Exploring tumor-normal cross-talk with TranNet: role of the environment in tumor progression

Bayarbaatar Amgalan[1], Chi-Ping Day[2], and Teresa M. Przytycka[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, Maryland, USA
[2]Laboratory of Cancer Biology and Genetics, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, Maryland, USA

[*]correspondence to przytyck@nih.gov

Abstract

There is a growing awareness that tumor-adjacent normal tissues used as control samples in cancer studies do not represent fully healthy tissues. Instead, they are intermediates between healthy tissues and tumors. The factors that contribute to the deviation of such control samples from healthy state include exposure to the tumor-promoting factors, tumor-related immune response, and other aspects of tumor microenvironment. Characterizing the relation between gene expression of tumor-adjacent control samples and tumors is fundamental for understanding roles of microenvironment in tumor initiation and progression, as well as for identification of diagnostic and prognostic biomarkers for cancers.

To address the demand, we developed and validated TranNet, a computational approach that utilizes gene expression in matched control and tumor samples to study the relation between their gene expression profiles. TranNet infers a sparse weighted bipartite graph from gene expression profiles of matched control samples to tumors. The results allow us to identify predictors (potential regulators) of this transition. To our knowledge, TranNet is the first computational method to infer such regulation.

We applied TranNet to the data of several cancer types and their matched control samples from The Cancer Genome Atlas (TCGA). Many predictors identified by TranNet are genes associated with regulation by the tumor microenvironment as they are enriched in G-protein coupled receptor signaling, cell-to-cell communication, immune processes, and cell adhesion. Correspondingly, targets of inferred predictors are enriched in pathways related to tissue remodelling (including the epithelial-mesenchymal Transition (EMT)), immune response, and cell proliferation. This implies that the predictors are markers and potential stromal facilitators of tumor progression. Our results provide new insights for the relationships between tumor adjacent control sample, tumor and the tumor environment. Moreover, the set of predictors identified by TranNet will provide a valuable resource for future investigations.

The TranNet method was implemented in python, source codes and the data sets used for and generated during this study are available at the Github site `https://github.com/ncbi/TranNet`.

1 Introduction

In multi-stage carcinogenesis theory, mutations accumulate to perturb the cell regulatory program, eventually causing cell transformation. These perturbations interact with other factors such as micro and macro environment or preexisting health conditions. Although mutations are central for the emergence of cancer, much of the understanding of the disease comes from studies of the cancer-related changes in gene expression. Cancer is characterized by dysregulated functions of many cellular processes including proliferation, cell-cell interactions, chromatin organization, DNA repair, and others. Yet many of these alterations are not mechanically linked to specific mutations, but are driven by changes in gene expression [1]. One of the emerging concepts is that cancer progression is facilitated by increased cell plasticity, which allows cancer cells to switch dynamically between a differentiated states in response to stress [2]. Cancer cell plasticity has been linked to the epithelial-to-mesenchymal transition (EMT) which has been shown to respond to microenvironmental signals or cancer therapy [3, 4, 5, 6, 7, 8]. In addition, it has been estimated that at least 25% of cancers are associated with chronic inflammation [9, 10]. Last but not least, a recent study suggested that the tumor microenvironment and the microenvironment of control sample are strongly dependent [11]. In fact, tumor-related alteration of adjacent tissue are believed to contribute to postoperative cancer recurrences that occur in up to a third of patients [12]. This is not a surprise, as tumors and their adjacent normal tissues also share some exogenous exposures including environmental factors, such as smoking, diet, and genetic variations. However could gene expression from a control sample inform on tumor state? Which molecular pathways and functions in tumor are associated with gene expression changes in normal tissue? What can they teach us about tumor progression?

To address these questions, we developed the Transition Network model (TranNet), a computational approach to study the relation between the gene expression patterns in matched normal and tumor tissues. Focusing on tumor genes (defined as genes differently expressed between tumor and control samples), TranNet infers a transition function from gene expression in a control sample to an estimate of gene expression in the matched tumor sample. Simultaneously, the method infers a set of genes that are predictors of this transition (Figure 1). These predictor genes might but do not have to regulate this transition and might instead reflect expression changes due to factors common to control and tumor samples. Functional analysis of the predictors and their target genes helps to shed light on possible sources of this relation.

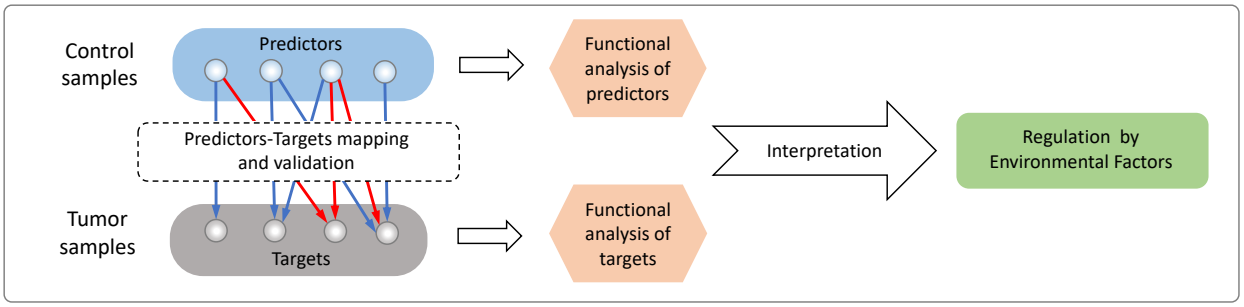


Figure 1: **Workflow of the analysis.** TranNet constructs a transition function from gene expression in control samples to gene expression in the matched tumor samples. Simultaneously, the method infers a set of predictors in the control tissue that used for computing the transition and target genes in the tumor tissue that are influenced by the predictors. After validating the method, functional analysis of the predictors and their target genes is used to shed light on possible sources of the relation between gene expression in control and tumor samples.

Computationally, TranNet utilizes a network construction based on a sparse estimation of partial correlation [13] that uses l_1 norm constraint to ensure the selection of the most informative predictors (Figure 2). Recognising that the expression of the genes which did not pass the p-value threshold to be included in the set of tumor genes might also contribute to gene expression in tumor, we extended the network by including additional nodes, principal components, representing the informative trends on the expression data of these genes. We refer to the genes and principal components whose activity in control samples influence gene expression in tumor tissue samples as *predictors*.

TranNet opens a new way to explore the relationship between gene expression in adjacent control samples and tumors. Our results indicate that the former can provide information about the latter and link, at least in part, inferred relations to tumor environment. As elaborated in the Discussion, not all inferred associations are assumed to be tumor drivers. Yet, as predictors of the expression of tumor genes, they are markers of other tumor-related processes including tumor-environment interaction and are important in this respect. Indeed, using TranNet we identified a set of genes that can serve as the predictors of tumor-environment relation as well as genes and pathways involved in this interaction. Taken together, this work offers a computational method to infer the relation between the gene expression patterns in matched normal and tumor tissues and provided a new understanding of the relation between tumor, normal, and the environment.

2 The Transition Network Model (TranNet)

An overview of the TranNet workflow is presented in Figure 2. Two matrices representing gene expressions in normal (A1) and tumor (A2) tissues of cancer patients are provided as the input. Tumor genes, defined as genes differentially expressed between control and tumor samples, are represented as network nodes while the expression of the rest of the genes is represented by meta-nodes corresponding to the principal components of the expression of these remaining genes. Each gene or principal component is represented by two nodes — one for each condition. The conceptual idea of including the principal components in the analysis is illustrated in (B). The transition matrix from normal to tumor state is obtained by solving the problem in (C) and the bipartite network formed by the transition matrix represents dependency flow from normal to tumor tissues (D). Finally, the strength of the total influence of a node in normal tissue on gene expression in tumor is quantified by regulatory potentials (RP) (E).

2.1 Inference of the transition network

Let X and Y be the matrices describing gene expression in tumor and control samples respectively for p patients. The first n columns of each matrix represent the expression of the n tumor genes followed by k meta nodes corresponding the principal components summarising gene expression trends of the remaining set of genes.

Specifically, x_{ij} and y_{ij} denotes the expression of gene g_j (or pc_{j-n} if $j > n$) in the ith patient in tumor and normal respectively where the "expression" of a principal component in patient i is explained below (see also Figure 2B right, for the description of differentially expressed (DE) principal component).

Assuming that X represents regulatory influences (or markers of such influences) and Y their targets, the weight matrix M describing transformations between them can be written as

$$X \times M \approx Y. \tag{1}$$

More precisely, for patient i, the expression value y_{if} of gene g_f or principal component pc_f in the patient's tumor tissue can be approximated by a linear combination of the expressions of the n genes and k meta nodes in the patient's normal tissue

$$\sum_{j=1}^{n+k} x_{ij} \cdot m_{jf} \approx y_{if}$$

where $m_{(\cdot)f} \in \mathbb{R}^{n+k}$ denotes the transition weights from the normal tissue expression of the n genes and k principal components to the tumor tissue expression of gene g_f or principal component pc_f. Thus, our goal is to minimize the least square error subject to a unit l_1 norm constraint on $m_{(\cdot)f}$ as follows:

$$\underset{m_{(\cdot)f} \in R^{n+k}}{\text{minimize}} \sum_{i=1}^{p} \left(\sum_{j=1}^{n+k} x_{ij} \cdot m_{jf} - y_{if} \right)^2, \quad \text{subject to} \quad \sum_{j=1}^{n+k} |m_{jf}| \leq 1. \tag{2}$$

The l_1 norm constraint on the edge weights allows for selecting the strongest transition effects on a given target node. Hence, this regularization acts to avoid over-fitting issues and only non-zero m_{jf}, selected for g_f, denotes the transition effect from g_j. In this setting, for every target node, the optimization problem

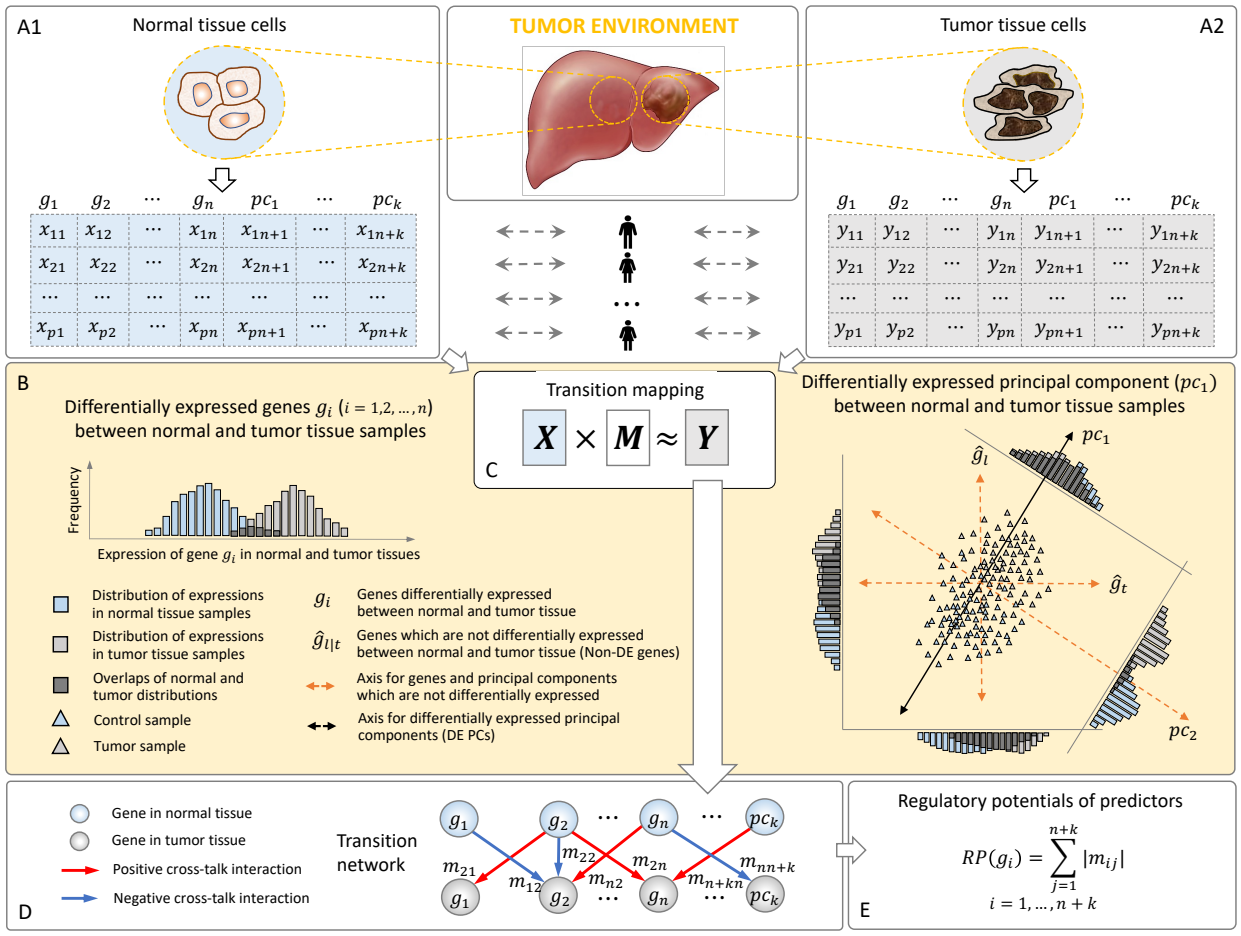


Figure 2: **Outline of the TranNet method.** Input matrices X and Y represent the expressions of genes in control (X, panel **(A1)**) and tumor (Y, panel **(A2)**) samples across p patients (rows). Genes (g_i, $i = 1, \ldots n$) which are differentially expressed between control and tumor samples (referred to as tumor genes) are represented by individual nodes (columns from 1 to n) and the rest of the genes are represented by principal components of their expression (columns from pc_1 to pc_k) preserving major trends of the expression. The principal components which are differentiated between normal and tumor tissue samples are considered as additional variables in the multi-variate analysis. Distribution trend of expression of a differentially expressed gene g_i is visualized in the left of the yellow panel. As illustrated in a 2-dimensional space in the right of the yellow panel, although neither $\hat{g}_l$ nor $\hat{g}_t$ is differentially expressed between control and tumor tissue samples, their joint distribution is differentiated along the axis of principal component pc_1 showing a similar trend to the DE genes while pc_2 is not differentiated **(B)**. The transition map is a linear operator defined by matrix M computed to minimise Y's representation error subject to a sparsity constraint explained in the main text **(C)**. The result is summarised as a bipartite network representing regulatory influences from control samples to tumors **(D)**. Regulatory potential of a gene represents its total contribution to the transition network **(E)**.

in Equation (2) (a constrained version of Equation (1)) searches over all possible combinations of transition effects from the predictornodes and selects the best combination with their optimal weights to explain the activity of the target node [13].

To explain the role of principal components we shall recall that a fundamental assumption for multivariate analysis is that there are no unobserved factors affecting both explanatory and response variables globally. The expressions of the genes that were not selected as tumor genes (DE genes) might have such influence. To adjust for such effects on the transition mapping, we include meta nodes (adjustment variables in [14]) that

represent the principal components of the expression data for these non-tumor genes (see Section 6). By the definition, principal components are vector representations of the general trends in a multi-dimensional data [15]. Here data points represent all the samples (control and tumor) and each point is a vector representing gene expression (Figure 2B right). Embedding this data using principal components as reference axis, we say that a given principal component is differentially expressed between the control and tumor samples if the coordinates of the tumor samples on the axis defined by this principal component are significantly larger or smaller than the coordinates of the control samples (Figure 2B right).

Finally, we comment on the optimization problem. Although the least square minimization in Equation (2) is convex, the l_1 norm constraint is non-smooth and derivative-based optimality conditions such as Lagrangian multipliers and Karush–Kuhn–Tucker (KKT) conditions are not, in general, directly applicable. While a coordinate descent algorithm with convex penalties [16] is commonly used to approximate this non-smooth problem, there is no optimal strategy for tuning parameters controlling the strength of the penalty term [17]. Similarly complementing the standard quadratic programming formulation, the soft shrinkage method decomposes the l_1 norm into 2^{n+k} inequality constraints [18]. However, handling 2^{n+k} constraints is not practical for the large-scale problem. To overcome this issue, we implemented the projected gradient method [19] that converges to the optimal solution for the non-smooth constrained problem.

2.2 Regulatory potentials of predictors

TranNet aims to identify predictors (genes and principal comments) whose expression in one condition (here normal) predicts gene expression in a different condition (here tumor). We model the information flow from normal to tumor tissue by the transition matrix in Equation (1), obtained by solving the non-smooth convex optimization problem in Equation (2). The total contribution of an individual gene or a principal component to the transition is captured by the non-zero transition weights from the corresponding node to its selected targets in the network. We define the *regulatory potential* (RP) of such predictor as the summation of the absolute weights of the outgoing edges from the node representing the corresponding gene/principal component in normal tissue.

$$RP(g_j) = \sum_{i=1}^{n+k} |m_{ji}|. \tag{3}$$

Thus, predictors can be prioritized with respect to their regulatory potential defined in Equation (3).

3 Validation of the TranNet model

We applied the TranNet model to the TCGA gene expression data for five different cancer types, focusing on solid tumors selected based on availability of matched control-tumor samples: Breast Cancer (BRCA), Lung Adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Prostate Adenocarcinoma (PRAD), and Liver Hepatocellular Carcinoma (LIHC) (see Section 6 for data processing). To validate TranNet model, we first show that the impact of gene expression in normal tissue on gene expression in tumor is higher than expected by chance (predictability). Next, we show that the network edges inferred by TranNet and are enriched in functional interactions. Subsequently, we analyse gene-wise prediction accuracy. These validations in BRCA are collected in Figure 3A, and extended figures including the validation results for all the five cancers are provided in Figure S1, S2 and S3 in Additional file 1. Finally, we confirm that the results are stable thus the conclusions of this study do not depend on the threshold used for selecting the gene set for analyzing (Table 2).

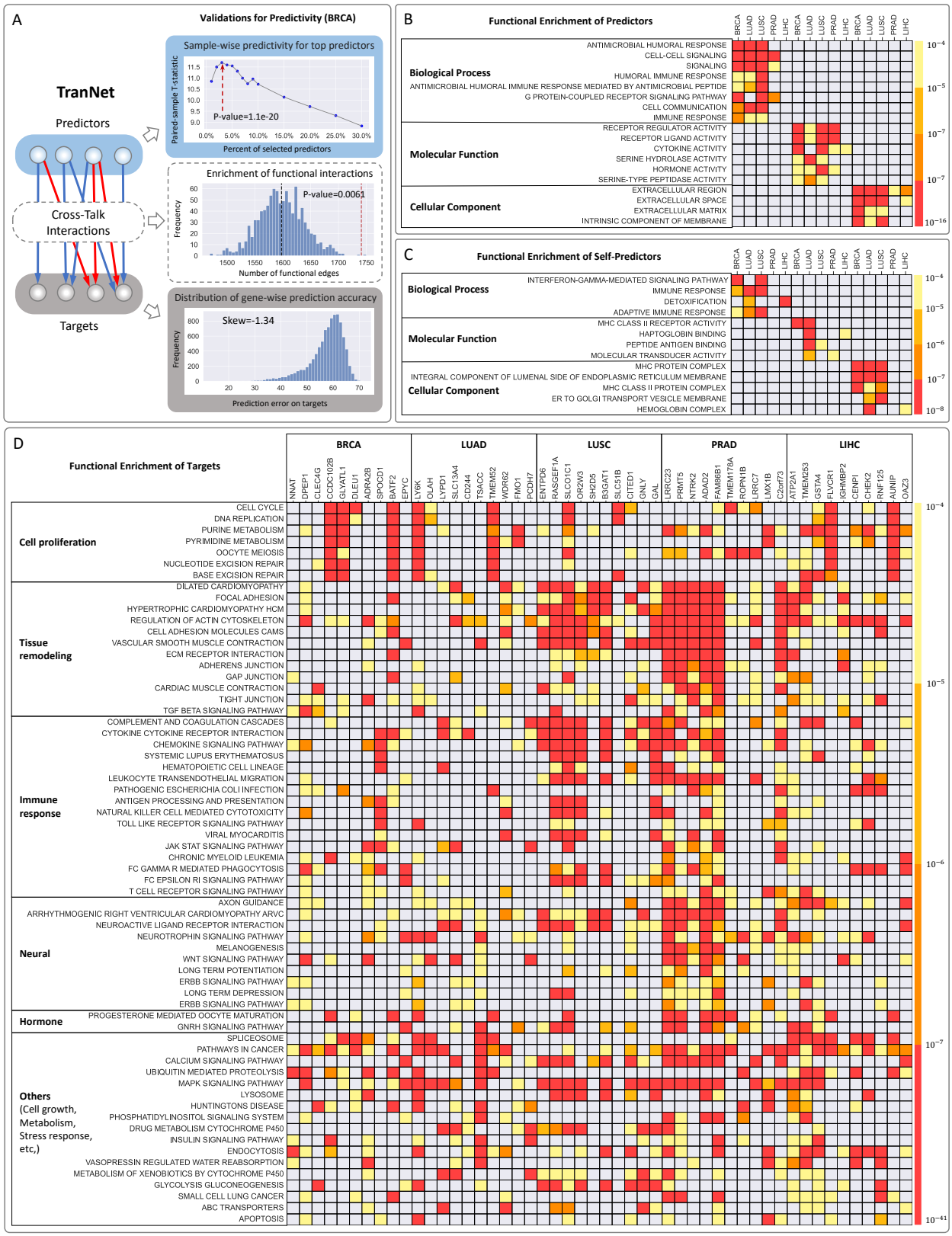

A
Validations for Predictivity (BRCA)
TranNet
Predictors
Cross-Talk Interactions
Targets
Sample-wise predictivity for top predictors
Paired-sample T-statistic
P-value=1.1e-20
Percent of selected predictors
Enrichment of functional interactions
Frequency
P-value=0.0061
Number of functional edges
Distribution of gene-wise prediction accuracy
Skew=-1.34
Prediction error on targets
B
Functional Enrichment of Predictors
BRCA LUAD LUSC PRAD LIHC
Biological Process
ANTIMICROBIAL HUMORAL RESPONSE
CELL-CELL SIGNALING
SIGNALING
HUMORAL IMMUNE RESPONSE
ANTIMICROBIAL HUMORAL IMMUNE RESPONSE MEDIATED BY ANTIMICROBIAL PEPTIDE
G PROTEIN-COUPLED RECEPTOR SIGNALING PATHWAY
CELL COMMUNICATION
IMMUNE RESPONSE
Molecular Function
RECEPTOR REGULATOR ACTIVITY
RECEPTOR LIGAND ACTIVITY
CYTOKINE ACTIVITY
SERINE HYDROLASE ACTIVITY
HORMONE ACTIVITY
SERINE-TYPE PEPTIDASE ACTIVITY
Cellular Component
EXTRACELLULAR REGION
EXTRACELLULAR SPACE
EXTRACELLULAR MATRIX
INTRINSIC COMPONENT OF MEMBRANE
C
Functional Enrichment of Self-Predictors
Biological Process
INTERFERON-GAMMA-MEDIATED SIGNALING PATHWAY
IMMUNE RESPONSE
DETOXIFICATION
ADAPTIVE IMMUNE RESPONSE
Molecular Function
MHC CLASS II RECEPTOR ACTIVITY
HAPTOGLOBIN BINDING
PEPTIDE ANTIGEN BINDING
MOLECULAR TRANSDUCER ACTIVITY
Cellular Component
MHC PROTEIN COMPLEX
INTEGRAL COMPONENT OF LUMENAL SIDE OF ENDOPLASMIC RETICULUM MEMBRANE
MHC CLASS II PROTEIN COMPLEX
ER TO GOLGI TRANSPORT VESICLE MEMBRANE
HEMOGLOBIN COMPLEX
D
Functional Enrichment of Targets
BRCA
LUAD
LUSC
PRAD
LIHC
Cell proliferation
CELL CYCLE
DNA REPLICATION
PURINE METABOLISM
PYRIMIDINE METABOLISM
OOCYTE MEIOSIS
NUCLEOTIDE EXCISION REPAIR
BASE EXCISION REPAIR
Tissue remodeling
DILATED CARDIOMYOPATHY
FOCAL ADHESION
HYPERTROPHIC CARDIOMYOPATHY HCM
REGULATION OF ACTIN CYTOSKELETON
CELL ADHESION MOLECULES CAMS
VASCULAR SMOOTH MUSCLE CONTRACTION
ECM RECEPTOR INTERACTION
ADHERENS JUNCTION
GAP JUNCTION
CARDIAC MUSCLE CONTRACTION
TIGHT JUNCTION
TGF BETA SIGNALING PATHWAY
Immune response
COMPLEMENT AND COAGULATION CASCADES
CYTOKINE CYTOKINE RECEPTOR INTERACTION
CHEMOKINE SIGNALING PATHWAY
SYSTEMIC LUPUS ERYTHEMATOSUS
HEMATOPOIETIC CELL LINEAGE
LEUKOCYTE TRANSENDOTHELIAL MIGRATION
PATHOGENIC ESCHERICHIA COLI INFECTION
ANTIGEN PROCESSING AND PRESENTATION
NATURAL KILLER CELL MEDIATED CYTOTOXICITY
TOLL LIKE RECEPTOR SIGNALING PATHWAY
VIRAL MYOCARDITIS
JAK STAT SIGNALING PATHWAY
CHRONIC MYELOID LEUKEMIA
FC GAMMA R MEDIATED PHAGOCYTOSIS
FC EPSILON RI SIGNALING PATHWAY
T CELL RECEPTOR SIGNALING PATHWAY
Neural
AXON GUIDANCE
ARRHYTHMOGENIC RIGHT VENTRICULAR CARDIOMYOPATHY ARVC
NEUROACTIVE LIGAND RECEPTOR INTERACTION
NEUROTROPHIN SIGNALING PATHWAY
MELANOGENESIS
WNT SIGNALING PATHWAY
LONG TERM POTENTIATION
ERBB SIGNALING PATHWAY
LONG TERM DEPRESSION
ERBB SIGNALING PATHWAY
Hormone
PROGESTERONE MEDIATED OOCYTE MATURATION
GNRH SIGNALING PATHWAY
Others (Cell growth, Metabolism, Stress response, etc,)
SPLICEOSOME
PATHWAYS IN CANCER
CALCIUM SIGNALING PATHWAY
UBIQUITIN MEDIATED PROTEOLYSIS
MAPK SIGNALING PATHWAY
LYSOSOME
HUNTINGTONS DISEASE
PHOSPHATIDYLINOSITOL SIGNALING SYSTEM
DRUG METABOLISM CYTOCHROME P450
INSULIN SIGNALING PATHWAY
ENDOCYTOSIS
VASOPRESSIN REGULATED WATER REABSORPTION
METABOLISM OF XENOBIOTICS BY CYTOCHROME P450
GLYCOLYSIS GLUCONEOGENESIS
SMALL CELL LUNG CANCER
ABC TRANSPORTERS
APOPTOSIS

Figure 3: **Validations and visualisation of the results. (A)** Sample-wise predictivity of TranNet was evaluated based on a leave-one-out test (top sub-panel, Section 3.1); Interactions inferred by TranNet was tested for enrichment of functional interactions (middle sub-panel, Section 3.2); Gene-wise prediction accuracy of TranNet was evaluated by the prediction errors (bottom sub-panel Section 3.1). **(B)** Enrichment of GO terms for the predictor genes: Rows correspond to enriched terms, while the color represents enrichment p-value. GO terms enriched in at least three cancers are shown. The complete lists of GO enrichment for the lists of predictorgenes sorted by their RP scores are provided as in Additional file 2. **(C)** Enrichment of GO terms for self-predictors defined as predictor genes with edges leading from a gene in control tissue to the same gene in tumor. GO terms enriched in at least two cancers are shown and the complete lists of GO terms enriched in such self-predicting genes are provided, for each cancer, in Additional file 4. **(D)** Kegg pathways enrichment for the targets of the top 10 regulators (Table 3) in each cancer type. Rows correspond to the predictor genes while columns represent the pathways enriched for more than 10 predictor genes over all the five cancers. The complete lists are combined in Additional file 5.

3.1 Expression of genes in control samples informs gene expression in tumor

The motivation for the TranNet model is the hypothesis that gene expression in control samples carry relevant information on gene expression in matched tumors. In order to confirm that this is indeed the case, we used a leave-one-out strategy (see Section 6) to test if the model can predict the expression of tumor genes from the expression of predictors better than expected by chance. The paired-sample T-test is used to compare the results (prediction errors) of two predictions: the first based on the real expression data and the second based on permuted data. In addition, we tested how prediction accuracy depends on the number of used predictors assuming that the markers are selected in the decreasing order of their RP values. The prediction accuracy varies with the selection of different percentages of the top RP scoring predictors (see Sample-wise predictability in Figure 3A and Figure S1 in Additional file 1, and Table 1). As we mentioned before, the distribution of regulatory potentials in all the five cancers is characterised by a small number of nodes with high regulatory potential and very large number of genes with low regulatory potential. We point out that despite very small percentage, the summations of regulatory potentials for the top predictors that are above the cut-offs (Table 1) were 42.2% (BRCA), 46.6% (LUAD), 50.7% (LUSC), 56.5% (PRAD) and 29.5% (LIHC) of the total regulatory potentials over all predictors (Additional file 6).

Cancer	BRCA	LUAD	LUSC	PRAD	LIHC
Number/Percentage of predictors (optimized set)	279/3%	347/4%	398/4%	300/4%	127/2%
Paired-sample T-statistic (optimized set)	11.7	11.1	8.3	8.7	9.0
Paired-sample P-value (optimized set)	1.1e-20	8.7e-16	5.6e-11	2.6e-11	4.5e-11
The number of predictors (top 30%)	2792	2607	2989	1919	2254
Paired-sample T-statistic (top 30%)	8.8	7.5	6.1	6.0	6.5
Paired-sample P-value (top 30%)	2.6e-14	4.9e-10	1.6e-07	2.7e-07	8.1e-08

Table 1: **Leave-one-out validation for the TranNet Model.** The paired-sample T-test was used to compare the sample-wise prediction errors on real and permutated response data. This test was performed for the different numbers of top RP scoring predictors. The selected predictors corresponding to the optimal cut-off are provided with their RP scores in Additional file 6. The bottom rows report result using top 30% of predictors. The top rows report the results for the cancer-specific optimised set of predictors (See Figure 3A and Figure S1 in Additional file 1).

As for the target genes, we note that the expressions are not predicted equally well for all the tumor genes. Importantly, the error distributions are highly non-symmetric and negatively skewed towards small errors as visualized in Figure 3A suggesting that the expression of only some tumor genes genes is related to their expression in control samples. The GO terms enriched for the gene lists sorted by the prediction errors on the genes (tumor tissue) are summarized in Additional file 7 and discussed in Section 4.2.

3.2 The transition edges inferred by TranNet are enriched in functional interaction edges.

TranNet infers a sparse bipartite graph modelling the transition from the expression of predictors to the expression of tumor genes in matched tumor samples. Since these edges encode potential regulatory influences (or markers associated with such influences), we expect that they should be enriched for functional interactions. To test this, we computed the overlap of the TranNet edges with the edges in the human functional interaction network [20] (see Section 6) and tested whether this overlap is larger than expected by chance. To this end, we constructed 1,000 random networks by permuting target nodes of the inferred transition network without changing the topology of the network. The enrichment of the functional interactions [20] for BRCA is provided in Figure 3A, and the performances for all the five cancers are summarized as Figure S2 in Additional file 1. These results confirm that TranNet infers relations consistent with functional interactions [20].

We note that, similar to the other biological networks, the distribution of both degrees and regulatory potentials in the TranNet networks is characterised by a large number of predictors with very small regulatory potentials and small proportion of predictors with high potentials (see Figure S4 in Additional file 1 and the table in Additional file 6).

3.3 Stability of TranNet

Since the networks inferred by TranNet are sparse (see the table in Figure S4 in Additional file 1) and inferred from expression data using some cut-offs (see Section 6), it is important to test if an alternative selection of tumor genes would not lead to very different results. To test this, we computed the transition matrix for an alternative definition of tumor (DE) and Non-DE genes where using a q-value cut-off of 0.001 in the T-test for differential expression. Then, we compared the optimal set of predictors for the main set with the same number of predictors computed for the alternative test. The Jaccard similarity and hyper-geometric test results between the two selected sets of predictors showed very high agreements as summarized in Table 2.

	Main	Alternative	Selected-predictors	Overlap	Jaccard Index	HG p-value
BRCA	9309	8172	279	228	0.8	≈ 0.0
LUAD	8691	7225	347	282	0.805	≈ 0.0
LUSC	9965	8635	398	317	0.79	≈ 0.0
PRAD	6399	4464	127	73	0.57	8.7e-125
LIHC	7514	5806	300	215	0.711	3.5e-322

Table 2: **Stability of TranNet.** Overlap in the selected predictors based on the networks constructed form two different selection of tumor genes. Main set is described in Table 4 in Section 6 and the Alternative set was selected with a different threshold as described in Section 3.3. We selected the same number of top predictors from each list based on the number of optimal predictors for the Main set. Overlap reports the number of common predictors, Jaccard Index: Jaccard Index similarity between the two sets of the highest scoring predictors and HG p-value is hyper-geometric p value for the enrichment of the predictors obtained with the Alternative set in the Main set.

4 Insights into the relation between gene expression in tumor and in matched control samples

After validating TranNet model, we investigated the properties of the inferred interactions between control and tumor tissue samples. As noted above, the gene expressions are not predicted equally well for all genes. However, the error distributions are highly non-symmetric and negatively skewed towards small errors as visualized in Figure 3A for BRCA (the error distributions for all the five cancers are provided as Figure S3 in Additional file 1). We started by investigating which biological pathways are enriched within the genes with better predictions. Towards this end we performed enrichment analysis for the gene list ranked by prediction

accuracy. Strikingly, the genes whose expression was predicted with higher accuracy were enriched with specific groups of GO categories related to: environment, cell communication, immune response, signalling, and cell cycle (see the complete list of enriched GO terms in Additional file 7). Particularly, the enriched pathways in BRCA included cell-to-cell signalling, ion transport, and regulation of hormone level. Pathways in LUAD and LUSC had many terms related to cilia as expected due to the known impact of smoking on the properties of ciliated cells [21]. Pathways in LUAD were generally related to cell cycle. Finally pathways enriched in LIHC included, in addition to the pathways related to cell cycle, pathways related to detoxification and taxis and thus related to liver-specific relation with environment [22].

Interestingly, the third principal components PC3 in LUAD and PRAD, are significantly differentiated between control and tumor samples with q-values 2.10E-05 for LUAD and 4.72E-05 for PRAD (see Table 4) and have relatively high regulatory potentials ranked 104-th in the LUAD predictors and 37-th in the PRAD predictors. This demonstrates that expression of genes that did not pass the significance level of differential expression have effects on the differentially expressed genes. As for the targets genes of these components, both are enriched with terms related to metabolism and DNA repair, and Spliceosome (see Table S1 in Additional file 1) while also containing cancer type specific two terms.

4.1 Predictor genes are enriched in pathways that imply the features of tumor microenvironment

To find the properties of the inferred predictor genes, we used GOrilla [23] to perform a GO enrichment analysis. For each cancer, we identified predictor genes, ranked them by their regulatory potential (RP) scores, and performed a ranked list enrichment analysis [23]. We considered all three levels: enrichment of biological process (P), molecular function (F) and cellular component (C) (Figure 3C).

The predictor genes for all the five cancer types are enriched in pathways that imply the feature of the microenvironment around the tumors. For the Biological Process, they include humoral immune response, cell-cell signaling and cell communication, and G-protein coupled receptor. Molecular functions were enriched with cell communication (receptor regulator activity and receptor ligand activity), hormone activity, and inflammation (cytokine activity, serine hydrolase activity, serine-type peptidase activity). With respect to the terms related to cellular component, all cancers were enriched in pathways related to extracellular space and membrane. When compared among individual cancer types, the predictor genes have shown distinct features of tumor microenvironment. This is illustrated by the top ten predictor genes in each cancer type (Table 3, the top 10 cut-off was selected based on 50% dropout in regulatory potential). Unprecedented 85% of these genes have been identified in literature as cancer related and generally associated with tumor progression. For BRCA, the predictors include inflammation-induced genes (CCDC102B [24], SPOCD1 [25], EPYC [26], GLYATL1, SLC6A9 [27]), immune cell markers (CLEC4G [28] and BATF4 [29]), and neurogenesis genes (ADRA2B [30] and NNAT [31]). For LUAD, in addition to inflammation-induced genes (LYPD1 [32]), immune cell markers (CD244 [33]), and neural genes (WDR62 [34]), the predictors also include invasiveness-related genes (ELMO1 [35], LY6K [36], TMEM52 [37], TSACC, PCDH7). Similarly, the predictors of LUSC include to inflammation-induced genes (ENTPD6 [38], SLC51B [39], GNLY, SLCO1C1 [40], and B3GAT1[41]), invasiveness-related genes (CITED1, OR2W3 [42], RASGEF1A [43]), and neural genes (GAL and SH2D5). Importantly, most of these genes are involved in chronic obstructive pulmonary disease (COPD). Taken together, these results suggested that BRCA, LUAD, and LUSC exhibited inflammatory neuroepithelial reactive stroma that are regulated by different mechanisms in immune responses and tissue regeneration. Also shown in Figure 3C, the predictors of PRAD and LIHC were enriched in distinct pathways from other cancer types. Their features are demonstrated by the function of top-ranked predictor genes. For PRAD, the majority of top-ten predictors are neurogenesis genes (NTRK2 [44], LRRC7, LMX1B, ROPN1B); others include genes involved in inflammatory response (TMEM178A) and androgen receptor regulation (PRMT5 [45]). This is relevant to the intense neuroepithelial interactions in prostate cancer. For LIHC, the predictors include genes responsive to DNA damage (AUNIP, CHEK2, CENPI) and hypoxia (FLVCR1), and two genes involved in liver disease such as fatty liver and hepatitis (GSTA4, ATP2A1 [46], RNF125 [47]). They reflect the metabolic modification in the tumor microenvironments.

BRCA		LUAD		LUSC		PRAD		LIHC	
NNAT	**[48]**	LY6K	**[49]**	ENTPD6	**[38]**	LRRC23	[50]	ATP2A1	[51]
DPEP1	**[52]**	OLAH	[53]	RASGEF1A	**[54]**	PRMT5	**[45]**	TMEM253	
CLEC4G	[55]	LYPD1	[56]	SLCO1C1	[57]	NTRK2	**[58]**	GSTA4	**[59]**
CCDC102B	**[24]**	SLC13A4	[60]	OR2W3	[61]	ADAD2	[62]	FLVCR1	**[63]**
GLYATL1	**[64]**	CD244	**[65]**	SH2D5	[66]	FAM86B1		IGHMBP2	[67]
DLEU1	**[68]**	TSACC		B3GAT1	[69]	TMEM178A		CENPI	[70]
ADRA2B	**[30]**	TMEM52	**[71]**	SLC51B	[72]	ROPN1B	[73]	CHEK2	**[74]**
SPOCD1	[75]	WDR62	**[76]**	CITED1	**[77]**	LRRC7	[50]	RNF125	**[78]**
BATF2	**[79]**	ELMO1		GNLY	**[80]**	LMX1B	**[81]**	AUNIP	[82]
EPYC	**[83]**	PCDH7	**[84]**	GAL		C2orf73		OAZ3	[85]

Table 3: **Top ten RP scoring genes.** The lists of the top 10 genes with the highest regulatory potentials (Equation (3)) in each cancer together with references to sample papers discussing their roles in cancer. In bold are references to papers that point to relations with the same cancer type. The regulatory potentials of the top predictors for the five cancers are provided at `https://github.com/ncbi/TranNet`.

4.2 Pathway enrichment of genes identified as targets of predictors revealed tumor-stroma interactions

To investigate how the microenvironments might regulate the tumors, we identified the pathways enriched (see Section 6 for the pathway enrichment) by the tumor genes ("targets"; Matrix Y in Figure 2B) associated with top-ranked predictor genes in Table 3. As shown in Figure 3D, the targets in BRCA and LUAD are more enriched in cell proliferation-related pathways (the first group in Figure 3D). In LUSC, the majority of targets are enriched in tissue remodeling and immune response pathways (the second and third groups in Figure 3D, respectively). In PRAD, in addition to immune response and tissue remodeling, the targets are mostly enriched in neural and hormone pathways (fourth and fifth groups in Figure 3D, respectively). In LIHC, the targets are enriched mostly in cell proliferation and tissue remodeling pathways. These results reflect the distinct signaling for tumor initiation and expansion in each cancer type. For example, studies have shown that the occurrence of LUSC and LIHC are frequently preceded with Chronic Obstructive Pulmonary Disease (COPD) and hepatitis, respectively, and the tissue injury and remodeling in such chronic inflammation promote tumorigenesis. Interestingly, the same types of pathways in tumors could be associated with different types of predictors, depending on cancer type. For example, cell proliferation-related pathways in tumors are associated with inflammation-related predictors for BRCA (CCDC102B and GLYATL1) and LUSC (SLCO1C1, SLC51B, and BATF2); with invasiveness predictors for LUAC (LY6K and TMEM52); with neurogenesis predictors (LRRC7, LMX1B, ROPN1B) in PRAD; and with hypoxic (FLVCR1) and DNA-damage response (AUNIP) predictors for LIHC. These observations suggest that tumor growth is stimulated by distinct microenvironmental factors in each cancer type, as summarised in 4. Taken together, our results also give important implication in the detection and diagnosis of cancers at early stage.

4.3 Predictors whose expression in control samples is predictive of expression in tumor have unique properties

Interestingly, we noted that the top predictors do not include genes that have edges leading from a gene in control tissue to the same gene in tumor. We refer to such genes as self-predictors and asked, if such self-predictors have properties distinct form the properties of the top predictors. To answer this question, in each cancer, we have identified all self predictors (see Additional file 3 for the list of self-predictors). Note that since all network edges are between differentially expressed genes only, this excludes genes whose expression is simply the same in normal and tumor tissues. Not surprisingly, gene expression in control and tumor samples for self-predictors tends to be significantly correlated (Additional file 3). Strikingly, in contrast to the predictors with high regulatory potential, GO enrichment analysis of self-predicting genes revealed relations to oxidation (including hemoglobin complex) and detoxification (Figure 3 C, Additional file 4). In particular self-predictors enrichment terms did not include terms directly related to tissue remodelling.

While both, the top predictors and self-predictors, contained terms related to immune response, the terms for self-predictive genes were more strongly focused on Major Histocompatibility Complex (MHC)-related immune response (Figure 3 C, and Additional file 4). Methods and data used for functional enrichment are described in Section 6.

5 Conclusions

The regulation of tumor progression is not fully understood. Recent studies demonstrated that control samples representing adjacent normal tissue in solid tumor studies differ form healthy tissue [11]. We hypothesised that genes whose expression in control samples is predictive of the expression in tumor samples might provide insights into this relation.

We developed and validated a new computational method, TranNet, which is able to identify predictor genes, predictive (and potentially facilitating) regulation of a tumor by the tumor environment (Fig 4). We found that the functional properties of the predictor genes are consistent with the mechanisms of regulation by the environment. In addition, many of the identified predictors are previously recognised markers of the microenvironment-mediated tumor progression. One limitation of this study is that while TranNet identifies predictors of such external influences on a tumor it does not establish causality.

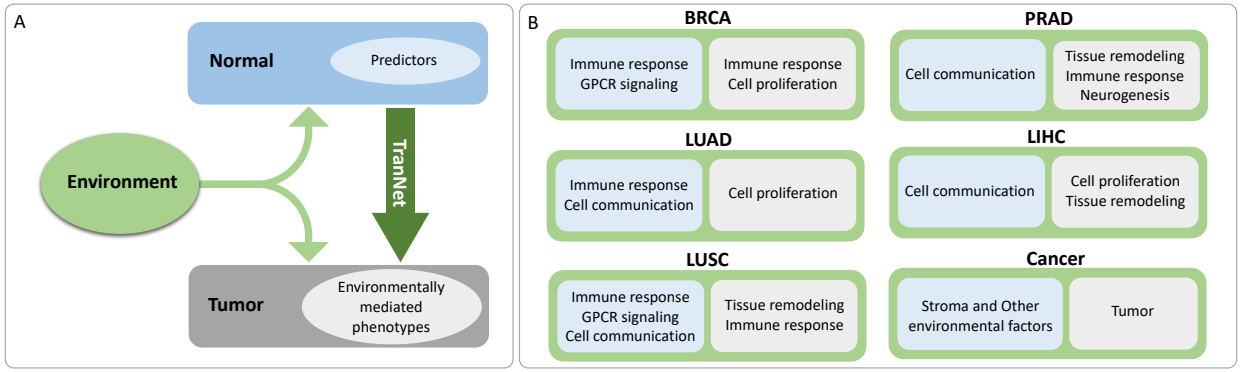


Figure 4: **Interpretations of the relations between tumor and normal tissues inferred by TranNet.** TranNet predictors (light blue) in normal tissue predict environment-related changes in tumor (light grey) **(A)**. Cancer-specific functional enrichment of predictor genes (light blue) and inferred environmentally modulated changes in tumor (light gray) **(B)**.

It is important to recognize that some external factors that might impact both normal tissues and tumors and do not necessarily cause tumor progression directly. Even though, many of these factors interact with and/or regulate tumor and are interesting in this respect COPID and hepatitis are likely examples. There is also an association between the incidence of a wide variety of malignancies and diabetes. Identification of such interactions is of fundamental importance for preventing cancer and development of novel treatments.

The concept of tumor environment, has many layers. Previous studies distinguished tumor's micro and macro environments to capture the effects of the closest neighborhood (such as normal cells, molecules, and blood vessels that surround and feed a tumor cell) and more distant and more distant influences respectively [86]. Here we also allow yet another level of influences: organismal environment and germline variations. Our approach provides a proof of principle that control samples can be used to gain insight into the impact of external factors, jointly referred here as the environment, on tumor and tumor progression. The enrichment of predictor genes in processes associated with immune response suggest that some of these predictors might be in fact sensors of common exposure to chronic infection in normal and tumor tissues. Indeed, it is recognized that cigarette smoking induces lung inflammation, leading to the changes in cellular composition and functions, such as a reduction in the number and properties of ciliated cells. In future studies, it would be interesting to extend the model to allow for a separation of different types of external contributions. In general, it might be difficult to fully deconvolute such processes, as environmental contributors are mostly unknown. However, some environmental factors are mutagenic and in such cases the exposure of the organism

to an environmental factor can be estimated by the strength of a corresponding mutational signature [87, 88]. These potential extensions aside, the results obtained by TranNet already provided new insights into the relation between tumor, adjacent control samples, and tumor environment.

6 Materials and Methods

Gene expression data: Batch effect removed TCGA gene expression data [89] were downloaded from GitHub repository (`https://github.com/mskcc/RNAseqDB`) on April 5th, 2022. Requiring matched normal and cancer samples, five types of EMT cancers: Breast Cancer (BRCA), Lung Adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Prostate Adenocarcinoma (PRAD), and Liver Hepatocellular Carcinoma (LIHC) were selected for the analysis and only patients that had both normal and tumor gene expressions were kept for each cancer type. We excluded genes which are expressed in neither normal (control) nor tumor samples, and the expression level of a gene is assumed as significant if 80% or more of its expression values over samples are greater than a threshold (see Figure S5 in Additional file 1). We then selected the genes that are differentially expressed between control and tumor samples, and expression data for the remaining genes (non-differentially expressed) are compressed in the principal components that preserve 99% of covariate information of the data [15]. From these principal components (PCs), only differentiated (differentially expressed) components are assumed as additional meta nodes in the network, representing the presence of potential covariates in the transition network, coming from the non-differentially expressed genes [14]. The conceptual idea for defining the differentially expressed principal components and for including them as additional variables in the presenting analysis is described in Section 2.1 and Figure 2B. Finally, activity (expression of genes and magnitudes of PCs) of each node is standardized with Z-score to bring dissimilar features on a similar scale. The data cohorts for all the five cancers are provided as Table 6.

Leave-one-out test: The leave-one-out test was performed as follows. For each patient i, remove the corresponding expression values $x_{i(\cdot)}$ (a row) and $y_{i(\cdot)}$ (a row) from the data X and Y. A transition matrix $M_{(-i)}$ is then computed from the remaining data (such as $X_{(-i)} \times M_{(-i)} \approx Y_{(-i)}$). Then, check if the prediction error on the true response $y_{i(\cdot)}$ is less than the error on the random response $\tilde{y}_{i(\cdot)}$, such that $\|x_{i(\cdot)} \cdot M_{(-i)} - y_{i(\cdot)}\|_{l_1} < \|x_{i(\cdot)} \cdot M_{(-i)} - \tilde{y}_{i(\cdot)}\|_{l_1}$, where $\tilde{y}_{i(\cdot)}$ is obtained as permuting $y_{i(\cdot)}$ ($\|\cdot\|_{l_1}$ is denoted as the absolute summation over the prediction errors on the genes and principal components). A T-test for paired samples was performed to check if this prediction error is less than the random error. The prediction accuracy varies depending on the number of selected predictors as shown in Figure 3A for BRCA, and Figure S1 for all the five cancers in Additional file 1. The number of predictors for each cancer was decided based on different cut-off percents of top genes and principal components with the highest regulatory potentials. The test results for the five cancers are summarized in Table 1.

Functional Network: We used a functional network [20] of human genes for disease studies downloaded from HumanNet v3 (`https://www.inetbio.org/humannet`). This data set contains 18,459 protein coding genes and 977,495 interactions inferred from various datasets and functional networks. The interactions corresponding to our genes sets are used as the edge validation sets in the analysis of the cancers. The characteristics of the gene expression data and edge validation set are summarized in Table 4.

Enrichment of GO terms: Enrichment of GO term biological processes, molecular functions and cellular components was performed using GOrilla [23] for either the sorted lists or selected subsets of genes.

Enrichment of Kegg pathways: Pathway enrichment was performed with a hypergeometric test followed by a Benjamin–Hochberg test for multiple comparison corrections using Kegg pathways [90]. Pathways enriched with $q - value < 0.01$ are reported in this study.

Availability of source codes and data: The TranNet method was implemented in python, source codes, data sets and supplemental tables used for and generated during this study are available at the Github site `https://github.com/ncbi/TranNet`.

#	DE Genes	Non-DE Genes	DE PCs (q value)	TranNet	Patients	HN Genes	HN Edges
BRCA	9308	4831	PC27(0.0015)	9309	105	8633	308985
LUAD	8688	5798	PC3,6,8(2.1E-05, 0.0056, 0.0099)	8691	57	8087	269821
LUSC	9965	4853	PC5(0.0003)	9965	50	9246	345514
PRAD	6398	8440	PC3(4.72E-05)	6399	47	5951	129963
LIHC	7514	6177		7514	40	7011	230952

Table 4: **Data summary.** DE genes (tumor genes): differentially expressed genes selected based on T-test q-value < 0.01; Non-DE Genes: genes whose expressions are not differentiated between control and tumor samples; DE PCs: principal components which are differentiated between control and tumor samples representing potential impacts of Non-DE Genes (T-test q-value) on the transition mapping; TranNet: Pairs of Nodes in the Transition Network (DE Genes + DE PCs); Patients: Patients control and tumor samples are available; HN Genes: genes in both DE Genes and HumanNet-FN; HN Edges: Edges within HN Genes in the functional network.

Additional Files

- Additional file 1 — Supplemental Information: Table S1, Figures S1, S2, S3, S4 and S5. (PDF)
- Additional file 2 — Lists of GO terms enriched for the gene list sorted based on their regulatory potential (PR) scores for each of the five types of cancers: Enrichment of biological process, molecular function and cellular component are provided in separate sheets in the same file (Excel).
- Additional file 3 — Lists of the self-predictor genes with their self-interaction weights, correlation between their normal and tumor tissue expressions and the corresponding p-values. (Excel)
- Additional file 4 — Lists of enriched GO terms for the genes whose normal samples are predictive for the corresponding tumor samples (self-predictors): Enrichment of biological process, molecular function and cellular component (Excel).
- Additional file 5 — Lists of enriched Kegg pathways for the target genes of the top ten RP scoring predictors: Enriched pathways and the corresponding p-values (Excel).
- Additional file 6 — Lists of the selected high scoring predictor genes and principal components with their regulatory potentials for the five cancers. (Excel)
- Additional file 7 — Lists of GO terms enriched for the gene list sorted based on prediction errors for each of the five cancers (Genes whose tumor expressions are predicted with smaller errors are listed in higher ranks in the list): Enrichment of biological process, molecular function and cellular component are provided in separate sheets in the same file (Excel).

Acknowledgement

We want to thank M.G. Hirsch for comments on the manuscript. This research was supported by the Intramural Research Program of the National Library of Medicine.

References

[1] R. Sager. Expression genetics in cancer: shifting the focus from DNA to RNA. *Proc Natl Acad Sci U S A*, 94(3):952–955, Feb 1997.

[2] V. da Silva-Diz, L. Lorenzo-Sanz, A. Bernat-Peguera, M. Lopez-Cerda, and P. oz. Cancer cell plasticity: Impact on tumor progression and therapy response. *Semin Cancer Biol*, 53:48–58, Dec 2018.

[3] S. Shen and J. Clairambault. Cell plasticity in cancer cell populations. *F1000Res*, 9, 2020.

[4] S. Yuan, R. J. Norgard, and B. Z. Stanger. Cellular Plasticity in Cancer. *Cancer Discov*, 9(7):837–851, Jul 2019.

[5] S. Huang. Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells. *Cancer Metastasis Rev*, 32(3-4):423–448, Dec 2013.

[6] S. Huang. Tumor progression: chance and necessity in Darwinian and Lamarckian somatic (mutationless) evolution. *Prog Biophys Mol Biol*, 110(1):69–86, Sep 2012.

[7] M. Wang, J. Zhao, L. Zhang, F. Wei, Y. Lian, Y. Wu, Z. Gong, S. Zhang, J. Zhou, K. Cao, X. Li, W. Xiong, G. Li, Z. Zeng, and C. Guo. Role of tumor microenvironment in tumorigenesis. *J Cancer*, 8(5):761–773, 2017.

[8] D. F. Quail and J. A. Joyce. Microenvironmental regulation of tumor progression and metastasis. *Nat Med*, 19(11):1423–1437, Nov 2013.

[9] F. Balkwill and A. Mantovani. Inflammation and cancer: back to Virchow? *Lancet*, 357(9255):539–545, Feb 2001.

[10] L. M. Coussens and Z. Werb. Inflammation and cancer. *Nature*, 420(6917):860–867, 2002.

[11] D. Aran, R. Camarda, J. Odegaard, H. Paik, B. Oskotsky, G. Krings, A. Goga, M. Sirota, and A. J. Butte. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat Commun*, 8(1):1077, 10 2017.

[12] P. Mehlen and A. Puisieux. Metastasis: a question of life or death. *Nat Rev Cancer*, 6(6):449–458, Jun 2006.

[13] B. Amgalan and H. Lee. DEOD: uncovering dominant effects of cancer-driver genes based on a partial covariance selection method. *Bioinformatics*, 31(15):2452–2460, Aug 2015.

[14] K. P. Jablonski, M. Pirkl, D. Ćevid, P. Bühlmann, and N. Beerenwinkel. Identifying cancer pathway dysregulations using differential causal effects. *Bioinformatics*, Dec 2021.

[15] M. Ringnér. What is principal component analysis? *Nat Biotechnol*, 26(3):303–304, Mar 2008.

[16] J. Kim, Y. Kim, and Y. Kim. A gradient-based optimization algorithm for lasso. *Journal of Computational and Graphical Statistics*, 17(4):994–1009, 2008.

[17] J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*, 33(1):1–22, 2010.

[18] D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

[19] E. M. Gafni and D. P. Bertsekas. Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization*, 22(6):936–964, 1984.

[20] C. Y. Kim, S. Baek, J. Cha, S. Yang, E. Kim, E. M. Marcotte, T. Hart, and I. Lee. HumanNet v3: an improved database of human gene networks for disease research. *Nucleic Acids Res*, 50(D1):D632–D639, 01 2022.

[21] P. L. Leopold, M. J. O'Mahony, X. J. Lian, A. E. Tilley, B. G. Harvey, and R. G. Crystal. Smoking is associated with shortened airway cilia. *PLoS One*, 4(12):e8157, Dec 2009.

[22] D. M. Grant. Detoxification pathways in the liver. *J Inherit Metab Dis*, 14(4):421–430, 1991.

[23] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:48, Feb 2009.

[24] J. Si, R. Guo, B. Xiu, W. Chi, Q. Zhang, J. Hou, Y. Su, J. Chen, J. Xue, Z. M. Shao, J. Wu, and Y. Chi. B Pathway. *Front Oncol*, 12:927358, 2022.

[25] K. Sato, T. Masuda, Q. Hu, T. Tobo, S. Kidogami, Y. Ogawa, T. Saito, S. Nambara, H. Komatsu, H. Hirata, S. Sakimura, R. Uchi, N. Hayashi, T. Iguchi, H. Eguchi, S. Ito, T. Nakagawa, and K. Mimori. Phosphoserine Phosphatase Is a Novel Prognostic Biomarker on Chromosome 7 in Colorectal Cancer. *Anticancer Res*, 37(5):2365–2371, May 2017.

[26] J. Zhang, S. Zhang, Y. Zhou, Y. Qu, T. Hou, W. Ge, and S. Zhang. KLF9 and EPYC acting as feature genes for osteoarthritis and their association with immune infiltration. *J Orthop Surg Res*, 17(1):365, Jul 2022.

[27] S. Carmans, J. J. Hendriks, K. Thewissen, J. Van den Eynden, P. Stinissen, J. M. Rigo, and N. Hellings. The inhibitory neurotransmitter glycine modulates macrophage activity by activation of neutral amino acid transporters. *J Neurosci Res*, 88(11):2420–2430, Aug 2010.

[28] A. Dominguez-Soto, L. Aragoneses-Fenoll, E. Martin-Gayo, L. Martinez-Prats, M. Colmenares, M. Naranjo-Gomez, F. E. Borras, P. Munoz, M. Zubiaur, M. L. Toribio, R. Delgado, and A. L. Corbi. The DC-SIGN-related lectin LSECtin mediates antigen capture and pathogen binding by human myeloid cells. *Blood*, 109(12):5337–5345, Jun 2007.

[29] S. Roy, R. Guler, S. P. Parihar, S. Schmeier, B. Kaczkowski, H. Nishimura, J. W. Shin, Y. Negishi, M. Ozturk, R. Hurdayal, A. Kubosaki, Y. Kimura, M. J. de Hoon, Y. Hayashizaki, F. Brombacher, and H. Suzuki. Batf2/Irf1 induces inflammatory responses in classically activated macrophages, lipopolysaccharides, and mycobacterial infection. *J Immunol*, 194(12):6035–6044, Jun 2015.

[30] E. M. Rivero, L. M. Martinez, C. D. Bruque, L. Gargiulo, A. Bruzzone, and I. A. Lüthy. Prognostic significance of α and β2-adrenoceptor gene expression in breast cancer patients. *Br J Clin Pharmacol*, 85(9):2143–2154, 09 2019.

[31] N. Nass, S. Walter, D. Jechorek, C. Weissenborn, A. Ignatov, J. Haybaeck, S. Sel, and T. Kalinski. High neuronatin (NNAT) expression is associated with poor outcome in breast cancer. *Virchows Arch*, 471(1):23–30, Jul 2017.

[32] X. W. Fu, P. F. Song, and E. R. Spindel. Role of Lynx1 and related Ly6 proteins as modulators of cholinergic signaling in normal and neoplastic bronchial epithelium. *Int Immunopharmacol*, 29(1):93–98, Nov 2015.

[33] L. Sun, X. Gang, Z. Li, X. Zhao, T. Zhou, S. Zhang, and G. Wang. Advances in Understanding the Roles of CD244 (SLAMF4) in Immune Regulation and Associated Diseases. *Front Immunol*, 12:648182, 2021.

[34] J. F. Chen, Y. Zhang, J. Wilde, K. C. Hansen, F. Lai, and L. Niswander. Microcephaly disease gene Wdr62 regulates mitotic progression of embryonic neural stem cells and brain size. *Nat Commun*, 5:3885, May 2014.

[35] H. Li, L. Yang, H. Fu, J. Yan, Y. Wang, H. Guo, X. Hao, X. Xu, T. Jin, and N. Zhang. i2 and ELMO1/Dock180 connects chemokine signalling with Rac activation and metastasis. *Nat Commun*, 4:1706, 2013.

[36] M. AlHossiny, L. Luo, W. R. Frazier, N. Steiner, Y. Gusev, B. Kallakury, E. Glasgow, K. Creswell, S. Madhavan, R. Kumar, and G. Upadhyay. Ly6E/K Signaling to TGFβ Promotes Breast Cancer Progression, Immune Escape, and Drug Resistance. *Cancer Res*, 76(11):3376–3386, 06 2016.

[37] K. C. Ehrlich, M. Lacey, and M. Ehrlich. Gene Families. *Epigenomes*, 4(1), Jan 2020.

[38] Z. Zhuang and C. Gao. Development of a Clinical Prognostic Model for Metabolism-Related Genes in Squamous Lung Cancer and Correlation Analysis of Immune Microenvironment. *Biomed Res Int*, 2022:6962056, 2022.

[39] T. Berg, T. ck, M. Olsson, J. rd, V. m, X. H. Zhou, J. Grunewald, L. Gustavsson, and M. Nord. Gene expression analysis of membrane transporters and drug-metabolizing enzymes in the lung of healthy and COPD subjects. *Pharmacol Res Perspect*, 2(4):e00054, Aug 2014.

[40] D. A. Ridder, M. F. Lang, S. Salinin, J. P. derer, M. Struss, C. Maser-Gluth, and M. Schwaninger. TAK1 in brain endothelial cells mediates fever and lethargy. *J Exp Med*, 208(13):2615–2623, Dec 2011.

[41] J. D. Trimarco, S. L. Nelson, R. R. Chaparian, A. I. Wells, N. B. Murray, P. Azadi, C. B. Coyne, and N. S. Heaton. Cellular glycan modification by B3GAT1 broadly restricts influenza virus infection. *Nat Commun*, 13(1):6456, Oct 2022.

[42] M. Gao, W. Kong, Z. Huang, and Z. Xie. Identification of Key Genes Related to Lung Squamous Cell Carcinoma Using Bioinformatics Analysis. *Int J Mol Sci*, 21(8), Apr 2020.

[43] D. Wuttig, B. Baier, S. Fuessel, M. Meinhardt, A. Herr, C. Hoefling, M. Toma, M. O. Grimm, A. Meye, A. Rolle, and M. P. Wirth. Gene signatures of pulmonary metastases of renal cell carcinoma reflect the disease-free interval and the number of metastases per patient. *Int J Cancer*, 125(2):474–482, Jul 2009.

[44] R. Chow, J. M. Wessels, and W. G. Foster. Brain-derived neurotrophic factor (BDNF) expression and function in the mammalian reproductive Tract. *Hum Reprod Update*, 26(4):545–564, Jun 2020.

[45] E. Beketova, J. L. Owens, A. M. Asberry, and C. D. Hu. PRMT5: a putative oncogene and therapeutic target in prostate cancer. *Cancer Gene Ther*, 29(3-4):264–276, 03 2022.

[46] J. Kong, S. Sun, F. Min, X. Hu, Y. Zhang, Y. Cheng, H. Li, X. Wang, and X. Liu. Integrating Network Pharmacology and Transcriptomic Strategies to Explore the Pharmacological Mechanism of Hydroxysafflor Yellow A in Delaying Liver Aging. *Int J Mol Sci*, 23(22), Nov 2022.

[47] H. Toyoda, T. Kumada, S. Kiriyama, M. Tanikawa, Y. Hisanaga, A. Kanamori, T. Tada, S. Kitabatake, and Y. Murakami. Association between hepatic steatosis and hepatic expression of genes involved in innate immunity in patients with chronic hepatitis C. *Cytokine*, 63(2):145–150, Aug 2013.

[48] W. Pieper, A. Ignatov, T. Kalinski, J. Haybaeck, P. Czapiewski, and N. Nass. The predictive potential of Neuronatin for neoadjuvant chemotherapy of breast cancer. *Cancer Biomark*, 32(2):161–173, 2021.

[49] M. M. Ali, M. Di Marco, S. Mahale, D. Jachimowicz, S. T. Kosalai, S. Reischl, L. Statello, K. Mishra, C. Darnfors, M. Kanduri, and C. Kanduri. LY6K-AS lncRNA is a lung adenocarcinoma prognostic biomarker and regulator of mitotic progression. *Oncogene*, 40(13):2463–2478, 04 2021.

[50] M. J. Santoni, R. Kashyap, L. Camoin, and J. P. Borg. The Scribble family in cancer: twentieth anniversary. *Oncogene*, 39(47):7019–7033, Nov 2020.

[51] G. Zhang, H. Shang, B. Liu, G. Wu, D. Wu, L. Wang, S. Li, Z. Wang, S. Wang, and J. Yuan. Predicts Poor Prognosis in Patients With Colorectal Carcinoma. *Front Genet*, 13:661348, 2022.

[52] A. R. Green, S. Krivinskas, P. Young, E. A. Rakha, E. C. Paish, D. G. Powe, and I. O. Ellis. Loss of expression of chromosome 16q genes DPEP1 and CTCF in lobular carcinoma in situ of the breast. *Breast Cancer Res Treat*, 113(1):59–66, Jan 2009.

[53] Q. Wu, D. Wang, Z. Zhang, Y. Wang, W. Yu, K. Sun, N. R. Maimela, Z. Sun, J. Liu, W. Yuan, and Y. Zhang. DEFB4A is a potential prognostic biomarker for colorectal cancer. *Oncol Lett*, 20(4):114, Oct 2020.

[54] C. Suzuki, K. Takahashi, S. Hayama, N. Ishikawa, T. Kato, T. Ito, E. Tsuchiya, Y. Nakamura, and Y. Daigo. Identification of Myc-associated protein with JmjC domain as a novel therapeutic target oncogene for lung cancer. *Mol Cancer Ther*, 6(2):542–551, Feb 2007.

[55] Y. Zhang, H. Wei, L. Fan, M. Fang, X. He, B. Lu, and Z. Pang. CLEC4s as Potential Therapeutic Targets in Hepatocellular Carcinoma Microenvironment. *Front Cell Dev Biol*, 9:681372, 2021.

[56] R. M. Burnett, K. E. Craven, P. Krishnamurthy, C. P. Goswami, S. Badve, P. Crooks, W. P. Mathews, P. Bhat-Nakshatri, and H. Nakshatri. Organ-specific adaptive signaling pathway activation in metastatic breast cancer cells. *Oncotarget*, 6(14):12682–12696, May 2015.

[57] Y. F. Xiao, B. S. Li, J. J. Liu, S. M. Wang, J. Liu, H. Yang, Y. Y. Hu, C. L. Gong, J. L. Li, and S. M. Yang. Role of lncSLCO1C1 in gastric cancer progression and resistance to oxaliplatin therapy. *Clin Transl Med*, 12(4):e691, 04 2022.

[58] C. M. Faltermeier, J. M. Drake, P. M. Clark, B. A. Smith, Y. Zong, C. Volpe, C. Mathis, C. Morrissey, B. Castor, J. Huang, and O. N. Witte. Functional screen identifies kinases driving prostate cancer visceral and bone metastasis. *Proc Natl Acad Sci U S A*, 113(2):E172–181, Jan 2016.

[59] D. L. White, D. Li, Z. Nurgalieva, and H. B. El-Serag. Genetic variants of glutathione S-transferase as possible risk factors for hepatocellular carcinoma: a HuGE systematic review and meta-analysis. *Am J Epidemiol*, 167(4):377–389, Feb 2008.

[60] M. L. Yang, J. H. Zhang, S. Li, R. Zhu, and L. Wang. SLC13A4 Might Serve as a Prognostic Biomarker and be Correlated with Immune Infiltration into Head and Neck Squamous Cell Carcinoma. *Pathol Oncol Res*, 27:1609967, 2021.

[61] S. Masjedi, L. J. Zwiebel, and T. D. Giorgio. Olfactory receptor gene abundance in invasive breast carcinoma. *Sci Rep*, 9(1):13736, 09 2019.

[62] J. Pan, Q. Dai, Z. Xiang, B. Liu, and C. Li. Three Biomarkers Predict Gastric Cancer Patients' Susceptibility To Fluorouracil-based Chemotherapy. *J Cancer*, 10(13):2953–2960, 2019.

[63] K. Zhang, Z. Zhao, J. Yu, W. Chen, Q. Xu, and L. Chen. LncRNA FLVCR1-AS1 acts as miR-513c sponge to modulate cancer cell proliferation, migration, and invasion in hepatocellular carcinoma. *J Cell Biochem*, 119(7):6045–6056, 07 2018.

[64] J. Wang, A. Shidfar, D. Ivancic, M. Ranjan, L. Liu, M. R. Choi, V. Parimi, D. B. Gursel, M. E. Sullivan, M. S. Najor, A. M. Abukhdeir, D. Scholtens, and S. A. Khan. Overexpression of lipid metabolism genes and PBX1 in the contralateral breasts of women with estrogen receptor-negative breast cancer. *Int J Cancer*, 140(11):2484–2497, Jun 2017.

[65] R. D. W. Vaes, K. Reynders, J. Sprooten, K. T. Nevola, K. M. A. Rouschop, M. Vooijs, A. D. Garg, M. Lambrecht, L. E. L. Hendriks, M. Rucevic, and D. De Ruysscher. Identification of Potential Prognostic and Predictive Immunological Biomarkers in Patients with Stage I and Stage III Non-Small Cell Lung Cancer (NSCLC): A Prospective Exploratory Study. *Cancers (Basel)*, 13(24), Dec 2021.

[66] Y. Zheng, P. Ming, C. Zhu, Y. Si, S. Xu, A. Chen, J. Wang, and B. Zhang. Hepatitis B virus X protein-induced SH2 domain-containing 5 (SH2D5) expression promotes hepatoma cell growth via an SH2D5-transketolase interaction. *J Biol Chem*, 294(13):4815–4827, 03 2019.

[67] J. Shen, M. B. Terry, M. D. Gammon, M. M. Gaudet, S. L. Teitelbaum, S. M. Eng, S. K. Sagiv, A. I. Neugut, and R. M. Santella. IGHMBP2 Thr671Ala polymorphism might be a modifier for the effects of cigarette smoking and PAH-DNA adducts to breast cancer risk. *Breast Cancer Res Treat*, 99(1):1–7, Sep 2006.

[68] B. Pang, S. Sui, Q. Wang, J. Wu, Y. Yin, and S. Xu. Upregulation of DLEU1 expression by epigenetic modification promotes tumorigenesis in human cancer. *J Cell Physiol*, 234(10):17420–17432, 08 2019.

[69] M. C. Ramello, N. G. Núñez, J. Tosello Boari, S. N. Bossio, F. P. Canale, C. Abrate, N. Ponce, A. Del Castillo, M. Ledesma, S. Viel, W. Richer, C. Sedlik, C. Tiraboschi, M. Muñoz, D. Compagno, A. Gruppi, E. V. Acosta Rodríguez, E. Piaggio, and C. L. Montes. T Cells Infiltrate Tumors and Are Expanded in Peripheral Blood From Breast Cancer Patients. *Front Immunol*, 12:713132, 2021.

[70] N. Ding, R. Li, W. Shi, and C. He. CENPI is overexpressed in colorectal cancer and regulates cell migration and invasion. *Gene*, 674:80–86, Oct 2018.

[71] Z. Wang, H. Pei, H. Liang, Q. Zhang, L. Wei, D. Shi, Y. Chen, and J. Zhang. Construction and Analysis of a circRNA-Mediated ceRNA Network in Lung Adenocarcinoma. *Onco Targets Ther*, 14:3659–3669, 2021.

[72] J. Cheng, Y. Li, X. Wang, Z. Dong, Y. Chen, R. Zhang, J. Huang, X. Jin, J. Yao, A. Ge, L. Song, Y. Lu, and Z. Zeng. Response Stratification in the First-Line Combined Immunotherapy of Hepatocellular Carcinoma at Genomic, Transcriptional and Immune Repertoire Levels. *J Hepatocell Carcinoma*, 8:1281–1295, 2021.

[73] W. Wu, M. Warner, L. Wang, W. W. He, R. Zhao, X. Guan, C. Botero, B. Huang, C. Ion, C. Coombes, and J. A. Gustafsson. Drivers and suppressors of triple-negative breast cancer. *Proc Natl Acad Sci U S A*, 118(33), Aug 2021.

[74] M. Lulli, L. Del Coco, T. Mello, C. Sukowati, S. Madiai, L. Gragnani, P. Forte, F. P. Fanizzi, A. Mazzocca, K. Rombouts, A. Galli, and V. Carloni. DNA Damage Response Protein CHK2 Regulates Metabolism in Liver Cancer. *Cancer Res*, 81(11):2861–2873, 06 2021.

[75] D. Liu, Y. Yang, A. Yan, and Y. Yang. SPOCD1 accelerates ovarian cancer progression and inhibits cell apoptosis via the PI3K/AKT pathway. *Onco Targets Ther*, 13:351–359, 2020.

[76] K. Shinmura, H. Kato, Y. Kawanishi, H. Igarashi, Y. Inoue, K. Yoshimura, S. Nakamura, H. Fujita, K. Funai, M. Tanahashi, H. Niwa, H. Ogawa, and H. Sugimura. WDR62 overexpression is associated with a poor prognosis in patients with lung adenocarcinoma. *Mol Carcinog*, 56(8):1984–1991, Aug 2017.

[77] G. Cantelli, J. L. Orgaz, I. Rodriguez-Hernandez, P. Karagiannis, O. Maiques, X. Matias-Guiu, F. O. Nestle, R. M. Marti, S. N. Karagiannis, and V. Sanz-Moreno. TGF-β-Induced Transcription Sustains Amoeboid Melanoma Migration and Dissemination. *Curr Biol*, 25(22):2899–2914, Nov 2015.

[78] T. Kodama, M. Kodama, N. A. Jenkins, N. G. Copeland, H. J. Chen, and Z. Wei. Ring Finger Protein 125 Is an Anti-Proliferative Tumor Suppressor in Hepatocellular Carcinoma. *Cancers (Basel)*, 14(11), May 2022.

[79] Y. Lin, X. Zhou, W. Peng, J. Wu, X. Wu, Y. Chen, and Z. Cui. Expression and clinical implications of basic leucine zipper ATF-like transcription factor 2 in breast cancer. *BMC Cancer*, 21(1):1062, Sep 2021.

[80] Z. Zhang, W. Ji, J. Huang, Y. Zhang, Y. Zhou, J. Zhang, Y. Dong, T. Yuan, Q. Yang, X. Ding, L. Tang, H. Li, J. Yin, Y. Wang, T. Ji, J. Fei, B. Zhang, P. Chen, and H. Hu. Characterization of the tumour microenvironment phenotypes in malignant tissues and pleural effusion from advanced osteoblastic osteosarcoma patients. *Clin Transl Med*, 12(11):e1072, Nov 2022.

[81] M. Meng and Y. C. Wu. LMX1B Activated Circular RNA GFRA1 Modulates the Tumorigenic Properties and Immune Escape of Prostate Cancer. *J Immunol Res*, 2022:7375879, 2022.

[82] C. Ma, W. Kang, L. Yu, Z. Yang, and T. Ding. AUNIP Expression Is Correlated With Immune Infiltration and Is a Candidate Diagnostic and Prognostic Biomarker for Hepatocellular Carcinoma and Lung Adenocarcinoma. *Front Oncol*, 10:590006, 2020.

[83] S. Duss, H. Brinkhaus, A. Britschgi, E. Cabuy, D. M. Frey, D. J. Schaefer, and M. Bentires-Alj. Mesenchymal precursor cells maintain the differentiation and proliferation potentials of breast epithelial cells. *Breast Cancer Res*, 16(3):R60, Jun 2014.

[84] Y. Chen, L. Shen, B. Chen, X. Han, Y. Yu, X. Yuan, and L. Zhong. in lung cancer. *Ann Transl Med*, 9(10):843, May 2021.

[85] P. C. T. Ngoc, S. H. Tan, T. K. Tan, M. M. Chan, Z. Li, A. E. J. Yeoh, D. G. Tenen, and T. Sanda. Identification of novel lncRNAs regulated by the TAL1 complex in T-cell acute lymphoblastic leukemia. *Leukemia*, 32(10):2138–2151, 10 2018.

[86] M. R. Rutkowski, N. Svoronos, A. Perales-Puchalt, and J. R. Conejo-Garcia. The Tumor Macroenvironment: Cancer-Promoting Networks Beyond Tumor Beds. *Adv Cancer Res*, 128:235–262, 2015.

[87] L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, G. Getz, S. G. Rozen, M. R. Stratton, L. B. Alexandrov, E. N. Bergstrom, A. Boot, P. Boutros, K. Chan, K. R. Covington, A. Fujimoto, G. Getz, D. A. Gordenin, N. J. Haradhvala, M. N. Huang, S. M. A. Islam, M. Kazanov, J. Kim, L. J. Klimczak, N. Lopez-Bigas, M. Lawrence, I. Martincorena, J. R. McPherson, S. Morganella, V. Mustonen, H. Nakagawa, A. W. Tian Ng, P. Polak, S. Prokopec, S. A. Roberts, S. G. Rozen, R. Sabarinathan, N. Saini, T. Shibata, Y. Shiraishi, M. R. Stratton, B. T. Teh, I. Vázquez-García, D. A. Wheeler, Y. Wu, F. Yousif, and W. Yu. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 02 2020.

[88] Y. A. Kim, E. Hodzic, B. Amgalan, A. Saslafsky, D. Wojtowicz, and T. M. Przytycka. Mutational Signatures as Sensors of Environmental Exposures: Analysis of Smoking-Induced Lung Tissue Remodeling. *Biomolecules*, 12(10), Sep 2022.

[89] Q. Wang, J. Armenia, C. Zhang, A. V. Penson, E. Reznik, L. Zhang, T. Minet, A. Ochoa, B. E. Gross, C. A. Iacobuzio-Donahue, D. Betel, B. S. Taylor, J. Gao, and N. Schultz. Unifying cancer and normal RNA sequencing data from different sources. *Sci Data*, 5:180061, 04 2018.

[90] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

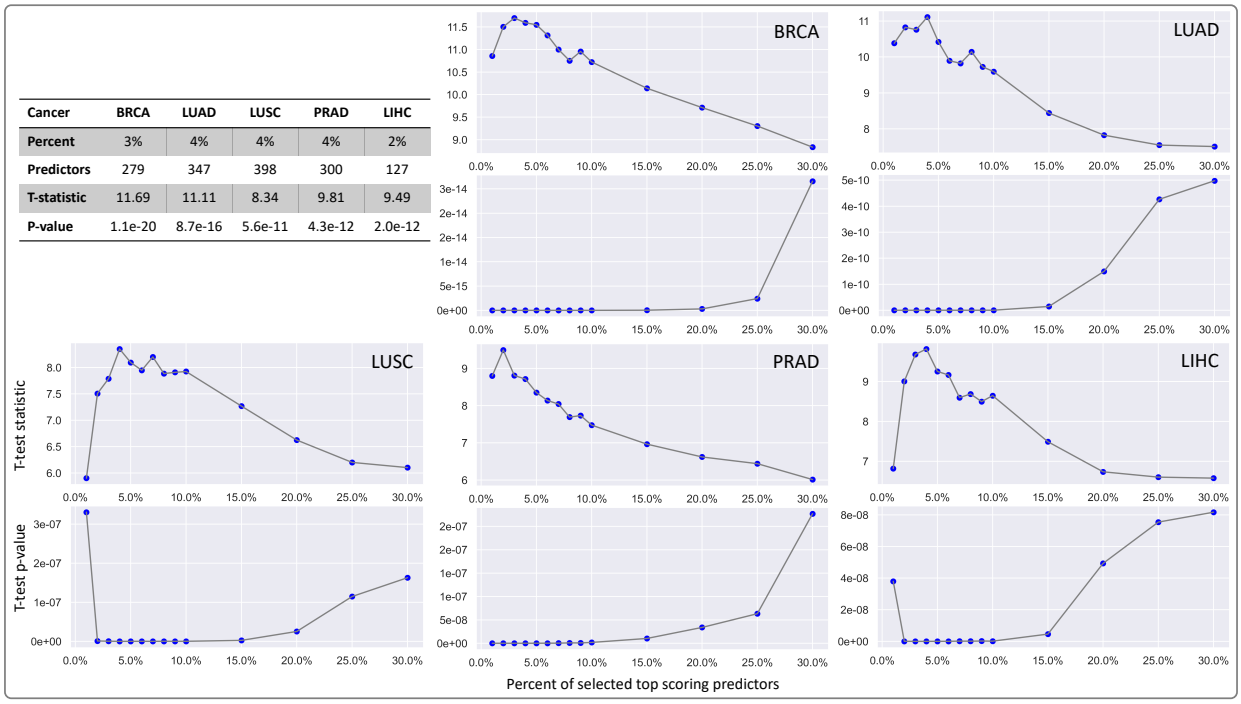
Supplemental Information

Supplementary Table

KEGG	LUAD (PC3)		PRAD (PC3)	
	p-value	q-value	p-value	q-value
Cell cycle	3.3E-15	4.3E-14		
DNA replication	6.6E-12	4.3E-11		
Pyrimidine metabolism	1.4E-08	6.3E-08		
Spliceosome	1.4E-05	4.8E-05	4.4E-07	8.8E-06
Oocyte meiosis	3.2E-05	8.5E-05		
Nucleotide excision repair	4.3E-05	9.2E-05	0.00176	0.0035
Purine metabolism	4.9E-05	9.2E-05		
Ether lipid metabolism			2.1E-05	0.00021
Mismatch repair	0.00014	0.00021		
Homologous recombination			0.00014	0.00021
Base excision repair	0.00029	0.00038		
Pathways in cancer	0.00131	0.00154	0.00254	0.00438
Vibrio cholerae infection			0.00046	0.00202
Mapk signaling pathway			0.00046	0.00202
Aminoacyl trna biosynthesis			0.00065	0.00202
Proximal tubule bicarbonate reclamation			0.00065	0.00202
Endocytosis			0.00070	0.00202
Adherens junction			0.00089	0.00223
Glycerophospholipid metabolism			0.00126	0.00280
Neuroactive ligand receptor interaction			0.00263	0.00438
Progesterone mediated oocyte maturation	0.00442	0.00479		
Basal cell carcinoma			0.00412	0.00625
Inositol phosphate metabolism			0.00464	0.00625
Calcium signaling pathway			0.00469	0.00625
Arginine and proline metabolism			0.00642	0.00802
VEGF signaling pathway			0.00780	0.00918

Table S1: **Functional enrichment for principal components.** Kegg pathways enriched for targets of differentially expressed principal components selected in the list of the high scoring predictors.

Supplementary Figures



Cancer	BRCA	LUAD	LUSC	PRAD	LIHC
Percent	3%	4%	4%	4%	2%
Predictors	279	347	398	300	127
T-statistic	11.69	11.11	8.34	9.81	9.49
P-value	1.1e-20	8.7e-16	5.6e-11	4.3e-12	2.0e-12

Figure S1: **Selection of the top predictors.** Paired-sample T-test was used for evaluating prediction power of TranNet. Depending on the number of the top RP scoring predictors, the prediction accuracy varies for each of the five cancers as visualized in the figures. The cut-off number describing the predictor set is decided based on the percents varying from 1% to 30%. For each of the selected sets of predictors, the leave-one-out test was performed over patients. As an illustrative example, for predicting the tumor expression $y_{i(\cdot)}$ of a patient i (105 patients for BRCA), the transition matrix $M_{(-i)}$ of size 279×9309 is inferred from the expressions of the top scoring 279 predictors in normal and the 9308 genes and 1 principal component in tumor (for 104 patients except for patient i). Then $M_{(-i)}$ is used as the transition mapping to predict the tumor tissue expression $y_{i(\cdot)} \in R^{9309}$ from the given normal expression $x_{i(\cdot)} \in R^{9309}$.

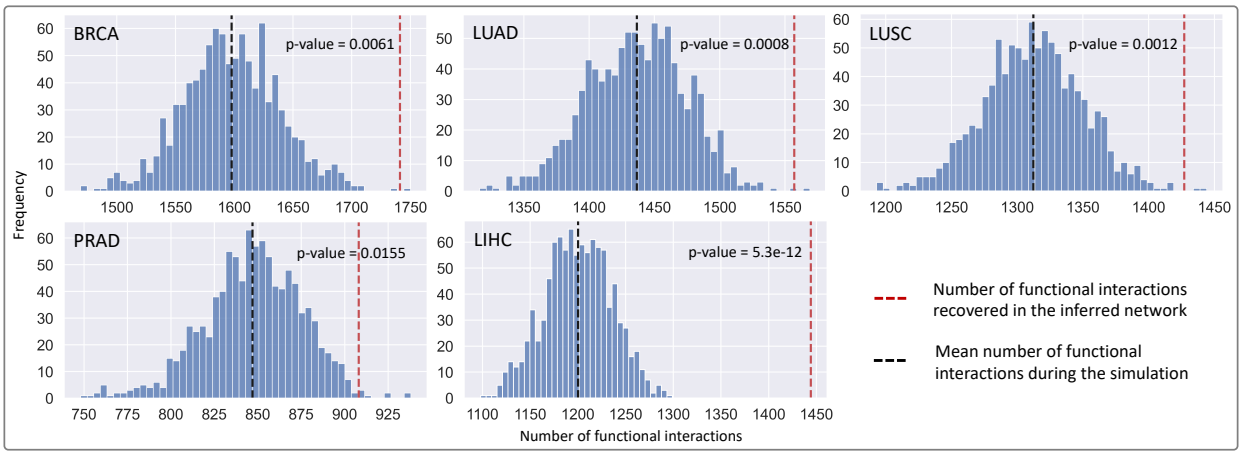
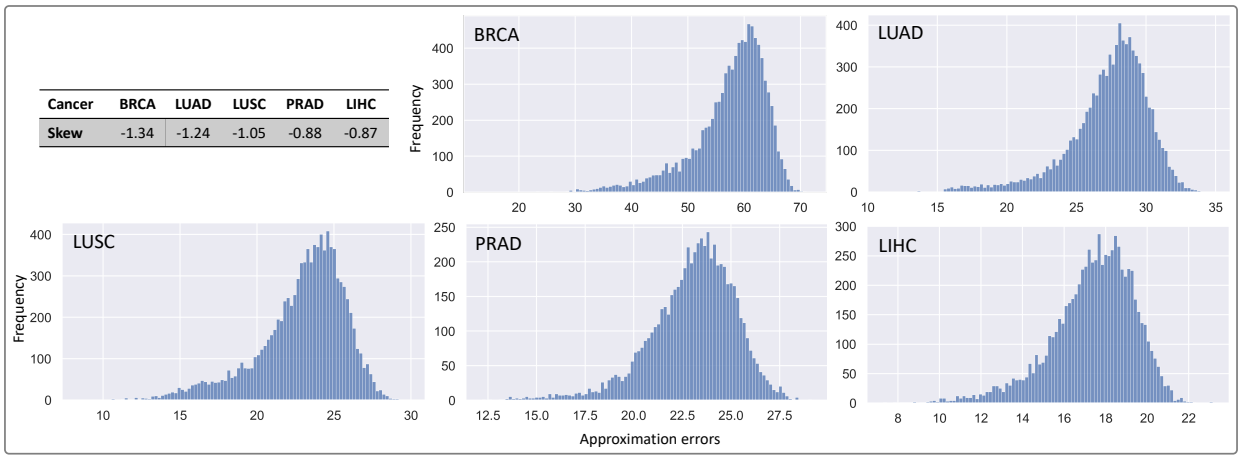


Figure S2: **Enrichment of the TranNet networks in functional interactions.** Horizontal axis represents the number of functional interactions recovered in randomized networks whereas vertical axis shows the corresponding frequencies. Doted-red line corresponds to the TranNet network. The p-value of the enrichment in functional interactions in the TranNet network relative to random and computed based on z-score test.



Cancer	BRCA	LUAD	LUSC	PRAD	LIHC
Skew	-1.34	-1.24	-1.05	-0.88	-0.87

Figure S3: **The distribution of the gene-wise prediction errors.** The horizontal axis denotes the error between the true expression and the predicted expression for a gene. In other words, it is the absolute summation over sample-wise errors for each gene. Vertical axis denotes the corresponding frequency.

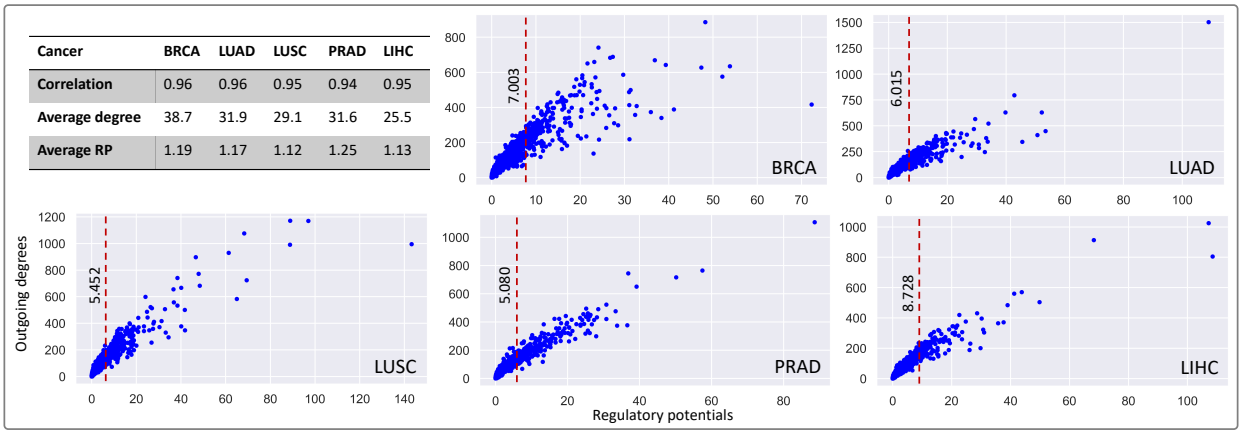


Figure S4: **Joint distribution of RP scores and outgoing degrees.** Scatter plots visualize the joint distribution of regulatory potentials (RP) (horizontal axis) and outgoing degrees (vertical axis) of the genes and principal components with nonzero regulatory potentials. Dotted-red line denotes the regulatory potential cut-off for selecting the high RP scoring predictors. Correlation: Pearson correlation between regulatory potentials and outgoing degrees of the genes. Average degree: average outgoing degree of the genes and principal components having nonzero regulatory potentials in the network. Average RP: average regulatory potential of the genes and principal components having nonzero regulatory potentials in the network.

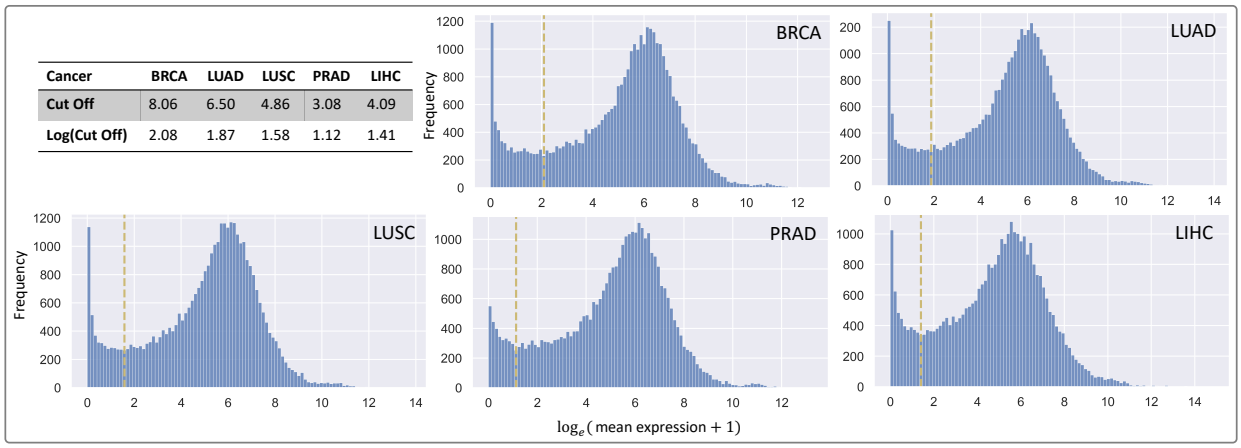


Figure S5: **Distribution of expression level.** The histograms visualize the distributions of log_e(mean expression + 1) for the five cancers. For each cancer, the expression level cut-off was described by the smallest frequency in the range between two peaks in the distribution.