ACCSCIENCE PUBLISHING

Tumor Discovery

ORIGINAL RESEARCH ARTICLE

Highly specific and sensitive gene panels for cancer screening: First application of only-normal and only-tumor genes

Gabriel Gil[1], Claudia Carricarte[2], Julio C. Drake-Pérez[3], Yasser Perera[4,5], and Augusto Gonzalez[1]*

[1]Department of Theoretical Physics, Institute of Cybernetics, Mathematics and Physics, Havana, Cuba

[2]Group of Computation, Faculty of Biology, University of Havana, Cuba

[3]Department of General Physics, Faculty of Physics, University of Havana, Cuba

[4]Biomedical Research Division, Center for Genetic Engineering and Biotechnology, Havana, Cuba

[5]China-Cuba Biotechnology Joint Innovation Center, Yongzhou Zhong Gu Biotechnology Co., Ltd, Yongzhou, Hunan, China

Abstract

The traditional paradigm of gene expression dysregulation emphasizes log-fold differential expression, with differentially expressed genes presumed to play key roles in relevant biological processes. In cancer, where normal tissue and tumors occupy non-overlapping regions in gene expression space, we propose an alternative and broader framework based on differentially expressed only-tumor genes (T-genes) and non-differentially dysregulated only-normal genes (N-genes). N-genes exhibit expression intervals found exclusively in normal samples, while T-genes display intervals exclusive to tumor samples. These N- and T-genes serve as markers that can be combined into small gene panels capable of perfectly discriminating between normal and tumor tissues. In most cases, these panels highlight biologically significant properties, such as altered glutamine metabolism in tumors. We provide an inventory of perfect gene panels for 12 cancer types, with potential applications in diagnostics and immunotherapy. Significance: Highly specific and sensitive combinatorial gene panels for the identification of 12 types of solid tumors in humans were derived from RNA sequencing expression profiles reported by The Cancer Genome Atlas network (https://www.cancer.gov/ccg/research/genome-sequencing/tcga). The corresponding software is available at the GitHub repository https://github.com/gabriel-gil/GenePan. This study revisits the concept of cancer-related gene expression dysregulation by introducing N-genes and T-genes as novel dysregulation patterns that can be leveraged in diagnosis, tumor classification, and therapeutic interventions.

Keywords: Cancer; Combinatorial gene panel; Expression dysregulation; Only-normal genes; Only-tumor genes

***Corresponding author:**
Augusto Gonzalez
(agonzale@icimaf.cu)

Citation: Gil G, Carricarte C, Drake-Pérez JC, Perera Y, Gonzalez A. Highly specific and sensitive gene panels for cancer screening: First application of only-normal and only-tumor genes. *Tumor Discov*. 2025;4(3):58-69. doi: 10.36922/TD025190035

Received: May 7, 2025

Revised: June 6, 2025

Accepted: June 20, 2025

Published online: July 17, 2025

1. Introduction

The Human Genome Project of the 1990s opened the door to many large-scale omics catalogs.[1] In the following decade, the field advanced further with the advent of

high-throughput microarrays and next-generation RNA sequencing (RNA-seq).[2] These technologies enabled the development of increasingly specialized databases with a focus on biomedical applications.[3] A prominent example is The Cancer Genome Atlas (TCGA), which provides potentially crucial information on cancer detection, treatment, and the fundamental biology of oncogenesis.[4,5] TCGA hosts extensive genomic, epigenomic, transcriptomic, and proteomic data on tumor and normal tissue samples for 33 cancer types.[6] All of this data are publicly available for mining and analysis in pursuit of discovering specific genetic markers and targets.[6] As expected, the current analyses of TCGA data reflect the scale and complexity of this experimental feat of collecting such a vast amount of data.[7,8] However, a definitive consensus on the most adequate set of genes for diagnosis and therapy remains elusive.

Gene discovery relevant to carcinogenesis and tumor progression is partially guided by the assessment of gene dysregulation based on both statistical and biological significance.[9] The paradigmatic kind of gene dysregulation is differential expression,[10] whereby a gene is expressed differently in a tumor compared to a normal tissue. Conventionally, differential expression is associated with cancer only when there is a marked deviation from normal expression levels, typically defined in terms of average values across tumor and normal samples. However, as emphasized by several authors,[11-14] framing gene expression dysregulation solely in terms of central tendency can hinder gene discovery in translational cancer research. Indeed, gene expression levels in tumor or normal tissue samples may differ in their variance or distribution, even when mean values remain unchanged. Consequently, the detection of differential dispersion[12,13] and differential distribution[14] provides a broader perspective on human cancer-related genes by addressing the shortcomings of standard differential expression protocols. Despite their important contributions, these alternative techniques often rest on distributional assumptions that may not reflect the regulatory dynamics of many genes, such as those involved in circadian rhythm control.[15] To the best of our knowledge, the field still lacks sufficiently flexible methods to detect diverse patterns of gene expression dysregulation beyond changes in central tendency.

In this context, we identify novel candidate genes for cancer therapy and diagnostics by applying an original non-parametric approach to gene expression profiles from the TCGA database. Rather than relying on uniform characterizations based on averages or specific distributional shapes, we explore gene-dependent definitions of normal and tumor-like expression using intervals that encompass either all normal or all tumor samples. This allowed us to identify genes that serve as classifiers without false positives or false negatives when distinguishing tumor and normal tissue within the training data. We refer to these as T-genes (differentially expressed only-tumor genes) and N-genes (non-differentially dysregulated only-normal genes). These genes are characterized by specific expression intervals that are exclusively populated by tumor and normal tissue samples, respectively. By combining N- or T-genes, we constructed compact gene panels – referred to as "perfect gene panels" – that perfectly discriminate between tumor and normal samples within the training data.

Our core procedure resembles formal concept analysis[16-27] and rough set theory (RST),[28-39] both with a growing number of applications in omics. The main scope of these techniques is to discover patterns (namely, formal concepts or rough sets) in multivariate data, where a set of attributes is made to correspond to a set of objects through a specific relation.[40,41] This is precisely the framework under consideration, with the following mapping: genes take the role of attributes, clinical samples correspond to objects, and gene expression profiles define the relation between them.[18] Our sets of N-genes and T-genes define both formal and attribute-oriented concepts,[40,41] where the extents of these concepts correspond to either tumor or normal samples, depending on the concept type. Moreover, the perfect gene panels align with the notion of a reduct in RST,[42-45] in the sense that none of their gene members can be removed without compromising the panel's ability to perfectly classify samples.

Perfect gene panels appear in various forms, depending on the location of tumor-exclusive or normal-exclusive intervals within the gene expression space. Some of these panels have a clear interpretation within the state-of-the-art taxonomy of driver genes, provided an interventionist proof of their causal power. For instance, certain panels feature a single gene whose over-expression signals a tumor – a behavior akin to oncogenes. Conversely, for other panels, a single non-silenced gene is an indication of a tumor-free sample, which fits our current understanding of tumor suppressor genes. Other panels may include cooperative tumor suppressor genes, oncogenes, and oscillatory genes.

In this paper, we explore 12 solid tumors among the 33 cancer types in TCGA. For each tissue analyzed, we identify perfect gene panels with potential applications in diagnosis and therapy. By design, perfect panels achieve zero false positives or false negatives within the training data. Notably, one T-gene panel for lung adenocarcinoma (LUAD) also demonstrated high sensitivity and specificity in an external dataset.

The remainder of the article is structured as follows. Section 2 (Materials and Methods) provides a detailed and thorough account of our methodology. In Section 3 (Results), we illustrate our workflow for a case in point (namely, LUAD) and summarize our findings for the other selected cancer types. In addition, we provide a validation analysis of our gene panels in different datasets. Section 4 (Discussion) addresses gene dysregulation as conceptualized in this study, highlighting how it enables us to better understand homeostasis and cancer. We further examine potential applications of the proposed gene panels and their role in tumorigenesis. Section 5 (Conclusion) summarizes the key findings and offers an outlook for future translational research based on this framework.

2. Materials and methods

2.1. Data

TCGA is a publicly accessible database of gene expression profiles drawn from cohort studies involving hundreds of normal tissue and solid tumor biopsy samples, classified by histopathological techniques.[6] Expression profiles were obtained through RNA-seq, capturing 60,483 genes per sample. TCGA reports gene expression values using the standard units of fragments per kilobase of transcript per million mapped reads. The size of the dataset varies with the cancer type and is consistently skewed toward tumor samples.

We selected 12 cancer types from TCGA for a systematic analysis (Table 1). These cancers manifest as solid tumors, particularly affecting the liver, breast, colon, head and neck, kidneys, lungs, prostate, stomach, thyroid, and uterus. We included five (out of six) of the most common cancer types (breast, lung, colon, prostate, and stomach), each with an incidence of over a million cases in 2020. Among the selected cancer types, there were also the most common causes of cancer death (lung, colon, liver, stomach, and breast), each accounting for over half a million deaths in 2020 (worldwide statistics reported by the World Health Organization[46]).

Table 1. Cancer types, the cancer genome atlas abbreviations, and the number of samples

Cancer types	Abbreviation	Normal samples	Tumor samples
Breast invasive carcinoma	BRCA	112	1,096
Colon adenocarcinoma	COAD	41	473
Head and neck squamous cell carcinoma	HNSC	44	502
Kidney renal clear cell carcinoma	KIRC	74	539
Kidney renal papillary cell carcinoma	KIRP	32	289
Liver hepatocellular carcinoma	LIHC	50	374
Lung adenocarcinoma	LUAD	59	535
Lung squamous cell carcinoma	LUSC	49	502
Prostate adenocarcinoma	PRAD	52	499
Stomach adenocarcinoma	STAD	32	375
Thyroid carcinoma	THCA	58	510
Uterine corpus endometrial carcinoma	UCEC	23	552

The selection of cancer types for our systematic study was motivated by the number of normal samples available in the data. For the cases under study, TCGA reports more than 20 normal samples per cancer type. Notably, achieving a reliable discrimination between normal and tumor tissues based on gene expression profiles required both normal and tumor samples to be adequately represented in the datasets.

2.2. Pre-processing of data

Gene expression distributions tend to be heavy-tailed, with many low-frequency outliers.[47] RNA-seq is known to be inaccurate at detecting low expression levels and may produce spurious null readings for genes that are nearly silenced.[48] To avoid artifacts associated with the low-expression region, we set all values below 0.1 fragments per kilobase of transcript per million mapped reads to zero. Moreover, we excluded all genes with non-zero expression in fewer than 5% of normal samples and fewer than 10% of tumor samples from the analysis.

2.3. Expression dysregulation patterns

We searched for genes that exhibit specific dysregulation patterns. In our framework, a gene conforms to a "differential expression" pattern if all normal samples express it in a certain manner (specified below), while a significant number of tumor samples exhibit a distinctly different expression. Conversely, a gene conforms to a "non-differential dysregulation" pattern if all tumor samples express it in a certain way, while a substantial number of normal samples express it differently. Non-differential dysregulation can be interpreted as the dual category of differential expression, achieved by swapping the roles of normal and tumor samples. By monitoring the expression values of a differentially expressed or non-differentially dysregulated gene, we can classify samples with no type I errors – i.e., no false positives for tumors in the case of differential expression and no false positives for normal samples in the case of non-differential dysregulation.

For simplicity, this study focuses on four types of gene sets, each named to reflect the classificatory potential of its individual gene members. Let x represent a class of samples,

either normal (N) or tumor (T). Genes that are only expressed above or below a threshold level for class x are referred to as "only x above" or "only x below," respectively. Specifically, we examined the "only-T-above," "only-T-below," "only-N-above," and "only-N-below" gene sets. By combining the "above" and "below" within the same class, we obtained the full sets of T-genes and N-genes. Notably, a single gene may simultaneously belong to both the only-T-above and only-N-below groups.

2.4. Data digitalization

We explicitly defined normal and tumor expression intervals for each gene. In each case, the populated expression space can be segmented into three regions: "N-only," "N-T," and "T-only" subintervals, which were associated with the ternary values −1, 0, and 1, respectively.

Figure 1 shows the distribution of expression values for *PYCR1*, *ALDH18A1*, and *TRIM27* genes in normal lung and LUAD samples. Notably, all three genes contain only-T intervals above the common N-T region. The number of tumor samples in the only-T interval is significant (above 90% of the tumor population). Thus, they may be included in the only-T-above set of genes.

These genes also show N-only intervals below the N-T region. However, the number of samples in the N-only intervals may not be sufficient to be included in the only-N-below class.

2.5. Statistically significant expression dysregulations

The significance of dysregulation patterns within the T-only and N-only sample subsets can be assessed using Fisher's exact test[49] to filter out genes exhibiting such patterns by chance.

Verifications show that with a p=0.01 and the sample sizes in Table 1, a dysregulation pattern is significant when observed in approximately 5% of normal samples (N-only subset) or 10% of tumor samples (T-only subset). We applied these thresholds, respectively, across all cancer types. This threshold justifies the exclusion of certain genes from analysis and explains why some genes identified in the previous subsection do not appear in the only-N-below set.

2.6. Expression dysregulation matrix

Gene expression profiles were encoded into a matrix where each column corresponded to a clinical sample and each row represented a significantly dysregulated gene. The matrix entries, derived from the prior data digitalization step, were assigned values of −1, 1, and 0, indicating

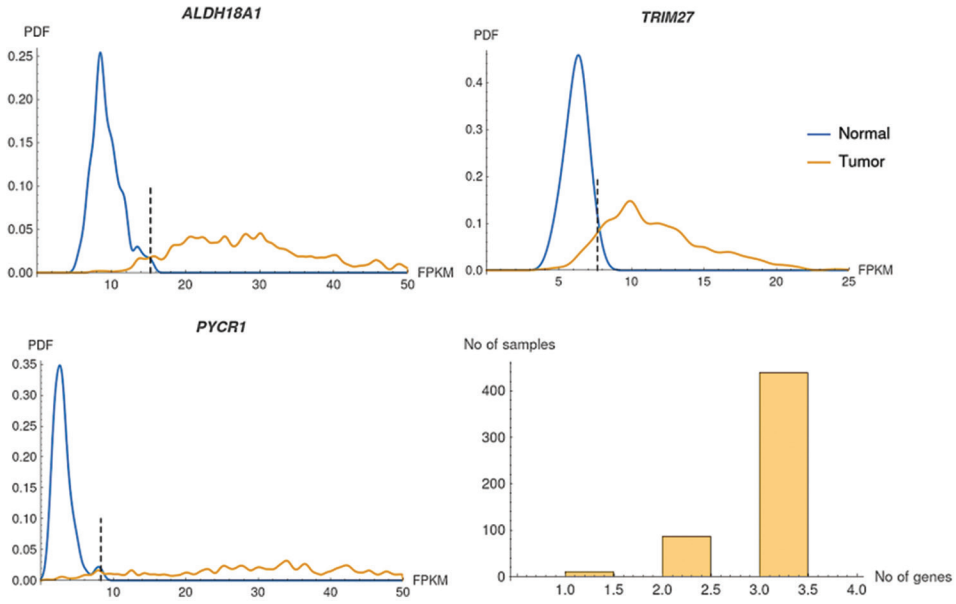

Figure 1. The Cancer Genome Atlas-Lung adenocarcinoma gene expression data for three "only-T-above" genes forming a perfect panel. Smooth probability density functions (PDF) are shown as solid lines, whereas the maximum of the normal set of values (the threshold) is marked by a dashed line. There are intervals for each gene common to both normal and tumor samples (expression values below the threshold), and "T-only" intervals populated only by tumor samples (expression above the threshold). The histogram shows that there is at least one dysregulated gene, i.e., with expression above the threshold, for each tumor sample; thus, the panel correctly classifies all of the normal samples with 0 dysregulated genes and all of the tumors, which show at least one dysregulated gene.
Abbreviation: FPKM: Fragments per kilobase of transcript per million mapped reads.

whether the gene's expression interval was N-exclusive, T-exclusive, or shared by N and T. This matrix structure provided all the necessary information for constructing perfect gene panels.

2.7. Perfect panels

Differentially expressed and non-differentially dysregulated genes often form large pools containing over a thousand members, which are impractical for real-world applications. In genetic-based hereditary risk assessment, diagnostics, and therapy, smaller gene panels (comprising 5 – 50 genes) are often preferred.[50]

Due to the low dimensionality of the gene expression data,[51] it is possible to extract compact panels from these large gene sets. In particular, panels can be designed to perfectly classify all normal and tumor samples collectively, with the additional requirement that removing any member from the panel would compromise this classification accuracy.

These panels can be identified using a concept similar to, but distinct from, reducts in RST,[42,43] which we termed a formal-concept reduct. To the best of our knowledge, this is the first presentation of formal-concept reducts,[44] although more stringent related concepts have been proposed by Zhang.[45]

Our algorithm for constructing perfect panels is based on progressively maximizing sensitivity. At each step, we iteratively add the differentially expressed genes that are most dysregulated in tumor samples not yet identified by the current panel (i.e., those samples where the included genes show no dysregulation), until all tumor samples are discovered.

The equivalent procedure involves iteratively adding the non-differentially dysregulated gene that most frequently exhibits normal regulation in the remaining undiscovered normal samples (i.e., those in which the genes already included are dysregulated) until all normal samples are discovered. If, at any iteration, there is gene selection ambiguity, we prioritize the most redundant candidate – i.e., the gene whose dysregulation pattern overlaps maximally with existing panel members across already classified samples. Further ambiguities are resolved by arbitrarily selecting the first candidate in the list.

Panels constructed this way are minimal: no gene can be removed without compromising perfect classification. However, they are not necessarily the smallest collection of genes achieving such goal nor are they necessarily unique. Modifying ambiguity-resolution criteria may give rise to different and/or smaller gene sets that can achieve perfect discrimination between normal and tumor samples, while remaining irredundant. In practice, these panels comprise 1 – 20 genes, making them suitable for cancer diagnostics.[50]

In the example considered in Section 2.4, the three-gene set constitutes a perfect panel for the only-T-above class. In its expression dysregulation matrix, normal samples show expression values of −1 or 0. Every tumor sample has at least one dysregulated gene (value 1) in the panel, as shown in the histogram of Figure 1. Thus, this panel exhibits no false negatives or false positives.

3. Results

First, we note that, in the average cancer type, only nearly 3% of the genes qualify as N-genes. The observation that more than one-third of the genome, and the vast majority of classifier genes, fall within the T-gene category aligns with cancer's characterization as a high-entropic state of gene regulatory networks[52,53] and is an indication of the abundance of potential genetic triggers for cancer.

Perfect panels constructed according to our procedure are summarized in Table 2. When no perfect panel exists, we reported the size of the minimal gene set that classifies the largest sample subset. Notably, only-T-above and only-T-below panels may include oncogenes and tumor suppressors, respectively. As shown in Table 2, all 12 cancer types exhibit perfect panels of both T-types.

Conversely, perfect panels with only-N-above or only-N-below genes appear irregularly in some tissues (Table 2). Specifically, breast invasive carcinoma (BRCA), head and neck squamous cell carcinoma, kidney renal clear cell carcinoma, kidney renal papillary cell carcinoma, LUAD, prostate adenocarcinoma, and thyroid carcinoma contain only only-N-above, uterine corpus endometrial carcinoma, colon adenocarcinoma (COAD), lung squamous cell carcinoma, and stomach adenocarcinoma contain both N-types, while liver hepatocellular carcinoma contains only only-N-below.

An inventory of perfect gene panels for the 12 types of cancer under study is presented in the Supplementary File. Notably, some cancer types can be perfectly classified using a single gene. This is the case for COAD with *SCARA5*, kidney renal papillary cell carcinoma with *UMOD*, and uterine corpus endometrial carcinoma with either *PLSCR4* or *TBC1D7*.

4. Discussion

4.1. Gene expression dysregulation

Dysregulation in gene expression can promote cancer.[54] Within this phenomenon, differential expression – where genes show altered expression in tumors versus normal tissues – represents the most extensively studied subset.[10]

Table 2. Summary of classifier genes per tissue

Set of genes	LIHC	BRCA	COAD	HNSC	KIRC	KIRP	LUAD	LUSC	PRAD	STAD	THCA	UCEC
Only-T-above	[a]3/23,986	[a]6/15,361	[a]2/17,536	[a]4/13,293	[a]4/22,654	[a]3/11,447	[a]3/20,274	[a]2/19,596	[a]8/8,093	[a]3/13,773	[a]5/5,744	[a]1/7,825
Only-N-above	11/40	[a]10/739	[a]1/876	[a]8/1,903	[a]3/780	[a]1/1,140	[a]5/613	[a]3/1,198	[a]14/1,415	[a]5/1,244	[a]11/794	[a]1/993
Only-T-below	[a]5/3,812	[a]6/6,701	[a]1/8,418	[a]5/2,093	[a]3/9,132	[a]1/10,263	[a]4/8,285	[a]2/9,404	[a]15/3,865	[a]5/1,499	[a]6/5,376	[a]1/7,443
Only-N-below	[a]5/1,246	12/682	[a]2/297	6/1,339	8/191	5/214	8/449	[a]3/985	15/915	[a]5/2,536	17/92	[a]1/506

Note: Each column identifies a cancer type based on The Cancer Genome Atlas terminology. Each row represents a different set of classifier genes (see main text for shorthand notation). Within each cell, we show the minimal number of genes that classify the largest number of samples, together with the total number of genes of the same sort. [a]marks the minimal gene sets that constitute perfect panels.
Abbreviations: BRCA: Breast invasive carcinoma; COAD: Colon adenocarcinoma; HNSC: Head and neck squamous cell carcinoma; KIRC: Kidney renal clear cell carcinoma; KIRP: Kidney renal papillary cell carcinoma; LIHC: Liver hepatocellular carcinoma; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; N: Normal; PRAD: Prostate adenocarcinoma; STAD: Stomach adenocarcinoma; T: Tumor; THCA: Thyroid carcinoma; UCEC: Uterine corpus endometrial carcinoma.

In cancer research, differential expression is often deemed significant only when the deviation from normal expression is substantial, consistent (i.e., always upregulated or downregulated), and present across most tumors.[55] A common practice is to define the lower and upper bounds of normal gene expression as ×0.5 and ×2 a reference level, respectively.[56] Therefore, a gene can be differentially expressed only if most tumors express either >×2 or <×0.5 the reference value. In turn, any such gene is considered differentially expressed when its expression level crosses the specific threshold above or below which most tumors are expressed.

From the outset, we contend that gene expression dysregulation comprises broader patterns than conventional differential expression. Certain dysregulation forms do not conform to either our definition of differential expression or the conventional one used in the field. As a result, these patterns are often overlooked in the analysis of gene expression data.

For example, consider a gene with bimodal expression distribution under normal conditions, such as those governed by circadian oscillations. If these oscillations are lost in tumor tissue, the gene may fall into the only-T-inside category. While such genes were identified through our data mining, they are not reported in this paper. Other underreported categories, like the only-T-outside genes, were also encountered.

Conversely, what we term as non-differential dysregulation, corresponding to N-genes, is typically overlooked. In our study, we focused on the only-N-above and only-N-below classes, although the only-N-outside and only-N-inside groups may likewise be present in specific tissues.

It is worth emphasizing that in single-cell RNA-seq expression analyses,[3] gene markers are routinely identified for individual cell types under normal conditions. However, to the best of our knowledge, this is the 1st time that markers are introduced specifically for whole normal tissue samples.

4.2. Panel validation

We provided two examples of panel validation using other datasets. The first involves the *SCARA5* gene in COAD. Microarray readings from Khamas *et al.*[57] demonstrate the perfect classification capability of *SCARA5* (data available at NCBI GEO,[58] accession GDS4382). Notably, this gene has also been independently identified as a biomarker for colorectal cancer.[59]

The second case concerns the LUAD dataset from a comprehensive study of a Chinese cohort,[60] which includes RNA-seq profiles from 51 tumor and 49 control samples. We evaluated the performance of our perfect only-T-above panel on this dataset. As shown in Figure S1, the genes *TRIM27*, *PYCR1*, and *ALDH18A1* fall within the only-T-above class, as they exhibit significantly populated T-exclusive intervals above the shared N–T expression range. The histogram in Figure S1 confirms that the panel remains perfect, achieving both maximal sensitivity and specificity in classification. However, within this particular cohort, the *TRIM27* gene proves redundant and can be removed without any loss in classification accuracy.

This finding raises an important question regarding the minimal number of genes required to assemble a perfect panel, and the extent to which that number remains robust to variations in cohort size.

4.3. The minimal number of genes needed to identify a tumor

The LUAD dataset[60] is particularly noteworthy, not only because its cohort differs markedly from that of TCGA but also due to its substantially smaller size, approximately an order of magnitude fewer samples. Specifically, the TCGA LUAD dataset comprises 59 normal and 535 tumor samples. This prompts the question: how does the number

of genes required for a perfect panel depend on the size of the tumor sample set?

The results, summarized in Figure S2, revealed that in the smaller external dataset, a single gene identifies 98% of the tumor samples, and the addition of a second gene completes the panel, achieving maximal sensitivity and specificity without requiring *TRIM27*. In contrast, for the larger TCGA dataset, the first gene alone covers only 95% of tumors, and the two-gene panel still leaves 1% of samples unclassified. In that case, *TRIM27* is necessary to achieve full classification. These observations suggest rare tumor variants emerge only in larger datasets. Their low frequency means that they are often absent in smaller cohorts, where simpler panels may suffice.

For illustration, a hypothetical cohort of 5,000 tumor samples is also considered in Figure S2. In that scenario, the 3-gene panel covers 99.7% of tumors, indicating that a fourth gene would likely be needed to achieve complete coverage. The figure also shows that saturation is reached very quickly: the number of classified tumor samples increases steeply with the addition of genes to the panel. This strongly supports our assertion that a small number of genes can effectively capture the global state of the Gene Regulatory Network, consistent with the effective reduced dimensionality of the tumor manifold.[51]

In summary, the expression distribution functions used to define the panels depend on the sample set size. When the sample size reaches the order of hundreds, the distribution appears "saturated," showing only minor changes when the number of samples is further increased.

This insight allowed us to evaluate how our panels would change with an increased number of normal samples. For instance, assuming that the distribution functions are saturated in BRCA (112 normal samples and 1094 tumor samples), we performed re-sampling to assess the performance of the six-gene only-T-above panel found for BRCA (Supplementary File) under highly imbalanced situations, such as 20 normal samples and 500 tumor samples. The results, shown in Figure S3, indicate that while the panel size tends to decrease in the reduced sets, notably, two genes from the original panel still classify more than 95% of samples in all cases.

Thus, we expect, for example, that the single-gene only-T-above panel found for uterine corpus endometrial carcinoma (23 normal samples) may change as the normal sample size grows, but the original gene will continue to cover at least 85% of the tumor samples.

It is worth noting that Figure S3 can also be interpreted as a form of validation of the six-gene only-T-above panel for BRCA across different experimental conditions.

4.4. Cancer diagnosis, tumor taxonomy, and gene therapy

Our construction of perfect gene panels follows a data-driven approach to gene expression profiles that do not require prior domain knowledge of the biological relevance of individual genes in a given tissue. These panels have an apparent value as candidate combinatorial biomarkers for diagnosis, which could be further enhanced by incorporating information about gene ontology and function into our data mining process.

In addition, the perfect T-gene panels could be leveraged in tumor taxonomy. Typically, tumor classification and the associated therapeutic decisions are made based on the most frequently mutated genes in a given tumor (for example, Ruiz-Cordero *et al.*,[61] for lung cancer). However, the classification is often incomplete, with a subset of tumors assigned to the so-called "wild-type" category, meaning that none of the genes in the reference panel are mutated. In our framework, any perfect T-gene panel enables a complete classification of tumors by providing the list of dysregulated genes in each tumor sample. Moreover, since multiple perfect panels may exist for a given tissue, tumors could be fully classified under different but equally valid criteria.

Consider, for example, the only-T-above panel for LUAD, examined above. Both *ALDH10A1* and *PYCR1* genes, related to glutamine metabolism, are known to play an important role in lung cancer.[62,63] The taxonomy based on this panel indicates that around 98 % of LUAD tumors rely on glutamine metabolism to foster cell proliferation and induce an immune-suppressive tumor microenvironment. In the remaining 2% of tumors, cell proliferation is regulated by *TRIM27* through the SIX homeobox 3-β-catenin signaling pathway.[64] These statements reflect the known role of these genes and their dysregulation frequencies in the tumor subpopulation. Nevertheless, further research is needed to validate these findings and translate them into therapeutic recommendations.

Moreover, N- and T-genes included in the perfect panels may have important applications in gene therapy. Consider, for instance, a gene belonging to both N- and T-groups, such as the *AGER* gene in LUAD. This gene is silenced in tumors and strongly expressed in normal samples. What happens if, through a transfection vector, its expression were shifted from the N-region to the T-region or vice versa? Such an experiment has already been conducted on cellular lines,[65] and the results indicate a significant change in the proliferation rate and invasion capacity of both tumor and normal cells. These astonishing results warrant further investigation.

4.5. Other challenges

Several other challenges remain, such as the role of low-expressed genes, the method's performance in highly heterogeneous tumors, and the possible impact of batch effects. In principle, our gene panels are robust against these concerns. Specifically, our N- and T-genes exhibit distinct expression intervals populated only by a significant fraction of N and T samples, respectively. They are not low-expressed genes. Regarding batch effects, the TCGA data used for panel discovery is largely free from such biases, as all samples were processed using a consistent technological framework and standardized procedures. For each tissue, there is a single batch of normal samples and a single batch of tumor samples. Concerning tumor heterogeneity, because the taxonomy derived from a panel is comprehensive, the panel should be capable of detecting tumors regardless of their mosaic composition or degree of heterogeneity.

5. Conclusion

We have shown that it is possible to construct a combinatorial gene panel that acts as a perfect biomarker for cancer. By monitoring the gene expression profile of the panel members, samples can be accurately classified as either normal or tumorous. In some cases, it is possible to classify a sample as tumorous based on the overexpression of a single gene. However, this represents just one example among various panel types, all of which are highly sensitive and specific.

Our study analyzed 12 cancer types from the TCGA database, encompassing many of the most prevalent cancers in the world. Panels are provided on a per-cancer-type basis, tailored to each specific context. A comprehensive inventory of these panels can be found in the Supplementary Information. Despite the fact that other panels combining classifier genes could be constructed, these are not discussed in the present paper.

While a single gene can have sufficient discriminative power in one tissue, other tissues require panels of up to nine genes to achieve the same level of accuracy. Figure S4 shows the relationship between panel length and the distance between the centers of the normal and tumor sample clusters in gene expression space.[66,67] It is evident that the shorter inter-cluster distances correspond to greater overlap between normal and tumor expression profiles, complicating classification and necessitating larger panels. The figure also suggests that the inter-cluster distance in gene expression space functions as a global tumor classifier, a factor often overlooked in tumor studies.

Our gene discovery framework extends beyond the paradigm of differential expression by introducing the concepts of N-genes and T-genes, characterized by gene expression intervals populated only by normal and tumor samples, respectively. The construction of perfect gene panels represents the first practical application of these concepts, which we anticipate can be translated into flexible and effective diagnostic tools.

In addition, this paper presents arguments supporting the use of perfect panels in tumor taxonomy and highlights their gene members as candidate targets of therapeutic applications. Other potential applications, such as early diagnosis and efficacy monitoring, alongside challenges, like technical standardization and cost considerations in clinical implementation, are particularly important and warrant further attention. Research in this direction is currently in progress.

Acknowledgments

The author, Augusto Gonzalez is grateful to Rolando Perez for a careful reading of the manuscript. The author, Gabriel Gil thanks Laura Azor, Fabiana Fuentes, Jorge Mato, and Karen Alfaro for critical comments and suggestions.

Funding

The research was supported by the Financial and International Projects Office of the Ministry of Sciences, Cuba (project PN692LH007-095).

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Gabriel Gil, Augusto Gonzalez
Formal analysis: Gabriel Gil, Julio C. Drake-Pérez
Investigation: All authors
Methodology: Gabriel Gil, Augusto Gonzalez
Writing–original draft: Gabriel Gil, Augusto Gonzalez
Writing–review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The data used was taken from TCGA public network (https://portal.gdc.cancer.gov/). Relevant software is available at the GitHub repository https://github.com/gabriel-gil/GenePan.

Further disclosure

Initial versions of the paper have been deposited in the bioRxiv preprint server (doi: 10.1101/2022.07.25.501449, 10.1101/2024.07.25.604730).

References

1. Collins FS, Morgan M, Patrinos A. The human genome project: Lessons from Large-scale biology. *Science.* 2003;300(5617):286-290.

   doi: 10.1126/science.1084564

2. Chu Y, Corey DR. RNA sequencing: Platform selection, experimental design, and data interpretation. *Nucleic Acid Ther.* 2012;22(4):271-274.

   doi: 10.1089/nat.2012.0367

3. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* 2017;9(1):75.

   doi: 10.1186/s13073-017-0467-4

4. The Cancer Genome Atlas Research Network, Weinstein J, Collisson E, *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Gen.* 2013;45(10):1113-1120.

   doi: 10.1038/ng.2764

5. Hutter C, Zenklusen JC. The cancer genome atlas: Creating lasting value beyond its data. *Cell.* 2018;173(2):283-285.

   doi: 10.1016/j.cell.2018.03.042

6. The Cancer Genome Atlas Research Network. *The Cancer Genome Atlas*. 2006. Available from: https://www.cancer.gov/tcga [Last accessed on 2025 Apr 15].

7. Cheng PF, Dummer R, Levesque MP. Data mining the cancer genome atlas in the era of precision cancer medicine. *Swiss Med Wkly.* 2015;145:w14183.

   doi: 10.4414/smw.2015.14183

8. Liñares-Blanco, J, Pazos, A, Fernandez-Lozano, C. Machine learning analysis of TCGA cancer data. *PeerJ Comput Sci.* 2021;7:e584.

   doi: 10.7717/peerj-cs.584

9. Li Q, Dai W, Liu J, Sang Q, Li YX, Li YY. Gene dysregulation analysis builds a mechanistic signature for prognosis and therapeutic benefit in colorectal cancer. *J Mol Cell Biol.* 2020;12(11):881-893.

   doi: 10.1093/jmcb/mjaa041

10. Ali HEA, Lung PY, Sholl AB, *et al.* Dysregulated gene expression predicts tumor aggressiveness in African-American prostate cancer patients. *Sci Rep.* 2018;8(1):16335.

    doi: 10.1038/s41598-018-34637-8

11. Mezlini AM, Das S, Goldenberg A. Finding associations in a heterogeneous setting: Statistical test for aberration enrichment. *Genome Med.* 2021;13(1):68.

    doi: 10.1186/s13073-021-00864-4

12. Le Priol C, Azencott CA, Gidrol X. Detection of genes with differential expression dispersion unravels the role of autophagy in cancer progression. *PLoS Comput Biol.* 2023;19(3):e1010342.

    doi: 10.1371/journal.pcbi.1010342

13. Li H, Khang TF. clrDV: A differential variability test for RNA-Seq data based on the skew-normal distribution. *PeerJ.* 2023;11:e16126.

    doi: 10.7717/peerj.16126

14. Roberts AGK, Catchpoole DR, Kennedy PJ. Identification of differentially distributed gene expression and distinct sets of cancer-related genes identified by changes in mean and variability. *NAR Genom Bioinform.* 2022;4(1):lqab124.

    doi: 10.1093/nargab/lqab124

15. Andreani TS, Itoh TQ, Yildirim E, Hwangbo DS, Allada R. Genetics of circadian rhythms. *Sleep Med Clin.* 2015;10(4):413-421.

    doi: 10.1016/j.jsmc.2015.08.007

16. Gebert J, Motameny S, Faigle U, Forst CV, Schrader R. Identifying genes of gene regulatory networks using formal concept analysis. *J Comput Biol.* 2008;15(2):185-194.

    doi: 10.1089/cmb.2007.0107

17. Choi V, Huang Y, Lam V, Potter D, Laubenbacher R, Duca K. Using formal concept analysis for microarray data comparison. *J Bioinform Comput Biol.* 2008;6(1):65-75.

    doi: 10.1142/s021972000800328x

18. Motameny S, Versmold B, Schmutzler R. Formal Concept Analysis for the Identification of Combinatorial Biomarkers in Breast Cancer. In: Medina R, Obiedkov S, editors. *Formal Concept Analysis. ICFCA 2008. Lecture Notes in Computer Science*. Vol. 4933. Berlin, Heidelberg: Springer; 2008. p 229–240.

    doi: 10.1007/978-3-540-78137-0_17

19. Amin II, Kassim SK, Hassanien A, Hefny HA. Formal Concept Analysis for Mining Hypermethylated Genes in Breast Cancer Tumor Subtypes. In: *12th International Conference on Intelligent Systems Design and Applications (ISDA)*. Kochi, India; 2012. p. 764-769.

    doi: 10.1109/ISDA.2012.6416633

20. Kaytoue-Uberall M, Duplessis S, Napoli A. Using Formal Concept Analysis for the Extraction of Groups of Co-expressed Genes. In: Le Thi HA, Bouvry P, Pham Dinh T, editors. *Modelling, Computation and Optimization in Information Systems and Management Sciences. MCO 2008. Communications in Computer and Information Science*. Vol. 14. Berlin, Heidelberg: Springer; 2008.

    doi: 10.1007/978-3-540-87477-5_47

21. Kaytoue M, Kuznetsov SO, Napoli A, Duplessis S. Mining gene expression data with pattern structures in formal concept analysis. *Inf Sci.* 2011;181(10):1989-2001.

    doi: 10.1016/j.ins.2010.07.007

22. González-Calabozo JM, Valverde-Albacete FJ, Peláez-Moreno C. Interactive knowledge discovery and data mining on genomic expression data with numeric formal concept analysis. *BMC Bioinform.* 2016;17(1):374.

    doi: 10.1186/s12859-016-1234-z

23. Singh PK, Kumar CA, Gani AA. Comprehensive survey on formal concept analysis, its research trends, and applications. *Int J Appl Math Comput Sci.* 2016;26(2):495-516.

    doi: 10.1515/amcs-2016-0035

24. Raza K. Formal concept analysis for knowledge discovery from biological data. *Int J Data Min Bioinform*. 2017;18(4):281.

    doi: 10.1504/IJDMB.2017.088138

25. Ferreira LM, Pinto CLN, Dias SM, Nobre CN, Zárate LE. Extraction of Conservative Rules for Translation Initiation Site Prediction Using Formal Concept Analysis. In: *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS).* Vol. 1. SciTePress; 2017. p. 265-271.

    doi: 10.5220/0006326202650271

26. Zhao M, Zhang S, Li W, Chen G. Matching biomedical ontologies based on formal concept analysis. *J Biomed Semantics.* 2018;9(1):11.

    doi: 10.1186/s13326-018-0178-9

27. Roscoe S, Khatri M, Voshall A, Batra S, Kaur S, Deogun J. Formal concept analysis applications in bioinformatics. *ACM Comput Surv.* 2023;55(8):1-40.

    doi: 10.1145/3554728

28. Maji P, Paul S. Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. *Int J Approx Reason.* 2011;52(3):408-426.

    doi: 10.1016/j.ijar.2010.09.006

29. Midelfart H, Komorowski J, Nørsett K, Yadetie F, Sandovik AK, Lægreid A. Learning rough set classifiers from gene expressions and clinical data. *Fundam Inform.* 2002;53(2):155-183.

    doi: 10.3233/FUN-2002-53204

30. Dai J, Xu Q. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Appl Soft Comput.* 2013;13(1):211-221.

    doi: 10.1016/j.asoc.2012.07.029

31. Li D, Zhang W. Gene selection using rough set theory. In Wang GY, Peters JF, Skowron A, Yao Y, editors. *Rough Sets and Knowledge Technology. RSKT 2006. Lecture Notes in Computer Science*. Vol. 4062. Berlin, Heidelberg: Springer; 2006. p. 778-785.

    doi: 10.1007/11795131_113

32. Mishra D, Dash R, Rath AK, Acharya M. Feature selection in gene expression data using principal component analysis and rough set theory. In: Arabnia HR, Tran QN, editors. *Software Tools and Algorithms for Biological Systems. Advances in Experimental Medicine and Biology*. Vol. 696. New York: Springer; 2011. p. 91-100.

    doi: 10.1007/978-1-4419-7046-6_10

33. Pati SK, Das AK, Ghosh A. Gene Selection Using Multi-objective Genetic Algorithm Integrating Cellular Automata and Rough Set Theory. In: Panigrahi BK, Suganthan, PN, Das S, Dash SS, editors. *Swarm, Evolutionary, and Memetic Computing. SEMCCO 2013. Lecture Notes in Computer Science*. Vol. 8298. Cham: Springer; 2013. p. 144-155.

    doi: 10.1007/978-3-319-03756-1_13

34. Zhang Q, Xie Q, Wang G. A survey on rough set theory and its applications. *CAAI Trans Intell Technol.* 2016;1(4):323-333.

    doi: 10.1016/j.trit.2016.11.001

35. Chen Y, Zhang Z, Zheng J, Ma Y, Xue Y. Gene selection for tumor classification using neighborhood rough sets and entropy measures. *J Biomed Inform.* 2017;67:59-68.

    doi: 10.1016/j.jbi.2017.02.007

36. Sun L, Zhang X, Xu J, Wang W, Liu R. A Gene selection approach based on the fisher linear discriminant and the neighborhood rough set. *Bioengineered*. 2018;9(1):144-151.

    doi: 10.1080/21655979.2017.1403678

37. Saha S, Roy S, Ghosh A, Dey KN. Gene-Gene Interaction Analysis: Correlation, Relative Entropy and Rough Set Theory Based Approach. In: *Bioinformatics and Biomedical Engineering: 6th International Work-Conference, IWBBIO 2018. Proceedings, Part II*. Granada, Spain: Springer-Verlag; 2018. p. 397-408.

    doi: 10.1007/978-3-319-78759-6_36

38. Patil S, Balmuri KR, Frnda J, Parameshachari BD, Konda S, Nedoma J. Identification of triple-negative breast cancer genes using rough set-based feature selection algorithm and ensemble classifier. *Hum Centric Comput Inf Sci.* 2022;12:54.

    doi: 10.22967/HCIS.2022.12.054

39. Majumder S, Thakran Y, Pal V, Singh K. Fuzzy and rough set theory based computational framework for mining genetic interaction triplets from gene expression profiles for lung adenocarcinoma. *IEEE/ACM Trans Comput Biol Bioinform.* 2022;19(6):3469-3481.

    doi: 10.1109/TCBB.2021.3120844

40. Duntsch N, Gediga G. Modal-style Operators in Qualitative Data Analysis. In: *2002 IEEE International Conference*

on Data Mining Proceedings. Maebashi City, Japan; 2002. p. 155-162.

doi: 10.1109/ICDM.2002.1183898

41. Lai H, Zhang D. Concept lattices of fuzzy contexts: Formal concept analysis vs. rough set theory. *Int J Approx Reason.* 2009;50(5):695-707.

doi: 10.1016/j.ijar.2008.12.002

42. Pawlak, Z. Rough sets. *Int J Comput Inf Sci.* 1982;11(5):341-356.

doi: 10.1007/BF01001956

43. Pawlak Z. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Dordrecht: Springer; 1991.

doi: 10.1007/978-94-011-3534-4

44. Jia X, Shang L, Zhou B, Yao Y. Generalized attribute reduct in rough set theory. *Knowl Based Syst*. 2016;91:204-218.

doi: 10.1016/j.knosys.2015.05.017

45. Zhang W. Attribute reduction theory and approach to concept lattice. *Sci China Ser F Inf Sci.* 2005;48(6):713-726.

doi: 10.1360/122004-104

46. World Health Organization. *Cancer*. Available from: https://www.who.int/news-room/factsheets/detail/cancer [Last accessed on 2025 April 15].

47. Bengtsson M, Ståhlberg A, Rorsman P, Kubista M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res.* 2005;15(10):1388-1392.

doi: 10.1101/gr.3820805

48. Sha Y, Phan JH, Wang MD. Effect of Low-expression Gene Filtering on Detection of Differentially Expressed Genes in RNA-seq Data. In: *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2015. p. 6461.

doi: 10.1109/EMBC.2015.7319872

49. Fang Z, Martin J, Wang Z. Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell Biosci.* 2012;2(1):26.

doi: 10.1186/2045-3701-2-26

50. Durães C, Pereira Gomes C, Costa JL, Quagliata L. Demystifying the discussion of sequencing panel size in oncology genetic testing. *Eur Med J.* 2022;7(2):68-77

doi: 10.33590/emj/22C9259

51. Gonzalez A, Leon DA, Perera Y, Perez R. On the gene expression landscape of cancer. *PLoS One*. 2023;18(2):e0277786.

doi: 10.1371/journal.pone.0277786

52. Mesa-Rodríguez A, Gonzalez A, Estevez-Rams E, Valdes-Sosa PA. Cancer segmentation by entropic analysis of ordered gene expression profiles. *Entropy (Basel).* 2022;24(12):1744.

doi: 10.3390/e24121744

53. Gonzalez A, Quintela F, Leon DA, Bringas Vega ML, Valdes-Sosa P. Estimating the number of available states for normal and tumor tissues in gene expression space. *Biophys Rep (NY).* 2022;2(2):100053.

doi: 10.1016/j.bpr.2022.100053

54. Bradner JE, Hnisz D, Young RA. Transcriptional addiction in cancer. *Cell.* 2017;168(4):629-643.

doi: 10.1016/j.cell.2016.12.013

55. Li Q, Dai W, Liu J, Sang Q, Li YX, Li YY. DysRegSig: An R package for identifying gene dysregulations and building mechanistic signatures in cancer. *Bioinformatics.* 2021;37(3):429-430.

doi: 10.1093/bioinformatics/btaa688

56. Dalman MR, Deeter A, Nimishakavi G, Duan ZH. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinform.* 2012;13(Suppl 2):S11.

doi: 10.1186/1471-2105-13-S2-S11

57. Khamas A, Ishikawa T, Shimokawa K, *et al.* Screening for epigenetically masked genes in colorectal cancer using 5-Aza-2'-deoxycytidine, microarray and gene expression profile. *Cancer Genomics Proteomics*. 2012;9(2):67-75.

58. Barrett T, Wilhite SE, Ledoux P, *et al.* NCBI GEO: Archive for functional genomics data sets--update. *Nucleic Acids Res.* 2013;41(D1):D991-D995.

doi: 10.1093/nar/gks1193

59. Liu J, Zheng ML, Shi PC, Cao YP, Zhang JL, Xie YP. SCARA5 is a novel biomarker in colorectal cancer by comprehensive analysis. *Clin Lab.* 2020;66(7).

doi: 10.7754/Clin.Lab.2019.191015

60. Xu JY, Zhang C, Wang X, *et al.* Integrative proteomic characterization of human lung adenocarcinoma. *Cell.* 2020;182(1):245-261.e17.

doi: 10.1016/j.cell.2020.05.043

61. Ruiz-Cordero R, Ma J, Khanna A, *et al.* Simplified molecular classification of lung adenocarcinomas based on EGFR, KRAS, and TP53 mutations. *BMC Cancer.* 2020;20(1):83.

doi: 10.1186/s12885-020-6579-z

62. Ren H, Ge DF, Yang ZC, Cheng ZT, Zhao SX, Zhang B. Integrated bioinformatics analysis identifies ALDH18A1 as a prognostic hub gene in glutamine metabolism in lung adenocarcinoma. *Discov Oncol.* 2025;16(1):1.

doi: 10.1007/s12672-024-01698-3

63. Zhang L, Zhao X, Wang E, Yang Y, Hu L, Xu H, Zhang B. PYCR1 promotes the malignant progression of lung cancer through the JAK-$STAT_3$ signaling pathway via PRODH-dependent glutamine synthesize. *Transl Oncol.*

2023;32:101667.

doi: 10.1016/j.tranon.2023.101667

64. Liu S, Tian Y, Zheng Y, Cheng Y, Zhang D, Jiang J, Li S. TRIM27 acts as an oncogene and regulates cell proliferation and metastasis in non-small cell lung cancer through SIX3-β-catenin signaling. *Aging (Albany NY).* 2020;12(24):25564-25580.

doi: 10.18632/aging.104163

65. Wang Q, Zhu W, Xiao G, Ding M, Chang J, Liao H. Effect of AGER on the biological behavior of non-small cell lung cancer H1299 cells. *Mol Med Rep.* 2020;22(2):810-818.

doi: 10.3892/mmr.2020.11176

66. Gonzalez A, Nieves J, Leon DA, Bringas Vega ML, Valdes Sosa P. Gene expression rearrangements denoting changes in the biological state. *Sci Rep.* 2021;11(1):8470.

doi: 10.1038/s41598-021-87764-0

67. Nieves J, Gonzalez A. The geometry of normal tissue and cancer gene expression manifolds. *Acta Biotheor.* 2024;72(3):9.

doi: 10.1007/s10441-024-09483-z