Deregulation cascades in carcinogenesis

Roberto Herrero, Jean Pierre Gomez, Gabriel Gil and Augusto Gonzalez

Institute of Cybernetics, Mathematics and Physics, Havana, Cuba

Carcinogenesis is one example of transition between two biological states involving strong gene expression rearrangements **[1]**. Normal and tumor states are understood as two different states of a Gene Regulatory Network **[2].** In terms of the normal state characteristics, the transition may be thought as a deregulation process **[3,4,5]**. According to a very abstract model for the transition, the available data on cancer risk support the idea of a single big jump in expression space **[6]**, which may be associated to a deregulation cascade. Here, we use the measured frequency distribution of gene deregulations, and concepts from the probabilistic theory of causation **[7]** in order to infer causal relations between pairs of genes and construct a deregulation causal network **[8]**. Then, the deregulated genes in each sample are organized according to the network in such a way that they show the deregulation cascade behind the transition from the normal to the tumor state. The explicit analysis for prostate adenocarcinoma is given. In most cases, the cascade happens to be unique and initiated by a single deregulation event. Using results from a companion paper **[9]**, we show that cascades conform classes which may be labeled by the deregulations in a predefined panel of 15 genes. The results of the paper may be checked in experiments with cellular lines or in animal models, and could have an impact on personalized cancer therapy .

The data

Gene-expression data for prostate adenocarcinoma (PRAD) was downloaded from the TCGA portal **[10,11]**. RNAseq measurements in FPKM format are reported for 52 normal tissue and 499 tumor samples. The way the data is processed is briefly described in the **Methods** section. For each sample, we end up with a vector of binary data: $\mathbf{e}_{\mathbf{sample}} = (e_1, e_2, \ldots, e_{60483})$, where for the

expression of gene i we have only two possibilities, either $e_i = 0$ (normal expression) or 1 (deregulated). The number 60483 is the measured number of genes in each TCGA sample.

The deregulation causal network

The set of binarized vectors is the basis for the construction of the deregulation causal network **[8]**. For the sake of consistency, we shall give a brief description of its construction.

The starting point is the data on frequencies of gene deregulations, ν_i , measured in the set of samples. Each gene is represented by a node in the network. A probabilistic causal sufficiency relation between genes i and j is represented by an oriented arc: $i \rightarrow j$. The conditions to draw such an arc in the network are the following:

$$0 \neq \nu_i \leq \nu_j \neq 1, \qquad \textbf{(1)}$$

$$H_{ij} = (\nu_{j|i} - \nu_j)/(1 - \nu_j) > H_0 > 0, \qquad \textbf{(2)}$$

where $\nu_{j|i}$ is the conditional probability of a deregulation at j given a deregulation at i, whereas H_{ij} is the Loevinger coefficient **[12, 13]**. Although seldom noticed, $H_{ij} > 0$ entails $\nu_{j|i} - \nu_j > 0$, which is the main postulate of the probabilistic theory of causation: causes increase the probability of their effects **[7, 14, 15]**. At difference from $\nu_{j|i} - \nu_j$, H_{ij} is largely independent from the frequency of putative effects and therefore, better suited as figure of merit of causation. We impose a more stringent condition (i.e., $H_{ij} > H_0 > 0$) in order to remove statistically insignificant instances of causal correlation, and to recover the inherent asymmetry of cause and effect, at least in the protypical case of $\nu i \neq \nu j$. Indeed, when $\nu i \neq \nu j$, $H_{ij} > H_0$ licenses the inference from a deregulation at i to a deregulation at j, but not the other way around, since $H_{ji} \leq H_0$. It is in this sense that we refer to causal sufficiency (i sufficient but not necessary to get j). Of course, here the inference is probabilistic and hence, allows for errors (instances in which i is on but still j off, even though $H_{ij} > H_0$). Our choice of $H_0 = 1/2$ amounts to get the effect more than half the times the cause is on, in the

limit of low frequency (i.e., $\nu i \approx \nu j \approx 0 \Rightarrow \nu_{j|i} > 1/2$). Therefore, we can interpret $H_{ij} > H_0$ as signaling a prima facie causal sufficiency relation between the involved genes **[8]**.

Some of the causal links ($i \rightarrow j$) discovered by condition **Eq. (2)** are spurious, i.e., arising from the correlation to a common cause k ($k \rightarrow i$ and $k \rightarrow j$). A Reichenbach-like test **[14]** help us to spot and remove such cases. Some other links are simply redundant via transitivity (i.e., $i \rightarrow k$ and $k \rightarrow j$). We prune them by appealing to Mokken test of double monotony **[13]**.

The result of applying the latter set of procedures to PRAD data **[8]** is a simplified causal network in the form of a directed acyclic graph with 46923 nodes and 10664000 arcs. This is a diluted network, with only around 0.5 % of arcs. The number of nodes was reduced, as compared to the original (60483), because there are trivial genes (zero frequency) and equivalent genes (same discretized expressions in all samples). The latter are compressed into a single node. The number of arcs was reduced from 69.6 to 10.6 millions due to the presence of spurious and transitive causal relations. Details may be found in Ref. **[8]**.

A taste of the causal network may be obtained from **Fig. 1**, where we draw the node degrees against their frequencies. The in-degree of the node is the number of arcs arriving to it, the out-degree is the number of arcs departing from it.

Nodes with high out-degrees belong to the region $\nu < 0.2$. Among them, there are 467 orphan nodes (in-degree = 0). For $\nu \geq 0.4$ there are only 19460 child-less nodes (out-degree = 0). We cannot assign a causal carcinogenic role to genes that belong to childless nodes, since there are no downstream genes that might be affected after them. These nodes are statistically but not causally correlated to cancer. At most, they can be markers but not triggers of cancer. The set of genes for

PRAD obtained from a Principal Component Analysis of the same data **[16]** belongs to this category.

Supplementary Table I contains a summary of node properties, that is frequencies, in- and out-degrees and the genes associated to each one. On the other hand, details of the causal network are included in **Supplementary Table II**. For each node, the children nodes are indicated. The number of children nodes is equal to the out-degree. The in-degrees and the parent nodes can be obtained by transposing this table.

Re-analizing the data in order to discover the deregulation cascades

Once we have at our disposal the deregulation causal network, we may re-analyze the data. For each sample, deregulated genes are organized according to the network. A summary of the results is presented in **Fig. 2**.

Each point in the figure corresponds to a sample. Normal and tumor samples are included. The y-axis represents the total number of deregulated nodes, whereas the x-axis gives the total number of nodes which are at the origins of deregulation cascades. That is, the node is deregulated but their causal parents are not. We call them Brownian deregulations in the assumption that they come from random processes in the cell (mutational or epigenetic). In contrast, other deregulations are assumed to be the result of propagation of the initial deregulation through the causal cascade. Notice that a Brownian node initiates a cascade only if its out-degree is not zero, and that there are many multi-cascade samples.

Two groups or clusters of samples are apparent in the figure. In the first one, a small number of Brownian deregulations per sample (2.3 in average) leads to a large number of total deregulations.

The cascade is very wide. Most of the samples are included in this group (45 normal and 418 tumor samples, see **Suppl. Table III** for details).

In the second group of samples (7 normal, 81 tumor, see **Suppl. Table III** for details), on the other hand, the number of Brownian deregulations per sample is much higher (445.7 in average, two orders of magnitude larger than in the first group). There are many small deregulation cascades, coming from these Brownian nodes.

Figure 2 shows that there is a threshold value for the number of deregulations, around 9000, above which there are only tumor samples. But there are many tumor samples below the threshold. On the other hand, the mean number of deregulations in normal samples is around 5200. But there are also many above this value. Thus, the number of deregulations alone can not be used to discriminate between samples. We find more useful to follow deregulations of genes belonging to the "normal concept" **[9]**.

The normal concept comprises genes which are deregulated in at least a tumor and in no normal sample at all – therefore, defining "normality" or homeostasis. In PRAD there are 15776 such nodes. A deregulation of any of them indicates a tumor sample. The lowest-frequency normal concept node in a cascade defines a trigger of cancer (we do not use the word driver because it is commonly related to mutated genes). In general, they are low in-degree (mean value 26), high out-degree (mean value 517) nodes, with an average frequency in the tumor subset of 3.3%.

There are 333 normal concept nodes out of the 1067 Brownian nodes which are deregulated in the first cluster of samples (left side of **Fig. 2**). Furthermore, 75% of the cascades in the tumor samples of this cluster start from a normal concept node. In the rest of the samples, a normal concept node is reached later in the causal evolution. Thus, we may state that in most tumors of this cluster, there is

a unique deregulation cascade, which initiates in a normal concept node. In other words, there is a single event initiating cancer.

We show in **Fig. 3** the deregulation cascades in two samples, one normal and one tumor. They have the lowest number of deregulations in the first group. Normal concept nodes are identified in the figure. Of course, there are not deregulated normal concept nodes among the 2220 deregulations in the normal sample. The Brownian node at the origin of the cascade is related to the BPIFB6 gene, which plays a role in the innate immune response, but not in carcinogenesis.

On the other hand, in the tumor sample the cascade starts in a very low frequency normal concept node, corresponding to MIR586, which has been identified to play an important role in gastric cancer **[17]**. Two additional Brownian nodes in this sample, corresponding to MIR4537 and TEKT4, have zero out-degree and can not lead to deregulation cascades. Notice that there are 33 normal concept nodes out of the 2182 deregulated nodes in the cascade. In the bottom panel of the figure, the causal dependencies in the normal concept subset of nodes are indicated. Only two causal steps are needed to reach any node in this subset, starting from MIR586.

In the second cluster of samples, there is a mean of 80 normal concept Brownian nodes per tumor samples, leading to many independent deregulation cascades. This significant increase, which may be qualified as a “deregulator phenotype”, could be related to a rise in the rate at which Brownian deregulations are generated. There are similar known phenomena at the mutational level, for example, in evolving populations of E. Coli **[18]** or in the appearance of genetic instabilities in colorectal tumors **[19]**. In both cases, the events are related to frameshift mutations in phosphatases: mutT in bacteria, and PTP genes in tumors. Whether the deregulator phenotype in PRAD has a mutational or epigenetic origin shall be further clarified.

We show in **Fig. 4** the deregulation cascades in the tumor sample with the lowest number of deregulations (3258) in Group II. Among them, there are 60 normal concept nodes. Note that there are many Brownian nodes initiating independent cascades.

In **Supplementary Table IV** the deregulated nodes in each sample are indicated.

Perfect classification of prostate tumors

Prostate cancer shows a very high heterogeneity, with a fraction of samples resisting any classification attempt **[20]**. In this section, we show that there is an exhaustive classification scheme deriving from a perfect normal concept panel of 15 nodes **[9]**.

By definition, normal concept nodes exhibit normal expression values in the normal samples. A perfect panel is a minimal set of nodes such that any tumor sample is characterized by the deregulation of at least one node in the panel. The nodes in the panels acts as markers for the deregulation cascades, no matter it is a single- or a multi-cascade. This fact is illustrated in **Figs. 3** and **4** for the specific tumor samples chosen as examples of Groups I and II. In the first one, deregulation of the panel gene CDKN1C, whereas in the Group II tumor sample the multi-cascade is marked by two genes: CFAP65 and CDKN1C.

We stress that there is a one to one correspondence between nodes and genes in the 15 nodes entering the panel. The genes are listed in additional columns of **Supplementary Table III**. The Table also shows their deregulations for each tumor sample, that is “the class” to which the sample belongs. An ordered representation of the table, including only tumor samples, is given in **Fig. 5**. Samples are distributed in rows and genes in columns. A deregulated gene is represented in black. The first three genes allow a rough classification of 426 samples. The remaining 73 samples are roughly characterized by the deregulation of one of the 12 genes completing the panel.

In short: Our classification scheme associates a 15-dimensional binary vector to each tumor sample plus an additional label indicating whether it belongs to the deregulator genotype (Group II) or not (Group I). That is, the panel provides a perfect molecular taxonomy of tumors. The biological significance of panel genes and the comparison with existing classifications of PRAD tumors **[20,21,22]** deserve further attention.

Discussion

In the paper, we sketched a method for the causal analysis of gene expression data. First, the causal network should be generated. The scheme for such a construction was delineated in Ref. **[8]** and exemplified in PRAD.

For each sample, the cascade of gene deregulations is identified. We observed that in PRAD most tumor samples show a unique cascade with an identifiable origin. This raises the question about the "trigger genes". For us, the trigger is the origin of the cascade in the subset of nodes belonging to the normal concept. This leads, for example, to the micro-RNA gene MIR586, usually not recognized as a driver in PRAD although our causal threshold of 0.5 indicates that there are strong causal relations in the scheme of **Fig. 3 c)**. In **Supplementary Table III** we list the possible origins of deregulation cascades for each sample in Group I.

Our definitions exclude from the normal concept most genes in Table I of **Ref. [23]**, i.e. exclude most accepted mutational drivers from our conception of triggers. Recall, for example, two genes widely recognized as PRAD drivers on the basis of mutation frequencies: SPOP and FOXA1 **[23]**. According to our **Supplementary Table I**, SPOP is a low-frequency gene with in-degree = 21, and out-degree = 1. The low frequency is a good property, but because of the low out-degree it may have limited causal impact. In addition, the gene is deregulated in two normal samples, thus it is not

in the normal concept class. On the other hand, FOXA1 is a very high frequency gene with in-degree = 1412 and out-degree = 0. It can not even be a tumor marker because it is deregulated in 56% of normal samples. This issue about mutational drivers and expression triggers deserves further attention.

In spite of limitations of this first approach to causality in gene-expression data, we feel we have a strong perfectible tool allowing the identification of the initiators of cancer and a scheme uniquely labeling PRAD tumors.

In principle, cascades observed in tumor samples could be reproduced in experiments with cellular lines or animal models. By identifying and acting on key genes in the cascade, one could destroy the molecular machinery sustaining the tumor. Both possibilities are promising and worth testing.

As mentioned above, in Ref. **[6]** we developed an abstract model of random jumps in gene expression space, allowing comparison with existing data on cancer risk. The results of that paper may be rephrased in present terms as the consistency of the risk data with a single deregulation cascade in carcinogenesis. The propagation of deregulations through the cascade shall involve additional steps, something reminding the multistep model of cancer **[23]**. Work along this direction is in progress.

[1] Gonzalez, A., Nieves, J., Leon, D. A., Bringas Vega, M. L. & Sosa, P. V. Gene expression rearrangements denoting changes in the biological state. Scientific Reports 11, 8470 (2021).

[2] Karlebach, G., Shamir, R. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol* **9**, 770–780 (2008). https://doi.org/10.1038/nrm2503

[3] Giancotti FG. Deregulation of cell signaling in cancer. FEBS Lett. 2014 Aug 19; 588(16):2558-70. doi: 10.1016/j.febslet.2014.02.005.

[4] Wang J, Liu Q, Shyr Y. Dysregulated transcription across diverse cancer types reveals the importance of RNA-binding protein in carcinogenesis. BMC Genomics. 2015;16 Suppl 7 (Suppl 7):S5. doi: 10.1186/1471-2164-16-S7-S5.

[5] Nath S, Ghatak D, Das P, Roychoudhury S. Transcriptional control of mitosis: deregulation and cancer. Front Endocrinol (Lausanne). 2015 May 5;6:60. doi: 10.3389/fendo.2015.00060.

[6] Roberto Herrero, Dario A. Leon, and Augusto Gonzalez (2022). A one dimensional parameter free model for carcinogenesis in gene expression space. Scientific Reports 12:4748. https://doi.org/10.1038/s41598-022-08502-8

[7] Patrick Suppes. A probabilistic theory of causality. North-Holland publishing company 1970.

[8] J.P. Gomez (2023). Non parametric causal discovery with applications to the bioinformatics of cancer. Diploma thesis. Havana University. arXiv preprint https://doi.org/10.48550/arXiv.2306.16520

[9] Gabriel Gil and Augusto Gonzalez (2022). Perfect genetic biomarkers for cancer from a fresh view of gene dysregulation. biorXiv preprint. https://doi.org/10.1101/2022.07.25.501449

[10] Katarzyna Tomczak, Patrycja Czerwiska, and Maciej Wiznerowicz. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn), 1A:6877, 2015. https://www.cancer.gov/tcga

[11] The Cancer Genome Atlas Research Network (2015). The Molecular Taxonomy of Primary Prostate Cancer. Cell 163, 10111025. http://dx.doi.org/10.1016/j.cell.2015.10.025

[12] Jane Loevinger (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. Psychological Bulletin 45.6, page 24.

[13] R.J. Mokken. A theory and procedure of scale analysis: With applications in political research. Mouton & Co 1971.

[14] Hans Reichenbach. The direction of time. University of California press 1971.

[15] Nancy Cartwright (1979). Causal laws and effective strategies. Nous, 13(4): 419-437.

[16] Perera Y, Gonzalez A, Perez R (2021). Principal Component Analysis of RNA-Seq Data Unveils a Novel Prostate Cancer-Associated Gene Expression Signature. Arch Can Res Vol.9 No.4: 002. https://www.acanceresearch.com/cancer-research/principal-component-analysis-of-rnaseq-data-unveils-a-novel-prostate-cancerassociated-gene-expression-signature.pdf

[17] Chao Liu, Jing Yang, Fengchi Zhu et al (2022). Exosomal circ 0001190 Regulates the Progression of Gastric Cancer via miR-586/SOSTDC1 Axis. Biochem Genet. 60(6):1895-1913. https://doi.org/10.1007/s10528-021-10180-6

[18] Jeffrey E. Barrick, Dong Su Yu2, Sung Ho Yoon, et. al. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia Coli. Nature 461, 1243. https://doi.org/10.1038/nature08480

[19] Sebastian Korff, Stefan M Woerner, Yan P Yuan et al (2008). Frameshift mutations in coding repeats of protein tyrosine phosphatase genes in colorectal tumors with microsatellite instability. BMC Cancer 8:329. https://doi.org/10.1186/1471-2407-8-329

[20] Kentaro Inamura (2018). Prostatic cancers: understanding their molecular pathology and the 2016 WHO classification. Oncotarget Vol. 9, (No. 18), pp: 14723-14737. https://doi.org/10.18632/oncotarget.24515

[21] Guocan Wang, Di Zhao, Denise J. Spring, and Ronald A. DePinho (2018). Genetics and biology of prostate cancer. GENES & DEVELOPMENT 32:11051140. http://www.genesdev.org/cgi/doi/10.1101/gad.315739.118

[22] Joshua Armenia, Stephanie A.M. Wankowicz, David Liu et. al. (2018). The long tail of oncogenic drivers in prostate cancer. Nat Genet. 50(5): 645651. https://doi.org/10.1038/s41588-018-0078-z

[23] Frank, Steven A. Dynamics of cancer : incidence, inheritance, and evolution. Princeton University Press 2007.

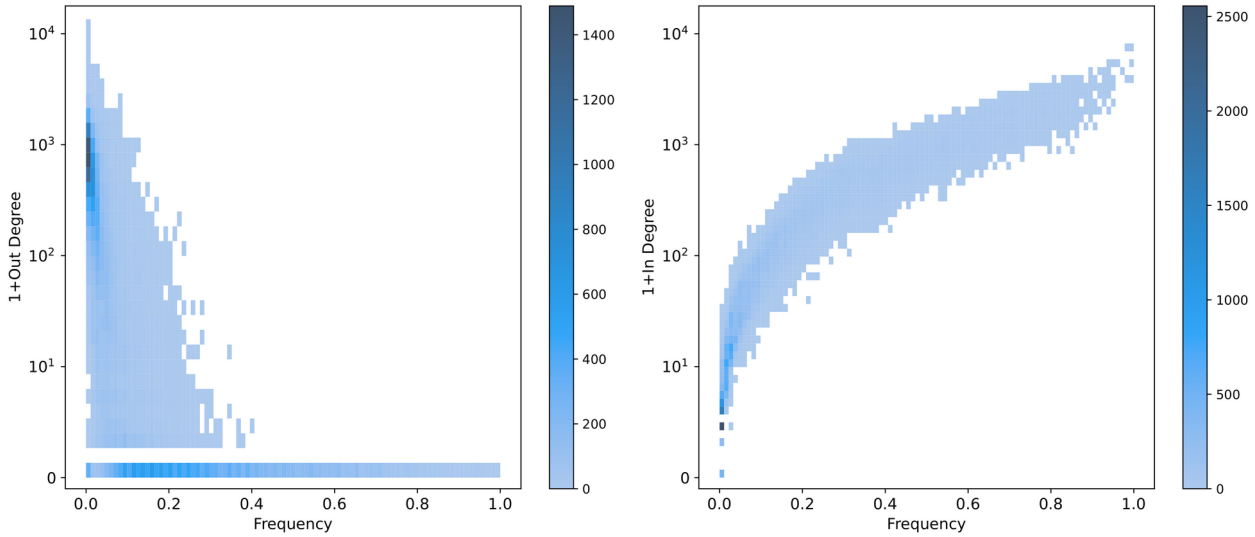

Figure 1: A characterization of the causal network for PRAD. Node in- and out-degrees against frequencies are drawn.

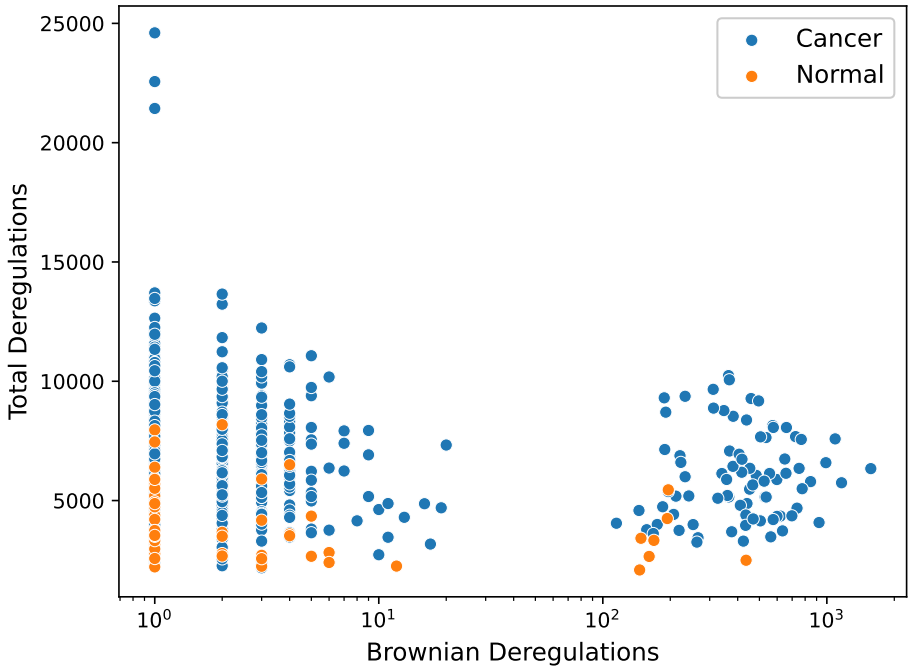

Figure 2: Total number of node deregulations in samples against the number of Brownian deregulations, i.e. nodes at the origins of deregulation cascades.

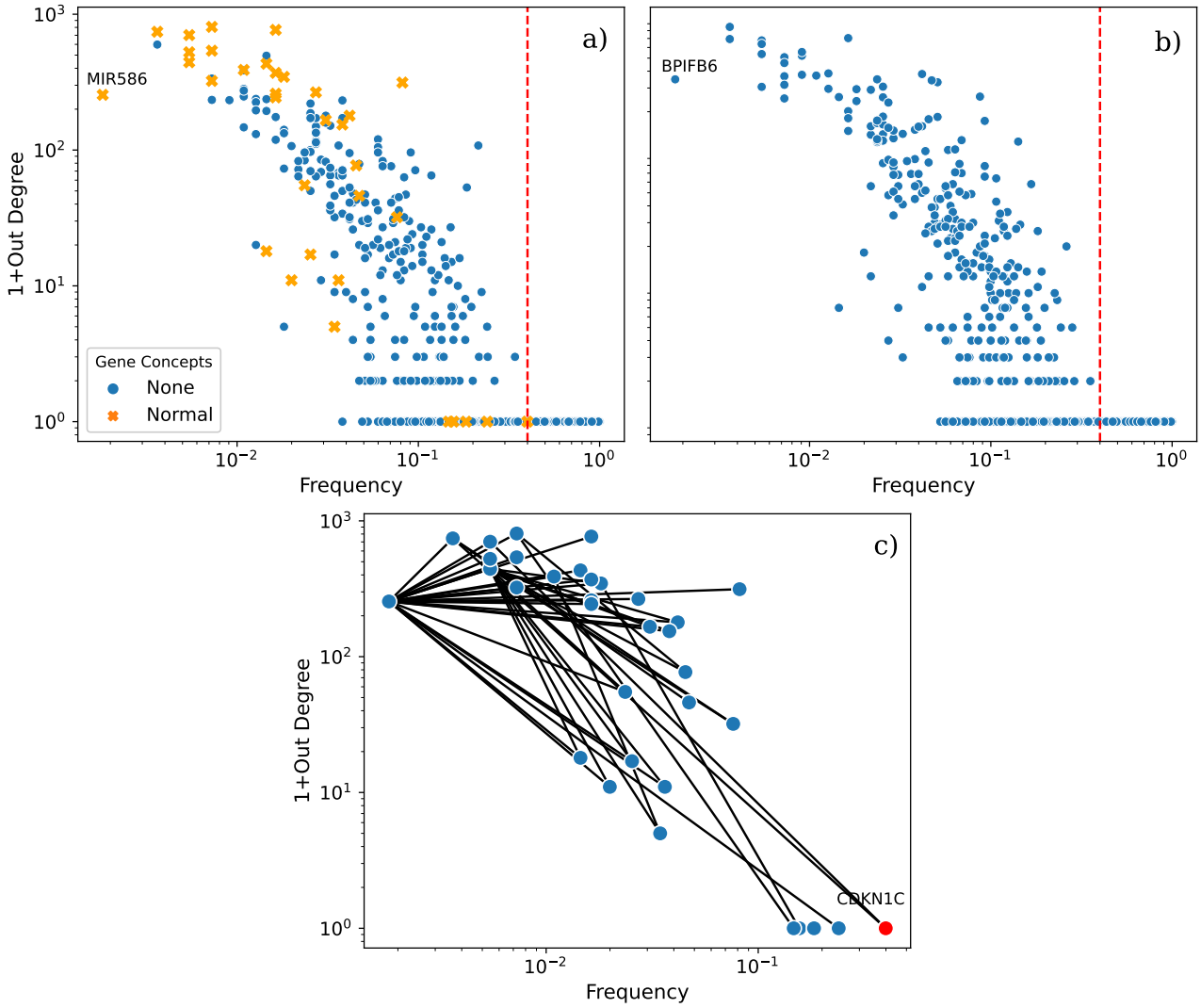

Figure 3: a) and b): Deregulation cascades in the tumor and normal samples of Group I with the lowest number of deregulated nodes. Genes at the origin of the cascades and normal concept nodes are indicated. Above the threshold frequency (vertical dashed line) only in-degree nodes exist. c): The causal cascade in the normal concept subset of the tumor sample. This cascade is labeled by a single deregulated gene (CDKN1C) of the perfect panel.

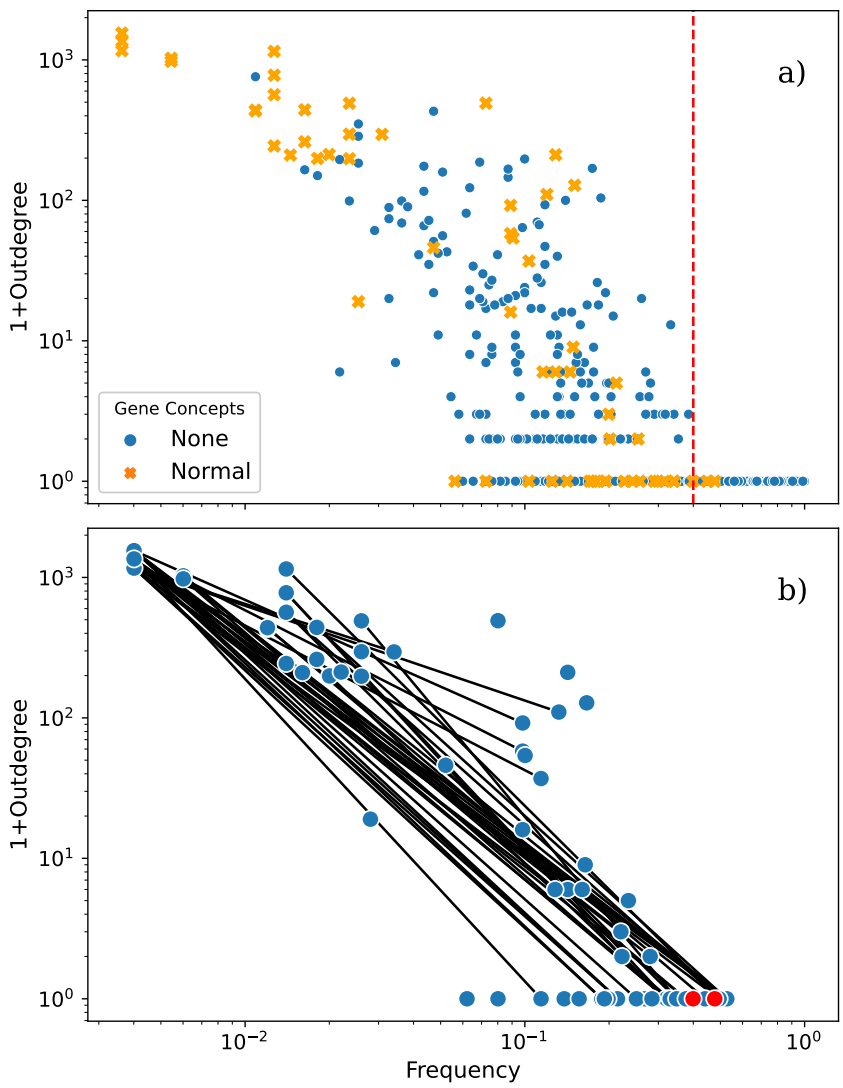

Figure 4: Similar to **Fig. 3**, but for the tumor sample with the lowest number of deregulations in Group II. We may think about a cascade with multiple initiators or a multi-cascade. The cascade in b) contains only normal concept nodes. It is labelled by two genes, CFAP65 and CDKN1C (marked in red), of the perfect panel.

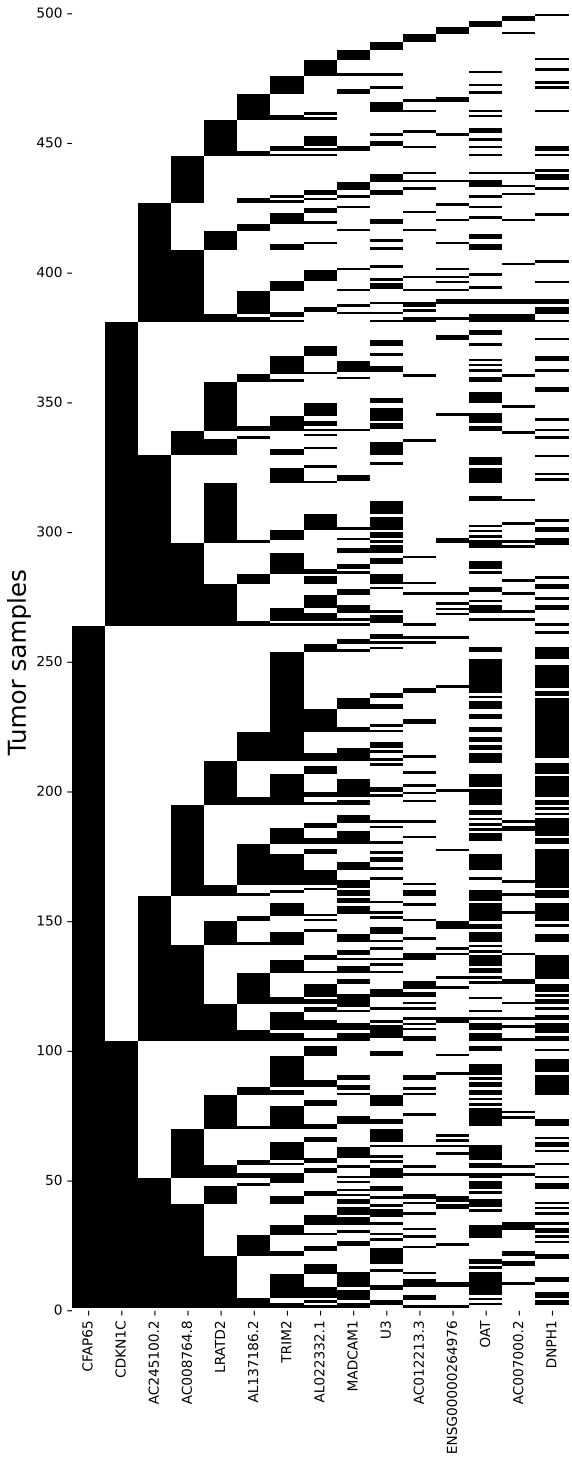

Figure 5: Classification of tumor samples according to the 15 normal concept genes in the perfect panel. Samples are organized as rows and genes as columns. Gene deregulations are represented in black.

Methods

Processing of TCGA data is described elsewhere **[8,24]**. For the sake of consistency, we provide some details. Geometric averages over normal samples define reference values for each gene, $e^{(ref)}$. Then the log-fold change of the expression of the i-th gene is defined as:

$$e_{fold}(i) = \log_2 (e_i / e_i^{(ref)}). \qquad \text{(M1)}$$

This expression is further discretized according to the following rule:

$$|e_{fold}(i)| < 1, \ \rightarrow e_i = 0 \text{ (normal expression)}, \qquad \text{(M2)}$$

$$|e_{fold}(i)| > 1, \ \rightarrow e_i = 1 \text{ (deregulated)}. \qquad \text{(M3)}$$

This gene-independent procedure is not the only way of defining gene deregulation **[9]**, but we shall adopt it in the present paper for simplicity.

As mentioned above, each sample is then represented by a vector of binary data: $\mathbf{e_{sample}} = (e_1, e_2, \ldots, e_{60483})$. These vectors are the basis for the construction of the causal deregulation network **[8]**, briefly sketched in the body of the paper.

The expression data for each sample is re-interpreted with the help of the deregulation network. A cascade (or multi-cascade) is defined in this way, as shown in **Figs. 3** and **4**.

Availability of data and code

The data we used, the Python routines and the main results are integrated in the following public GitHub repository: https://github.com/RobertoHePe/Causal_GE. In particular, **Supplementary Tables I – IV** may be found there.

[24] Gonzalez A, Leon DA, Perera Y, Perez R (2023). On the gene expression landscape of cancer. PLoS ONE 18(2): e0277786. https://doi.org/10.1371/journal.pone.0277786

Acknowledgments. Authors acknowledge the Office of External Activities of the Abdus Salam Centre for Theoretical Physics for support. The research is carried on under a project of the Nuclear Energy and Advanced Technologies Agency (AENTA), Cuba.

Author's contributions

G.G. and J.P.G. ideated the algorithm, implemented it and computed the causal network for PRAD. A.G. introduced the idea of data causal re-analysis and wrote an initial draft. R.H. re-analyzed PRAD data and is in charge of the GitHub repository. All authors analyzed and interpreted the results, contributed to the manuscript and approved the final version.

Competing interests.

The authors declare that they have no competing interests.