The tissue differentiation and cancer manifolds in gene and protein expression spaces

J Nieves[1] **and A Gonzalez**[2]

[1] Faculty of Physics, University of Havana, Cuba

[2] Institute of Cybernetics, Mathematics and Physics, Havana, Cuba

E-mail: joan.nieves@estudiantes.fisica.uh.cu

September 2021

Abstract. It is well known that, for a particular tissue, the homeostatic and cancer attractors are well apart both in gene expression and in protein expression spaces. By using data for 15 tissues and the corresponding tumors from The Cancer Genome Atlas, and for 49 normal tissues and 20 tumors from The Human Protein Atlas, we show that the set of normal attractors are also well separated from the set of tumors. Roughly speaking, one may say that there is a cancer progression axis orthogonal to the normal tissue differentiation and cancer manifolds. This separation suggests that therapies targeting common genes, which define the cancer axis, may be effective, irrespective of the tissue of origin.

1. Introduction

In gene expression (GE) space, the homeostatic and cancer phenotypes for a given tissue span distant regions allowing an easy classification of samples [1, 2]. The transition between these two radically different states of the tissue is understood to proceed discontinuously, through the rearrangement of the expressions of thousands of genes [3]. A similar situation takes place in protein expression (PE) space [4].

On the other hand, different tissues in the human body are also well separated in GE or PE spaces. We understand that this separation has a epigenetic origin [5]. The set of all human tissues span a tissue differentiation manifold, which experiences only minor changes during the individual lifespan because of its homeostatic nature.

In the present paper, we show that the set of cancer attractors also span a manifold in GE or PE spaces well apart from the tissue differentiation manifold. As a consequence, we may also define a main direction indicating cancer progression. This direction may slightly differ from the vector connecting the normal and cancer attractors for any specific tissue, and should be more likely related to genes playing an important role in any cancer.

We use RNA-seq GE data for 15 tissues and their corresponding tumors, obtained from The Cancer Genome Atlas portal (TCGA, https://www.cancer.gov/tcga) [6], and

immuno-histo-chemistry PE data for 49 normal tissues and 20 tumors coming from the Human Protein Atlas (HPA, https://www.proteinatlas.org) [7]. The data is processed by means of a Principal Component Analysis (PCA) technique [8]. 3D plots arising from the first 3 principal components allow the identification of the tissue differentiation and cancer manifolds and an approximate description of the cancer direction.

2. The tissue differentiation and cancer manifolds in GE space.

The PCA technique allows a distinct separation between normal samples and tumors, as it can be seen, for example, in **Supplementary Figure I** for Liver Hepatocellular Carcinoma [2]. Normal and tumor samples are grouped in clouds, the centers of which define the homeostatic and cancer attractors, respectively.

We compute center positions (mean geometric averages of sample clouds) for the 15 tissues and their corresponding tumors, listed in **Supplementary Table I**, and apply PCA to obtain the three main axes of maximal variance (see **Supplementary Table II**). Data, Python routines and results may be found in the GitHub repository described below.

The results are presented in figure 1, top panel, where a clear separation between normal and tumor centers is apparent. There is a region or manifold for normal tissues, which may be understood in terms of tissue differentiation, but there is also a region or manifold for tumors. Figure 1 bottom panel draws a schematics for this separation, where manifolds are represented by surfaces and a cancer axis is defined by the orthogonal vector.

It is quite interesting that genes consistently over- or under-expressed in all of these 15 tumors play also an important role in the unitary vectors along the principal component axes [2]. In other words, they play a role both in cancer development and tissue differentiation. This is the case, for example, of the UBE2C gene, which high expression values are related to low cell differentiation and tumors [9, 10], or of ADH1B, a gene participating in the oxidative metabolism of ethanol, which role in cancer is recognized [11], but its participation in basic tissue differentiation is less understood [12].

In order to get at least a rough description of the cancer axis we compute the vector connecting the center of the normal region with the center of the tumor region. The five over-expressed genes with the highest contribution to this vector are the following: MMP11, CST1, UBE2C, MYBL2 and COL10A1. MMP11 and UBE2C are consistently over-expressed in the 15 tissues under study [2]. On the other hand, the five under-expressed (silenced) genes with the highest contribution to the vector are the following: ADH1B, C7, DES, PI16 and DPT. ADH1B and C7 are consistently under-expressed in the set of 15 tissues [2]. In other words, pancancer genes play a crucial role in the main cancer direction. This analysis is limited, involving only a few genes, but illustrative.

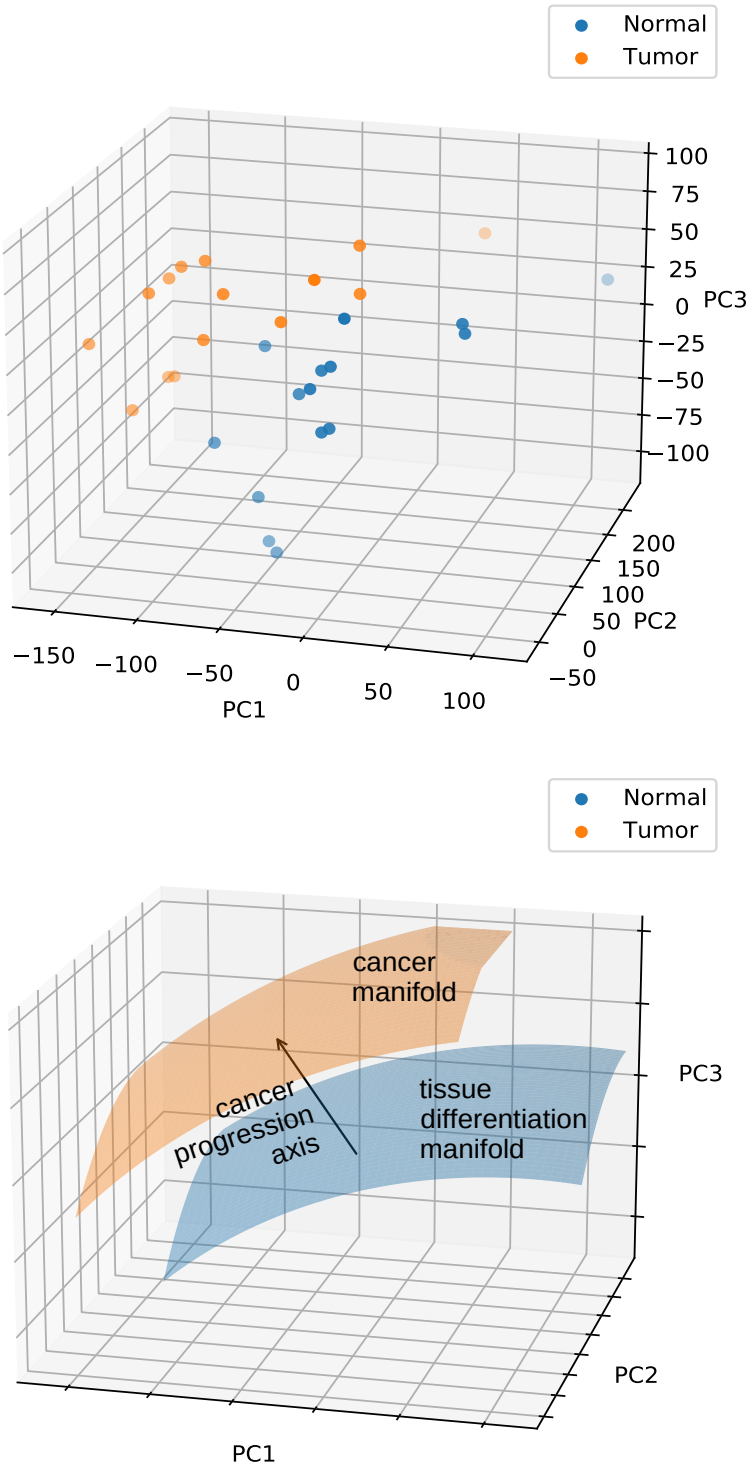


Figure 1: Top: Principal component analysis of the TCGA gene expression data for 15 tissues. The centers of the clouds of samples for each tissue are depicted. Bottom: Schematics of the topology of gene expression space. Normal tissues are distributed along the tissue differentiation manifold whereas cancer attractors are also grouped. There is a cancer progression axis connecting both regions.

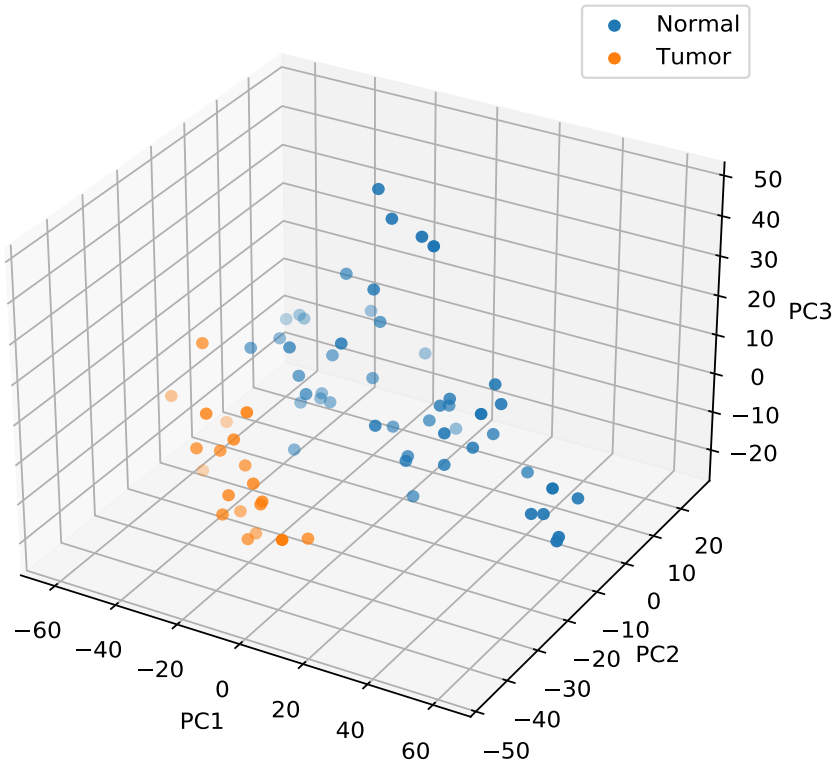


Figure 2: PCA results for the protein expression data of the HPA (49 normal tissues, 20 tumors). As in the previous figure, tissue differentiation and cancer regions are apparent.

3. The tissue differentiation and cancer manifolds in PE space.

Protein expression data coming from immuno-histo-chemistry measurements are available in the HPA database. The expression is discretized in four categories: *High*, *Medium*, *Low* and *Not Detected*. We translate these categories into numerical values in order to apply PCA (see Methods section below). 10166 proteins are consistently expressed in 49 normal tissues and 20 tumors, listed in **Supplementary Tables III and IV**. The results for the first three principal components are listed in the tables and represented in figure 2.

The tissue differentiation and cancer manifolds are also apparent in PE space. They are mainly separated along the PC2 direction. The five genes with the most significant amplitudes in the unitary vector along PC2 are: KLC1 (+), ALPK3 (+), DIDO1 (-), CCR9 (+) and GATA4 (+). The notation (+) and () refers to over- and under-expressed genes. The over-expressed genes play known roles in cancer, and the silenced DIDO1 in apoptosis. Other genes, like UBE2C, are well below in the ranking (in the 785th place). We assume that this difference is related to the discrete nature of the data, not accounting properly for the magnitude of the differential expression, although it could also be related to regulation of translation.

4. Discussion.

By applying a PCA technique to GE data for 15 normal tissue tumor pairs from the TCGA portal, and to PE data for 49 normal tissues and 20 tumors from the HPA website, we checked that normal tissues span a tissue differentiation manifold, and tumors a cancer manifold, which are clearly separated one from the other.

Conceptually, one can understand why normal tissues are grouped in GE or PE space. They have a common genetic origin, and differentiation comes from epigenetic events. Pluripotent stem cells undergo a sequence of transformations, leading to tissues, which may be seen as the result of reaching local attractor states of the global regulatory network governing the genoma [13].

With regard to tumors, one expect a cancer attractor related to each one of the homeostatic attractors of the normal tissues of origin [14]. However, there is not an obvious reason for these cancer attractors to be grouped. In our understanding, the best hypothesis to explain this fact is their common evolutionary origin, that is the idea that cancer is an atavistic state of metazoans, leading to modern multicellular organisms [15].

There are other manifestations of cancer universality in spite of its diversity. The known hallmarks, for example [16, 17]. The Pareto character of the tails in the gene expression distribution functions is also a general characteristic [3], but it is mainly related to complex behaviour [18]. In addition, we checked that the volumes of the basins of attraction for all tumors in GE space are very similar, differently from normal tissues, which exhibit a wider variability [19].

Finally, there is the issue related to the global cancer progression axis and the relevant genes contributing to it. It is hard to imagine a therapy targeting a gene such as UBE2C, which participates in so many important pathways. However, the fact that precisely this kind of genes delineates the border between the normal and cancer manifolds undoubtedly constitute a valuable information, and should have practical consequences in the near future.

5. Methods

The method we use to process the TCGA data is described elsewhere [2, 3]. First, we shift the FPKM data by adding a constant 0.1 in such a way that a expression value below 0.1 will never lead to a significant differential expression. Then, cloud centers are computed by geometric averaging over normal and tumor samples, respectively. Each of them represents a point in the 60,483 dimensional abstract gene expression space. As a result, we get 15 points for normal tissues and 15 for their corresponding tumors. Next, we compute a common reference value for each gene, e_{ref}, by geometrically averaging over the normal tissues. Fold variations are defined in logarithmic scale: $e_{fold} = Log[2, e/e_{ref}]$, such that under- and over-expression values enter the covariance matrix in a symmetrical way. The matrix generated from these 30 points is diagonalized.

The first three principal components account for 22, 16 and 12 % of the total variance, respectively. These 3 components are used to draw figure 1.

With regard to the approximate vector defining the cancer axis, we use the expression $Log[2, e_{tumor}/e_{ref}]$, where e_{tumor} is the geometric average of the tumor centers.

On the other hand, for the PE analysis we use the normal tissue and pathology data from the HPA. This data is discretized appearing only in four expression categories: *High*, *Medium*, *Low* and *Not Detected*. We assume that these categories may be translated into three values of the efold variable: +1, 0, -1, and -1, respectively. When various lectures are provided for a single tissue we take the average. A total of 10166 proteins are consistently expressed in 49 normal tissues and 20 tumors. This is the dimension of the covariance matrix. The first three principal components emerging after diagonalization account for 41 % of the total variance. They are used to draw figure 2.

6. References

[1] Alon U, Barkai N, Notterman D A, Gish K, Ybarra S, Mack D and Levine A J 1999 *Proceedings of the National Academy of Science* **96** 6745–6750 ISSN 0027-8424

[2] Gonzalez A, Perera Y and Perez R 2020 On the gene expression landscape of cancer (*Preprint* `https://arxiv.org/abs/2003.07828`)

[3] Gonzalez A, Nieves J, Leon D A, Vega M L B and Sosa P V 2021 *Scientific Reports* **11** ISSN 2045-2322

[4] Oh J M C, Brichory F, Puravs E, Kuick R, Wood C, Rouillard J M, Tra J, Kardia S, Beer D and Hanash S 2001 *PROTEOMICS* **1** 1303–1319 ISSN 1615-9853

[5] Gersbach C A 2013 *Engineered Proteins for Controlling Gene Expression* (Elsevier) chap 13, pp 125–138 2nd ed

[6] Tomczak K, Czerwiska P and Wiznerowicz M 2015 *Współczesna Onkologia* **1A** 68–77 ISSN 1428-2526

[7] Uhlen M, Zhang C, Lee S, Sjstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, Edfors F, Sanli K, von Feilitzen K, Oksvold P, Lundberg E, Hober S, Nilsson P, Mattsson J, Schwenk J M, Brunnstrm H, Glimelius B, Sjblom T, Edqvist P H, Djureinovic D, Micke P, Lindskog C, Mardinoglu A and Ponten F 2017 *Science* **357** ISSN 0036-8075

[8] Lever J, Krzywinski M and Altman N 2017 *Nature Methods* **14** 641–642

[9] Ben-Eliezer I, Pomerantz Y, Galiani D, Nevo N and Dekel N 2015 *The FASEB Journal* **29** 4670–4681

[10] Dastsooz H, Cereda M, Donna D and Oliviero S 2019 *International Journal of Molecular Sciences* **20** 2228 ISSN 1422-0067

[11] Jelski W and Szmitkowski M 2008 *Clinica Chimica Acta* **395** 1–5 ISSN 0009-8981

[12] Netanely D 2006 *Gene expression analysis of mesenchymal stem cell differentiation and leukemic over expression of tissue specific genes* Master's thesis Weissmann Institute of Science Rehovot, Israel URL `https://www.weizmann.ac.il/complex/compphys/sites/complex.compphys/files/uploads/group/group_papers1/dvir_n_msc_theis.pdf`

[13] Huang S, Eichler G, Bar-Yam Y and Ingber D E 2005 *Physical Review Letters* **94** 128701 ISSN 0031-9007

[14] Huang S, Ernberg I and Kauffman S 2009 *Seminars in Cell & Developmental Biology* **20** 869–876 ISSN 1084-9521

[15] Davies P C W and Lineweaver C H 2011 *Physical Biology* **8** 015001 ISSN 1478-3975

[16] Hanahan D and Weinberg R A 2000 *Cell* **100** 57–70

[17] Hanahan D and Weinberg R A 2011 *Cell* **144** 646–674 ISSN 0092-8674
[18] Sornette D 2009 Probability Distributions in Complex Systems *Encyclopedia of Complexity and Systems Science* (Springer New York) pp 7009–7024
[19] Quintela F and Gonzalez A 2020 Estimating the number of available states for normal and tumor tissues in gene expression space (*Preprint* `https://arxiv.org/abs/2005.02271v1`)

7. Acknowledgments

A.G. acknowledges 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 Cuban Nuclear Energy and Advanced Technologies Agency (AENTA).

8. Authors contributions

A.G. conceived and coordinated the work. J.N. processed GE and PE data and actualized the GitHub repository. Both authors analyzed and interpreted the results, contributed to the manuscript and approved the final version.

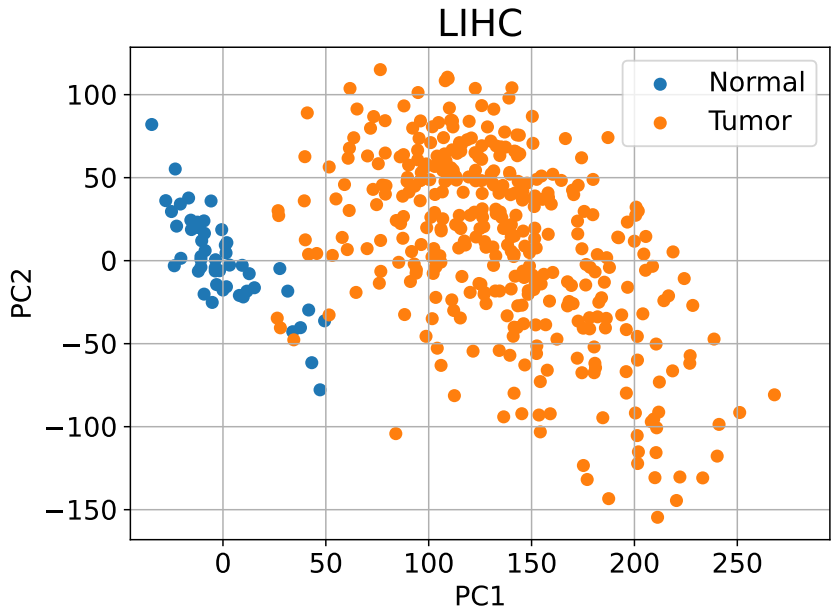
9. Competing interest

The authors declare that they have no competing interests.

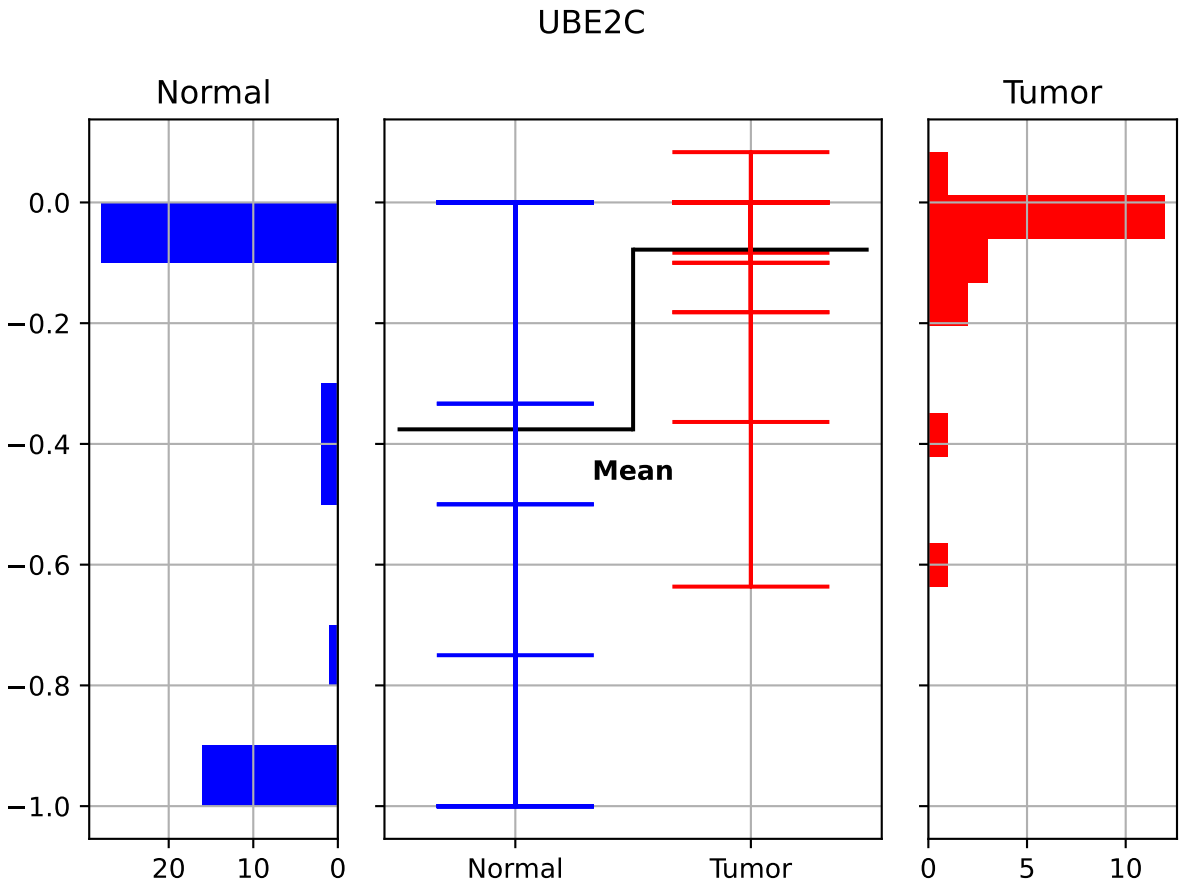
10. Availability of data and materials

The information about the data we used, the procedures and results are integrated in a public repository that is part of the project "Processing and Analyzing Mutations and Gene Expression Data in Different Systems": https://github.com/DarioALeonValido/evolp. The TCGA data is replicated in path ../evolp/databases_generated/tcga-hpa/ and the HPA data in ../evolp/databases_external/tcga-hpa/. To process each data set we include specific scripts in ../evolp/tcga-hpa/. The final result of the script for TCGA data is the top panel of figure 1, and the result of script for HPA data is the figure 2.

11. Supplementary Information



Suppl Figure I: An example taken from Ref. [2]: The (PC1,PC2) plane for the GE data in LIHC (liver hepato-cellular carcinoma). Normal and tumor samples are grouped. Cloud centers define the homeostatic and cancer attractors. The first is at the origin of coordinates by construction. The second is located along the PC1 axis at a distance of around 135 units from the origin [3].



Suppl Figure II: Center: The PE values for the UBE2C protein in different tissues. Mean values, drawn in black, are guides to the eye. The y axis is the fold variation in logarithmic scale. Histograms are drawn in the side panels. The x axes in the histograms correspond to the number of tissues. [3].

Suppl Table I: The TCGA abbreviations for the 15 tissues under study.

Abbreviation	Cancer type
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
COAD	Colon adenocarcinoma
ESCA	Esophageal carcinoma
HNSC	Head and and neck squamous cell carcinoma
KIRC	Kidney clear cell carcinoma
KIRP	Kidney papillary cell carcinoma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
STAD	Stomach adenocarcinoma
THCA	Thyroid carcinoma
UCEC	Uterine corpus endometrial carcinoma

Suppl Table II: Coordinates along the first three PC axes of the centers of normal and tumor sets of samples in GE space for the 15 tissues taken form the TCGA portal. In general, the transition from a normal tissue to its tumor involves displacement along the three coordinates.

Tissue	Normal			Tumor		
	PC1	PC2	PC3	PC1	PC2	PC3
BLCA	-20.51	-12.76	6.70	-102.25	26.85	68.06
BRCA	-2.17	-57.66	25.27	-65.80	-12.38	65.00
COAD	-51.98	34.61	-106.41	-121.17	60.57	-14.05
ESCA	-77.12	5.88	-35.51	-156.08	13.67	17.90
HNSC	-54.21	36.47	19.26	-114.12	41.14	55.71
KIRC	76.81	-9.42	52.41	11.44	4.92	67.09
KIRP	74.24	-4.81	56.83	7.01	21.54	92.03
LIHC	111.81	234.51	20.71	43.73	190.23	57.54
LUAD	7.88	-52.87	0.19	-77.89	-12.59	35.30
LUSC	3.31	-52.82	-2.61	-115.45	3.28	56.74
PRAD	0.04	-39.94	31.52	-29.06	-20.65	53.23
READ	-42.80	18.33	-107.33	-118.46	64.27	-14.47
STAD	-52.80	14.37	-70.89	-131.54	19.36	-24.06
THCA	19.04	-61.85	71.54	-5.02	-36.39	85.82
UCEC	8.45	-52.04	38.35	-84.06	14.20	76.80

Suppl Table III: Coordinates along the first three PC axes of the PE values for the 49 normal tissues from the HPA portal.

Tissue	PC1	PC2	PC3	Tissue	PC1	PC2	PC3
adipose tissue	-39.06	11.70	-6.88	oral mucosa	-3.98	2.38	28.71
adrenal gland	26.30	2.74	-9.15	ovary	-31.56	3.01	-3.26
appendix	3.66	-9.35	-0.79	pancreas	3.03	-4.05	-6.82
bone marrow	-11.64	4.59	8.04	parathyroid gland	6.52	-3.74	-16.06
breast	-17.74	1.67	-0.33	placenta	13.07	-0.91	4.55
bronchus	25.54	-0.37	8.23	prostate	5.38	-8.34	-5.60
caudate	-25.93	4.11	-8.76	rectum	44.18	-8.24	-0.68
cerebellum	-18.23	2.80	-7.04	salivary gland	7.07	-0.86	2.80
cerebral cortex	-26.04	3.50	-8.07	seminal vesicle	12.69	-6.58	-5.35
cervix, uterine	-5.66	1.05	17.42	skeletal muscle	-17.01	16.78	-8.68
colon	5.92	1.88	-2.34	skin 1	-17.85	5.41	11.50
duodenum	46.06	-9.02	-11.13	skin 2	2.52	1.13	30.98
endometrium 1	-15.61	-5.47	-2.09	small intestine	45.14	-8.14	-11.36
endometrium 2	-18.97	-3.34	-1.29	smooth muscle	-26.97	13.78	-3.71
epididymis	17.71	-4.04	0.13	soft tissue 1	-33.03	9.53	-4.04
esophagus	9.24	-1.28	30.81	soft tissue 2	-34.74	9.87	-4.15
fallopian tube	17.88	-6.94	4.60	spleen	-29.78	-2.72	-1.35
gallbladder	44.89	-2.42	-6.33	stomach 1	38.90	-6.09	-9.37
heart muscle	-5.24	10.98	-17.98	stomach 2	37.32	-8.03	-8.16
hippocampus	-27.72	2.91	-9.28	testis	-2.01	0.61	-6.83
kidney	3.37	0.40	-6.97	thyroid gland	17.91	2.35	-1.42
liver	-25.80	-1.76	-15.74	tonsil	-6.95	-7.24	10.08
lung	-13.82	4.32	-3.96	urinary bladder	24.77	-7.25	10.34
lymph node	-18.88	-6.54	4.46	vagina	-8.60	4.66	31.52
nasopharynx	23.74	0.54	10.79				

Suppl Table IV: Coordinates along the first three PC axes of the PE values for the 20 tumors from the HPA portal. By comparing with the corresponding normal tissues one may notice that, in this case, the transition from normal to tumor involves mainly a displacement along the PC2 axis.

Cancer type	PC1	PC2	PC3
breast cancer	-4.53	-24.75	-0.16
carcinoid	-12.87	-19.66	-7.08
cervical cancer	-16.05	-22.45	8.47
colorectal cancer	7.57	-31.04	-0.22
endometrial cancer	-9.84	-26.58	-3.70
glioma	-23.86	-11.22	-1.82
head and neck cancer	-8.83	-22.27	10.95
liver cancer	-9.46	-20.67	-10.79
lung cancer	-18.29	-22.12	2.25
lymphoma	-28.17	-15.76	1.09
melanoma	-11.09	-20.83	0.05
ovarian cancer	-6.26	-23.82	1.03
pancreatic cancer	-4.42	-24.35	-1.46
prostate cancer	-7.77	-24.14	-8.31
renal cancer	-27.32	-13.29	-9.09
skin cancer	-22.78	-15.93	13.61
stomach cancer	-9.80	-24.99	-0.96
testis cancer	-13.20	-22.50	3.99
thyroid cancer	3.31	-23.28	-7.05
urothelial cancer	-10.58	-22.48	6.51