Differentiating between tumor biopsies and normal cells in the TCGA dataset using targetable cell-surface gene sets

Nicholas Lorig-Roach

November 29, 2017

Abstract

The cancer genome atlas project (TCGA) represents one of the largest sources of mRNA-seq data collected from human tumors. The expression profile of CD, receptor tyrosine kinase, and nuclear hormone receptor transcripts within this data set was investigated using a number of methods including DESeq and t-SNE based clustering. These gene families mediate cell-to-cell communication, adhesion, immune recognition, and hormone response among many roles. By identifying unique expression patterns in TCGA tumor cohorts, mechanisms these cancers use to co-opt host resources, evade the immune system, and proliferate may be identified. Highly expressed members of these cell-surface protein families could also serve as targets for immunotherapies or even small molecule drugs. Here it is shown that TCGA tumor cohorts can be distinguished from TCGA normals using this small, targetable set of 486 genes. Initial insights into the genes within this set that may represent therapeutic targets are discussed.

1 Introduction

The classical view of cancer genomics is summed up well in Vogelstein's review 'Cancer Genome Landscapes'[(1)] – the focus has been on identifying highly mutated genes, and in particular, so-called driver genes that are the catalysts for a tissue's transformation into a tumor. Cancer genome sequencing projects like the Cancer Genome Atlas strove to provide the data needed to identify these drivers, and now it is not difficult to identify highly mutated genes among the cancer biopsies sequenced as part of the TCGA project.[(2)] However, it is still a challenge to correlate even the most probable 'driver' genes to expected survival and a bigger challenge still to choose an appropriate therapy based on a patient's cancer phenotype. Gene transcription data now accessible via mRNA-seq experiments allows another view into the 'life' of a cancer, specifically its transcriptome and by proxy its proteome, which can inform us of the effects of the mutations observed via genome sequencing. The author would argue that a tumor's proteome is a better indicator of what sort of therapy might effectively kill it.

One way to take advantage transcriptome/proteome data is to find highly expressed cell surface proteins that can be used to identify tumors such that targeted therapies like antibodies can be trained to find and eliminate tumor cells.[(3)] This strategy has increased in popularity dramatically since the 1990s: up to 2001, approximately 30 human mAbs were clinically evaluated – now the cumulative number of clinically investigated antibodies is greater than 150 with the majority developed for cancer indications.[(4)] Despite these advantages, mAbs still face significant hurdles including immunogenicity of the antibodies themselves[(5)], high production costs, and challenges in formulation and administration. The three basic mechanisms by which antibody therapeutics act is by interruption of a protein function, recruitment of the effector arm of the immune system (e.g. NK cells, granulocytes, complement proteins), or by carrying a 'warhead' to a target.[(6)] Currently this methodology has been successful primarily in leukemias and lymphomas, where the antibody is targeted to B-cell specific lectins like CD22, but solid tumors have also been successfully targeted like in the HER2-targeting Herceptin for HER2 positive breast cancers [(6)].

Here the transcription of 'clusters of differentiation' (CD), receptor tyrosine kinase (RTK), and nuclear hormone receptor (nH) genes are investigated as a proof of concept for methods of identifying tumors with unusual protein expression. Both the CD and RTK gene families are cell surface proteins of which at least one member is already the target of an antibody therapeutic or other therapy in the clinic or in development (e.g. CD20 and HER2;[(7)] VEGF[(7)] and PTK7[(8)]). The CD genes are a family of proteins closely tied to the immune system containing numerous immunoglubulin containing proteins, lectins, selectins, and

other proteins involved in cell-cell recognition and communication.[9] RTKs can have dramatic effects on protein expression in response to cytokines, hormones, and growth factors and are known to be important in angiogenesis, development, and oncogenic processes.[10] The nuclear hormone receptor gene set was added to aid in differentiating between androgen dependent tumors like breast cancer. The transcription of these genes was evaluated using TCGA mRNA-seq data processed with the TOIL pipeline[11] that used RSEM[12] to quantify gene expression. Gene expression data from normal cell biopsies and primary tumors was clustered using t-SNE, which showed generally robust separation between tumors and normal cells, indicating the CD/RTK/nH gene set is suitable to differentiate between some cancers and the normal cells in TCGA cohorts. The genes driving the separation of primary tumors and normal cells were assessed using a number of techniques including analysis with DESeq2.[13]

2 Methods

The mRNAseq data used in this study is available at the TOIL hub of the UCSC Xena browser (https://toil.xenahubs.net). Specifically, RSEM expected counts data from the TCGA Pan-cancer cohort was used. A metadata table covering most of these samples is available at Synapse (https://www.synapse.org/#!Synapse:syn7248855). To prepare data for analyses, a Python program was written that can remove genes with low variance and subset the bulk RSEM data table by cohort, sample type, and gene set. The output are csv files containing the data matrix subset accompanied by a row-matched metadata subset for easy import into programs like DESeq2. The Python package sklearn was used for the t-SNE based clustering, which was incorporated into a program that takes the data & metadata output generated previously and plots 2D representations of the computed similarity between samples' gene transcription.

Prior to analysis with DESeq2, RSEM's expected count values were rounded to the nearest integer and all entries were incremented by 1 pseudocount. Here, a single cohort's primary tumor samples were compared to all normals with more than 20 samples of the same cohort. An example DESeq run is shown below, where 'reduced_data.csv' and 'col_data.csv' correspond to the data and metadata matrices, respectively.

```
> dd <- read.csv("reduced_data.csv", header = TRUE, row.names = 1, check.names = False)
> dd <- round(dd)
> dd <- dd + 1
> col_data <- read.csv("col_data.csv")
> ddsmat <- DESeqDataSetFromMatrix(countData= dd, colData= col_data, design= ~ sample_type)
> dds <- DESeq(ddsmat, parallel = TRUE)
> results <- results(dds, parallel = TRUE)
> sum(res$padj < 0.01, na.rm = TRUE)
[1] 138
> res_ordered <- results[order(results$padj),]
> res_ordered <- as.data.frame(res_ordered)[1:138,]
> write.csv(res_ordered, file = "deseq_results.csv")
```

After exporting the significant (p-adj < 0.01) differentially expressed genes, they were sorted by greatest-fold increase in gene expression in the primary tumor sample type to show genes that may be targetable or used to identify tumors. In cohorts like BRCA where there are both 'TP' and 'NT' sample types, it is possible to run deseq with the parameter 'design $= \sim$ sample_type + disease' where disease corresponds to cohort.

See Table S1 for a full list of TCGA cohorts and the number of samples within each sample type. CD and RTK, and nuclear hormone receptor genes used in this study were ripped from HUGO's gene database, simply using all genes in the respective categories and removing any overlap between categories.[9].

3 Results

A subset of TCGA tumor cohorts were analyzed using t-SNE based dimensionality reduction to assess whether the CD-RTK-nH gene set would be effective for differentiating between primary tumors and TCGA

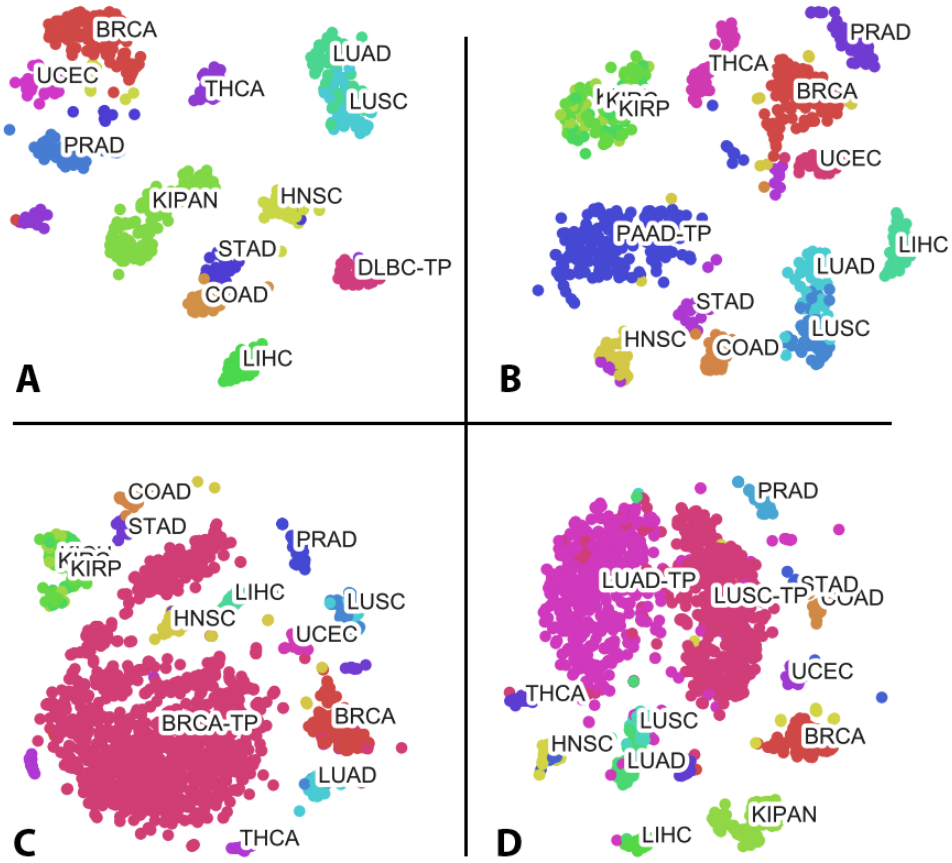

Figure 1: Plots showing t-SNE based dimensionality reduction using the CD-RTK-nH gene set with TCGA normals and A: DLBC primary tumors, B: pancreatic adenocarcinoma tumors, C: breast adenocarcinoma tumors, and D: lung adenocarcinoma and squamous cell carcinomatumors . t-SNE was performed using the scikit-learn python package with a euclidean-based distance matrix, perplexity = 30, and default settings for remaining parameters. Plotted using matplotlib and seaborn, labels were applied automatically based on the coordinates of the median value for a given cohort:sample-type.

normals. For this trial analysis, DLBC, PAAD, BRCA, and the lung carcinomas LUAD and LUSC were chosen because these cancers are either high occurence (lung, breast), high lethality (pancreatic), or act as controls where genes in this gene set are already used clinically to identify or treat disease (DLBC). All four t-SNE analyses (figure 1) showed excellent separation from TCGA normals to the author's naive eye. These same data were also investigated with DESeq2 to potentially identify the genes withine the CD-RTK-nH gene set that drove the separation between samples in the t-SNE analyses (Table 1 and Table 2). Genes with *log*2 fold change higher than 1 are shown for the DLBC cohort and higher than 0.45 for the PAAD, BRCA and LUSC cohorts. In these DESeq analyses, Fc receptor like proteins and TNF-related proteins (like CD70 and TNFRS- family genes) appeared most frequently. Figure 2 plots expression of a subset of these 'hits' across all TCGA cohorts, using a visualization provided at firebrowse.org. [(14)]

4 Discussion

It is exciting to see how well the primary tumors in Fig. 1 were separable from TCGA normal using t-SNE with this small gene set. However, it is unclear from that analysis whether the gene features that allow these

Cohort-Sample type	Gene	log_2 fold change	p-adj	Gene description
DLBC-TP	CD79B	7.82	0.0E+00	CD79b molecule
	CD19	7.63	1.9E-140	CD19 molecule
	FCRL3	7.42	6.0E-176	Fc receptor like 3
	MS4A1	7.25	1.0E-103	membrane spanning 4-domains A1
	FCRL2	7.03	1.0E-122	Fc receptor like 2
	TNFRSF13B	6.78	2.7E-113	TNF receptor superfamily member 13B
	IL21R	6.59	1.6E-261	interleukin 21 receptor
	FCER2	6.45	1.2E-92	Fc fragment of IgE receptor II
	LILRA4	6.38	4.0E-144	leukocyte immunoglobulin like receptor A4
	CD70	6.32	6.0E-132	CD70 molecule
	TNFRSF13C	6.3	3.3E-125	TNF receptor superfamily member 13C
	CD72	6.29	0.0E+00	CD72 molecule
	CD79A	6.11	7.6E-72	CD79a molecule
	FCRL5	5.96	1.5E-91	Fc receptor like 5
	CD22	5.48	2.1E-96	CD22 molecule
	CD27	5.31	9.4E-112	CD27 molecule
	TLR10	5.28	8.4E-120	toll like receptor 10
	CD37	5.25	2.0E-180	CD37 molecule
	FCRL1	5.24	7.4E-67	Fc receptor like 1
PAAD-TP	ADAM8	3.31	0.0E+00	ADAM metallopeptidase domain 8
	SEMA7A	2.91	4.0E-205	semaphorin 7A (John Milton Hagen blood group)
	CEACAM6	2.83	1.2E-26	carcinoembryonic antigen related cell adhesion molecule 6
	MST1R	2.72	2.8E-98	macrophage stimulating 1 receptor
	TNFSF11	2.58	5.0E-51	TNF superfamily member 11
	PLAUR	2.25	1.4E-106	plasminogen activator, urokinase receptor
	MUC1	1.84	6.2E-35	mucin 1, cell surface associated
	CD70	1.79	2.1E-49	CD70 molecule
	ITGB4	1.75	4.7E-43	integrin subunit beta 4
	IL2RA	1.67	1.5E-71	interleukin 2 receptor subunit alpha
	CEACAM5	1.6	3.3E-07	carcinoembryonic antigen related cell adhesion molecule 5
	CCR8	1.55	6.1E-61	C-C motif chemokine receptor 8
	IL2RG	1.52	2.1E-49	interleukin 2 receptor subunit gamma
	TNFSF4	1.36	6.5E-48	TNF superfamily member 4
	LAIR2	1.35	2.1E-29	leukocyte associated immunoglobulin like receptor 2
	HMMR	1.32	4.0E-25	hyaluronan mediated motility receptor
	CD72	1.29	2.6E-52	CD72 molecule
	CD55	1.29	6.6E-35	CD55 molecule (Cromer blood group)
	FUT3	1.25	7.2E-09	fucosyltransferase 3 (Lewis blood group)

Table 1: "Highly expressed" CD/RTK/nH genes identified using DESeq2 from the TCGA cohorts DLBC and PAAD primary tumors. DESeq was run comparing the cohorts 'TP' samples to 'NT' samples from all cohorts with at least 20 normal samples. Gene descriptions from HUGO.[9]

Cohort-Sample type	Gene	log_2 fold change	p-adj	Gene description
LUSC -TP	NR0B1	6.94	0.0E+00	nuclear receptor subfamily 0 group B member 1
	TNFRSF18	3.55	6.5E-302	TNF receptor superfamily member 18
	NR5A1	3.12	1.7E-145	nuclear receptor subfamily 5 group A member 1
	TFRC	3.1	0.0E+00	transferrin receptor
	HMMR	2.85	7.7E-267	hyaluronan mediated motility receptor
	NR2E1	2.74	6.0E-118	nuclear receptor subfamily 2 group E member 1
	CD70	2.37	4.5E-126	CD70 molecule
	SLC7A5	2.33	1.8E-114	solute carrier family 7 member 5
	TNFRSF9	2.19	2.0E-190	TNF receptor superfamily member 9
	CCR8	2.16	2.0E-228	C-C motif chemokine receptor 8
	EPHA8	2.16	6.9E-82	EPH receptor A8
	ADAM8	2.15	1.3E-265	ADAM metallopeptidase domain 8
	FZD10	2.15	9.4E-68	frizzled class receptor 10
	ITGB4	2.08	1.4E-135	integrin subunit beta 4
	IL2RA	2.04	1.9E-199	interleukin 2 receptor subunit alpha
	FCRL5	2.02	2.5E-81	Fc receptor like 5
	CD109	1.96	2.8E-132	CD109 molecule
	ATP1B3	1.87	2.3E-150	ATPase Na+/K+ transporting subunit beta 3
	CLEC7A	1.77	6.7E-136	C-type lectin domain containing 7A
	TNFRSF13C	1.76	1.7E-78	TNF receptor superfamily member 13C
	ADAM17	1.67	4.4E-217	ADAM metallopeptidase domain 17
BRCA-TP	RET	4.08	0.0E+00	ret proto-oncogene
	EPHA8	3.92	2.6E-272	EPH receptor A8
	ESR1	3.63	8.8E-261	estrogen receptor 1
	BMPR1B	3.4	1.8E-162	bone morphogenetic protein receptor type 1B
	TNFRSF18	2.92	5.3E-258	TNF receptor superfamily member 18
	HMMR	2.69	0.0E+00	hyaluronan mediated motility receptor
	CCR8	2.64	2.1E-301	C-C motif chemokine receptor 8
	FLT3	2.33	1.3E-151	fms related tyrosine kinase 3
	PGR	1.96	1.4E-61	progesterone receptor
	TNFSF4	1.93	0.0E+00	TNF superfamily member 4
	ERBB2	1.85	1.2E-129	erb-b2 receptor tyrosine kinase 2
	ADAM8	1.83	3.0E-295	ADAM metallopeptidase domain 8
	MUC1	1.74	1.1E-96	mucin 1, cell surface associated
	NR2E3	1.74	6.1E-112	nuclear receptor subfamily 2 group E member 3
	IFITM1	1.7	2.8E-181	interferon induced transmembrane protein 1
	ERBB3	1.68	1.4E-195	erb-b2 receptor tyrosine kinase 3
	NR5A1	1.67	3.0E-56	nuclear receptor subfamily 5 group A member 1
	IGF1R	1.67	4.0E-106	insulin like growth factor 1 receptor
	RARA	1.6	3.0E-198	retinoic acid receptor alpha

Table 2: "Highly expressed" CD/RTK/nH genes identified using DESeq2 from the TCGA cohorts LUSC and BRCA primary tumors. DESeq was run comparing the cohorts 'TP' samples to 'NT' samples from all cohorts with at least 20 normal samples. Gene descriptions from HUGO. [9]

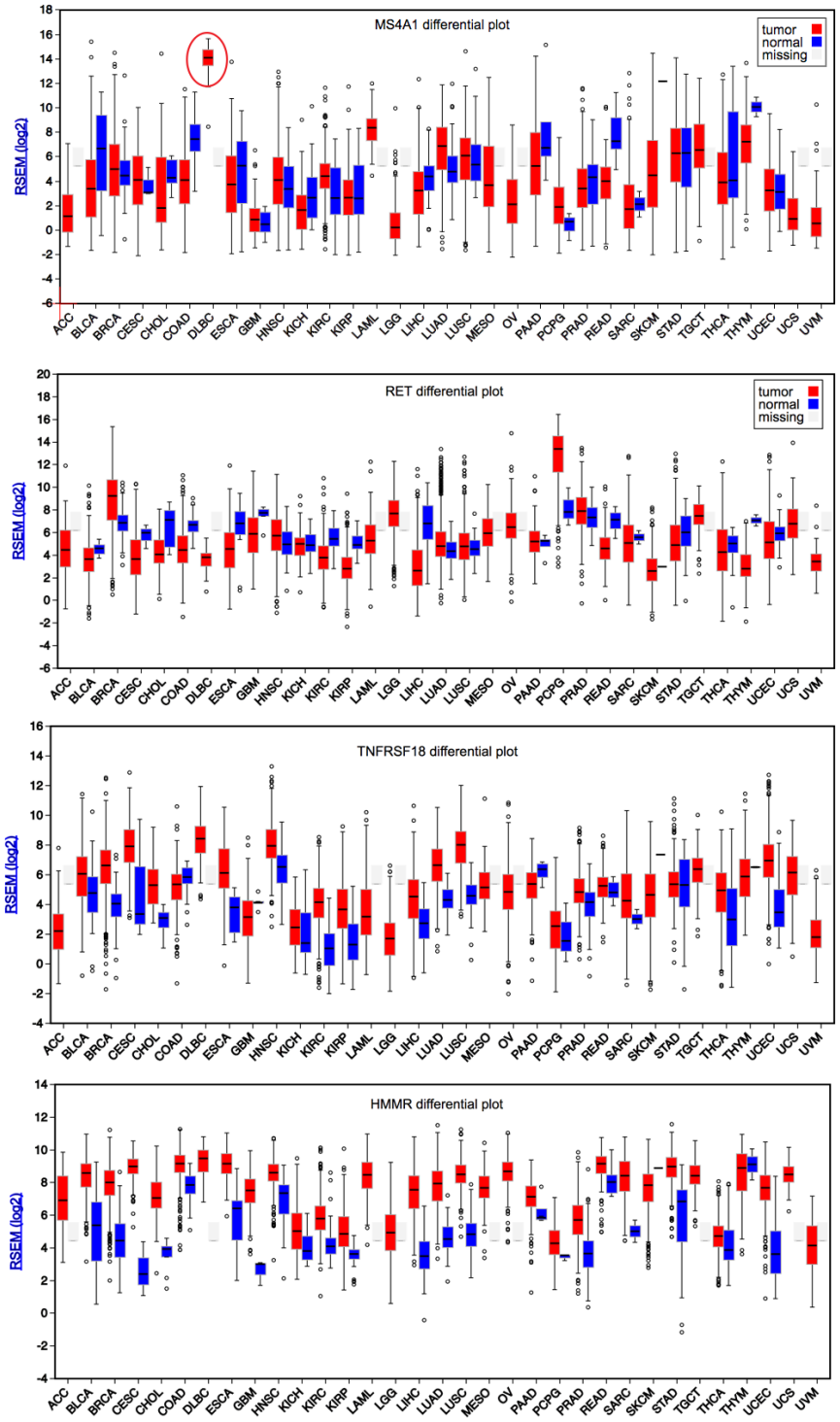

Figure 2: Expression *log*2 histograms across all TCGA cohorts (excluding duplicate 'pan' cohorts e.g. KIPAN, COADREAD) for MS4A1 (CD20), RET, TNFSF18, and HMMR. Made at firebrowse.org/viewGene.html.

samples to be distinguished from each other could be viable therapeutic targets. An ideal target would have significantly higher expression among tumors, giving a wide therapeutic window. Based on the significant (p ¡ 0.01) genes identified with DESeq, there are nearly equivalent number of up- and down-regulated genes between normals and DLBC (174 vs 171), PAAD (72 vs 79), LUSC (66 vs 72), and BRCA (23 vs 22). Despite many genes showing significantly higher expression, only the DLBC tumor cohort had genes with more than 2-fold greater expression than the aggregate normals.

The DLBC cohort in many ways acted as a positive control for this analysis, where currently used antibody therapeutic targets like MS4A1 (more commonly CD20) stands out from both normal cells in these cohorts and the other tumor cells (table **??**, figure 2). Among these genes and cohorts, no aggregate primary tumor expression data had a distribution that paralleled hits like MS4A1 in DLBC, where nearly the entirety of the e-$log2$ distribution was outside of the upper percentile of all other cohorts. In light of this, the majority of the genes identified in this work may not be suitable for targeted therapies. However, the expression trends are significant enough to warrant further investigation into the biology backing the observed trends.

Other semi-validating hits include TNFR family members: at least one sub-type is highly expressed in each cohort assessed here. TNF is a known mediator of inflammation and angiogenesis and a potent modulator of immune cell activation with a known role in cancer progression.[15] The carcinoembryonic antigen molecules (CEACAM5/6) were also over-expressed in the pancreatic solid tumors assessed with DESeq, another family of molecules known to be associated with cancers.[16] CEACAMs are named due to their expression during fetal development, though some expression, particularly in epithelial cells, can be observed into adulthood. However, these proteins are reportedly (and according to the data analyzed here) markers of tumorogenesis and are thought to enable tumor growth and metastasis by stimulating release of tumorogenic cytokines from immune and other cell types.[16] In the BRCA cohort, many known players appear including RET and estrogen and progesterone receptors, in addition to numerous other highly significant leads.

One of the main flaws with the mRNA-seq data used in this work is the lack of normal cell biopsies for many tissue types, particularly CNS, immune system, and germ line (i.e. cervical, ovarian, testicular) tissues. This calls into question many of the genes deemed significant in the DESeq analyses, however it is clear that it is possible to distinguish between the normals available here and DLBC, PAAD, BRCA and lung primary tumors. The TOIL pipeline has also run the entirety of the GETx[17] mRNAseq data with the same RSEM based quantification methodology, which may be a suitable source of additional normals. Their suitability can be assessed by clustering GTEx normals with those available through TCGA – if the normals do not cluster well then additional primary data may be needed for comprehensive expression analyses.

Another important point to recall is that these values are based on transcript abundance which may not accurately reflect the state of the proteome. Additional experiments are required to verify these transcripts are being translated before any could be considered targets. However, the data here suggest that rapid and accurate clustering of tumor and normal cells can be achieved using a modestly sized, biologically relevant gene set. This is useful for analysis of patient biopsies and could perhaps be used to aid in early identification of cancers.

Though the work presented here is immature, further development of differential transcription analysis and gene family based phenotyping appears to be a worthwhile pursuit in uncovering novel cancer cell biology and perhaps even new therapeutic targets.

5 Acknowledgments

Many thanks to Josh Stuart and John Vivian for both curating the data used in this work, and pointing me toward it! .

	sample type							
cohort	TP	TR	TB	TAP	TM	TAM	NT	description
BRCA	1093	-	-	-	7	-	112	Breast AC
KIPAN	889	-	-	2	-	-	129	Pan-kidney cohort
COADREAD	623	2	-	-	1	-	51	Colorectal AC
STES	599	-	-	-	1	-	46	Stomach and Esophageal C
UCEC	545	1	-	-	-	-	35	Uterine Corpus Endometrial C
LGG	516	14	-	-	-	-	-	Brain Lower Grade Glioma
HNSC	520	-	-	-	2	-	44	Head and neck SCC
LUAD	515	2	-	-	-	-	59	Lung AC
LUSC	501	-	-	-	-	-	51	Lung SCC
THCA	501	-	-	-	8	-	59	Thyroid C
PRAD	497	-	-	-	1	-	52	Prostate AC
BLCA	408	-	-	-	-	-	19	Bladder Urothelial C
LIHC	371	2	-	-	-	-	5	Liver hepatocellular C
CESC	304	-	-	-	2	-	3	Cervical SCC & endocervical AC
OV	303	4	-	-	-	-	-	Ovarian serous cyst-AC
SARC	259	3	1	-	-	-	2	Sarcoma
PAAD	178	4	-	-	1	-	4	Pancreatic AC
LAML	-	-	173	-	-	-	-	Acute Myeloid Leukemia
TGCT	150	-	-	6	-	-	-	Testicular Germ Cell Tumors
THYM	120	-	-	-	-	-	2	Thymoma
MESO	87	-	-	-	-	-	-	Mesothelioma
ACC	79	-	-	-	-	-	-	Adrenocortical AC
UCS	57	-	-	-	-	-	-	Uterine Carcinosarcoma
DLBC	48	-	-	-	-	-	-	Diffuse Large B-cell Lymphoma
CHOL	36	-	-	-	-	-	9	Cholangiocarcinoma

Table S1: Summary of TCGA mRNAseq data used in this study. Counts for each sample type reflect those served by the Firebrowse python API as of March 2017 and do not include filtered samples (e.g replicates or redactions).[14] This 'level 3' Illummina HiSeq/Ga2 data is preprocessed with the RSEM[12] method and served as log_2 transformed floats. When available, 'pan' cohorts like KIPAN, COADREAD, and STES were used rather than their component cohorts. AC = adenocarcinoma; C = carcinoma; SCC = squamous cell carcinoma.

References

[1] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer Genome Landscapes. *Science*, 339(6127):1546–1558, March 2013.

[2] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna M. Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature genetics*, 45(10):1113–1120, October 2013.

[3] Andrew M. Scott, Jedd D. Wolchok, and Lloyd J. Old. Antibody therapy of cancer. *Nature Reviews Cancer*, 12(4):278–287, April 2012.

[4] Aaron L. Nelson, Eugen Dhimolea, and Janice M. Reichert. Development trends for human monoclonal antibody therapeutics. *Nature Reviews Drug Discovery*, 9(10):767–774, October 2010.

[5] Ian Shieh, Danielle Leiske, and Ankit Patel. Dynamics of interfacial adsorption and self-association of antibody therapeutics. *Abstracts of Papers, 247th ACS National Meeting and Exposition (Dallas, TX, United States)*, March 2014.

[6] George J. Weiner. Building better monoclonal antibody-based therapeutics. *Nature Reviews Cancer*, 15(6):361–370, June 2015.

[7] Asher Mullard. 2015 FDA drug approvals. *Nature Reviews Drug Discovery*, 15(2):73–76, February 2016.

[8] Marc Damelin, Alexander Bankovich, Jeffrey Bernstein, Justin Lucas, Liang Chen, Samuel Williams, Albert Park, Jorge Aguilar, Elana Ernstoff, Manoj Charati, Russell Dushin, Monette Aujay, Christina Lee, Hanna Ramoth, Milly Milton, Johannes Hampl, Sasha Lazetic, Virginia Pulito, Edward Rosfjord, Yongliang Sun, Lindsay King, Frank Barletta, Alison Betts, Magali Guffroy, Hadi Falahatpisheh, Christopher J. O'Donnell, Robert Stull, Marybeth Pysz, Paul Escarpe, David Liu, Orit Foord, Hans Peter Gerber, Puja Sapra, and Scott J. Dylla. A PTK7-targeted antibody-drug conjugate reduces tumor-initiating cells and induces sustained tumor regressions. *Science Translational Medicine*, 9(372):eaag2611, January 2017.

[9] Kristian A. Gray, Bethan Yates, Ruth L. Seal, Mathew W. Wright, and Elspeth A. Bruford. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Research*, 43(Database issue):D1079–1085, January 2015.

[10] Mark A. Lemmon and Joseph Schlessinger. Cell signaling by receptor-tyrosine kinases. *Cell*, 141(7):1117–1134, June 2010.

[11] John Vivian, Arjun Arkal Rao, Frank Austin Nothaft, Christopher Ketchum, Joel Armstrong, Adam Novak, Jacob Pfeil, Jake Narkizian, Alden D. Deran, Audrey Musselman-Brown, Hannes Schmidt, Peter Amstutz, Brian Craft, Mary Goldman, Kate Rosenbloom, Melissa Cline, Brian O'Connor, Megan Hanna, Chet Birger, W. James Kent, David A. Patterson, Anthony D. Joseph, Jingchun Zhu, Sasha Zaranek, Gad Getz, David Haussler, and Benedict Paten. Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, 35(4):314–316, April 2017.

[12] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

[13] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, 2014.

[14] Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run, 2016.

[15] Remco van Horssen, Timo L. M. ten Hagen, and Alexander M. M. Eggermont. TNF-alpha in Cancer Treatment: Molecular Insights, Antitumor Effects, and Clinical Utility. *The Oncologist*, 11(4):397–408, April 2006.

[16] Bernhard B. Singer, Inka Scheffrahn, Robert Kammerer, Norbert Suttorp, Suleyman Ergun, and Hortense Slevogt. Deregulation of the CEACAM Expression Pattern Causes Undifferentiated Cell Growth in Human Lung Adenocarcinoma Cells. *PLOS ONE*, 5(1):e8747, January 2010.

[17] Genotype-tissue expression (gtex) project – http://www.gtexportal.org/home/gene/ensg00000109956.8.