Published online 25 September 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D1031–D1038***
doi: 10.1093/nar/gkx850

TissGDB: tissue-specific gene database in cancer

Pora Kim[1], Aekyung Park[1,2], Guangchun Han[1], Hua Sun[1], Peilin Jia[1] and Zhongming Zhao[1,3,*]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [2]College of Pharmacy and Research Institute of Life and Pharmaceutical Sciences, Sunchon National University, Suncheon 57922, Korea and [3]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Received August 05, 2017; Revised September 11, 2017; Editorial Decision September 12, 2017; Accepted September 15, 2017

ABSTRACT

Tissue-specific gene expression is critical in understanding biological processes, physiological conditions, and disease. The identification and appropriate use of tissue-specific genes (TissGenes) will provide important insights into disease mechanisms and organ-specific therapeutic targets. To better understand the tissue-specific features for each cancer type and to advance the discovery of clinically relevant genes or mutations, we built TissGDB (Tissue specific Gene DataBase in cancer) available at http://zhaobioinfo.org/TissGDB. We collected and curated 2461 tissue specific genes (TissGenes) across 22 tissue types that matched the 28 cancer types of The Cancer Genome Atlas (TCGA) from three representative tissue-specific gene expression resources: The Human Protein Atlas (HPA), Tissue-specific Gene Expression and Regulation (TiGER), and Genotype-Tissue Expression (GTEx). For these 2461 TissGenes, we performed gene expression, somatic mutation, and prognostic marker-based analyses across 28 cancer types using TCGA data. Our analyses identified hundreds of TissGenes, including genes that universally kept or lost tissue-specific gene expression, with other features: cancer type-specific isoform expression, fusion with oncogenes or tumor suppressor genes, and markers for protective or risk prognosis. TissGDB provides seven categories of annotations: TissGeneSummary, TissGene-Exp, TissGene-miRNA, TissGeneMut, TissGeneNet, TissGeneProg, TissGeneClin.

INTRODUCTION

Tissue-specific gene expression is dynamic and complex, but it is crucial in understanding biological processes, physiological conditions and disease. The identification and analysis of tissue-specific genes (TissGenes) in combination with other biomedical data will provide important insights into disease mechanisms and organ-specific therapeutic targets. Since disease and physiological condition are often associated with a specific tissue, the appropriate use of TissGene expression will substantially reduce false discoveries in biomedical research. With the exponential growth of biomedical data recently, such as cancer genomics, many studies have searched for alterations of cancer genes across multiple cancer types (e.g. pan-cancer studies). However, such studies and related clinical investigations are often performed without considering tissue-specificity in each cancer type, leading to both high false positive and negative discoveries. For example, pan-cancer studies with umbrella or basket trials using big cancer datasets were based on the assumption of existing common genetic alterations across multiple cancer types (1–4). Indeed, several driver mutations in one cancer type also exist in other cancer types like BRAF V600E in melanoma, colorectal cancer, thyroid cancer, non-small-cell lung cancer, and hairy cell leukemia with frequencies ∼ 50%, ∼10%, ∼35%, ∼4% and ∼100%, respectively (5). However, the BRAF inhibitor Vermurafenib was not successful in non-melanoma cancers that have the BRAF V600E mutation (6). This implies that the same oncogenic gene plays different roles (e.g. BRAF V600E is actionable in melanoma but likely has passenger role in other cancer types) in different cancer types, which are highly tissue-specific. These different clinical outcomes across multiple cancers might be inferred through a tissue-specificity gene resource, in which gene expression measure is an appropriate way to detect tissue-specificity of the cancer genes. Accordingly, curation and characterization of tissue-specific genes at the molecular level will be useful for better understanding oncogene's roles across tissue-based cancers, leading to the enhanced therapeutic strategies in precision oncology (7).

Thus far, many researchers have studied gene- and protein-expression and gene-gene networks of tissue-specific genes for a better understating of the molecular details of the different tissues in a healthy human body

[*]To whom correspondence should be addressed. Tel: +1 713 500 3631; Email: zhongming.zhao@uth.tmc.edu

(8,9). Other researchers have investigated tissue-specific mutations of cancer genes (2,3,10–12). However, until now, a systematic annotation of the alterations of TissGenes in cancer or other diseases has not been available. For example, the transmembrane protease, serine 2 gene (*TMPRSS2*), one of prostate cancer (PRAD) specific genes annotated in TissGDB, could demonstrate the role of a tissue-specific gene in the oncogenic process. The translocation between the prostate tissue-specific, androgen-inducible gene *TMPRSS2* and the proto-oncogene *ERG* leads to amplified proliferation of the prostate tissue cells. This tumorigenesis might occur through the androgen receptor dependent environment, which is related to prostate cancer initiation and progression by *TMPRSS2* and the key regulation of cell proliferation by *ERG*.

Here, we built TissGDB, the tissue-specific gene annotation database in cancer, aiming to provide a resource or reference for cancer and related disease studies in the context of tissue specificity. This paper introduces TissGDB (Tissue specific Gene DataBase in cancer), the web interface, and its applications. Our database includes features of all human tissue-specific genes based on large cancer data sets through systematic bioinformatics analyses. Therefore, it will be a unique resource for broad biomedical research communities.

DATABASE OVERVIEW

We collected and curated 2461 TissGenes across 22 tissue types, which matched the 28 cancer types from The Cancer Genome Atlas (TCGA) project (13), from three representative tissue-specific gene expression resources: The Human Protein Atlas (HPA) (8), Tissue-specific Gene Expression and Regulation (TiGER) (14), and Genotype-Tissue Expression (GTEx) (13). (Figure 1A). For these 2461 TissGenes, we performed gene expression, somatic mutation, and prognostic marker-based analyses across 28 cancer types using TCGA data through seven categories of annotations: TissGeneSummary, TissGeneExp, TissGene-miRNA, TissGeneMut, TissGeneNet, TissGeneProg, and TissGeneClin (Figures 1B and 2). The main features of the TissGDB annotations are summarized as follows. (i) The TissGeneExp information category shows multiple bar plots for gene and isoform expression using TCGA and GTEx data with colors distinguishing different cancer types and different gene isoforms. From these plots, we identified 294 and 209 TissGenes that could universally keep (TissGenesKTS) or lose (TissGenesLTS) tissue-specific gene expression across the cancer types, respectively. (ii) The TissGene-microRNA information category provides the significantly anti-correlated microRNAs (miRNAs) for each TissGene among the 28 cancer types using Spearman's Rank Correlation method. (iii) In the TissGeneMut information category, we present a lollipop plot of nonsynonymous single-nucleotide variants (nsSNVs) on the amino acid sequence with different colored circles for cancer types. By examining the extent and patterns of copy number variation (CNV) in TissGenesKTS, we identified 201 TissGenes whose ratio, which is defined as the number of samples with copy number gained versus the number of samples with copy number lost, is at least two across multiple cancer types. By investigating TCGA fusion genes involving TissGenes, we found 447 fusion genes among the 350 TissGenes and 341 oncogenes or tumor suppressor genes (TSGenes). (iv) The TissGeneProg information category summarizes the results of survival analyses with Kaplan-Meier and Forest plots based on the log-rank test and Cox regression analysis using overall survival and relapse free survival outcomes. From these survival analyses, we found 152 protective and 56 risk TissGenes. (v) Through the TissGeneClin information category, we identified 705 FDA-approved drugs that target 144 TissGenes (5.85%). We also found 1844 TissGenes (74.9%) that are reported to be associated with 6979 different IDs of diseases based on DisGeNet. The details of the data and analysis processes are described in a later section.

Table 1 summarizes the statistics of 2461 TissGenes and among them, 546 highly confident TissGenes (we named them as cTissGenes—those genes that were identified in all three tissue-specific gene expression resources). All entries and annotation data are available for browsing and downloading on the TissGDB web site with unique and efficient visualization (http://zhaobioinfo.org/TissGDB).

DATA INTEGRATION AND ANNOTATIONS

Creation of tissue-specific gene list

To create TissGenes, we used three representative resources: HPA, TiGER and GTEx. From the HPA and TiGER resources, we retrieved 2543 and 4899 genes, respectively (Figure 1A). These two resources applied the fold change of the expression values of each human gene as their criterion for selecting TissGenes across multiple tissues. They selected tissue-specific genes that have at least 5-fold higher FPKM level in one tissue compared to all other tissues. Here, we adopted their annotation of tissue-specific gene list. After matching each tissue type with the cancer types of TCGA data (Supplementary Table S1), 2050 and 3090 TissGenes were selected from HPA and TiGER resources, respectively. To identify TissGenes from GTEx data, we generated the gene list by investigating *z*-scores ranging from 1.0 to 4.0 based upon the expression levels of the genes. Here, the *z*-score equal to *N* represents that more than *N* standard deviations greater than the mean expression in all tissues. For the appropriate number of genes, we set the threshold of the *z*-score value as 3.0 in the GTEx data. This criterion resulted in 11 223 genes across 32 tissues of the GTEx data. After matching each tissue type with the cancer types of TCGA data, 6039 genes were selected as TissGenes from GTEx data. In summary, across 22 tissue types matching the 28 cancer types, 2050, 3090 and 6039 TissGenes were selected from HPA, TiGER and GTEx, respectively. Through a union of these genes, we had 8172 unique human genes. Among them, 2461 genes overlapped by at least two tissue-specific gene expression resources, we selected them as TissGenes to ensure the reliability of tissue specificity of the genes (Supplementary Table S2).

Manual curation of PubMed articles

For the 546 highly confident TissGenes (i.e., cTissGenes), a literature query of PubMed was performed in June 2017

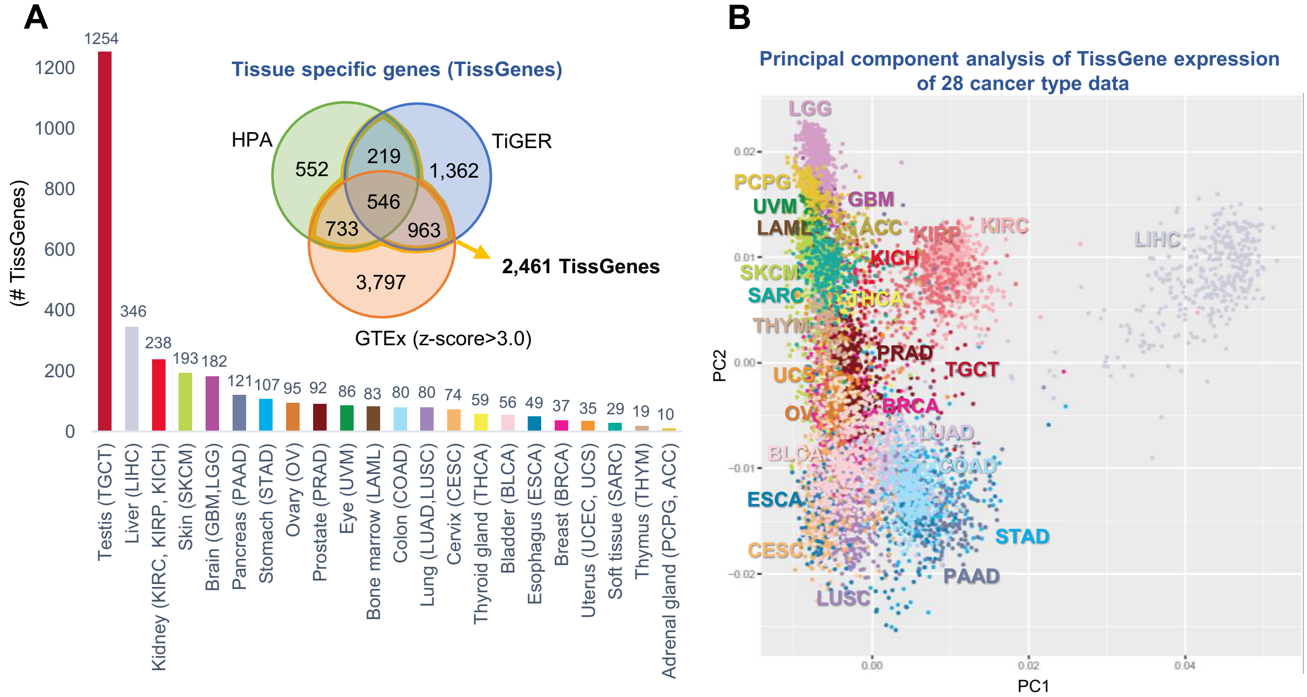
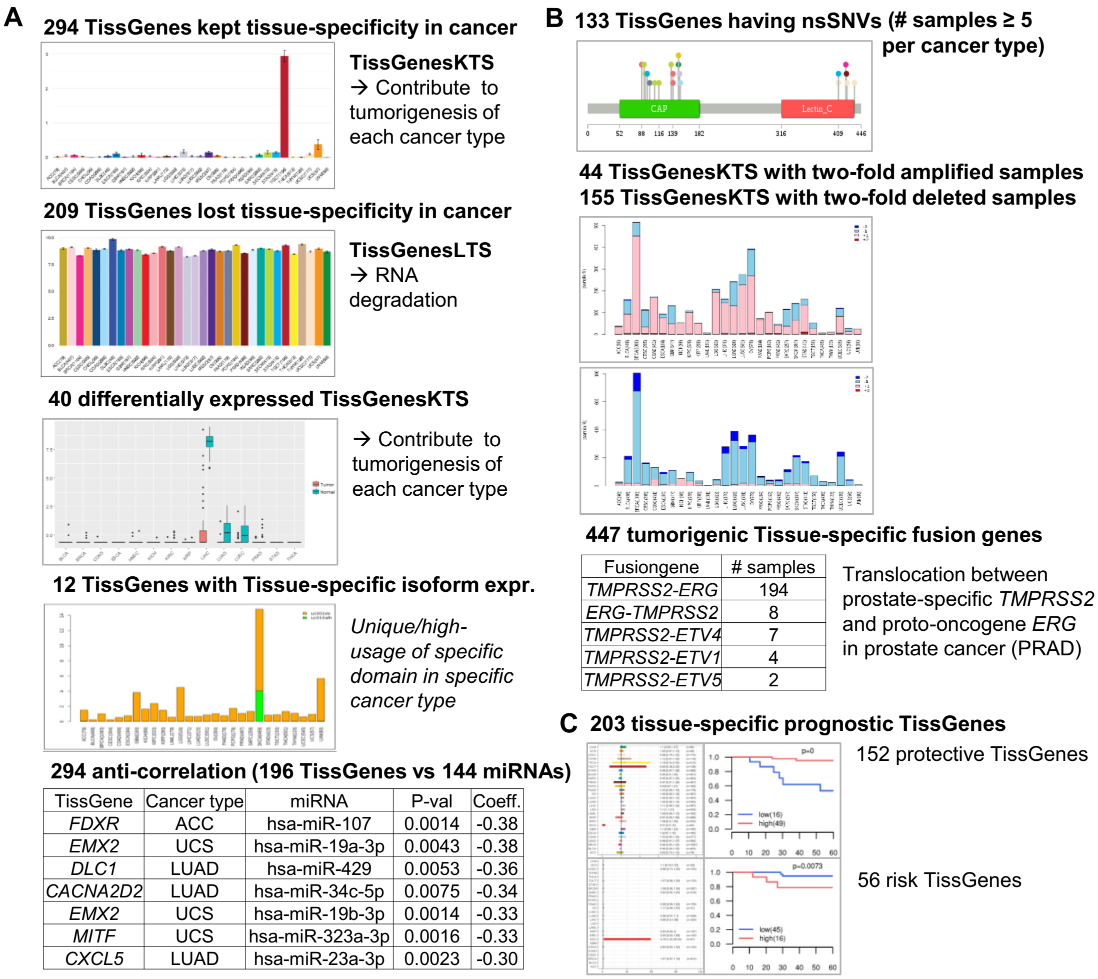


Figure 1. Overview of TissGDB. (**A**) Venn-diagram of TissGenes among the three representative tissue-specific gene expression resources. We selected 2461 TissGenes overlapped in at least two out of three resources. Among them, 546 genes were overlapped in all three tissue-specific gene expression resources (confident TissGenes, cTissGenes). (**B**) The overall expression distribution of TissGenes among the 28 cancer types using the principal component analysis (PCA) method.

Table 1. Annotation entry statistics for TissGenes and cTissGenes

Data type	# entries	# TissGenes[a]	# cTissGenes[b]
		Total 2461 (%)	Total 546 (%)
Tissue specific genes	# genes		
HPA[c]	2050	1498 (60.9%)	546 (100.0%)
TiGER[d]	3090	1728 (70.2%)	546 (100.0%)
GTEx[e]	6039	2242 (91.1%)	546 (100.0%)
Cancer genes			
CCG[f]	4050	443 (18.0%)	87 (15.9 %)
Expression	# genes		
TCGA[g]	20 530	2444 (99.3%)	546 (100.0%)
GTEx	56 318	2461 (100.0%)	546 (100.0%)
Mutation	# genes		
TCGA	39 571	2461 (100.0%)	546 (100.0%)
Copy number variation	# genes		
TCGA	24 776	2461 (100.0%)	546 (100.0%)
Fusion gene	# genes		
ChimerDB3.0[h]	10 713	1393 (56.6%)	293 (53.7%)
TCGA data Fusion Portal[i]	7765	718 (29.2%)	155 (28.4%)
Survival analysis	# clin.info		
TCGA	11 896	2461 (100.0%)	546 (100.0%)
Molecule	# molecules		
DrugBank[j]	8206 drugs	218 (8.9%)	61 (11.2%)
UniProt[k]	2374 proteins	2446 (99.4%)	545 (99.8%)
Phenotype	# phenotype		
DisGeNet[l]	15 094 disease ID	1844 (74.9%)	434 (79.5%)

[a] Tissue specific genes (TissGenes). [b] Confident TissGenes (cTissGenes). [c] The Human Protein Atlas. [d] Tissue-specific gene expression and regulation (TiGER). [e] Genotype-Tissue Expression. [f] Catalogue of cancer genes. [g] The Cancer Genome Atlas. [h] ChimerDB3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. [i] TCGA fusion gene data portal. [j] Related drug with the TissGenes from DrugBank database. [k] The Universal Protein Resource (UniProt). [l] Gene-level disease annotation from DisGeNet database.



TissGene	Cancer type	miRNA	P-val	Coeff.
FDXR	ACC	hsa-miR-107	0.0014	-0.38
EMX2	UCS	hsa-miR-19a-3p	0.0043	-0.38
DLC1	LUAD	hsa-miR-429	0.0053	-0.36
CACNA2D2	LUAD	hsa-miR-34c-5p	0.0075	-0.34
EMX2	UCS	hsa-miR-19b-3p	0.0014	-0.33
MITF	UCS	hsa-miR-323a-3p	0.0016	-0.33
CXCL5	LUAD	hsa-miR-23a-3p	0.0023	-0.30

Fusiongene	# samples
TMPRSS2-ERG	194
ERG-TMPRSS2	8
TMPRSS2-ETV4	7
TMPRSS2-ETV1	4
TMPRSS2-ETV5	2

Figure 2. Overview of TissGene annotations in TissGDB. (**A**) Gene expression-based annotations of TissGenes. (**B**) Somatic mutation-based annotations of TissGenes. (**C**) Prognostic-based annotations of TissGenes.

using the search expression that applied to each TissGene. Using *TMPRSS2* as an example, it is '((*TMPRSS2* [Title/Abstract]) AND tissue [Title/Abstract]) AND specific [Title/Abstract])'. After manual review of the abstracts of over 1000 articles, we found 189 genes (∼34.6%) had literature evidence (196 articles) that support the tissue- or cancer type-specificity of these TissGenes. Building on these search results, we created a classification system for the genes in the database to roughly measure reliability. Class A consists of genes with literature evidence and is part of the cTissGenes. Class B consists of only cTissGenes without additional evidence. The remaining genes belong to Class C. There were 189, 358 and 1914 genes in classes A, B and C, respectively.

Expression data preparation

Gene expression data (HiSeqV2) were downloaded from TCGA (December 2016). We used pan-cancer normalized values log2(normalized read count+1) from IlluminaHiSeq_RNASeqV2. Isoform expression data were downloaded from the Broad Institute GDAC FireBrowse portal (http://firebrowse.org/). GTEx RPKM gene expression V6 was downloaded from the GTEx portal (http://gtexportal.org).

Calculation of anti-correlation between mRNA and miRNA expression

We obtained the conserved human miRNA-target gene interaction information from TargetScan (release 7.1, June 2016) (15). miRNA expression data were obtained from TCGA (December 2016). We removed miRNAs with NA values in over 50% of the samples and selected the miRNAs with an expression value greater than 10 reads per million (RPM) in >10% of the samples in each cancer type. A gene-miRNA correlation coefficient was calculated using the Spearman's Rank Correlation method. We defined a miRNA as significantly anti-correlated if it had *P*-value <0.05 and coefficient 0.3.

SNV and CNV data

Somatic gene-level non-silent mutation data including nonsense, missense, frame-shift insertions and deletions (indels), splice site mutations, stop codon read-throughs, changes of start codon, and inframe indels were downloaded from TCGA (December 2016). Thresholded gene-level copy number variation (CNV) data estimated using the GISTIC2 method were also downloaded from TCGA (December 2016).

Fusion gene information

30 001 and 7992 fusion transcript candidates and their related information were downloaded from the ChimerDB 3.0 (http://ercsb.ewha.ac.kr/fusiongene, December 2016) (16) and TCGA fusion Data Portal (http://54.84.12.177/PanCanFusV2/, December 2014) (17), respectively. For the ChimerDB fusion transcripts, we found that 13 729 were derived from TCGA samples. By union of the 13 729 and 7992 fusion genes, we obtained 21 724 unique fusion genes. We overlapped each of the 13 729 fusion genes from ChimerDB and the 7992 fusion genes from TCGA fusion Data Portal with 2461 TissGenes. This resulted in 3622 and 1151 fusion genes including 1393 and 718 TissGenes from ChimerDB3.0 and TCGA fusion Data Portal, respectively.

Co-expressed protein interaction network (CePIN)

The protein interaction network (PIN) reported in our previous study (18) included 113 473 unique protein-protein interaction pairs connecting 13 579 protein-coding genes. To build the CePIN, we calculated the Pearson's Correlation Coefficient (PCC) for each gene–gene pair. Co-expressed network figures were drawn using the igraph package in R (19). For each gene, the top 20 neighbor genes with the highest PCC values were kept in the network to reflect the genetic signals.

Survival analysis and data preparation

Based on gene expression and survival outcomes [overall survival (OS) and relapse free survival (RFS)], we identified prognostic TissGenes in each cancer type. To achieve this, we used log-rank test and Cox proportional hazards regression. In the log-rank test, patients were divided into two groups as high and low expression groups. We used three cutoff values (25, 50 or 75 percentiles of gene expression) and the cutoff value showing the most significant statistical *P*-value was used for the result of each gene. We present the results of the log-rank test as Kaplan–Meier survival curves. The hazard ratios (HRs) obtained from the Cox proportional hazards regression are presented with 95% confidence intervals as a forest plot for each cancer type.

Drug and disease information

Drug-target interactions (DTIs) were extracted from DrugBank (April 2017) (20) with the duplicated DTI pairs excluded. All drugs were grouped using Anatomical Therapeutic Chemical (ATC) classification system codes. Disease-gene information was extracted from a database of gene-disease associations (DisGeNet, June 2016) (21).

Database architecture

The TissGDB system is based on a three-tier architecture: client, server, and database. It includes a user-friendly web interface, Perl's DBI module and MySQL database. This database was developed on MySQL 3.23 with the MyISAM storage engine.

WEB INTERFACE AND ANALYSIS RESULTS

Gene expression category (TissGeneExp)

This category presents the landscape of gene expression for the TissGenes across 22 normal tissue-types and 28 cancer-types based on GTEx and TCGA data. The expression is displayed with both gene- and isoform-levels. First, we investigated the overall expression distribution of TissGenes among the 28 cancer types using the principal component analysis (PCA) method. As shown in Figure 1B, the cancer types that are of the same tissue type were clustered together like brain (LGG and GBM), adrenal gland (PCPG and ACC), kidney (KICH, KIRP and KIRC) and lung (LUAD and LUSC). We also observed that samples of liver cancer (LIHC) showed a distinct pattern separate from the samples of other cancer types. Figure 2 shows the overview of TissGene annotation results. From the gene expression bar plot across 28 cancer-types for each TissGene, we identified 294 TissGenes that may universally keep tissue-specific gene expression across the cancer types (TissGenesKTS, Figure 2A). We also identified 209 TissGenes that may universally lose tissue-specific gene expression across 28 cancer types (TissGenesLTS). The number of TissGenesKTS and TissGeneLTS across 28 cancer types are presented in Figure 3A. To infer the active pathways of these genes, we performed gene set enrichment tests for the 294 TissGenesKTS for each cancer type and 209 TissGenesLTS in pan-cancer using online tool WebGestalt (WEB-based Gene SeT AnaLysis Toolkit) (22). We applied adjusted *P*-value (i.e. *q*-value) <0.05, hypergeometric test followed by multiple test correction using Benjamini–Hochberg's method, as implemented in WebGestalt). As shown in Figure 3B, TissGenesKTS were enriched in the pathways relevant to their tumorigenesis in each cancer type, while TissGenesLTS were enriched in RNA degradation related pathways (Figure 3C).

Next, we manually curated and selected TissGenes showing cancer type specific isoform expression. Here, we found nine and three TissGenes with high expression of specific isoforms and unique expression of specific isoforms in their assigned cancer type, respectively. These nine TissGenes were *CFHR1*, *DDX4*, *G6PC2*, *GFAP*, *HAPLN2*, *IGLL1*, *IQCF1*, *IQCF2*, *SLC13A1*, and *SLC22A2* and the three TissGenes were *DCT*, *MASP2* and *TG*. For example, dopachrome tautomerase, encoded by gene *DCT*, is involved in the formation of the photo-protective skin pigment eumelanin (23). It has a specific isoform (UCSC known gene id: uc010afh) that contains a tyrosinase domain in skin cutaneous melanoma (SKCM). Tyrosinase is an oxidase and rate-limiting enzyme controlling production of melanin. This result provided cancer type-specific isoforms, which are clinically important.

We next examined differentially expressed genes (DEGs) in each cancer type. A DEG is defined by $|\log_2$(fold

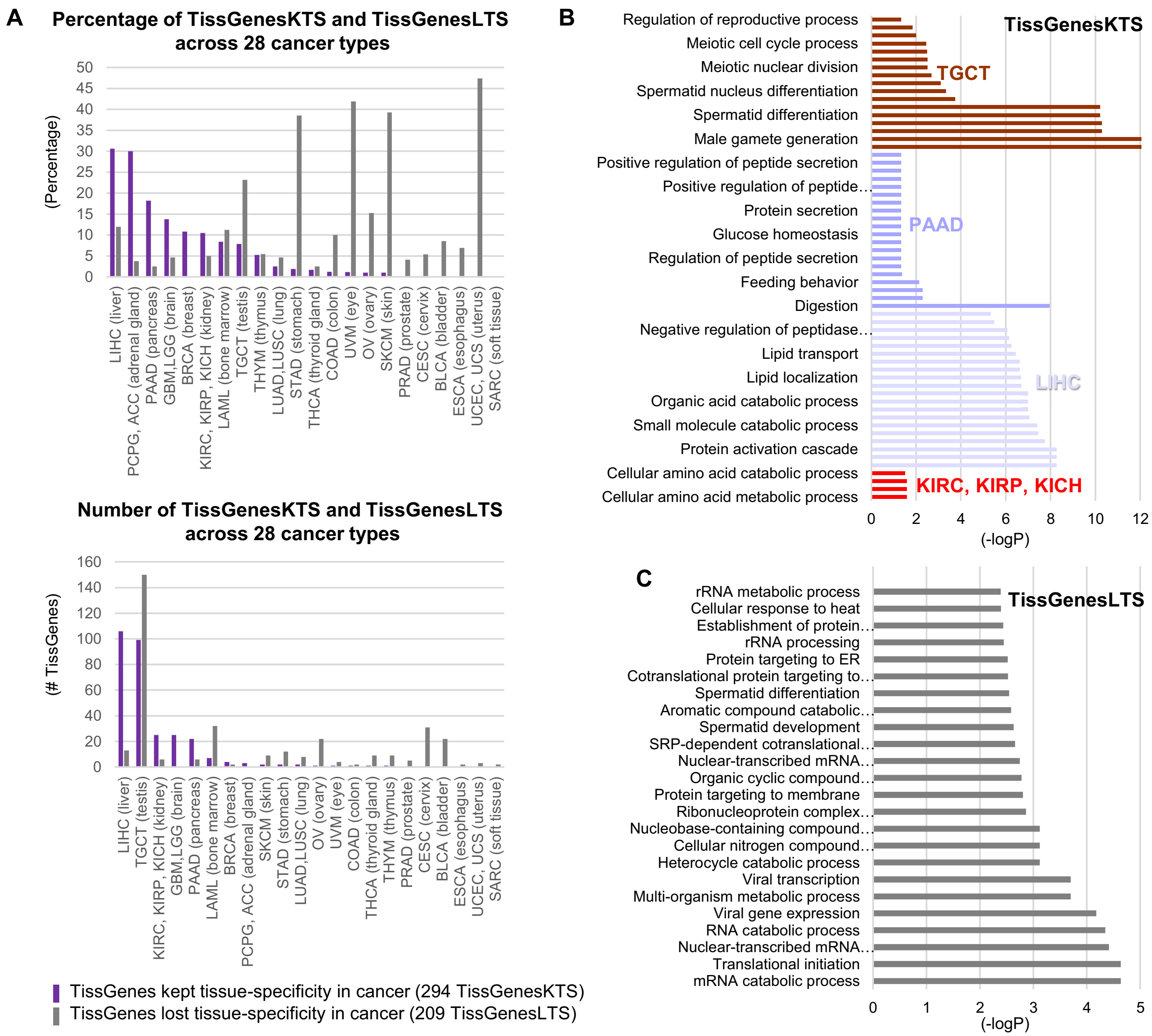


Figure 3. TissGenes that keep or lose tissue-specificity in cancer. From the gene expression patterns of all TissGenes across 28 cancer types, we identified 294 TissGenesKTS and 209 TissGenesLTS. (**A**) The percentage and number of TissGenesKTS and TissGenesLTS across 28 cancer types. (**B**) Enriched biological processes of TissGenesKTS per cancer type. (**C**) Enriched biological processes of TissGenesLTS.

change)| > 1 and *q*-value < 0.05 (*t*-test followed by multiple test correction using Benjamini–Hochberg's method). We found 113 TissGenes, which were differentially expressed in their specific cancer types. Interestingly, among the 113 TissGenes, 40 (35%) overlapped with the 294 TissGenesKTS. Among these 40 genes, only one gene, *POTEG*, was up-regulated in PRAD. The other 39 genes were down-regulated; they were significantly enriched in 'organic anion transport', 'organonitrogen compound catabolic process', and 'sodium-independent organic anion transport' (WebGestalt, *q*-value < 0.05, hypergeometric test followed by multiple test correction using Benjamini-Hochberg's method, Supplementary Table S3). This strong bias toward down-regulation of TissGenes might be related to some more active functions of cancer genes during the progression of cancer.

miRNA category (TissGene-miRNA)

In this category, we provide the significantly anti-correlated miRNAs for each TissGene based on the miRNAs-target human genes pair information predicted by TargetScan. We calculated Spearman's Rank Correlation coefficient and found that 196 TissGenes had significant anti-correlation with 144 miRNAs with a correlation coefficient value less than –0.3 (Supplementary Table S4). Among these, only six TissGene-miRNA correlations were from the assigned cancer type for each gene. The six gene-miRNA pairs with the specific cancer types are *CACNA2D2*-miR-34c-

5p (LUAD), *CXCL5*-miR-23a-3p (LUAD), *DLC1*-miR-429 (LUAD), *EMX2*-miR-19a/b-3p (UCS), *FDXR*-miR-107 (ACC) and *MITF*-miR-323a-3p (UCS). Two correlations were supported by previous studies: *CACNA2D2* upregulation in has-miR-34a/b/c-deficient adenomas (24) and miR-429 promoting the proliferation of lung adenocarcinoma via targeting *DLC-1* (25).

Mutation category (TissGeneMut)

This category provides somatic mutation annotations of each TissGene including non-synonymous single-nucleotide variants, copy number variations, and gene fusions in the subcategories such as TissGeneSNV, TissGeneCNV, and TissGeneFusion. From the TissGeneSNV part, one can retrieve information about how frequently the TissGenes were mutated across 28 cancer types. Specifically, we used a lollipop plot of nsSNVs on the amino acid sequence with different colored circles for each cancer type to illustrate the landscape of nsSNVs across multiple cancer types. Overall, there were 133 TissGenes mutated in at least five samples per cancer type. The TissGeneCNV part shows the number of samples that have CNV across 28 cancer types. By searching the fold change of the number of samples between copy number gained and lost ($|\log_2(\mathrm{FC})| > 1$), we found 44 TissGenes that had more than a 2-fold difference between the number of samples with copy number gain versus loss and 155 TissGenes having the opposite measure. This is consistent with a biased DEG result toward more down-regulation of TissGenesKTS. Interestingly, the 44 genes were enriched in various metabolic or biosynthetic processes (WebGestalt, q-value < 0.05, hypergeometric test followed by multiple test correction using Benjamini–Hochberg's method, Supplementary Table S5). In contrast, the 155 genes that had more than a 2-fold difference between the numbers of samples with copy number loss versus gain were enriched in anion transport and germ cell development (Supplementary Table S6). The TissGeneFusion part shows the fusion genes involving TissGenes. Through overlapping 2461 TissGenes with ∼22 000 fusion genes from ChimerDB3.0 and TCGA fusion Data Portal, we identified 1393 TissGenes involved in 2662 fusion genes. Interestingly, 146 and 255 TissGenes were fused with 115 oncogenes and 255 tumor suppressor genes, which resulted in 168 and 327 fusion genes, respectively (Supplementary Table S7). For example, the chromosomal rearrangement between prostate tissue-specific, androgen-inducible gene *TMPRSS2* and proto-oncogene *ERG* occurring in 30–50% of prostate cancer patients. This is a typical example of an alteration of a tissue-specific gene for tumorigenesis. These candidates might be helpful to understand tissue specific factors. The tissue-specific gene fusion events in pan-cancer will provide an important reference for investigators in broad cancer research.

Prognostic information category (TissGeneProg)

To identify prognostic TissGenes, we performed a log-rank test and Cox regression analysis based on survival outcome information (OS and RFS) in 28 cancer types. In Cox regression analysis with 2461 TissGenes, the expression of 1956 and 1783 TissGenes showed significant associations with overall survival and relapse free survival outcomes, respectively (P-value < 0.05). According to the HRS from Cox regression analysis, we defined a gene as a 'protective' TissGene if the increased expression of the gene was statistically associated with prolonged survival (HR < 1.0), or as a 'risk' TissGene if the increased expression was associated with poor survival (HR > 1.0). Based on the ranks of the HR from Cox regression with OS or RFS, we first obtained two sets of the top 1000 protective and risk TissGenes without consideration of tissue specificity: one set of 1000 protective and risk TissGenes from Cox regression with OS and the other set from that with RFS. From each set of the top 1000 TissGenes, we collected only the TissGenes that matched to their assigned cancer types, i.e., prognostic TissGenes. Subsequently, we obtained the final 152 protective and 57 risk prognostic TissGenes by combining two lists of prognostic TissGenes from the results with OS and RFS (Supplementary Table S8). Among these, kidney cancers (KICH, KIRC, and KIRP) showed the largest number of prognostic TissGenes, including 52 protective and 18 Risk TissGenes. A previous study also revealed large prognostic markers in KIRC (26). We performed enrichment analysis with the kidney-specific prognostic TissGenes in terms of biological function (WebGestalt, q-value < 0.05, hypergeometric test followed by multiple test correction using Benjamini–Hochberg's method). The 52 protective TissGenes were enriched with genes involved in 'distal tubule development' pathway (q-value of 1.24×10^{-8}) and 'urogenital system development' pathway ($q = 4.68 \times 10^{-6}$). However, no significant enriched term was obtained with the 18 risk TissGenes. From the results, we could infer that the TissGenes involved in kidney development were related to the prognosis in kidney cancers and high expression of the genes might lead to better prognosis in kidney cancer patients.

Pharmacological information and disease information categories (TissGeneClin)

This category includes two subcategories: TissGeneDrug and TissGeneDisease. TissGeneDrug provides TissGene related pharmacological information from DrugBank. Overall, TissGDB includes 8206 drugs targeting 218 TissGenes and 705 approved drugs targeting 144 TissGenes. Among the 218 TissGenes targeted by drugs, only 71 genes (32.6%) were cancer genes and the others (67.4%) were non-cancer genes. The TissGeneDisease part shows the related disease information for each gene from a database of gene-disease associations (DisGeNet). 1844 genes out of all 2461 TissGenes were reported to be associated with 6979 different types of diseases. Overall, only 443 genes (18.0%) out of all the TissGenes overlapped with the known cancer genes from the Catalogue of Cancer Genes (27).

DISCUSSION AND FUTURE DIRECTION

TissGDB is the first database that systematically annotates tissue-specific genes in pan-cancer and it can be extended to other diseases when data becomes available. To serve broad biomedical research communities, we will continuously update and curate TissGenes routinely by checking new tissue-specific gene or protein expression data. One near-term update is to obtain the related data from the International Cancer Genome Consortium (ICGC). The ICGC has more cancer types or subtypes than TCGA. As shown in the 447 fusion genes between 350 TissGenes and 341 oncogenes or TSGenes, their genomic relation, interaction, linked aberration between cancer genes and TissGenes, and products (chimeric proteins) are important for studying their potential roles in tumorigenesis and as possible molecular targets. We will investigate and extend the methods to find the significant relationship between cancer genes and TissGene in the near future. The easy-to-use website provides multiple annotation results to researchers and facilitates comprehensive studies of TissGenes. Thus, TissGDB will be useful for many investigators in pathology, cancer genomics and precision medicine, drug and therapeutic research, among others.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the members of Bioinformatics and Systems Medicine Laboratory (BSML) for valuable discussion. We thank Dr Timothy O'Brien and Mr. Hanyang Li for improving the English of the manuscript draft and the content of the website. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

FUNDING

National Institutes of Health [R01LM012806, R21CA196 508] (in part). Funding for open access charge: Chair Professorship for Precision Health Funds (to Dr. Zhao) from the University of Texas Health Science Center at Houston.

Conflict of interest statement. None declared.

REFERENCES

1. Schneider,G., Schmidt-Supprian,M., Rad,R. and Saur,D. (2017) Tissue-specific tumorigenesis: context matters. *Nat. Rev. Cancer*, **17**, 239–253.
2. Schaefer,M.H. and Serrano,L. (2016) Cell type-specific properties and environment shape tissue specificity of cancer genes. *Sci. Rep.*, **6**, 20707.
3. Ciriello,G., Miller,M.L., Aksoy,B.A., Senbabaoglu,Y., Schultz,N. and Sander,C. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127–1133.
4. Zehir,A., Benayed,R., Shah,R.H., Syed,A., Middha,S., Kim,H.R., Srinivasan,P., Gao,J., Chakravarty,D., Devlin,S.M. *et al.* (2017) Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.*, **23**, 703–713.
5. Hall,R.D. and Kudchadkar,R.R. (2014) BRAF mutations: signaling, epidemiology, and clinical experience in multiple malignancies. *Cancer Control*, **21**, 221–230.
6. Hyman,D.M., Puzanov,I., Subbiah,V., Faris,J.E., Chau,I., Blay,J.Y., Wolf,J., Raje,N.S., Diamond,E.L., Hollebecque,A. *et al.* (2015) Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations. *N. Engl. J. Med.*, **373**, 726–736.
7. Cheng,F., Zhao,J., Fooksa,M. and Zhao,Z. (2016) A network-based drug repositioning infrastructure for precision cancer medicine through targeting significantly mutated genes in the human cancer genomes. *J. Am. Med. Inform. Assoc.*, **23**, 681–691.
8. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
9. Kosti,I., Jain,N., Aran,D., Butte,A.J. and Sirota,M. (2016) Cross-tissue analysis of gene and protein expression in normal and cancer tissues. *Sci. Rep.*, **6**, 24799.
10. Bleeker,F.E., Lamba,S., Rodolfo,M., Scarpa,A., Leenstra,S., Vandertop,W.P. and Bardelli,A. (2009) Mutational profiling of cancer candidate genes in glioblastoma, melanoma and pancreatic carcinoma reveals a snapshot of their genomic landscapes. *Hum. Mutat.*, **30**, E451–E459.
11. Zhao,J., Cheng,F. and Zhao,Z. (2017) Tissue-specific signaling networks rewired by major somatic mutations in human cancer revealed by proteome-wide discovery. *Cancer Res.*, **77**, 2810–2821.
12. Shen,Q., Cheng,F., Song,H., Lu,W., Zhao,J., An,X., Liu,M., Chen,G., Zhao,Z. and Zhang,J. (2017) Proteome-scale investigation of protein allosteric regulation perturbed by somatic mutations in 7,000 Cancer Genomes. *Am. J. Hum. Genet.*, **100**, 5–20.
13. Carithers,L.J. and Moore,H.M. (2015) The Genotype-Tissue Expression (GTEx) project. *Biopreserv. Biobank*, **13**, 307–308.
14. Liu,X., Yu,X., Zack,D.J., Zhu,H. and Qian,J. (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.
15. Agarwal,V., Bell,G.W., Nam,J.W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, doi:10.7554/eLife.05005.
16. Lee,M., Lee,K., Yu,N., Jang,I., Choi,I., Kim,P., Jang,Y.E., Kim,B., Kim,S., Lee,B. *et al.* (2017) ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. *Nucleic Acids Res.*, **45**, D784–D789.
17. Yoshihara,K., Wang,Q., Torres-Garcia,W., Zheng,S., Vegesna,R., Kim,H. and Verhaak,R.G. (2015) The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*, **34**, 4845–4854.
18. Kim,P., Cheng,F., Zhao,J. and Zhao,Z. (2016) ccmGDB: a database for cancer cell metabolism genes. *Nucleic Acids Res.*, **44**, D959–D968.
19. Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research, Inter. *Comp. Syst.*, **1695**, 1–9.
20. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
21. Pinero,J., Bravo,A., Queralt-Rosinach,N., Gutierrez-Sacristan,A., Deu-Pons,J., Centeno,E., Garcia-Garcia,J., Sanz,F. and Furlong,L.I. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
22. Wang,J., Vasaikar,S., Shi,Z., Greer,M. and Zhang,B. (2017) WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.*, doi:10.1093/nar/gkx356.
23. Ainger,S.A., Yong,X.L., Wong,S.S., Skalamera,D., Gabrielli,B., Leonard,J.H. and Sturm,R.A. (2014) DCT protects human melanocytic cells from UVR and ROS damage and increases cell viability. *Exp. Dermatol.*, **23**, 916–921.
24. Jiang,L. and Hermeking,H. (2017) miR-34a and miR-34b/c suppress intestinal tumorigenesis. *Cancer Res.*, **77**, 2746–2758.
25. Xiao,P., Liu,W. and Zhou,H. (2016) miR-429 promotes the proliferation of non-small cell lung cancer cells via targeting DLC-1. *Oncol. Lett.*, **12**, 2163–2168.
26. Han,G., Zhao,W., Song,X., Ng,P.K.S., Karama,J.A., Jonasch,E., Mills,G.B., Zhao,Z., Ding,Z. and Jia,P. (2017) Unique protein expression signatures of survival time in kidney renal clear cell carcinoma through a pan-cancer screening. *BMC Genomics*. **18**, S9.
27. Cheng,F., Jia,P., Wang,Q., Lin,C.C., Li,W.H. and Zhao,Z. (2014) Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Mol. Biol. Evol.*, **31**, 2156–2169