Dynamics of N-genes and T-genes in cancer

Gabriel Gil and Augusto Gonzalez

Institute of Cybernetics, Mathematics and Physics, Havana

Abstract.

The normal-tissue and tumor attractors are far apart in gene-expression space **[1,2]**. As a consequence, there are genes with specific gene-expression intervals characteristic only of normal tissues or tumors **[3]**. We call them N-genes and T-genes, respectively. They are not only perfect markers, but are also responsible of anchoring the state in the vicinity of the corresponding attractor. According to the higher degree of disorder of tumors **[4,5]**, there are 10-times more T-genes than N-genes **[3]**. Carcinogenesis, which is the progressive motion from the N-attractor to the T-attractor, can also be viewed as a process of deactivating the N-genes marking and holding the normal tissue and setting new T-markers and anchors for the tumor. In prostate, for example, the process starts with around 500 activated N-genes in any healthy normal tissue that may be dramatically reduced to a few N-genes before the transition to the tumor occurs. The available gene expression data **[6,7]** shows that even 2 or 3 active N-genes are enough for holding the normal state condition (at least from a pathological perspective). The process is expected to evolve in the form of consecutive deregulation cascades **[8]**, in which a few initial cascades inactivate most of the active N-genes, and a final cascade inactivates the remaining N-markers and sets the new T-markers for the tumor. The whole process shares similarities with the multistep model of cancer **[9]**.

N-genes, T-genes and O-genes.

The N- and T-attractors are far apart in gene-expression space. Indeed, using the 60,000 reported genes and log-fold coordinates in order to compute the distance **[1,2]**, this distance seems to be around 94.7 in prostate adenocarcinoma (PRAD) **[10]**. Normal samples are distributed in a kind of halo around the center of the attractor. The r.m.s. radius of this halo is 132.6. Thus, it is somehow like a limit cycle **[11]**, sustaining circadian oscillations **[12]**. On the other hand, tumor samples form also a halo around the center of the T attractor, with r.m.s radius of 145.6.

Let us introduce a coordinate x along the axis joining the centers of the N- and T-attractors. The x-axis is nearly orthogonal to both the normal tissue and tumor hyper-surfaces **[10]**. Values of x around 0 are characteristic of the normal samples whereas values around 94.7 are characteristic of tumors. The x coordinate is a metagene **[13]**, which contains the contribution of many genes. Thus, there should be many coordinates (genes) with expression intervals characteristic only of normal samples or tumors, we call them N-genes and T-genes, respectively. In the same way, coordinates "orthogonal" to the x axis may have expression intervals accessible to both normal samples and tumors. We call them O-genes (O for others).

The EPHA10 gene is an example of T-gene in PRAD. In **Fig. 1a** we show the expression histograms in normal and tumor samples. The FPKM normal and tumor intervals for the expressions are (0.04, 0.88) and (0.08, 5.25), respectively. Thus, there is an interval, (0.88, 5.25), characteristic only of tumor samples. We write e = 0 for normal and tumor samples in the common interval, e = 1 for tumor samples which expression is above 0.88, and include the gene in the "tumor above" group. Notice that e = 1 for this gene is an indication of a tumor, whereas e = 0 for all normal and for some tumor samples. When e = 1 we say the gene is T-activated. We refer the reader to **Ref. [3]** for details and a description of other categories.

On the other hand, for N-genes there is a distinct interval for normal samples. In this case, we write e = 0 for expression values of both tumor and normal samples in the common interval, and e = -1 for normal samples in the normal-only characteristic interval. We show in **Fig. 1b** paired histograms for the SEPTIN10 gene in PRAD. As it is apparent, we may include this gene in the “normal above” group. Values of expression above 22.26 seem to be incompatible with the tumor state and are characteristic of normal samples.

Notice that this latter gene exhibits also a tumor-only interval, in which the expression is below 8. Thus, it is a normal-above gene when the expression is above 22.26, or a tumor-below gene when the expression is below 8. It could serve as an example of gene marking the normal state which, in a transition, may be first reduced to the common interval of expression, and finally to the T-only interval if it becomes a marker of the tumor state.

N- and T-genes are markers of the states and they can be used to construct small gene panels perfectly classifying samples **[3]**.

But N- and T-genes may be also understood as holders of the states. In the sense that a set of activated N-genes, for example, indicate that we are in the vicinity of the normal attractor. To move the sample away from the attractor, we should first inactivate the anchors. We should come back to this point below when analyzing transitions.

It was already noticed **[3]** that there are plenty of T-genes, and much less N-genes. This is a consequence of the higher disorder of tumors. And indeed, in PRAD we found 12123 and 1254, respectively. 576 genes are in both N- and T- categories. These numbers are approximate and depend on the sample sets. Indeed, to include a gene in a group, we require the normal-only or tumor-only intervals to be populated by enough samples, and establish a statistical significance

criterion by means of a simple exact Fisher test **[3]**. Let n be the number of samples in the normal-only or tumor-only gene expression interval. Requiring that n shall not be the result of chance, with a p-value of 0.01 the test leads to:

$$n > (2 \ln 10) / \ln [(N+T) / N], \qquad (1)$$

or

$$n > (2 \ln 10) / \ln [(N+T) / T], \qquad (2)$$

for N-genes and T-genes, respectively. In **Eqs. (1,2)**, N and T correspond to the number of normal and tumor samples. In the TCGA data for PRAD **[6,7]**, N ~ 50, T ~ 500, thus n / N > 4% and n/T > 9%. In this paper, we use the stricter thresholds of 5% and 10% to define the N- and T-genes in PRAD.

The number of N- and T-genes should reach saturation values for large enough sets of samples. There will always be much more T-genes than N-genes.

Let us summarize. There are T-genes, which discretized expression may take the value e = 0 (the common interval for N and T samples) or e = 1 (an interval characteristic only of tumors). In this latter case, we say the gene is T-activated and the sample is, of course, a tumor. In the same way, there are N-genes, which expression may take values e = 0 or e = -1, in the latter case we say the gene is N-activated. It may occur only in a normal sample. Besides, there are genes which expression may take values e = 1, 0 and -1. That is, they may transit form N-activation to T-activation and vice-versa. Additionally, there are genes which expression is always e = 0. Their expression intervals are accessible to both normal samples and tumors. These are the O-genes (O for other).

The number of N-markers is reduced in the process of somatic evolution of a normal tissue.

First, we show in **Fig. 2a)** the number of activated N-genes in the 52 normal samples of the PRAD data. Notice that this number ranges from 2 to about 700. The data shows that what we qualify as "normal" are actually samples in very different stages of somatic evolution. More healthy conditions are probably associated to larger numbers of activated N-genes, 500 or more, whereas the number is reduced in the course of somatic evolution. Thus, in the available set of normal samples, there are some close to the normal state, and others close to the tumor attractor.

The x-axis in **Fig. 2a)** is the distance in gene-expression space from the normal sample to the center of the tumor attractor. As mentioned, the r.m.s. radius of the tumor cloud seems to be around 145. The distances from normal samples to the center of the tumor cloud ranges from 120 to 230. Healthy samples, with more than 500 activated N-genes are far from the tumor attractor, whereas samples experiencing a strong somatic evolution, leaving only a few activated N-genes, could be even inside the tumor basin of attraction. It is remarkable that 2 or 3 activated N-genes may hold the normal-state condition, even after more than 500 genes have been deactivated.

Figs. 2b) and **c)** show similar results for lung adenocarcinoma (LUAD, **[14]**) and liver hepatocarcinoma (LIHC, **[15]**). The number of N-genes in these tissues, with the same threshold of 0.05, are 638 and 1275, respectively. The r.m.s. radii of their respective tumor clouds are marked by dashed lines, as in **Fig. 2a**.

Let us stress that we are inferring the somatic evolution in normal samples from the observed distribution of N-genes. This resembles an ergodic argument **[16]**. We will use the argument below in the analysis of T-genes.

Conceptualizing the transition.

Carcinogenesis is the transition from the N-condition to the T-condition in a small tissue portion. We understand it as a process in which the sample is moved from the vicinity of the N-attractor to the vicinity of the T-attractor. To this end, the N-genes holding the state should gradually be inactivated. Finally, when all of the holders are down, new T-markers arise and the transition occurs.

A schematic picture is presented in **Fig. 3**. The total set of genes is contained in a rectangle. Inside the red ellipse there are the T-genes, and inside the blue one, the N-genes. The initial healthy normal state (**Fig. 3a**) has a set of activated N-markers (blue circles). The rest of the genes with e = 0 are not drawn.

The number of activated N-genes in the normal sample is reduced as a result of perturbations. Perturbations to the normal state come in the form of somatic mutations **[17]**, epigenetic alterations **[18]** or external influences **[19]**. They may provoke deregulation cascades **[20]** in the regulation network. In Refs. **[8,21]** we inferred the deregulation network directly from the gene expression data.

Then, assume a cascade is produced. The attractor is a stable state and may support small perturbations. If the cascade reaches a subset of the activated N-genes it may modify them, moving the expressions to the interval common to N and T samples, that is reducing the expression to e = 0. The reached genes stop acting as holders of the N state, and a step towards the tumor is taken. In the process, part of the information on the initial normal state (the initially activated N-genes) is lost. We represent it in **Fig. 3b)** as a change in the color of the circles, from blue (N-activated) to white. Notice that the expression of a gene can not be moved towards T-activation intervals if there remain activated N-genes holding the normal state.

Thus, the first cascades prepare the transition by inactivating most of the N-genes. But there should be a cascade that realizes the transition, where no N-genes are left and a set of activated T-genes mark the new tumor. This is represented in **Fig. 3c)**. The whole process is conceptually similar to the multi-step model of cancer **[9]**. Notice that there could be genes, that were normal tissue markers, and become converted by the last cascade to tumor markers, i.e. transit from e = -1 to e = 1. We shall come back to this point below.

New T markers and a more complex structure can be further gained by the tumor in the course of clonal evolution **[22], Fig 3d).**

Clonal evolution of tumors.

The gene expression data allows us to draw pictures for T-genes similar to **Fig. 2**. We show them in **Fig. 4**. The number of activated T-genes in tumor samples is drawn as a function of the distance to the normal attractor. This number ranges from 100 to around 6000 in PRAD. The figure shows tumor samples exhibiting very different stages of clonal evolution. The ones with relatively small number of activated T-genes could be even near the normal attractor (its r.m.s. radius is around 130). They probably correspond to recently formed tumors. Evolved tumors, in which spontaneous T-cascades freely develop, may show higher numbers of activated T-genes.

Figs. 4b) and **c)** show similar results for LUAD and LIHC, respectively. The main difference is that all tumor samples are outside the normal basin of attraction.

The trace of the normal tissue in the tumor.

We mentioned that in the cascade realizing the transition there could possibly be genes transiting directly from N-activation to T-activation. They were in the normal sample and are present in the tumor sample, but with a different regulation. In a way, they are a trace of the normal tissue in the

tumor. In PRAD, these genes should belong to the set of 576 genes that exhibit both types of regulation. We look at the existing 54 paired tumor and normal samples, that is, taken from the same patient, and write a Fisher test to avoid the possibility that the inherited genes may come from pure chance.

Let us stress that the paired normal sample is not exactly the sample leading to the tumor, thus its somatic evolution could be different. In addition, the clonal evolution of the tumor could add many T-markers by chance. Thus, it is not strange that only a few normal-tumor pairs satisfy the Fisher test. We show in **Tab. I** the paired samples with the lowest p-values.

N sample id	T sample id	N-genes	T-genes	Coincidences	p-values
TCGA-HC-7740-11A	TCGA-HC-7740-11A	220	13	8	5.1E-02
TCGA-HC-7740-11A	TCGA-HC-7740-01B	220	111	34	1.6E-02
TCGA-EJ-7783-11A	TCGA-EJ-7783-01A	204	54	15	5.7E-02
TCGA-EJ-7123-11A	TCGA-EJ-7123-01A	181	139	30	1.0E-03
TCGA-HC-7211-11A	TCGA-HC-7211-01A	252	141	62	7.8E-02
TCGA-EJ-7327-11A	TCGA-EJ-7327-01A	5	59	2	7.5E-02
TCGA-EJ-7315-11A	TCGA-EJ-7315-01A	47	106	2	3.0E-03
TCGA-G9-6356-11A	TCGA-G9-6356-01A	19	15	2	7.3E-02
TCGA-CH-5767-11B	TCGA-CH-5767-01A	4	150	3	5.2E-02
TCGA-HC-8258-11A	TCGA-HC-8258-01A	278	49	26	9.3E-02
TCGA-HC-8258-11A	TCGA-HC-8258-01B	278	90	54	5.0E-03
TCGA-EJ-7781-11A	TCGA-EJ-7781-01A	125	85	3	5.0E-07
TCGA-EJ-7125-11A	TCGA-EJ-7125-01A	51	125	7	5.2E-02
TCGA-HC-7737-11A	TCGA-HC-7737-01A	246	35	18	7.8E-02
TCGA-HC-7752-11A	TCGA-HC-7752-01A	12	298	9	6.4E-02
TCGA-HC-7738-11A	TCGA-HC-7738-01A	259	119	74	1.0E-05

Table I. Paired normal and tumor samples in PRAD evidencing genes that change from N-activation to T-activation in the transition to the tumor. Only 5 pairs have p-values below 0.01.

An illustrative example is contained in the last row of the table. The normal sample is relatively healthy, with a high number of N-markers. Recall that there are 576 genes in the common group. The tumor sample, on the other hand, is relatively “young”, with regard to its clonal evolution, with

only 119 T-markers, from which 74 seem to come from the large cascade realizing the transition to the tumor.

It is remarkable that we may find at least 5 normal-tumor pairs in the data with a p-value below 0.01 exhibiting inherited genes.

Discussion.

We stressed the significance of N- and T-genes in the process of carcinogenesis and the evolution of tumors. The normal tissue starts with a set of N-markers which are partially deactivated in the course of somatic evolution as a result of subsequent deregulation cascades.

Thus, the activated N-genes in a given normal sample are in fact markers, anchors and, when deactivated, delineate the way or trend to the tumor.

It is interesting to check which are the main genetic pathways related to the N-genes. A Reactome analysis **[23]** of the 1254 N-genes in PRAD (we include the genes in the Supplementary file **Ngenes_PRAD.txt**) shows that the main three over-represented pathways are Polo-like kinase mediated events (R-HSA-156711), Cell-extracellular matrix interactions (R-HSA-446353) and Smooth Muscle Contraction (R-HSA-445355). In particular, the first one is recognized as playing a major role in prostate tumorigenesis **[24]**.

Similar analysis in LUAD and LIHC shows that the N-genes in these tissues are mainly involved in the control of the cell cycle.

In the present perspective, the information about the trend to the tumor is extracted from the N-genes in normal samples. The final tumor is somehow independent of the formation history, that is

the genetic information about the initial normal state is partially lost in the course of somatic evolution. The distribution of T-genes in the tumor population is, in fact, the actual characterization of the tumor attractor. We show in **Fig. 5** the frequencies of N-genes and T-genes in PRAD. It is worth studying the more frequently activated T-genes in different tumors. Of course, there are T-genes coming from N-genes with different regulation. The idea of targeting the PLK1 gene in PRAD **[25]** is related to this fact.

Summarizing, we may say that by comparing the expression profiles in normal samples and tumors the N- and T-genes are identified. Changes in the activated N-genes of a particular sample describe the process of somatic evolution towards the tumor. The analysis of paired normal-tumor samples allows us to extract information about the cascade realizing the transition. Changes in the activated T-genes of a tumor sample describe the process of clonal evolution of tumors. We stress that in the differential expression approach, one compares the initial normal state with the final tumor state and all of these processes are mixed and treated as a single one.

It is remarkable that, in spite of the fact that an evolved tumor is marked by thousands of T-genes, a panel of a few genes may label any tumor. In PRAD, for example, the minimal panel contains 9 genes, in LUAD – 2 **[3,26]**. It is also worth studying the tumor taxonomies induced by different gene panels.

There remain many interesting questions to be answered. One of these questions is whether in a given tissue there are sets of N-genes frequently reached by deregulation cascades. In such case, a series of steps towards cancer is established, as the canonical Vogelstein model of colon cancer suggests **[27]**. On the other hand, a crucial question is whether a Boolean regulatory network [20], where there are only two states for the genes: activated or deactivated, may actually describe the

dynamics. As we argued in the paper, characteristic expression intervals are important and a 3-state model is needed. Research along some of these directions is currently in progress.

References.

[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). https://doi.org/10.1038/s41598-021-87764-0

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

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

[4] Mesa-Rodríguez, A.; Gonzalez, A.; Estevez-Rams, E.; Valdes-Sosa, P.A. Cancer Segmentation by Entropic Analysis of Ordered Gene Expression Profiles. Entropy 2022, 24, 1744. https://doi.org/10.3390/e24121744

[5] Augusto Gonzalez, Frank Quintela, Dario A. Leon, Maria Luisa Bringas-Vega, and Pedro A. Valdes-Sosa. Estimating the number of available states for normal and tumor tissues in gene expression space. Biophysical Reports 2, 100053 (2022). https://doi.org/10.1016/j.bpr.2022.100053

[6] Katarzyna Tomczak, Patrycja Czerwiska, and Maciej Wiznerowicz. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn), 1A:6877, 2472015. https://www.cancer.gov/tcga

[7] 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

[8] Roberto Herrero, Jean Pierre Gomez, Gabriel Gil and Augusto Gonzalez. Deregulation cascades in carcinogenesis. BiorXiv preprint 2023.12.28.573550 (2023). https://doi.org/10.1101/2023.12.28.573550

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

[10] Nieves J, Gonzalez A. The Geometry of Normal Tissue and Cancer Gene Expression Manifolds. Acta Biotheor. 2024 Jul 9;72(3):9. https://doi.org/10.1007/s10441-024-09483-z

[11] Yanqian Ye, Sui-lin Cai. Theory of Limit Cycles. Translations of mathematical monographs, Vol. 66. American Mathematical Soc., 1986.

[12] Marc D. Ruben et al. A database of tissue-specific rhythmically expressed human genes has potential applications in circadian medicine. Sci. Transl. Med. 10, eaat8806 (2018). https://doi.org/10.1126/scitranslmed.aat8806

[13] Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, et al. Gene expression predictors of breast cancer outcomes. The Lancet. 2003; 361(9369):1590–1596. https://doi.org/10.1016/S0140-6736(03)13308-9

[14] The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014). https://doi.org/10.1038/nature13385

[15] The Cancer Genome Atlas Research Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell 169, 1327–1341, 2017. http://dx.doi.org/10.1016/j.cell.2017.05.046

[16] I. P. Cornfeld, S. V. Fomin, Y. G. Sinai. Ergodic Theory. Grundlehren der mathematischen Wissenschaften, Vol. 245. Springer Science & Business Media, 2012.

[17] Martincorena, I. and Campbell, P.J., 2015. Somatic mutation in cancer and normal cells. *Science*, *349* (6255), pp.1483-1489. https://doi.org/10.1126/science.aab4082

[18] Ilango S, Paital B, Jayachandran P, Padma PR, Nirmaladevi R. Epigenetic alterations in cancer. Frontiers in Bioscience-Landmark. 2020 Mar 1;25(6):1058-109. https://doi.org/10.2741/4847

[19] Paolo Boffetta, Fredrik Nyberg. Contribution of environmental factors to cancer risk. British Medical Bulletin, Volume 68, Issue 1, December 2003, Pages 71–94, https://doi.org/10.1093/bmp/ldg023

[20] Harris, S.E., Sawhill, B.K., Wuensche, A. and Kauffman, S. (2002), A model of transcriptional regulatory networks based on biases in the observed regulation rules. Complexity, 7: 23-40. https://doi.org/10.1002/cplx.10022

[21] 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

[22] Greaves, M., Maley, C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012). https://doi.org/10.1038/nature10762

[23] Marc Gillespie, Bijay Jassal, Ralf Stephan, et al. The reactome pathway knowledgebase 2022, Nucleic Acids Research, 2021, gkab1028, https://doi.org/10.1093/nar/gkab1028

[24] X. Shawn Liu, Bing Song, Bennett D. Elzey, et. al. Polo-like Kinase 1 Facilitates Loss of Pten Tumor Suppressor-induced Prostate Cancer Formation. THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 286, NO. 41, pp. 35795–35800, October 14, 2011. https://doi.org/10.1074/jbc.C111.269050

[25] Xiaoqi Liu. Targeting Polo-Like Kinases: A Promising Therapeutic Approach for Cancer Treatment. Translational Oncology Volume 8, June 2015, Pages 185 – 195. https://doi.org/10.1016/j.tranon.2015.03.010

[26] Gabriel Gil and Augusto Gonzalez (2024). The minimal number of genes needed to identify a tumor. bioRxiv preprint https://doi.org/10.1101/2024.07.25.604730

[27] Eric R. Fearon, Bert Vogelstein. A genetic model for colorectal tumorigenesis. Cell, Volume 61, Issue 5, 1990, Pages 759-767. https://doi.org/10.1016/0092-8674(90)90186-I

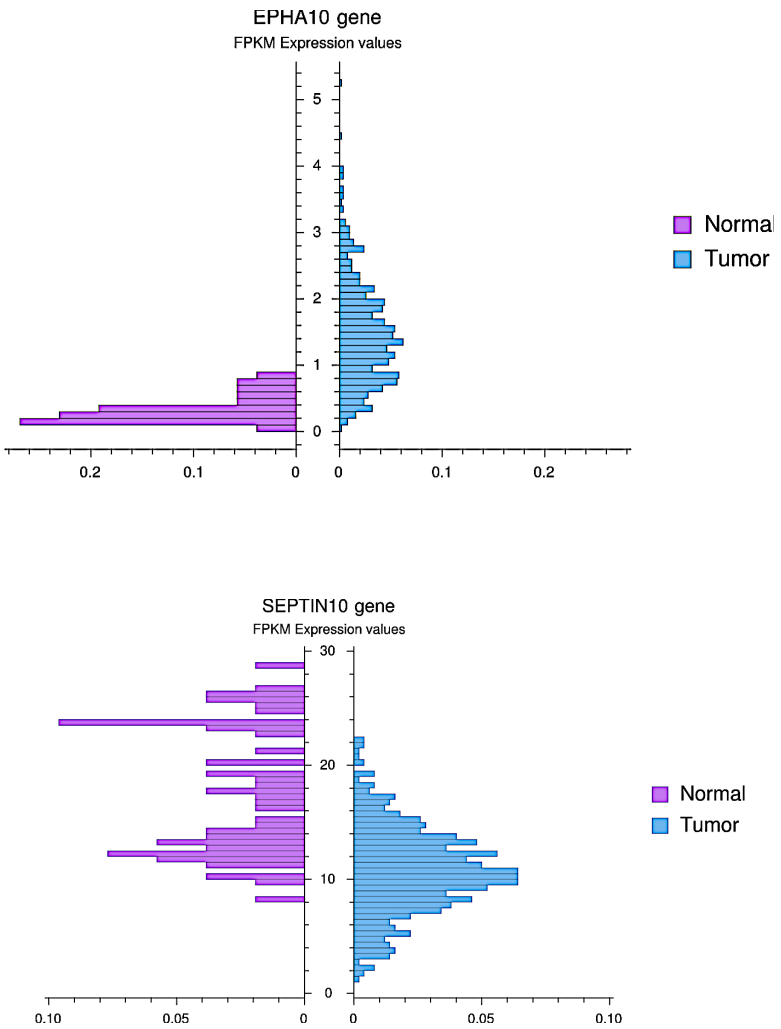

Fig. 1. Paired probability histograms of FPKM expression values of EPHA10 (a tumor-above gene) and SEPTIN10 gene in PRAD. The latter can be included in the normal-above group, or in the tumor-below group, indistinctly.

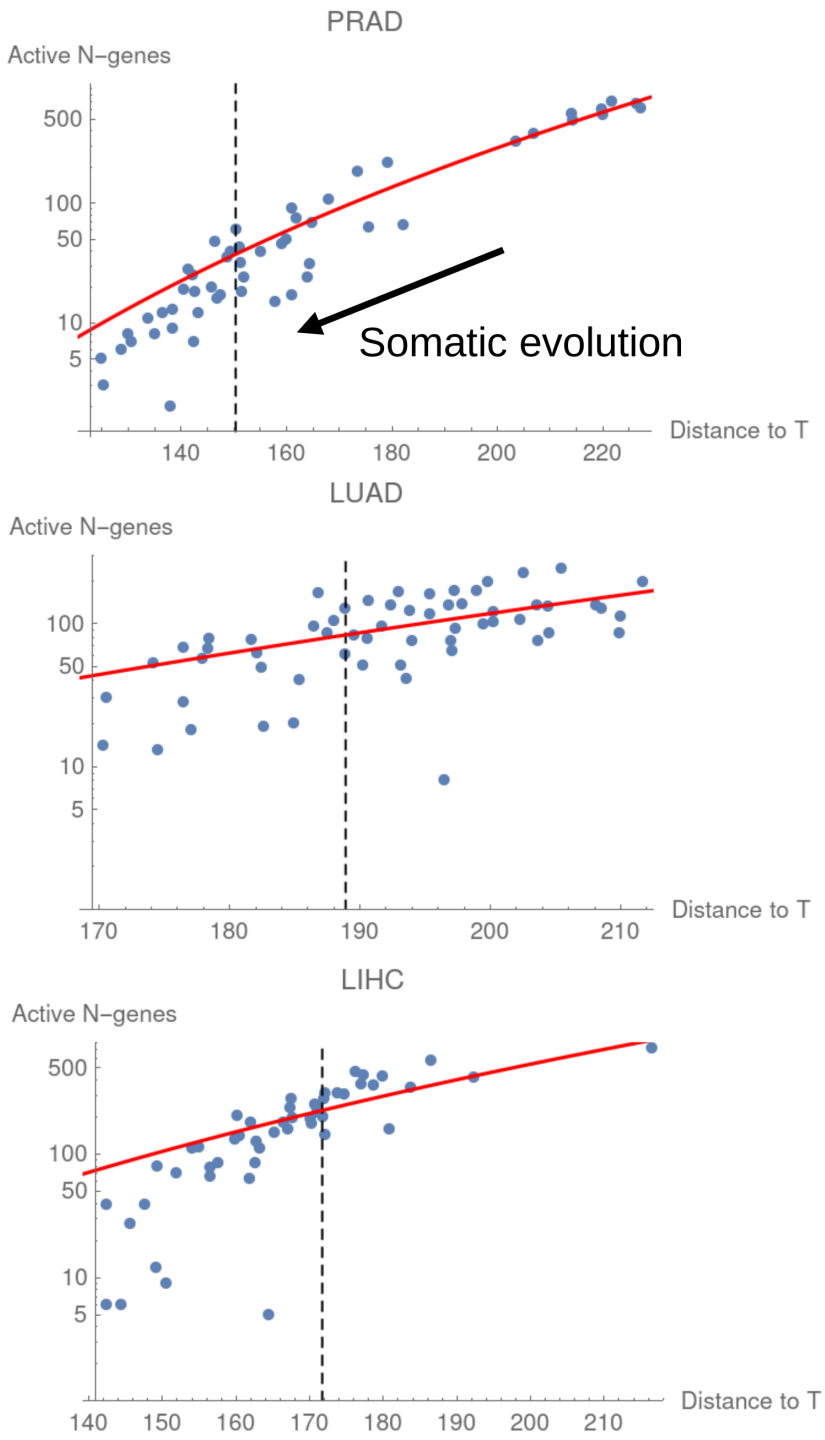

Fig. 2. Upper panel: Number of activated N-genes in normal samples as a function of the distance from the sample to the center of the tumor attractor in PRAD. The dashed line signals the r.m.s. radius of the tumor attractor. The red line is a guide to the eye. Lower panels: Similar figures in LUAD and LIHC.

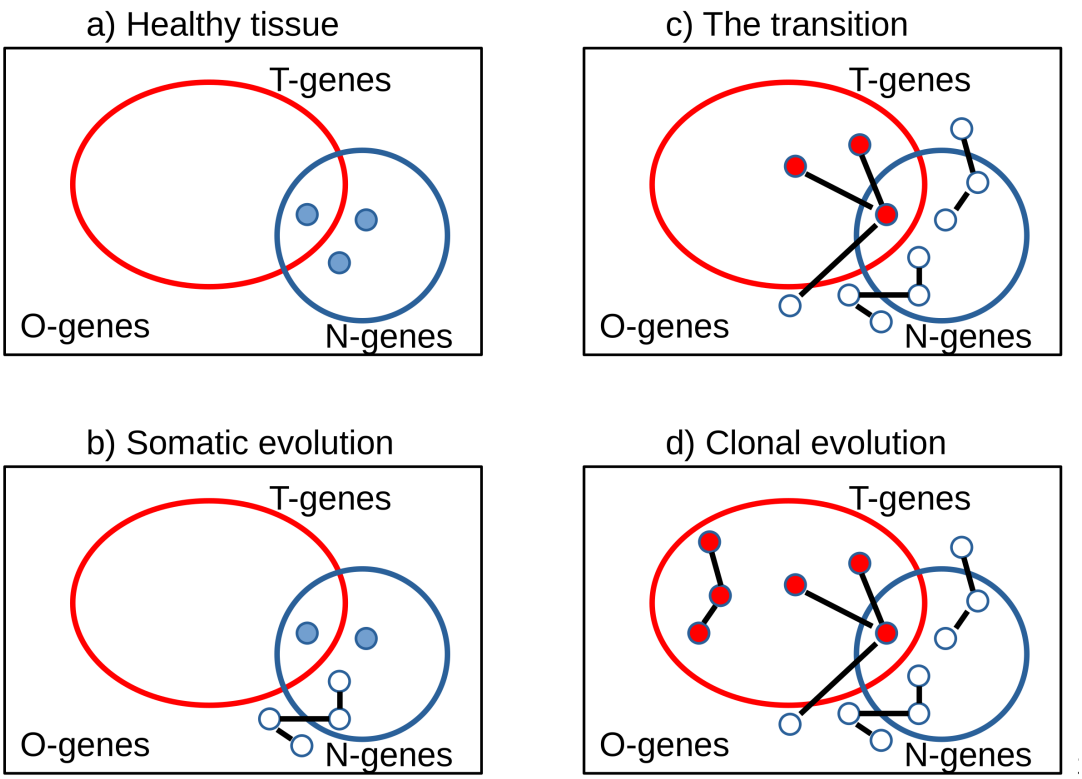

Fig. 3. Schematic representation of the process of carcinogenesis. Activated N- and T-genes are represented as blue and red circles, respectively. Empty circles are genes with e = 0 expression. Deregulation cascades are represented as genes joined by lines. **a)** The initial normal state, **b)** somatic evolution, **c)** the cascade realizing the transition, and **d)** the posterior clonal evolution of the tumor.

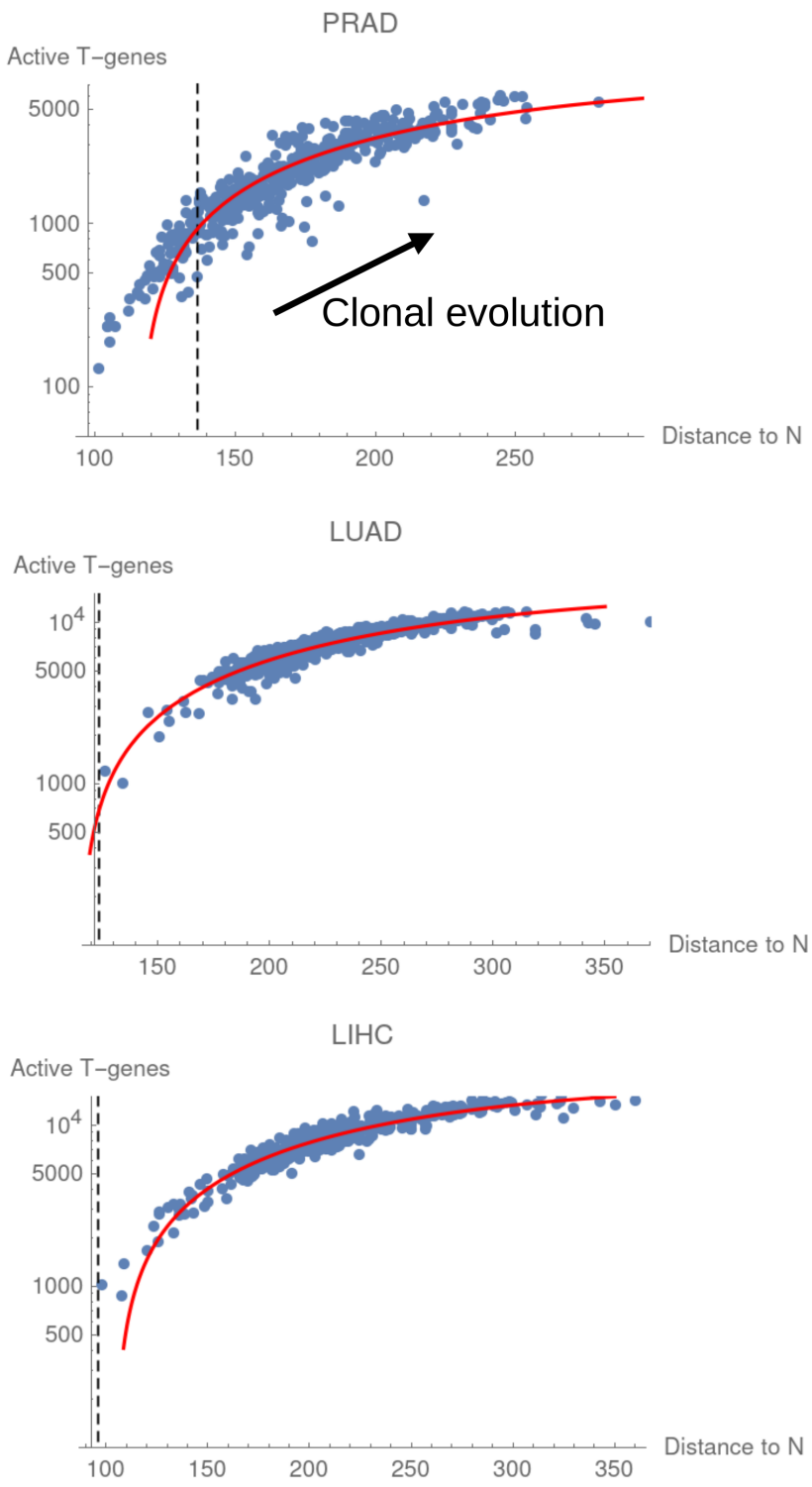

Fig. 4. Upper panel: Number of activated T-genes in tumor samples as a function of the distance to the center of the normal attractor in PRAD. The dashed line signals the r.m.s. radius of the normal attractor. The red line is a guide to the eye. Lower panels: Similar figures in LUAD and LIHC.

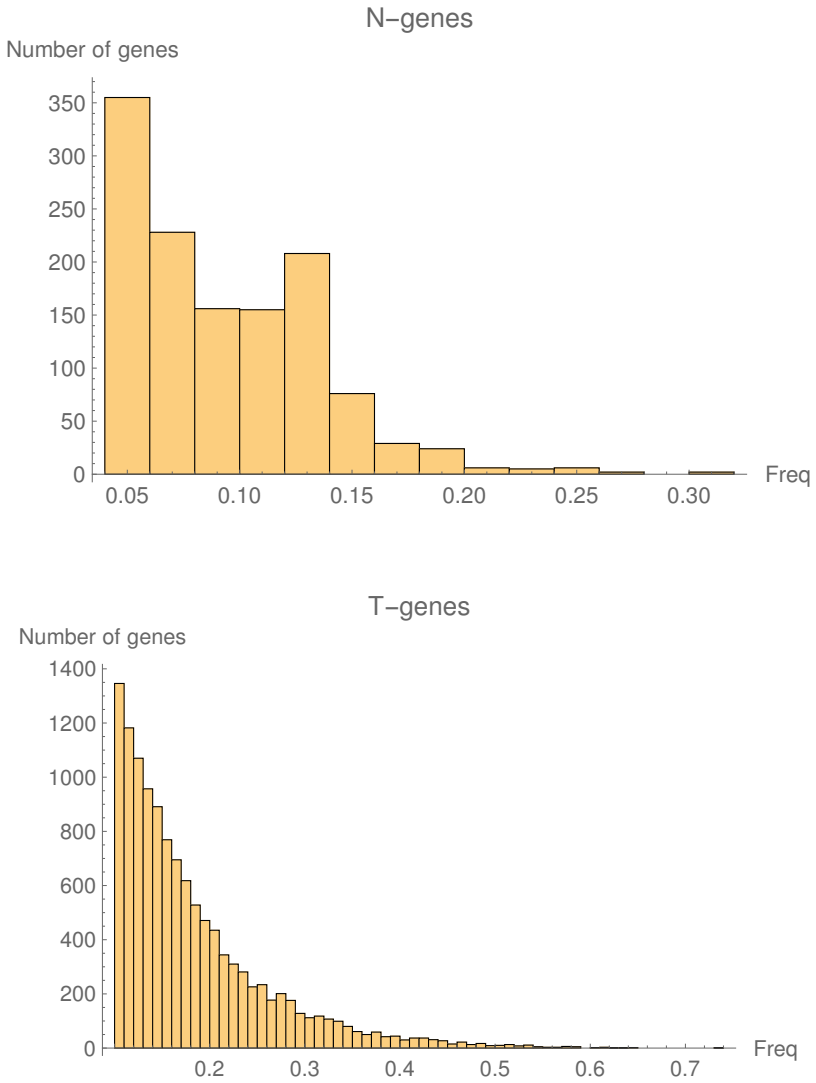

Fig. 5. Frequency histograms of N- and T-genes in PRAD. The frequencies are computed in their respective sub-populations.

Methods.

The methods used in the paper are extremely simple and do not deserve an extensive description.

We define the extrema of the expression intervals for normal and tumor samples and, by comparing them we define the “tumor above”, etc groups. Fisher tests, **Eqs. (1,2)**, are used in order to check that the number of samples found in tumor-only intervals, for example, is significant. A detailed inventory of all the possible groups is provided in **Ref. [3]**.

The list of N-genes in PRAD in Ensembl notation is supplied in the Supplementary file **Ngenes_PRAD.txt.**

We would like to clarify some terminology. In Refs. **[3,8,26]** we call the N-genes “tumor-concept genes”. An, indeed, they take values e = 0 for all tumors, defining in this way the class of tumors. We had, for example, the tumor-concept normal-above group. In the present paper, we simplify the notation and write normal-above group. All the genes with a characteristic interval for normal samples are grouped into the N-genes concept. In the same way, “normal-concept genes”, which specify a characteristic interval for tumor samples are simply named T-genes.

As mentioned, distances in gene expression space with log-fold coordinates are measured using the common euclidean metrics. In order to avoid difficulties with the logarithm we add a small shift of 0.1 to all FPKM data. Details may be found in Refs. **[1,2,10]**.

Availability of data and code

The data is taken from The Cancer Genome Atlas (TCGA) public network (https://portal.gdc.cancer.gov/). We illustrate the analysis with data for prostate cancer (PRAD, **[7]**),

lung adenocarcinoma (LUAD, **[14]**), and liver hepatocarcinoma (LIHC, **[15]**). The code is so simple that it is not necessary to provide it.

Acknowledgments. Authors acknowledge the Office of External Activities of the Abdus Salam Centre for Theoretical Physics for support. The research is carried on under the project PN692LH007-095 of the Financial and International Projects Office of the Ministry of Sciences, Cuba. A.G. is grateful to Rolando Perez for a careful reading of the manuscript.

Author's contributions

G.G. ideated the algorithm for the causal deregulation network and introduced the concepts of N-genes and T-genes. A.G. advanced the interpretation of N-genes and T-genes in oncogenesis and in the clonal evolution of tumors. Both 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.