The geometry of normal tissue and cancer manifolds

Joan Nieves and Augusto Gonzalez
Institute of Cybernetics, Mathematics and Physics, Havana, Cuba

May 17, 2023

Abstract

Recently, it was shown that both the normal states and their respective tumors span disjoint manifolds in gene expression space. By processing gene expression data we show that, in general, for the vectors defining the centers of cloud samples: 1. The closest tumor to a given normal state is its respective tumor, 2. Two normal states define quasi-orthogonal directions, 3. A tumor may have a projection onto its corresponding normal state, but it is quasi-orthogonal to all others and 4. The cancer manifold is roughly obtained by translating the normal tissue manifold along an orthogonal direction defined by the cancer projection axis.

1 Introduction

In gene expression (GE) space, the normal homeostatic and cancer phenotypes for a given tissue span distant regions allowing an easy classification of samples [1, 2]. A similar situation takes place in protein expression (PE) space [3]. 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 [4].

On the other hand, different normal 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 normal tissues span a normal

tissue manifold, which experiences only minor changes during the individual lifespan because of its homeostatic nature.

In Refs. [2, 6], it was shown that the set of cancer attractors also span a manifold in GE space, well apart from the normal tissue manifold. The atavistic origin hypothesis for tumors [7] was used to argue that they should be grouped. A main direction indicating cancer progression was defined. This direction may be more likely related to genes playing an important role in any cancer.

In the present paper, we elaborate on the geometry of normal tissue and cancer manifolds in GE space. Previously, we have shown that geometric magnitudes are biologically meaningful. For example, the distance between two tumor centers shows an inverse dependence on the number of common differentially expressed genes [2].

We use RNA-seq GE data for 15 tissues and their corresponding tumors, obtained from The Cancer Genome Atlas portal (TCGA) [8]. For the vectors defining the centers of the clouds of normal and tumor samples in each tissue, the following statements are proven: 1. The closest tumor to a given normal state is its respective tumor, 2. Two normal states define quasi-orthogonal directions, 3. A tumor may have a projection onto its corresponding normal state, but it is quasi-orthogonal to all others, and 4. The cancer manifold is roughly obtained by translating the normal tissue manifold along an orthogonal direction defined by the cancer projection axis.

2 The normal tissue and cancer manifolds

First, let us define the variables. In Table 1, we list the studied tissues and the number of normal and tumor samples in each one. The data includes the expression of 60483 genes, thus each sample is defined by an expression vector of dimension 60483. The i-component is given by:

$$e_{fold}(i) = \log_2(e_i/e_i^{ref}). \tag{1}$$

$e_{fold}(i)$ gives the logarithmic fold variation of gene i with respect to the reference level e_i^{ref}. We shall use the same reference for all tissues, constructed by geometric averaging over all normal samples. $e_{fold}(i) > 0$ means that the gene is over-expressed with regard to the reference, whereas $e_{fold}(i) < 0$ means that the gene is under-expressed or silenced.

From sample vectors, we define tissue vectors by simply averaging over tissue normal or tumor samples, respectively: $\vec{e}_{\text{tissue}}(\text{normal})$, and $\vec{e}_{\text{tissue}}(\text{tumor})$. They may be understood as the centers of sample clouds. Indeed, a Principal Component Analysis (PCA) [9] of the data for liver hepatocellular carcinoma (LIHC), for example, shows a 2D cartoon of samples in GE space [2], as it is apparent in Fig. 1. Cloud centers are well defined. We shall work with these centers as tissue representatives.

In Ref. [2], we show that tumor centers define a tumor manifold, well separated from the normal tissue manifold. A 3D representation, obtained from the PCA of all studied tissues, is given in Fig. 2. A schematics of results is drawn in the right panel. Manifold are well separated and a cancer progression axis, connecting the centers of both manifolds, may be defined. Actual distances or angles, however, are not correctly represented in this reduced dimensionality diagram.

We first show that all vectors connecting a normal tissue to its respective tumor has a positive projection onto the cancer projection axis. In other words, the schematics shown in Fig. 2 is preserved in higher dimensions. The exact formulation of this statement is the following:

$$[\vec{e}_{\text{tissue}}(\text{tumor}) - \vec{e}_{\text{tissue}}(\text{normal})] \cdot [\langle \vec{e}_{\text{tissue}}(\text{tumor}) \rangle - \langle \vec{e}_{\text{tissue}}(\text{normal}) \rangle] > 0. \quad (2)$$

Where $< ... >$ means average over tissues. The results for the normalized projections, i.e. dividing by the norms of both vectors, are listed in the second column of Table 2.

2.1 The closest tumor to a given normal state is its proper tumor

As mentioned above, actual distances are not properly represented in 2D or 3D PCA projections. The distance from a normal state N to a tumor state T' is defined as:

$$\text{dist}(N, T') = \sqrt{\sum_i |\log_2(e_i^{t'}/e_i^n)|^2}, \quad (3)$$

where the sum runs over genes, and $e_i^{t'}$ and e_i^n are, respectively, tissue expressions of gene i in T' and N. Thus, the distance is an L2-mean of the log-fold differential expressions when N is taken as reference.

Distances computed from Eq. (3) are given in the fourth and fifth columns of Table 2. It shows that, in general, the closest tumor to a given normal tissue is its proper tumor. Exceptions to this rule are tumors in the same organ (KIRC - KIRP, LUSC - LUAD, READ - COAD), and the ESCA - STAD pair, which are not in the same organ, but physiologically close.

This statement may be rephrased as follows: Differential expressions for a tumor are the smallest when the proper normal state is taken as reference.

Notice that distances are sharply distributed around the mean, 142.01, with the exceptions of PRAD and THCA, the closest tumors.

2.2 A pair of normal states define quasi-orthogonal directions in GE space

We shall compute the angles $\theta_{NN'}$ defined from the scalar product between two normal tissue vectors. Figure 3a) shows an histogram of $\theta_{NN'}$ values. Angles are distributed around 0.50 π with the exception of the three pairs of tissues in the same organ, and the pair esophagus - stomach, mentioned above.

Thus, two normal tissue vectors define quasi-orthogonal directions. The meaning of such a statement is as follows: gene expressions in two normal tissues are not correlated. That is, genes expressed in the first tissue are not related to those expressed in the second one.

It is interesting to compute also the norm of normal state vectors, that is the distance from the global reference to the center of the normal sample cloud in a given tissue. All normal states radii are within one standard deviation from the mean: 154.63 ± 36.06. The exceptions are BLCA and LIHC, for which Norm(BLCA) = 109.29, and Norm(LIHC) = 270.80. These are the closest and the farthest from the reference tumors, respectively.

According to these results we may say that the normal tissue vectors span a sector of a nearly spherical hyper-surface, where pairs of vectors are quasi-orthogonal.

2.3 A tumor tissue vector may have projection onto its normal state, but it is quasi-orthogonal to the rest.

The mean value of angles $\theta_{TN'}$ between tissue vectors corresponding to T and N' is again 0.50 π, but the distribution is much more broad than for $\theta_{NN'}$. The results are drawn in Fig. 3b).

In general, loosely speaking we may say that there are non-zero projections of T onto its corresponding N and other tissues in the same organ, but T is quasi-orthogonal to the rest of normal tissues.

One may check whether the angle $\theta_{TT'}$ between a pair of tumor tissue vectors, measured in the cancer manifold, that is taken as reference the mean of all tumors, is similar to the angle $\theta_{NN'}$ between their corresponding normal tissues. In Fig. 3c) we draw the histogram for the difference. Indeed, angles are roughly preserved. For the mean difference, we obtain: $(0.001 \pm 0.056)\pi$. The outlayers are related to tumors in the same organ.

A final question concerns the orthogonality of the cancer projection axis, $\langle \vec{e}_{\text{tissue}}(\text{tumor})\rangle - \langle \vec{e}_{\text{tissue}}(\text{normal})\rangle$ to the normal tissue manifold. In the third column of Table 2, we show the projections onto each of the normal tissue vector. It is apparent that the cancer progression axis roughly defines an orthogonal direction. For the mean angle, we obtain 0.50 π.

In general, we may say that tumor state vectors are also distributed on a sector of a hyper-surface, which is roughly obtained by translating the normal tissue manifold along the orthogonal direction defined by the cancer progression axis.

3 Conclusions

We use the vectors defining cloud centers of GE data in order to compute geometrical properties of normal tissue and tumor manifolds. The 2D cartoon of Fig. 2, right panel, is shown to hold in higher dimensions. Normal states are evenly distributed on a sector of a hyper-sphere. Any pair of vectors is quasi-orthogonal. Tumors also lie on a manifold, which is roughly obtained by translating the normal tissue manifold along the cancer progression axis.

We notice that quasi-orthogonality of normal tissues could be related to preserving the identity or avoiding errors in embryonary development. Other geometric properties described in the paper may also have biological

significance.

It is interesting to look at these geometrical results from the point of view of the atavistic theory of tumors. Cooperation between unicellular organisms leading to primitive metazoans could probably be related to a process of mixing and orthogonalization of their genome. This hypothesis may be perhaps checked in the simplest multicellular organisms. The evolution towards advanced pluricellularity could then proceed through an orthogonal direction in GE space, a kind of inverse of the cancer progression axis.

GE data may also be used in order to characterize the groups of normal tissues and tumors by means of other measurements. For example, the degree of disorder may be described by the entropy of the expression profile [10]. Tumors are shown to be more entropic than normal tissues. A related measurement comes from the estimation of the volume of the basins of attraction for both normal and tumor attractors in GE space [11]. The number of tumor genotypes is shown to be exponentially greater than that of normal state genotypes. Work along these directions is in progress.

Acknowledgments The authors acknowledge the Cuban Agency for Nuclear Energy and Advanced Technologies (AENTA) and the Office of External Activities of the Abdus Salam Centre for Theoretical Physics (ICTP) for support.

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.

Competing interests The authors declare that they have no competing interests.

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 for the cloud centers (tissue mean geometric averages of genes) is replicated in the path ../evolp/databases generated/tcga-hpa/ and the HPA data in ../evolp/databases external/tcga-hpa/. To perform PC analysis of each data set we include specific scripts in ../evolp/tcga-hpa/. The final result of the script for TCGA data is the left panel of figure 2, and

the result of script for HPA data is the figure 4.

Angles are trivially computed from the GE vectors.

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Science, 96(12):6745–6750, 1999.

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

[3] Jane M. C. Oh, Franck Brichory, Eric Puravs, Rork Kuick, Chris Wood, Jean Marie Rouillard, John Tra, Sharon Kardia, David Beer, and Samir Hanash. A database of protein expression in lung cancer. PROTEOMICS, 1(10):1303–1319, 2001.

[4] Augusto Gonzalez, Joan Nieves, Dario A. Leon, Maria Luisa Bringas Vega, and Pedro Valdes Sosa. Gene expression rearrangements denoting changes in the biological state. Scientific Reports, 11(1):8470, 2021.

[5] Bo Zhang, Yan Zhou, Nan Lin et. al. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M & M algorithm. Genome Research 23:1522–1540 (2013). http://www.genome.org/cgi/doi/10.1101/gr.156539.113

[6] Joan Nieves and Augusto Gonzalez. The tissue differentiation and cancer manifolds in gene and protein expression spaces. BiorXiv preprint 2021. https://doi.org/10.1101/2021.08.20.457160

[7] P C W Davies and C H Lineweaver. Cancer tumors as metazoa 1.0: tapping genes of ancient ancestors. Physical Biology, 8(1):015001, 2011.

[8] Katarzyna Tomczak, Patrycja Czerwiska, and Maciej Wiznerowicz. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn), 1A:68–77, 2015. https://www.cancer.gov/tcga

[9] Jake Lever, Martin Krzywinski, and Naomi Altman. Principal component analysis. Nature Methods, 14(7):641–642, 2017.

[10] 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

[11] 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.

[12] Mathias Uhlen, Cheng Zhang, Sunjae Lee, Evelina Sjstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, Muhammad Arif, Zhengtao Liu, Fredrik Edfors, Kemal Sanli, Kalle von Feilitzen, Per Oksvold, Emma Lundberg, Sophia Hober, Peter Nilsson, Johanna Mattsson, Jochen M. Schwenk, Hans Brunnstrm, Bengt Glimelius, Tobias Sjblom, Per-Henrik Edqvist, Dijana Djureinovic, Patrick Micke, Cecilia Lindskog, Adil Mardinoglu, and Fredrik Ponten. A pathology atlas of the human cancer transcriptome. Science, 357(6352):660, 2017. https://www.proteinatlas.org

A The normal tissue and cancer manifolds in PE space

For the sake of consistency, we show separation of the normal tissue and cancer manifolds in protein expression space.

Protein expression data coming from immuno-histo-chemistry measurements are available in the Human Protein Atlas (HPA) database [12]. We use the normal tissue and pathology data.

Expression values are classified in one of four categories: High, Medium, Low and Not Detected. We assume that these categories may be translated into three values of the e_{fold} 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 in the PCA. The first three principal components emerging after diagonalization account for 41 % of the total variance. They are used to draw Figure 4.

The tissue differentiation and cancer manifolds are also apparent in PE space. They are mainly separated along the PC2 direction. The difference with regard to GE data could be related to the discrete nature of the PE data and also to regulation of translation.

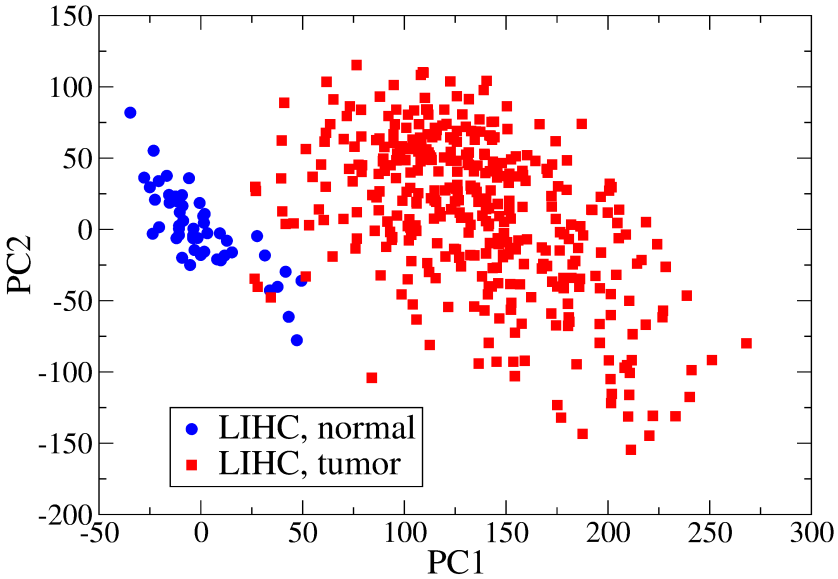

Figure 1: Reduced dimensionality representation of normal and tumor samples in LIHC.

TCGA notation	Cancer type	Normal samples	Tumor samples
BLCA	Bladder urothelial carcinoma	19	414
BRCA	Breast invasive carcinoma	112	1096
COAD	Colon adenocarcinoma	41	473
ESCA	Esophageal carcinoma	11	160
HNSC	Head and neck squamous cell carcinoma	44	502
KIRC	Kidney clear cell carcinoma	72	539
KIRP	Kidney papillary cell carcinoma	32	289
LIHC	Liver hepatocellular carcinoma	50	374
LUAD	Lung adenocarcinoma	59	535
LUSC	Lung squamous cell carcinoma	49	502
PRAD	Prostate adenocarcinoma	52	499
READ	Rectum adenocarcinoma	10	167
STAD	Stomach adenocarcinoma	32	375
THCA	Thyroid carcinoma	58	510
UCEC	Uterine corpus endometrial carcinoma	23	552

Table 1: The TCGA abbreviations and the number of samples in the 15 tissues under study.

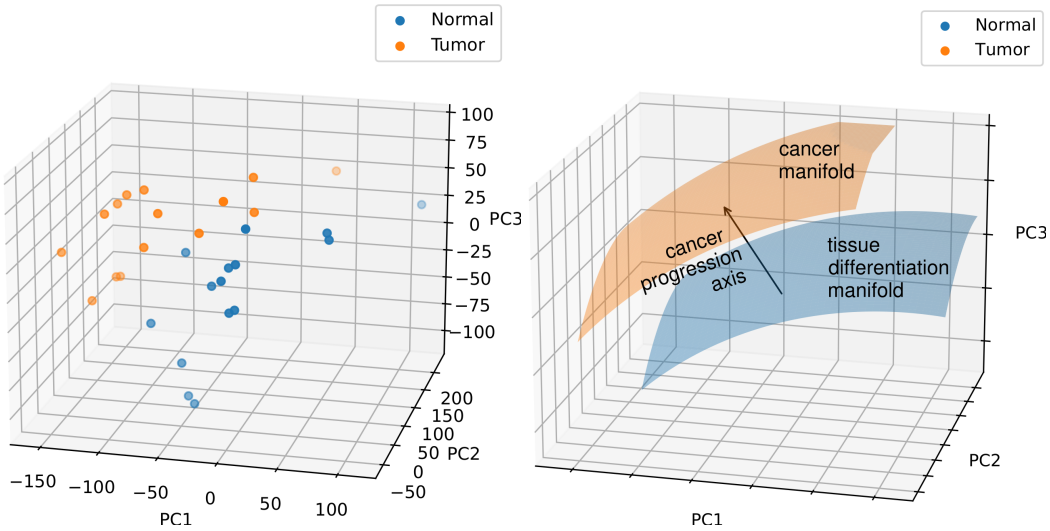

Figure 2: Left: Principal component analysis of the TCGA gene expression data for 15 tissues. The centers of the clouds of samples for each tissue are depicted. Right: Schematics of the topology of gene expression space. Normal tissues are distributed along the tissue differentiation manifold whereas cancer attractors are also grouped. There is a cancer progression axis connecting the centers of both regions.

Tissue	Normalized projection	$\theta(N)/\pi$	Min(dist(N,T'))	dist(N,T)
BLCA	0.80	0.48	140.97	140.97
BRCA	0.65	0.51	137.53	137.53
COAD	0.68	0.48	156.05	156.05
ESCA	0.67	0.43	136.77 (STAD)	139.69
HNSC	0.61	0.43	123.57	123.57
KIRC	0.42	0.56	169.91 (KIRP)	173.02
KIRP	0.52	0.54	163.58	163.58
LIHC	0.51	0.52	135.72	135.72
LUAD	0.68	0.54	147.34	147.34
LUSC	0.67	0.51	142.52 (LUAD)	194.99
PRAD	0.50	0.49	92.68	92.68
READ	0.71	0.50	166.22 (COAD)	168.13
STAD	0.71	0.48	137.46	137.46
THCA	0.32	0.49	113.00	113.00
UCEC	0.67	0.51	171.43	171.43

Table 2: 2nd column: The normalized projections over the cancer projection axis of vectors connecting a given normal tissue to its tumor. 3rd column: The angle between a normal tissue vector and the cancer progression axis. 4th column: Minimal distances between a given normal tissue and the set of tumors. 5th column: The distance between a given normal tissue and its proper tumor

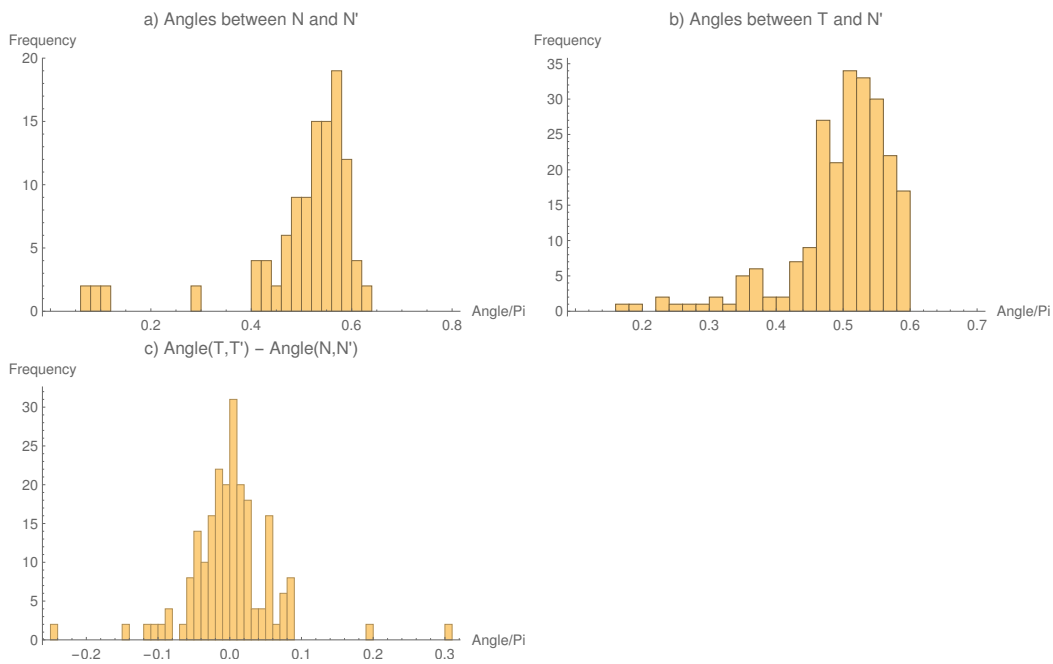

Figure 3: a) Frequency histogram for the angle between two normal tissue vectors. The average value for the angle is 0.50 π. b) The same for the angle between tissue vectors corresponding to T and N'. The average value for the angle is again 0.50 π. c) The same for the difference between the angles $\theta_{TT'}$ and $\theta_{NN'}$, measured in their respective manifolds. The average value for this difference is 0.001 π.

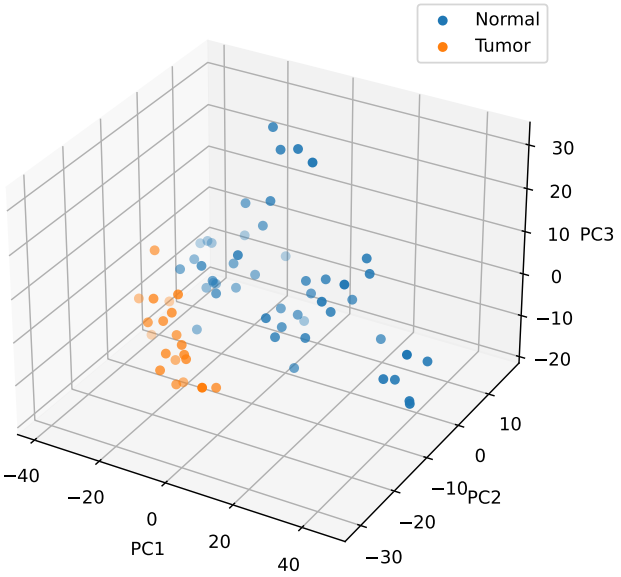

Figure 4: PCA results for the protein expression data of the HPA (49 normal tissues, 20 tumors). As in Fig. 2, normal tissue and cancer regions are apparent.