number of inconsistent inferences.

For GeneSigNet, we first evaluated the directionality inference of the SPCS step alone. It performed significantly better than the random selection ($p <$ 1.0e-05, see GeneSigNet SPCS in Table 1). The performance of GeneSigNet improved when the GSN was refined with the higher moment-based strategy and the normalization technique. GeneSigNet provided 316 consistent directions out of 518 directed edges recovered from the ChEA interactions (61.4%, $p =$ 3.1e-07) in the BRCA analysis, and 310 consistent directions out of 543 directed edges (57%, $p =$ 5.4e-04) in the LUAD analysis (see GeneSigNet in Table 1).

For GENIE3, we first determined directions by selecting the edge with a higher weight, resulting in a worse than random performance (see GENIE3 naïve in Table 1). We hypothesized that this behaviour is due to assigning the edge weight to all pairs of nodes irrespective of the inferred interaction strength. Subsequently, we modified this approach and considered only edges that are above a certain threshold. The threshold (0.0037 and 0.0017 for BRCA and LUAD, respectively) was selected to maximize the performance of GENIE3 in terms of the percent of consistent directions. The modified GENIE3 provided 162 consistent directions out of 278 directed edges recovered (58%, $p =$ 3.4e-03) in BRCA analysis and 370 consistent directions out of 664 directed edges (55%, $p =$ 1.8e-03) recovered in LUAD analysis (see GENIE3 modified in Table 1).

Overall the results indicate that the strategy used by GeneSigNet provides a better prediction of interaction direction than the competing approach.

2.2 Insights from the analysis of breast cancer data

Mutational signatures in breast cancer and construction of the GSN We utilized BRCA data collection obtained from ICGC which includes 266 cancer samples providing both whole genome sequencing data and gene expression data (for details, see Section 4.1 in Materials and Methods). The genomes harbor mutations mainly contributed by 6 COSMIC mutational signatures – SBS1, 2, 3, 5, 8, and 13. We further refined the mutational signatures based on mutation density and sample correlations. The mutations in BRCA are characterized by occurrences of short highly mutated regions whose origin is believed to be different than sparse mutations [6,28,11,29,30]. The information available from whole genome sequencing allows for distinguishing these two types of mutation patterns and to treat such dense and sparse mutation regions differently. The post-processing of mutational signatures resulted in 6 signature groups that we use for subsequent analysis to construct the GSN – SBS1, APOBEC-C (clustered SBS2 and SBS13 corresponding to APOBEC hypermutation), APOBEC-D (SBS2 corresponding to disperse APOBEC mutations), DSB (SBS3 and clustered SBS8), SBS5, and SBS8D (dispersed SBS8). In addition to gene expressions and exposures of mutational signatures, we included a node indicating the binary status of homologous recombination deficiency (HRD) as it is assumed to lead to specific patterns of mutational signatures in BRCA [31]. We applied GeneSigNet to construct a GSN for genes, mutational signatures, and HRD status, and to find relations between these features.

GeneSigNet uncovers mutagenic processes consistent with current knowledge Many relations uncovered with GeneSigNet are consistent with our current knowledge on mutational signatures, confirming the validity of our method. In particular, it is well appreciated that homologous recombination (HR) plays an

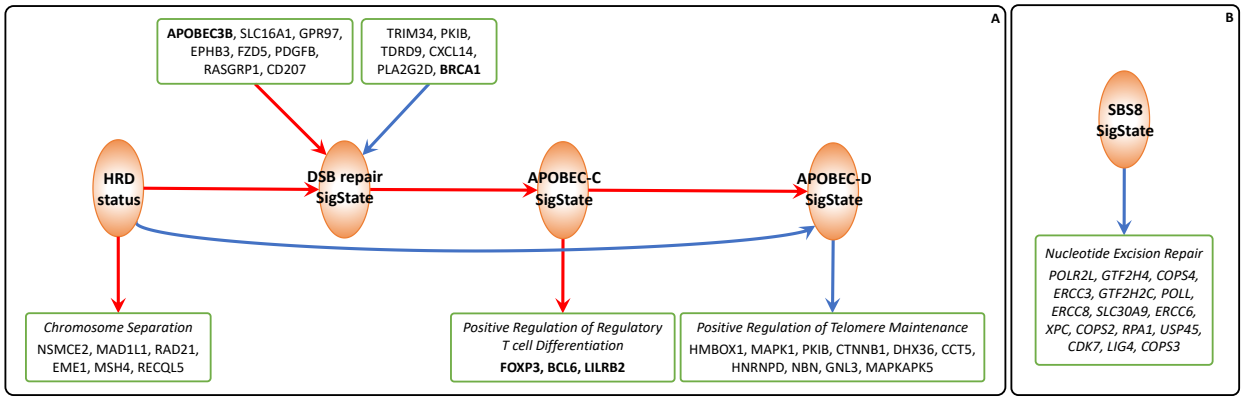

Fig. 3. (A) The HRD status dominantly contributes to the base substitution load in the DSB repair SigState in the presence of negative and positive effects from specific genes. Then, the increased DSB repair mutations produces carry-over effects on APOBEC-C and APOBEC-D SigStates while the other SigStates and genes are contributing to the flow. **(B)** SBS8 mutations are linked to the deficiency of the nucleotide excision repair. An extended network including genes and GO terms associated with SigStates is provided in Figure S1 in Supplementary Information.

important role in the double-strand break (DSB) repair mechanism and that HR deficiency is associated with the DSB signature [32]. Indeed, our network correctly predicted a strong positive influence from HRD status to the DSB signature (Fig. 3A). In addition, GeneSigNet identified the known negative impact of BRCA1 expression on the DSB signature which is also consistent with the role of BRCA1 in HRD [32]. Furthermore, GeneSigNet captured the impact of HRD on chromosome separation, reflecting the role of homologous recombination in maintaining genomic stability [33,34], and identified the association of APOBEC-D with telomere maintenance, consistent with the well recognized role of APOBEC mutagenesis in replication [35,36].

Interestingly, our method linked SBS8 to the nucleotide excision repair (NER) pathway (Fig. 3B). The etiology of this signature has remained unknown until a recent experimental study linked it to the NER pathway as well [24]. This demonstrates the power of the GeneSigNet method to uncover non-obvious relationships.

Untangling the interactions between APOBEC and DSB processes Previous studies speculated that APOBEC related mutational signatures can arise in multiple different scenarios. First, double-strand breaks (DSB) created by the homologous recombination deficiency (HRD) provide mutational opportunities for APOBEC enzymes to act on the ssDNA regions, resulting in clustered APOBEC mutations [37,38,30]. In another scenario, a recent study attributed APOBEC-mediated hypermutations to the normal activity of mismatch repair which also involves creating ssDNA regions, generating "fog" APOBEC mutations [29]. The complex interplay between APOBEC activities and other DNA repair mechanisms is yet to be elucidated.

Focusing on the interactions of APOBEC signatures with the other SigStates and genes, we observe that GeneSigNet supports a positive influence of the DSB on APOBEC-C SigState, consistent with the assumption that double-strand breaks provide an opportunity for APOBEC mutations. Additionally, our analysis reveals that the expression level of the APOBEC3B enzyme is associated with the strength of the DSB signature. Indeed, a previous study proposed that APOBEC3 proteins are recruited to DSB sites to participate in the DSB repair process [5]. Thus, DSB contributes to an increase in APOBEC-C strength by two different mechanisms: (i) increased mutation opportunity due to ssDNA created by DSB and (ii) increased mutation probability due to increased APOBEC3B expression. Note that increased APOBEC expression would also increase APOBEC mutations in the "fog" regions proposed in [29].

On the other hand, the activity of APOBEC-D is positively influenced by APOBEC-C activity, without direct relation to DSB. In fact, GeneSigNet inferred a negative influence from HR status to APOBEC-D SigState, confirming different mutagenic processes are involved in clustered and dispersed APOBEC mutations (Fig. 3A).

APOBEC hypermutation activates regulatory T cells – implications for immunotherapy Interestingly, GO enrichment analysis of the genes associated with APOBEC mutational signatures (genes influenced by APOBEC-C SigState) revealed significant enrichment in positive regulation of regulatory T

cell differentiation (Fig. 3A). Tumor cells with mutated DNA are likely to produce tumor-associated neoantigens (mutated peptides presented at their surface) that allow the immune system to recognize and destroy tumor cells. However, cells carrying a high mutation burden often develop mechanisms of immune tolerance involving activation of regulatory T cells (Tregs) to protect themselves from the destruction [39,40]. Tregs, a subtype of T cells that suppress the immune response, are important for maintaining cell homeostasis and self-tolerance but can also interfere with anti-tumor immune response [41]. The top three genes (FOXP3, BCL6, and LILRB2) positively influenced by APOBEC-C signature are all related to such inhibitory mechanism to immune response [42,43,44]. FOXP3 is a transcriptional regulator playing a crucial role in the inhibitory function of Tregs. BCL6 is also essential for the stability of Tregs that promotes tumor growth. LILRB2 is a receptor for class I MHC antigens and is involved in the down-regulation of the immune response and the development of immune tolerance.

Patients with cancers displaying a high mutation burden can benefit from immunotherapy [45]. In particular, the APOBEC mutational signature was identified as a potential predictive marker for immunotherapy response in some cancers [46,47]. However, an increased number of Tregs in a tumor may lead to resistance to immune checkpoint inhibitors [48,49]. Thus, our finding suggests that a combined strategy targeting Tregs in addition to immune checkpoint inhibitors would be most beneficial for a better outcome in APOBEC hypermutated breast cancer tumors.

2.3 Insights from the analysis of lung adenocarcinoma data

We next analyzed lung adenocarcinoma (LUAD) data using 466 cancer samples from the TCGA project. The exposure levels of 6 COSMIC mutational signatures (SBS1, 2, 4, 5, 13, and 40) present in the exome sequencing data were integrated with the RNAseq expression data of 2433 genes belonging to the DNA metabolic and immune system processes in GO terms to uncover influence between signatures and genes (see Materials and Methods for details on the lung cancer data).

GeneSigNet uncovers immune response due to smoking Two prominent mutational signatures in LUAD, SBS4 and SBS5, are assumed to result from exogenous causes [6]. SBS4 is associated specifically with exposure to cigarette smoking in lungs. SBS5 is known to accompany the smoking signature but it is also present in many other cancer types. Previous studies suggested that cigarette smoking stimulates an inflammatory response [4]. Consistent with these findings, the genes identified by GeneSigNet as influenced by SBS4 and SBS5 SigStates are indeed enriched with immune response genes (Fig. 4). In addition, tobacco

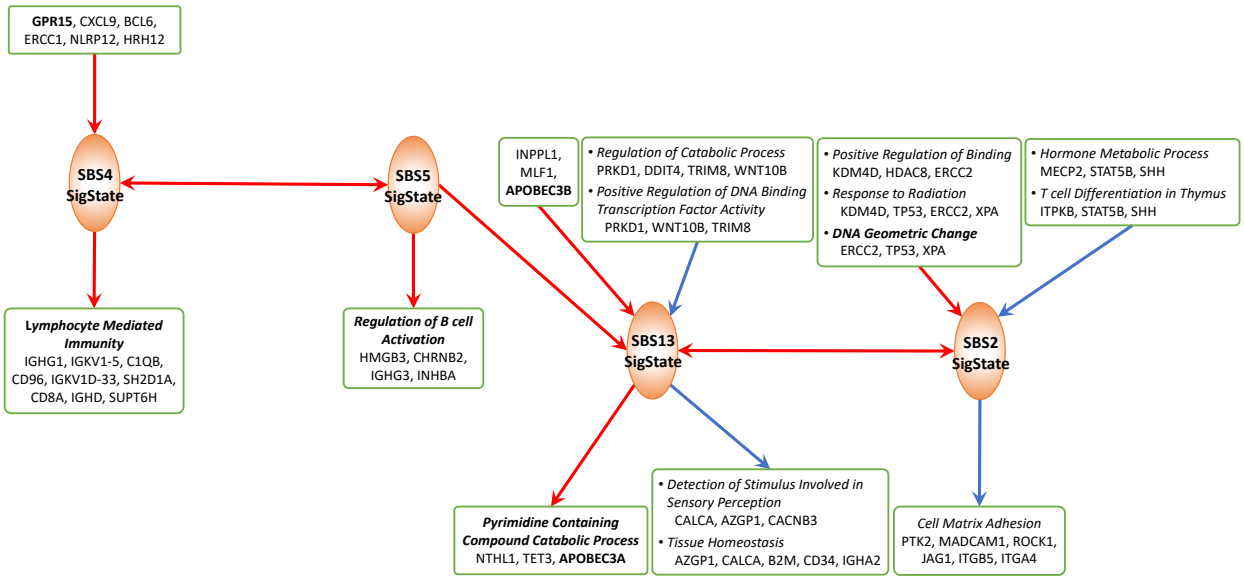

Fig. 4. Co-occurrence of the two SigStates corresponding to the smoking-related signatures SBS4 and SBS5 influences the SigStates corresponding to the APOBEC related signatures SBS13 and SBS2. An extended network including genes and GO terms associated with SigStates is provided in Figure S2 in Supplementary Information.

smoking is known to induce GPR15-expressing T cells; although the exact role of GPR15 in response to smoking is yet to be elucidated [50]. Consistent with previous studies, GeneSigNet inferred a strong association between GPR15 and SBS4 (without resolving the direction). In addition, the results of GeneSigNet suggest that GPR15 is involved in the negative regulation of several genes related to chemotaxis, including IL10, a cytokine with potent anti-inflammatory properties, and has a positive impact on lymphocyte migration and leukocyte mediated cytotoxicity (Fig. 5).

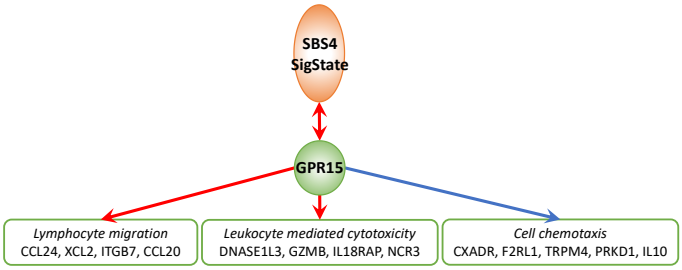

Fig. 5. SBS4-induced GPR15 expression contributes to the activation of immune responses.

GeneSigNet also identified the influence of the signatures SBS4 and SBS5 on two APOBEC signatures – SBS2 and SBS13. The APOBEC signatures are associated with immune response and this relationship is consistent with the previously proposed immune activation due to smoking exposure [51]. Finally, GeneSigNet correctly captured the association of SBS13 (consequently SBS2) with the expressions of APOBEC3B and APOBEC3A enzymes, and also identified the association of SBS13 with pyrimidine related catabolic processes, potentially reflecting the fact that SBS13 involves a pyrimidine to pyrimidine mutation (Fig. 4).

GeneSigNet points to the role of DNA geometric changes for APOBEC signature SBS2 As discussed earlier, APOBEC can only act on single-stranded DNA (ssDNA). Interestingly, one of the GO terms associated with SBS2 SigState identified by GeneSigNet is DNA geometric change (Fig. 4). DNA geometric changes are local changes of DNA conformation such as bulky DNA adducts (a type of DNA damage due to exposure to cigarette smoke) or DNA secondary structures such as ZDNA, cruciform, or quadruplex. Indeed, these structures often involve the formation of ssDNA regions which, in turn, provide mutation opportunities for APOBEC enzymes [52,53,54]. The formation of DNA secondary structures is often associated with DNA supercoiling - a form of DNA stress that is resolved by Topoisomerase 1 (TOP1). Interestingly, GeneSigNet identified a negative influence of TOP1 expression on one of the genes (XPA) contributing to this GO term. This suggests a relation between DNA stress mediated by TOP1 and APOBEC activity.

3 Discussion

Elucidating the nature of mutagenic processes and their interactions with cellular processes is of fundamental importance for understating cancer etiology and guiding cancer therapy. Here, we propose GeneSigNet, a new network-based approach which infers the relation between gene expression and the strength of mutation patterns (signature exposures) allowing to uncover the relations between signatures and processes involved in DNA repair and immune response among other cellular processes. Recognizing the limitations of the previous clustering-based approach, GeneSigNetrelies on a construction of a sparse directed network. For each node (gene or SigState), it selects a sparse set of incoming edges representing dominating incoming effects so that their combination explains the activity of the node. Aiming to capture the most direct influences, the method utilizes sparse partial correlation coefficients. In general, the inference of direction of influential relations from statistical dependencies is highly challenging and GeneSigNet provides an important step toward this direction that is independent of the specific application considered in this study.

Overall the relations discovered by GeneSigNet are consistent with the current knowledge, boosting the confidence in the method's applicability. In addition, GeneSigNet provided several new biological insights concerning the relation between mutagenic processes and other cellular processes. For example,

the uncovered relation between APOBEC hypermutation and activation of regulatory T-Cell can have an important implication in immunotherapy. We note that focusing on a sparse set of edges reduces the power of GO enrichment analysis and requires more specific biological knowledge for interpreting the results. Yet, this potential disadvantage is compensated by the compelling mechanistic insights provided by the method.

4 Materials and Methods

4.1 Breast cancer data

The normalized gene expression data for 266 breast cancer (BRCA) patients were downloaded from Table S7 in [55,56]. Gene expression profiles for 2,204 genes involved in either DNA metabolic or immune response processes of the Gene Ontology (GO) database were selected for the analysis.

For mutational signatures, somatic mutation data were downloaded from the ICGC data portal (https://daco.icgc.org, release 22). The 3,479,652 point mutations were assigned to mutational signatures using SigMa [11]. SigMa divided all mutations into two groups, close-by **C**lustered and **D**ispersed mutations, and assigned these to 12 COSMIC v2 signatures which were previously identified as active in BRCA (Signatures 1, 2, 3, 5, 6, 8, 13, 17, 18, 20, 26 and 30; https://cancer.sanger.ac.uk/cosmic/signatures). From the signatures classified by SigMa as described above, signature phenotype profiles 1D, 2C/D, 3C/D, 5D, 8C/D, and 13C/D that had exposure levels of at least 10% within each group were selected for further analysis (the numbering refers to the COSMIC signature index and C/D denotes signatures attributed to clustered and dispersed mutations). Examining their correlation patterns among patients, some of the signatures were grouped as follows: Signatures 3C/D and 8D were combined into DSB (double-stranded DNA break repair) related signatures, and Signatures 2C and 13C/D into APOBEC related signatures. The remaining signatures are treated separately, resulting in Signature 1, 2D, 5, APOBEC, DSB. A log transformation was consequently performed on exposures of each signature to make its distribution shape closer to a bell curve of normality.

Furthermore, we included binary information of homologous recombination deficiency as an additional variable in the analysis. The binary alteration information was obtained by aggregating functional inactivation information for BRCA1/BRCA2 and 16 other HR genes as provided in Supplementary Tables 4a and 4b of Davies *et al.* [31]. The positive entries were assigned a real value of 4.218 in the SPCS model (Section S2 in Supplementary Information) with the hyperparameter search for the best performance in terms of the means of minimum least square errors and maximum Pearson correlation between responses and predictions over all nodes.

4.2 Lung adenocarcinoma data

The expression data (RNA-seq) of the lung adenocarcinoma (LUAD) from The Cancer Genome Atlas (TCGA) project were downloaded from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) on 2020-06-05. Normalization and variance-stabilizing transformation (vst) of HTSeq count data were performed using DESeq2. Tumor and normal samples were split into different groups and only one sample per donor was kept in each group.

The TCGA LUAD exome mutation spectra were downloaded from Synapse (accession number: syn11801889) and decomposed into COSMIC v3 signatures SBS1, SBS2, SBS4, SBS5, SBS13, SBS40, and SBS45 using the quadratic programming (QP) approach available in the R package SignatureEstimation [57]. Only signatures predominantly active in lung cancer (signatures that were present in at least 5% of samples and were responsible for at least 1% of mutations) were considered based on the initial sample decomposition provided by Alexandrov et al. [2] (Synapse accession number: syn11804065). Signature SBS45 is likely a sequencing artifact so it was omitted from further analyses presented in this study. The same log transformation used in BRCA analysis was performed on signature exposure data here.

We analyzed 466 tumor samples that had both gene expression and mutational signature exposure data available. We analyzed 2,433 genes belonging to the DNA metabolic process and immune system process in GO terms (genes that are not expressed in at least 10% of the samples were omitted). The gene expression and mutational signature exposure data were combined to form an input data matrix.

Acknowledgements

This study was supported by the Intramural Research Programs of the National Library of Medicine (NLM), National Institutes of Health, USA. We thank Jan Hoinka for language editing and proofreading of the manuscript.

References

1. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*, 3(1):246–259, Jan 2013.
2. L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, G. Getz, S. G. Rozen, M. R. Stratton, L. B. Alexandrov, E. N. Bergstrom, A. Boot, P. Boutros, K. Chan, K. R. Covington, A. Fujimoto, G. Getz, D. A. Gordenin, N. J. Haradhvala, M. N. Huang, S. M. A. Islam, M. Kazanov, J. Kim, L. J. Klimczak, N. Lopez-Bigas, M. Lawrence, I. Martincorena, J. R. McPherson, S. Morganella, V. Mustonen, H. Nakagawa, A. W. Tian Ng, P. Polak, S. Prokopec, S. A. Roberts, S. G. Rozen, R. Sabarinathan, N. Saini, T. Shibata, Y. Shiraishi, M. R. Stratton, B. T. Teh, I. Vázquez-García, D. A. Wheeler, Y. Wu, F. Yousif, and W. Yu. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 02 2020.
3. Y. A. Kim, M. D. M. Leiserson, P. Moorjani, R. Sharan, D. Wojtowicz, and T. M. Przytycka. Mutational Signatures: From Methods to Mechanisms. *Annu Rev Biomed Data Sci*, 4:189–206, 07 2021.
4. L. B. Alexandrov, Y. S. Ju, K. Haase, P. Van Loo, I. Martincorena, S. Nik-Zainal, Y. Totoki, A. Fujimoto, H. Nakagawa, T. Shibata, P. J. Campbell, P. Vineis, D. H. Phillips, and M. R. Stratton. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, 11 2016.
5. R. Nowarski and M. Kotler. APOBEC3 cytidine deaminases in double-strand DNA break repair and cancer promotion. *Cancer Res*, 73(12):3494–3498, Jun 2013.
6. Y. A. Kim, D. Wojtowicz, R. Sarto Basso, I. Sason, W. Robinson, D. S. Hochbaum, M. D. M. Leiserson, R. Sharan, F. Vadin, and T. M. Przytycka. Network-based approaches elucidate differences within APOBEC and clock-like signatures in breast cancer. *Genome Med*, 12(1):52, 05 2020.
7. A. Viel, A. Bruselles, E. Meccia, M. Fornasarig, M. Quaia, V. Canzonieri, E. Policicchio, E. D. Urso, M. Agostini, M. Genuardi, E. Lucci-Cordisco, T. Venesio, A. Martayan, M. G. Diodoro, L. Sanchez-Mete, V. Stigliano, F. Mazzei, F. Grasso, A. Giuliani, M. Baiocchi, R. Maestro, G. Giannini, M. Tartaglia, L. B. Alexandrov, and M. Bignami. A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine*, 20:39–49, Jun 2017.
8. J. Kim, K. W. Mouw, P. Polak, L. Z. Braunstein, A. Kamburov, D. J. Kwiatkowski, J. E. Rosenberg, E. M. Van Allen, A. D'Andrea, and G. Getz. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet*, 48(6):600–606, 06 2016.
9. Xueqing Zou, Michel Owusu, Rebecca Harris, Stephen P Jackson, Joanna I Loizou, and Serena Nik-Zainal. Validating the concept of mutational signatures with isogenic cell models. *Nature communications*, 9(1):1–16, 2018.
10. S. Volinia, T. Druck, C. A. Paisie, M. S. Schrock, and K. Huebner. The ubiquitous 'cancer mutational signature' 5 occurs specifically in cancers with deleted FHIT alleles. *Oncotarget*, 8(60):102199–102211, Nov 2017.
11. D. Wojtowicz, I. Sason, X. Huang, Y. A. Kim, M. D. M. Leiserson, T. M. Przytycka, and R. Sharan. Hidden Markov models lead to higher resolution maps of mutation signature activity in cancer. *Genome Med*, 11(1):49, 07 2019.
12. D. Silverbush and R. Sharan. Network orientation via shortest paths. *Bioinformatics*, 30(10):1449–1455, May 2014.
13. A. Gitter, J. Klein-Seetharaman, A. Gupta, and Z. Bar-Joseph. Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res*, 39(4):e22, Mar 2011.
14. A. Vinayagam, U. Stelzl, R. Foulle, S. Plassmann, M. Zenkner, J. Timm, H. E. Assmus, M. A. Andrade-Navarro, and E. E. Wanker. A directed protein interaction network for investigating intracellular signal transduction. *Sci Signal*, 4(189):rs8, Sep 2011.
15. D. Silverbush and R. Sharan. A systematic approach to orient the human protein-protein interaction network. *Nat Commun*, 10(1):3015, 07 2019.
16. J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial Correlation Estimation by Joint Sparse Regression Models. *J Am Stat Assoc*, 104(486):735–746, Jun 2009.
17. D. Yu, J. Lim, X. Wang, F. Liang, and G. Xiao. Enhanced construction of gene regulatory networks using hub gene information. *BMC Bioinformatics*, 18(1):186, Mar 2017.

18. N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–620, 2000.
19. Y. Yuan, C. T. Li, and O. Windram. Directed partial correlation: inferring large-scale gene regulatory network through induced topology disruptions. *PLoS One*, 6(4):e16835, Apr 2011.
20. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, 5(9), Sep 2010.
21. R. Opgen-Rhein and K. Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst Biol*, 1:37, Aug 2007.
22. Y. Kang, D. Thieffry, and L. Cantini. Evaluating the Reproducibility of Single-Cell Gene Regulatory Network Inference Algorithms. *Front Genet*, 12:617282, 2021.
23. V. A. Huynh-Thu and P. Geurts. dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Sci Rep*, 8(1):3384, 02 2018.
24. M. Jager, F. Blokzijl, E. Kuijk, J. Bertl, M. Vougioukalaki, R. Janssen, N. Besselink, S. Boymans, J. de Ligt, J. S. Pedersen, J. Hoeijmakers, J. Pothof, R. van Boxtel, and E. Cuppen. Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. *Genome Res*, 29(7):1067–1077, 07 2019.
25. W. Wiedermann and J. Sebastian. Direction Dependence Analysis in the Presence of Confounders: Applications to Linear Mediation Models Using Observational Data. *Multivariate Behav Res*, 55(4):495–515, 2020.
26. Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.
27. A. Lachmann, H. Xu, J. Krishnan, S. I. Berger, A. R. Mazloom, and A. Ma'ayan. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, 26(19):2438–2444, Oct 2010.
28. F. Supek and B. Lehner. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell*, 170(3):534–547, Jul 2017.
29. D. Mas-Ponte and F. Supek. DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. *Nat Genet*, 52(9):958–968, 09 2020.
30. B. J. Taylor, S. Nik-Zainal, Y. L. Wu, L. A. Stebbings, K. Raine, P. J. Campbell, C. Rada, M. R. Stratton, and M. S. Neuberger. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife*, 2:e00534, Apr 2013.
31. H. Davies, D. Glodzik, S. Morganella, L. R. Yates, J. Staaf, X. Zou, M. Ramakrishna, S. Martin, S. Boyault, A. M. Sieuwerts, P. T. Simpson, T. A. King, K. Raine, J. E. Eyfjord, G. Kong, A. Borg, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, A. L. Børresen-Dale, J. W. Martens, P. N. Span, S. R. Lakhani, A. Vincent-Salomon, C. Sotiriou, A. Tutt, A. M. Thompson, S. Van Laere, A. L. Richardson, A. Viari, P. J. Campbell, M. R. Stratton, and S. Nik-Zainal. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med*, 23(4):517–525, Apr 2017.
32. R. Prakash, Y. Zhang, W. Feng, and M. Jasin. Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. *Cold Spring Harb Perspect Biol*, 7(4):a016600, Apr 2015.
33. M. E. Moynahan and M. Jasin. Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nat Rev Mol Cell Biol*, 11(3):196–207, Mar 2010.
34. X. Li and W. D. Heyer. Homologous recombination in DNA repair and DNA damage tolerance. *Cell Res*, 18(1):99–113, Jan 2008.
35. V. B. Seplyarskiy, R. A. Soldatov, K. Y. Popadin, S. E. Antonarakis, G. A. Bazykin, and S. I. Nikolaev. APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res*, 26(2):174–182, Feb 2016.
36. N. Kanu, M. A. Cerone, G. Goh, L. P. Zalmas, J. Bartkova, M. Dietzen, N. McGranahan, R. Rogers, E. K. Law, I. Gromova, M. Kschischo, M. I. Walton, O. W. Rossanese, J. Bartek, R. S. Harris, S. Venkatesan, and C. Swanton. DNA replication stress mediates APOBEC3 family mutagenesis in breast cancer. *Genome Biol*, 17(1):185, 09 2016.
37. C. J. Sakofsky, N. Saini, L. J. Klimczak, K. Chan, E. P. Malc, P. A. Mieczkowski, A. B. Burkholder, D. Fargo, and D. A. Gordenin. Repair of multiple simultaneous double-strand breaks causes bursts of genome-wide clustered hypermutation. *PLoS Biol*, 17(9):e3000464, 09 2019.
38. K. Chan and D. A. Gordenin. Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. *Annu Rev Genet*, 49:243–267, 2015.
39. T. A. Chan, M. Yarchoan, E. Jaffee, C. Swanton, S. A. Quezada, A. Stenzinger, and S. Peters. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann Oncol*, 30(1):44–56, 01 2019.
40. S. Venkatesan, R. Rosenthal, N. Kanu, N. McGranahan, J. Bartek, S. A. Quezada, J. Hare, R. S. Harris, and C. Swanton. Perspective: APOBEC mutagenesis in drug resistance and immune escape in HIV and cancer evolution. *Ann Oncol*, 29(3):563–572, 03 2018.

41. A. Facciabene, G. T. Motz, and G. Coukos. T-regulatory cells: key players in tumor immune escape and angiogenesis. *Cancer Res*, 72(9):2162–2171, May 2012.
42. A. Y. Rudensky. Regulatory T cells and Foxp3. *Immunol Rev*, 241(1):260–268, May 2011.
43. Y. Chung, S. Tanaka, F. Chu, R. I. Nurieva, G. J. Martinez, S. Rawal, Y. H. Wang, H. Lim, J. M. Reynolds, X. H. Zhou, H. M. Fan, Z. M. Liu, S. S. Neelapu, and C. Dong. Follicular regulatory T cells expressing Foxp3 and Bcl-6 suppress germinal center reactions. *Nat Med*, 17(8):983–988, Jul 2011.
44. H. M. Chen, W. van der Touw, Y. S. Wang, K. Kang, S. Mai, J. Zhang, D. Alsina-Beauchamp, J. A. Duty, S. K. Mungamuri, B. Zhang, T. Moran, R. Flavell, S. Aaronson, H. M. Hu, H. Arase, S. Ramanathan, R. Flores, P. Y. Pan, and S. H. Chen. Blocking immunoinhibitory receptor LILRB2 reprograms tumor-associated myeloid cells and promotes antitumor immunity. *J Clin Invest*, 128(12):5647–5662, 12 2018.
45. L. M. Sholl, F. R. Hirsch, D. Hwang, J. Botling, F. Lopez-Rios, L. Bubendorf, M. Mino-Kenudson, A. C. Roden, M. B. Beasley, A. Borczuk, E. Brambilla, G. Chen, T. Y. Chou, J. H. Chung, W. A. Cooper, S. Dacic, S. Lantuejoul, D. Jain, D. Lin, Y. Minami, A. Moreira, A. G. Nicholson, M. Noguchi, M. Papotti, G. Pelosi, C. Poleri, N. Rekhtman, M. S. Tsao, E. Thunnissen, W. Travis, Y. Yatabe, A. Yoshida, J. B. Daigneault, A. Zehir, S. Peters, I. I. Wistuba, K. M. Kerr, and J. W. Longshore. The Promises and Challenges of Tumor Mutation Burden as an Immunotherapy Biomarker: A Perspective from the International Association for the Study of Lung Cancer Pathology Committee. *J Thorac Oncol*, 15(9):1409–1424, 09 2020.
46. S. Wang, M. Jia, Z. He, and X. S. Liu. APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. *Oncogene*, 37(29):3924–3936, 07 2018.
47. D. L. Faden, F. Ding, Y. Lin, S. Zhai, F. Kuo, T. A. Chan, L. G. Morris, and R. L. Ferris. APOBEC mutagenesis is tightly linked to the immune landscape and immunotherapy biomarkers in head and neck squamous cell carcinoma. *Oral Oncol*, 96:140–147, 09 2019.
48. R. Saleh and E. Elkord. Treg-mediated acquired resistance to immune checkpoint inhibitors. *Cancer Lett*, 457:168–179, 08 2019.
49. D. R. Principe, L. Chiec, N. A. Mohindra, and H. G. Munshi. Regulatory T-Cells as an Emerging Barrier to Immune Checkpoint Inhibition in Lung Cancer. *Front Oncol*, 11:684098, 2021.
50. S. Kõks and G. Kõks. Activation of GPR15 and its involvement in the biological effects of smoking. *Exp Biol Med (Maywood)*, 242(11):1207–1212, 06 2017.
51. P. A. Patriarca, W. H. Foege, and T. A. Swartz. Progress in polio eradication. *Lancet*, 342(8885):1461–1464, Dec 1993.
52. F. Kouzine, D. Wojtowicz, L. Baranello, A. Yamane, S. Nelson, W. Resch, K. R. Kieffer-Kwon, C. J. Benham, R. Casellas, T. M. Przytycka, and D. Levens. Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. *Cell Syst*, 4(3):344–356, 03 2017.
53. X. Zou, S. Morganella, D. Glodzik, H. Davies, Y. Li, M. R. Stratton, and S. Nik-Zainal. Short inverted repeats contribute to localized mutability in human somatic cells. *Nucleic Acids Res*, 45(19):11213–11221, Nov 2017.
54. W. M. Guiblet, M. A. Cremona, R. S. Harris, D. Chen, K. A. Eckert, F. Chiaromonte, Y. F. Huang, and K. D. Makova. Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Res*, 49(3):1497–1516, 02 2021.
55. S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman, S. Morganella, M. R. Aure, O. C. Lingjærde, A. Langerød, M. Ringnér, S. M. Ahn, S. Boyault, J. E. Brock, A. Broeks, A. Butler, C. Desmedt, L. Dirix, S. Dronov, A. Fatima, J. A. Foekens, M. Gerstung, G. K. Hooijer, S. J. Jang, D. R. Jones, H. Y. Kim, T. A. King, S. Krishnamurthy, H. J. Lee, J. Y. Lee, Y. Li, S. McLaren, A. Menzies, V. Mustonen, S. O'Meara, I. Pauporté, X. Pivot, C. A. Purdie, K. Raine, K. Ramakrishnan, F. G. Rodríguez-González, G. Romieu, A. M. Sieuwerts, P. T. Simpson, R. Shepherd, L. Stebbings, O. A. Stefansson, J. Teague, S. Tommasi, I. Treilleux, G. G. Van den Eynden, P. Vermeulen, A. Vincent-Salomon, L. Yates, C. Caldas, L. van't Veer, A. Tutt, S. Knappskog, B. K. Tan, J. Jonkers, Å. Borg, N. T. Ueno, C. Sotiriou, A. Viari, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, J. W. Martens, A. L. Børresen-Dale, A. L. Richardson, G. Kong, G. Thomas, and M. R. Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 06 2016.
56. S. et al Nik-Zainal. Author Correction: Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 566(7742):E1, 02 2019.
57. Xiaoqing Huang, Damian Wojtowicz, and Teresa M Przytycka. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics*, 34(2):330–37, 2018.
58. M Tsagris. Bayesian network learning with the PC algorithm: an improved and correct variation. *Applied Artificial Intelligence*, 33(2):101–123, Oct 2019.
59. M. Scutari, C. Vitolo, and A. Tucker. Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation. *Statistics and Computing*, 9(5):1095–1108, Feb 2019.
60. Finn V Jensen and Thomas Dyhre Nielsen. *Bayesian networks and decision graphs*, volume 2. Springer, 2007.

61. Yasunori Fujikoshi, Vladimir V Ulyanov, and Ryoichi Shimizu. *Multivariate statistics: High-dimensional and large-sample approximations*, volume 760. John Wiley & Sons, 2011.
62. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
63. Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.
64. T. Chu, C. Glymour, R. Scheines, and P. Spirtes. A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, 19(9):1147–1152, Jun 2003.
65. Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.
66. Yadolah Dodge and Iraj Yadegari. On direction of dependence. *Metrika*, 72(1):139–150, 2010.

Supplementary Information

S1 Supplementary Methods

In the construction of a Gene-Signature Network (GSN), the nodes corresponding to genes and SigStates are considered as random variables and the activity of these variables consists of expressions of genes and exposures of mutational signatures over samples (Fig. 2A). We first describe a sparse estimation of partial correlation (SPCS) to initialize a partially directed network (Fig. 2B). Then, a higher moment-based strategy is adopted to decide the causality direction of bidirected edges in the network (Fig. 2C). Finally, a matrix normalization, alternate scaling [26], is performed to bring the total incoming and outgoing effects of each node to the same range.

The idea of SPCS is to decompose the problem of inferring a network of n nodes into n subproblems of variable selection. For each subproblem, a single node (gene or SigState) is considered as the focused target and the weights of incoming effects from the other $n-1$ nodes are obtained by minimizing a least square error function subject to an l_1 norm constraint.

In addition, to decide the direction of dependency between two variables (two nodes having effects on each other in the GSN), we first remove confounding effects due to the presence of the other $n-2$ variables (the other $n-2$ nodes in GSN) and utilize the higher moment-based strategy.

Finally, we apply alternate scaling, a matrix normalization method that iteratively maps rows and columns of a matrix onto the a unit space (l_1 norm ball with unit radius). An overview of the method workflow is presented in Section 2.1 and Figure 2 in the main text.

S1.1 A sparse partial correlation selection (SPCS)

Correlation networks are widely used to explore and visualize dependencies in high-dimensional data. However, without assuming prior knowledge, an ordinary correlation itself provides no means to distinguish between causal and affected factors in underlying causal processes [21].

Bayesian networks can be used to infer causal relations of nodes representing their local conditional dependencies via a directed acyclic graph (DAG) [18,58,59]. Alternative to constructing a DAG, directed partial correlation (DPC) [19] and regression tree based GENIE3 [20] have been proposed to uncover conditional dependencies in observed data. However, learning the structure of Bayesian networks from large data is known to be computationally challenging [59]. In addition, these networks are always acyclic and thus they do not support feedback loops [60]. DPC and GENIE3 return a complete list of interactions with non-zero weights of connectivity strengths, hence generating fully-connected networks in which the the choice of an optimal confidence threshold is left open. Other methods such as the sparse partial correlation estimation (SPACE) [16] and its extension (ESPASE) [17], consider a penalized regression approach to construct the gene regulatory network. Although both methods utilize a sparse variable selection, their estimations provide symmetric weight matrices representing undirected-weighted networks.

To address these issues, we modeled causal dependencies as sparse partial correlation coefficients which are obtained by minimizing a least square error subject to the unit l_1 norm ball. Inspired by the theoretical foundations for approximating partial correlations (see Box 1), SPCS selects the best combination of a small number of explanatory factors that, under the conditional dependency assumption, explains the activity of each node in the GSN (Fig. 2B).

Box 1: Partial correlations can be approximated by regression coefficients

Consider the ordinary correlation ρ_{12} between two random variables v_1 and v_2. If v_1 and v_2 are correlated with $n-2$ other variables $v_3, v_4, \ldots, v_n$, we may regard ρ_{12} as a mixture of a direct correlation between v_1 and v_2 and an indirect portion due to the presence of other variables correlating with v_1 and v_2. The partial correlation measuring the direct portion of the total correlation can be defined as a correlation between v_1 and v_2 after removing effects due to other variables by a linear regression and the least square linear regression coefficients are proportional to the partial correlation coefficients [61].

The entire network is represented as a weighted, directed graph, $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, where a set of nodes $\mathbb{V}$ represents genes and SigStates, and a set of edges $\mathbb{E}$ represents the relationships among the nodes.

Let $\mathbb{I} = \mathbb{I}_g \cup \mathbb{I}_s$ denote the index set for n variables representing the types of nodes in $\mathbb{V}$, where $\mathbb{I}_g$ and $\mathbb{I}_s$ are the index sets for the two types of nodes corresponding to m genes and $n-m$ SigStates, respectively. The nodes have observational activities over samples, and a $p \times n$ matrix $X = \{x_{ij}\}$ represents the data consisting of expressions of m genes and exposures of $n-m$ mutational signatures across p samples (patients). Assuming the incoming effects on a focused variable f from its upstream covariates, the observed value x_{if}, corresponding to i-th sample, can be approximated as the following affine combination

$$x_{if} \approx \sum_{k \in \mathbb{I}_g \setminus \{f\}} x_{fk} w_{kf} + \sum_{k \in \mathbb{I}_s \setminus \{f\}} x_{fk} w_{kf} + w_{0f} = \sum_{k \in \mathbb{I} \setminus \{f\}} x_{fk} w_{kf} + w_{0f} \tag{S1}$$

where $\mathbb{I} \setminus \{f\}$ denotes the index set of $n-1$ variables except for the focused response variable f, and $w_{*f} \in R^{n-1}$ denotes the contribution weights to the activation of f from the other $n-1$ variables. Thus, our goal is to find the minimum of the least square error function subject to a unit l_1 norm constraint on w_{*f} as following

$$\begin{aligned} \underset{w_{*f} \in R^{n-1}, w_{0f} \in R}{\text{minimize}} \quad & \sum_{i=1}^{p} \left(x_{if} - \sum_{k \in \mathbb{I} \setminus \{f\}} x_{ik} w_{kf} - w_{0f} \right)^2 \\ \text{subject to} \quad & \sum_{k \in \mathbb{I} \setminus \{f\}} |w_{kf}| \leq 1 \end{aligned} \tag{S2}$$

where w_{0f} denotes the intercept adjusting the fitness between the response variable and its prediction. For a focused node (affected) f, the weight vector $w_{*f} = (w_{1f}, w_{2f}, \dots w_{f-1f}, w_{f+1f}, \dots w_{nf})^T$, a solution of the problem in Equation (S2), represents the weights of incoming effects from the other $n-1$ nodes (causal factors) in the network. The l_1 norm constraint is used to avoid over-fitting issues and allow only dominant incoming effects as causal factors. Thus, a non-zero w_{kf} $(k \neq f)$, selected for the node f, denotes the partial correlation coefficient representing the potential effect of the node k on the node f. In this setting, focusing on the activity of every single node in the network, the optimization problem in Equation (S2) considers all possible combinations of incoming effects from the other $n-1$ nodes and selects the best combination with their optimal influence weights to explain the focused activity under conditional dependency assumption. Therefore, the causality relationship between nodes k and f is estimated in the presence of the other $n-2$ variables. Note that the inequality constraint in Equation (S2) is proposed to provide a flexible estimation of total weights of incoming effects for variable f. It is based on the assumption that the total incoming effects on a gene or SigState can be different from others depending on its responses to regulatory mechanisms.

S1.2 Solving SPCS model

An accurate solution of the problem in Equation (S2) is critical for the robust estimation of causality flows in the Gene-Signature Network. Although the least square error function in Equation (S2) is convex, the l_1 norm sparsity constraint is non-differentiable and derivative-based techniques such as Lagrange multipliers and Karush-Kuhn-Tucker (KKT) conditions are not directly applicable due to the non-smoothness. Another attempt to resolve such an issue is to decompose the inequality of the l_1 norm into 2^n inequality constraints [62]. However, biological networks are often large in scale and it is practically difficult to accurately minimize such a large scale objective function over the exponential number of constraints within a reasonable time. Thus, to solve the non-smooth constrained optimization, we rewrite the initial formulation in Equation (S2) as an unconstrained form with a penalty term

$$\underset{w_{*f} \in R^{n-1}, w_{0f} \in R}{\text{minimize}} \quad \sum_{i=1}^{p} \left(x_{if} - \sum_{k \in \mathbb{I} \setminus \{f\}} x_{ik} w_{kf} - w_{0f} \right)^2 + \lambda_f \cdot \sum_{k \in \mathbb{I} \setminus \{f\}} |w_{kf}| \tag{S3}$$

where a tuning parameter λ_f controls the strength of the penalty term, chosen for a focused variable f to provide a balance between the least square error term and the l_1 norm constraint in the formulation in Equation (S2). For a given f, the criterion to define a reasonable value of λ_f is not trivial, in general, due to

the incompleteness of information linked to biological relevance. In fact, the exact relationship between the radius of the l_1 norm ball in Equation (S2) and the tuning parameter λ_f in Equation (S3) is data-dependent. Therefore, it is reasonable to use a data-driven strategy for choosing λ_f. Akaike information criterion (AIC) is a statistical technique that provides the relative quality of statistical models for a given data by combining the maximum likelihood estimation of fitness with the number of parameters for inference [63]. The AIC is used in this work to decide the value of λ_f providing a solution with reasonable total incoming effect on f from its dominating factors.

S1.3 Higher moment-based strategy for causality direction

The solution to the problem in Equation (S2) may provide similar weights in both directions for some pairs of variables due to the presence of effects from confounding factors and noise in addition to their real dependencies. This uncertainty may require a complimentary analysis to decide the direction of causal relations for these pairs.

One way to decide causality relations is to perform perturbation experiments. However, optimization of experimental design to predict which combination of perturbations allows to discover causality flows in a given network topology is often challenging and costly [64]. Hence, revealing causality directions by analyzing purely observational data has become a special focus of network biology [65]. Under a confounder-free assumption, higher moment statistics [66] indicate causality direction between two dependent variables from purely observational data. Alternative to the directionality decision for bivariate distributions, a confounder model [25] was recently designed to assign causality directions for several factors under a standard dependency assumption. By combining the key ideas of the two methods, we propose a higher moment-based strategy to decide causality directions for the bidirected edges in the network obtained using the SPCS. The idea is to generate a bivariate distribution for a focused pair by removing confounding effects from their observed values, and then decide the causality direction between them using higher moment statistics on the corresponding residuals (Fig. 2C).

Specifically, for a pair of variables having similar effects on each other in W, we first calculate their residuals by removing effects due to the presence of the other $n-2$ variables [61]. Upon the removal, the pair of residuals can be assumed to follow a bivariate distribution and only the dependency between the focused pair remains in their residuals. Thus, the causal variable can be distinguished from the affected by comparing the higher moments of the two residual distributions. Under the confounder-free assumption, the causal factor is closer to normality than the affected and the skewness and kurtosis are the higher moment statistics used to measure close-normality of the residual distributions.

Let x_{*j} and x_{*f} be j-th and f-th columns of the given data matrix X representing the activities of the variables j and f respectively. Then, the residuals corresponding to the variables j and f can be obtained as

$$\begin{aligned} r_j &= x_{*j} - \sum_{k \in \mathbb{I} \setminus \{j,f\}} x_{*k} w_{kj} \\ r_f &= x_{*f} - \sum_{k \in \mathbb{I} \setminus \{f,j\}} x_{*k} w_{kf} \end{aligned} \tag{S4}$$

where r_j and r_f are column vectors of size p, representing the residuals of variables j and f after removing effects due to the other $n-2$ variables besides j and f. As described in [66], the causal factor is closer to the normal distribution than the affected under a confounder-free assumption, and the higher moment statistics, skewness and kurtosis can be used to measure the close-normality of the residual distributions. Comparing the distribution shapes of r_j and r_f, we assign a causality direction between j and f as following

$$Edge(j,f) = \begin{cases} j \to f \text{ and } w_{fj} := 0, & \text{if } |w_{jf}| > |w_{fj}| \text{ and } |\delta_j| > |\delta_f| \text{ and } |\gamma_j| > |\gamma_f| \\ j \leftarrow f \text{ and } w_{jf} := 0, & \text{if } |w_{jf}| < |w_{fj}| \text{ and } |\delta_j| < |\delta_f| \text{ and } |\gamma_j| < |\gamma_f| \\ j \leftrightarrow f, & \text{otherwise} \end{cases} \tag{S5}$$

where $\delta_j = E[(r_j - \mu_{r_j})^3]/\sigma^3_{r_j}$ and $\delta_f = E[(r_f - \mu_{r_f})^3]/\sigma^3_{r_f}$ describe the skewness while $\gamma_j = E[(r_j - \mu_{r_j})^4]/\sigma^4_{r_j} - 3$ and $\gamma_f = E[(r_f - \mu_{r_f})^4]/\sigma^4_{r_f} - 3$ describe the kurtosis of residuals r_j and r_f, respectively, where $\mu_.$ and $\sigma_.$ are the mean and standard deviation of the respective variables. In practice, we accept

the direction with stronger weight if one direction provides a stronger weight than the threshold τ while the opposite direction provides a weaker weight compared to τ. The edge with the smaller weight is consequently removed from the network. The higher moment-based strategy in Equation (S5) is used to decide the direction if $|w_{fj}| \geq \tau$ and $|w_{jf}| \geq \tau$.

We explored different thresholds for edge weight cut-off τ to obtain the best set of directed edges. The optimal threshold value was chosen ($\tau = 0.0391$ for BRCA and $\tau = 0.0521$ for LUAD) to maximize the fraction of consistent directions in the set of recovered edges.

Complementary to the partial correlation measurement, the higher moment-based strategy provides a causality direction between two correlated variables based on their distribution shapes and locations if the dependency direction is not solved by the SPCS. Particularly, this strategy is proposed to remove false directed-edges from the initial partial correlation network.

S1.4 Normalization of incoming and outgoing effects

The magnitudes of the causal effects in the network may provide valuable information to prioritize the candidate associations of genes and SigStates with underlining biological processes because the edge weights denote the contribution scores from causal factors to their affected targets. Hence, it is reasonable to bring the total incoming and outgoing effects of nodes to the same range in the GSN. The total incoming effect on each node was attempted to be normalized into the l_1 norm constraint in Equation (S2). However, the total outgoing effects are free from normalization. Moreover, an additional update described in Equation (S5) performed to remove edges from the initial network obtained by the SPCS model in Equation (S2).

We adopted a matrix normalization technique, alternate scaling [26], to rescale columns and rows of the weight matrix W into the unit l_1 norm ball. This procedure begins with rows in which each is mapped into the unit ball. Then do the same operation on columns, then on rows, and so on, until the sequence of matrices converges. The absolute difference of two consequence updates, by rows (W_{rows}) and by columns ($W_{columns}$), is used as the convergence criterion such that $\|W_{rows} - W_{columns}\|_F < 10^{-15}$. In the cases of the BRCA and LUAD analysis, the convergence was achieved after only 5 and 6 iterations, respectively. Note that this normalization increases the sparsity of the GSN since every column and row of W is iteratively mapped onto the l_1 norm space ($\|w_{k*}\|_{l_1} \leq 1$ for k-th row and $\|w_{*k}\|_{l_1} \leq 1$ for k-th column of W) which rescales the weights to lower values, even assigning zero weights to weak associations during the iterative procedure.

S2 Supplementary Figures

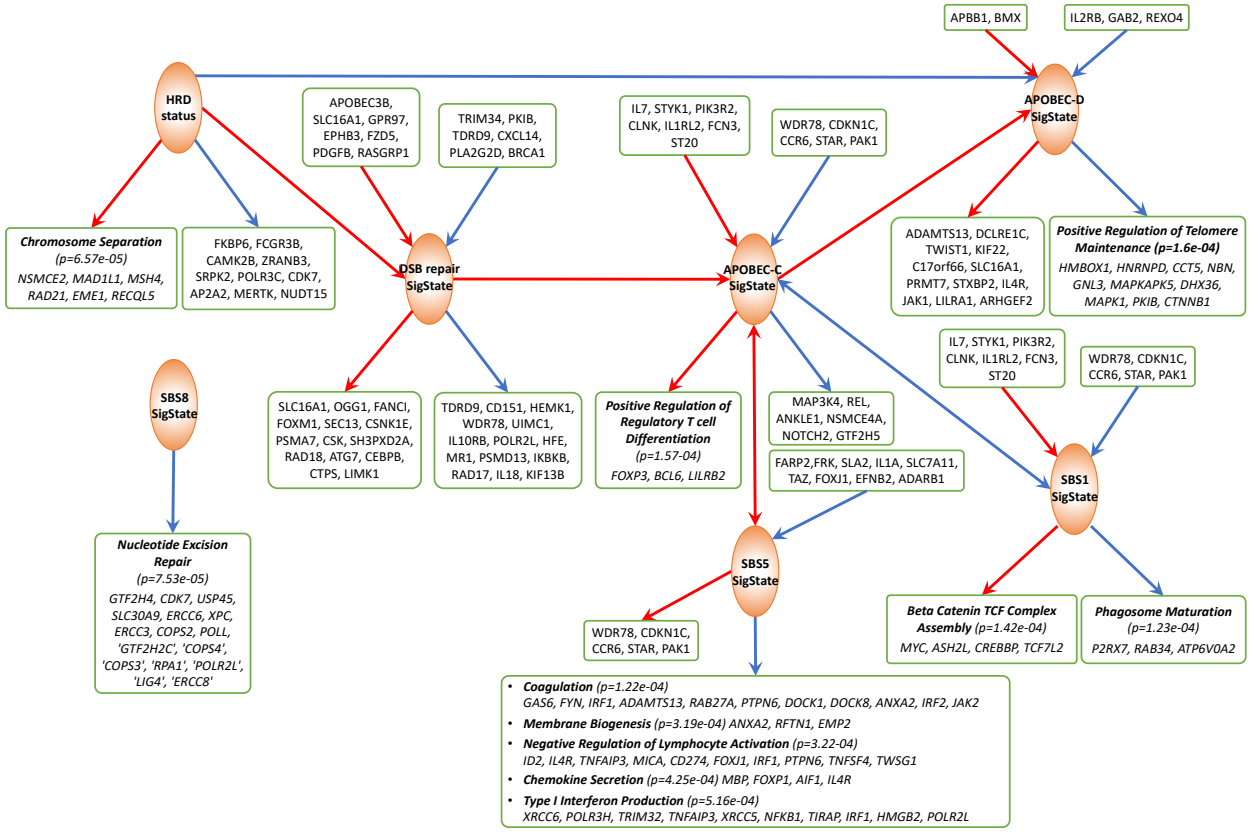

Fig. S1. Information flow over the SigStates in BRCA is initiated by activation of the HRD related mutations which gives arise to base substitution load in the DSB repair SigState. Consequently, the increased DSB mutations then produce a carry-over effect on the other SigStates in its downstream. Statistically significant GO terms (q-value < 0.01), enriched for the sets of genes whose expression status were affected by SigStates are shown as corner-rounded rectangle nodes. Genes in the upstream or downstream of SigStates are shown in the rectangles if the corresponding partial correlations are strong ($|w_{ij}| \geq 0.01$) even if the corresponding GO terms are not strongly significant (q-value>0.01). The upstream and downstream sets for SigStates and the corresponding GO terms enriched for the genes in those sets are available as a resource at the GeneSigNet Github repository (`https://github.com/ncbi/GeneSigNet`).

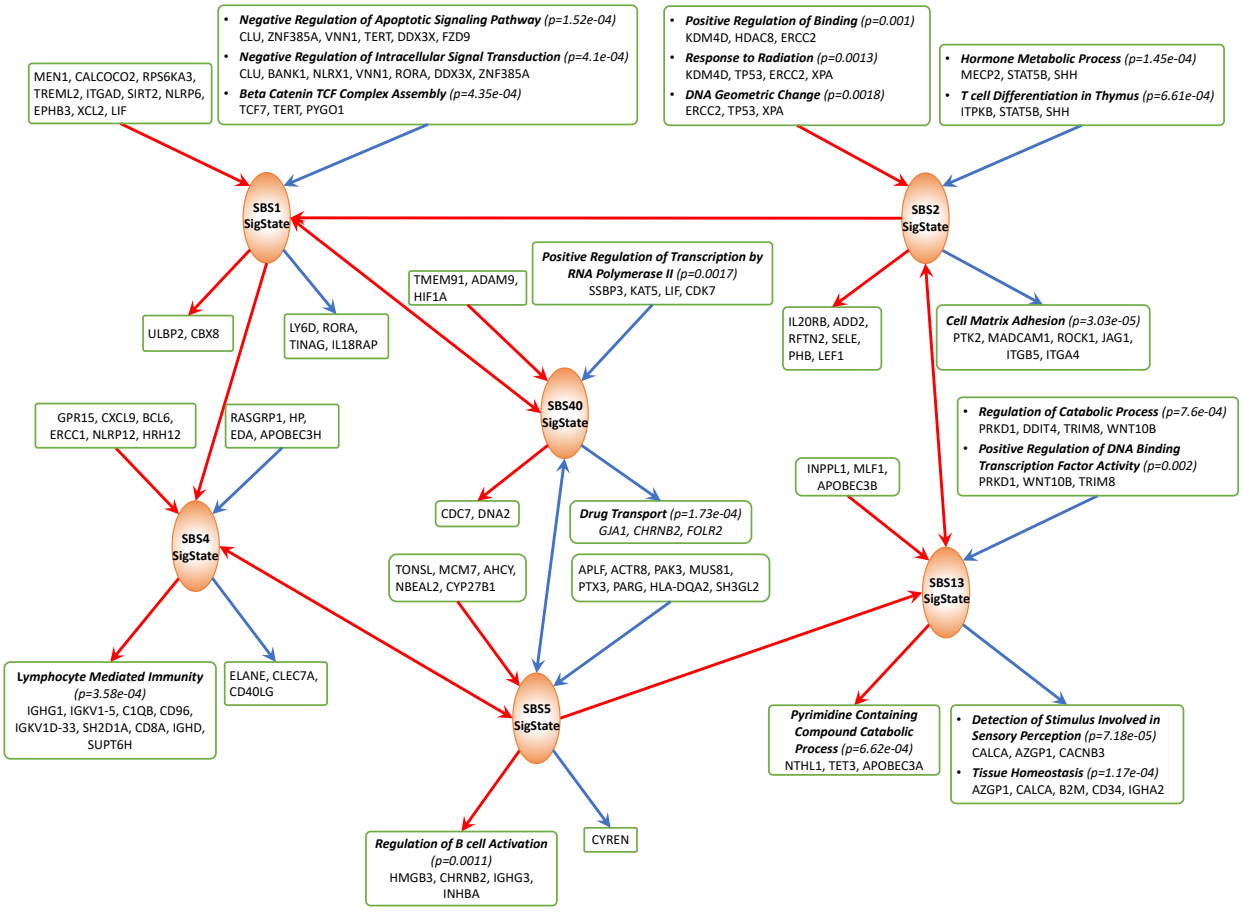

Fig. S2. Information flow over the SigStates in LUAD analysis. Statistically significant GO terms (q-value < 0.01), enriched for the sets of genes whose expression status were influencing in or affected by SigStates are shown as rounded rectangle nodes. Set of genes in upstream or downstream of SigStates are shown the rectangles if the corresponding partial correlations are strong ($|w_{ij}| \geq 0.01$) even if the corresponding GO terms are not strongly significant (q-value>0.01). The upstream and downstream sets for SigStates and the corresponding GO terms enriched for the genes in those sets are available as a resource at the GeneSigNet Github repository (`https://github.com/ncbi/GeneSigNet`).