A Pan-Cancer Analysis of Gene Expression Data Across 32 Tissues

Abrar Albahrani, Hossein ArianNejad, Shuvomoy Chowdhury, Amanda Vander Wal, Jordan Wheeler

December 12, 2018

Abstract

Cancer is among the leading causes of death worldwide, with estimate of more than one million newly diagnosed cancer cases and over 600,000 patients' deaths by 2018 in the United States alone[1]. Thus, the study of cancer genomic data is highly valuable as it would aid in deeper understanding cancer at the genomic level and possibly find novel genomic targets for early diagnosis and treatments. The difference between normal and tumor tissue is well-established in various studies on genomic data. Gene expression profiling is a method to measure activity and expression of genes in tumor and normal tissues, resulting in the identification of increase or decrease in gene expression between tumor and normal tissues. These genes that are differentially expressed between tumor and normal tissues are significant targets in cancer research. In this study, we evaluate 32 distinct tumor tissues from over 30,000 data files gathered from 10,672 patients from The Cancer Genome Atlas project (TCGA). Our research aims to find and identify the key patterns that differentiate between tumor and normal tissues through visualizations and data analysis using R.

Introduction

Understanding cancer genomics is crucial to identifying novel methods for early diagnosis, which leads to enhancing outcomes of treatment. In recent years, researchers have focused on genes when studying cancers, as well as other diseases, i.e. investigating the effect of genes activity on fundamental functions of cells and tissue. This approach is expected to yield a more specific comprehension of cancer biology. Two classes of genes that have an established role in progression of tumors are: tumor suppressors and oncogenes. In normal body tissues, tumor suppressor genes are actively expressed to repair DNA and suppress tumor progression as their name suggest, if it happens to occur. Oncogene genes, on the other hand, are known to have a tendency to allow a cancerous cell to escape cell programed death. Genes in both of these classes are often mutated in the case of cancer tissue and have different expressions than a normal tissue. We select three tumor suppressors: TP53, BRCA1, and BRCA2 , and three oncogenes: KRAS, MYC, and ABL1, according to their important role in tumorigenesis[2].

Background

The genes selected for this research have different roles in a cell and they need to be studied individually. Therefore, to have a better understanding of what is each genes function and in what situations they have been observed the most, a concise literature review for each of these genes have been presented. The chromosomal locations of these genes is provided in Figure 1.

[1] https://www.cancer.gov/about-cancer/understanding/statistics
[2] https://www.ncbi.nlm.nih.gov/books/NBK9894/

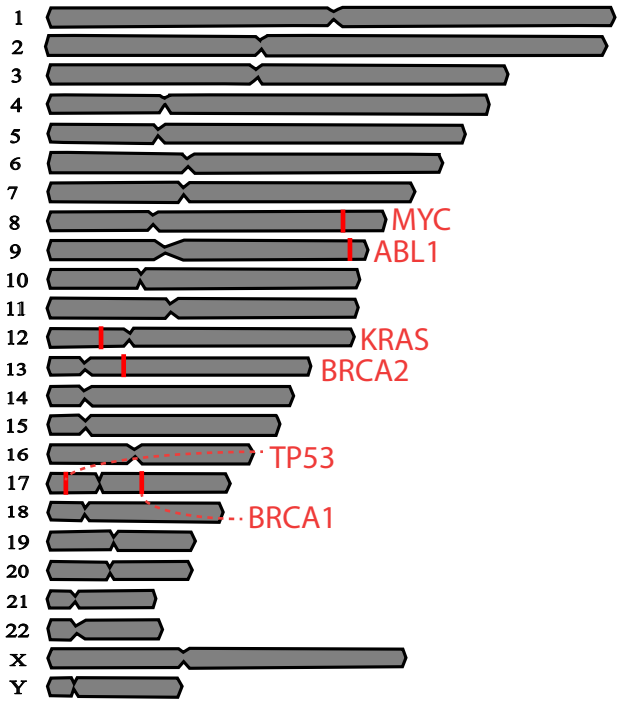

Figure 1: Chromosomal map showing location of selected genes: ABL1, BRCA1, BRCA2, KRAS, MYC, and TP53.

BRCA1 and BRCA2

According to National Institute of Health (NIH), normal function for BRCA1 and BRCA2 genes is to provide instruction for generating proteins that act as tumor suppressor. These proteins help prevent cells from growing and dividing in a rapid uncontrolled way. BRCA1 and BRCA2 proteins are involved with the repairing of the damaged DNA and by helping to repair DNA, theses two proteins plays a critical role maintaining the stability of a cell's genetic information. The BRCA1 protein can also regulate the activity of the other genes and to carry out this function, it needs to interacts with many other proteins, including other tumor suppressors and proteins that regulate cell division[3]. Regarding to BRCA2, researchers suspect that it has additional function inside cells. For example, it might help the with some steps in cell division[4]. Mutations of the BRCA1 and BRCA2 tumor suppressor genes often lead to the increase in chance of breast and ovarian cancers for females[5]. These mutations can be hereditary and passed down from either the father or the mother. Mutation in BRCA1 and BRCA2 are responsible for about 25% of the risk for familial breast cancer and therefore 5-10% of all the breast cancer cases. Also there is a concurrence that mutations in genes BRCA1, BRCA2 and TP53 are responsible for on average 16-20 % of the risk of the familial breast cancer[6].

TP53

The TP53 provides instructions to a protein (p53) that regulates the cell cycle. Mutations of the TP53 gene have been found in more than half of human cancers making it the most common target of gene mutations

[3]https://ghr.nlm.nih.gov/gene/BRCA1
[4]https://ghr.nlm.nih.gov/gene/BRCA2
[5]https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet#q1
[6]https://www.ncbi.nlm.nih.gov/pubmed/28985766

in human cancers[7]. Research done by de Moura Gallo CV. et. al.[8] suggest that, Mutations in the TP53 tumor suppressor gene plays a significant role in cancer risk, as impaired p53 function may contribute to the multistep process of carcinogenesis. On the other hand, research done by Xiao et. al[9] suggest no relation was found between TP53 mutation and survival rate. They suggested that TP53 mutation is a potential negative predictor of cancer in metastatic melanoma treated with CTLA-4 blockade.

KRAS

The KRAS gene is responsible for making a protein called K-Ras and this protein is responsible for relaying signals from outside the cell to the cell nucleusˆ [https://ghr.nlm.nih.gov/gene/KRAS]. KRAS is the most frequently mutated oncogene in cancer and its mutation is commonly associated with resistance to therapy and in many cases the the KRAS mutation is generally associated with shorter overall survival[10].

MYC

The MYC gene is a Proto-oncogene which means it's a normal gene which, when altered by mutation becomes an oncogene that contributes to cancer[11]. MYC's major role is to regulate and modify the expression of other genes or proteins inside a cell. Amplification of this gene is frequently observed in numerous human cancers. MYC genes have been observed in cervix, colon, breast, lung and stomach cancers[12].

ABL1

Based on NIH definition, the ABL1 gene is responsible for instructions for making a protein involved in many processes in cells throughout the body. ABL1 belongs to oncogenes category which means when mutated, they have the potential to cause normal cell to become cancerous[13].

Efforts in several journal articles and databases are made to report tumor suppressor expression, yet none report a pan-cancer gene expression study on all 32 tissues from The Cancer Genome Atlas (TCGA). In this study, we report analysis of gene expression across 32 tissues and focus on a selected tumor suppressors and oncogenes. We expect to find novel patterns of expression among our data to classify tissues accordingly.

Data Overview

The Cancer Genome Atlas (TCGA) is a resource of genomic data for 33 types of cancer. It contains collection of tumor and normal (tumor adjacent) tissues from patients. TCGA uses abbreviated names to refer to the tissue types (Table 1). In this study, analyze 32 of the tissues provided by TCGA. These tissues are derived mainly from more than 20 organs (Figure 2), capturing a wide variety tissues to be investigated. Most of the tissues belong to the digestive system, as it include more organs. The reason underlying an organ having multiple tissues is because of the diverse type of cells found in an organ, and different cell types are shown to have distinct characteristics as tumor tissues[14].

[7] https://www.ncbi.nlm.nih.gov/books/NBK9894/
[8] https://www.ncbi.nlm.nih.gov/pubmed/15878142?dopt=Abstract
[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6020711/
[10] https://www.ncbi.nlm.nih.gov/pubmed/28958387
[11] https://ghr.nlm.nih.gov/gene/MYC
[12] https://www.ncbi.nlm.nih.gov/gene/17869
[13] https://ghr.nlm.nih.gov/gene/ABL1
[14] https://www.ncbi.nlm.nih.gov/books/NBK9963/

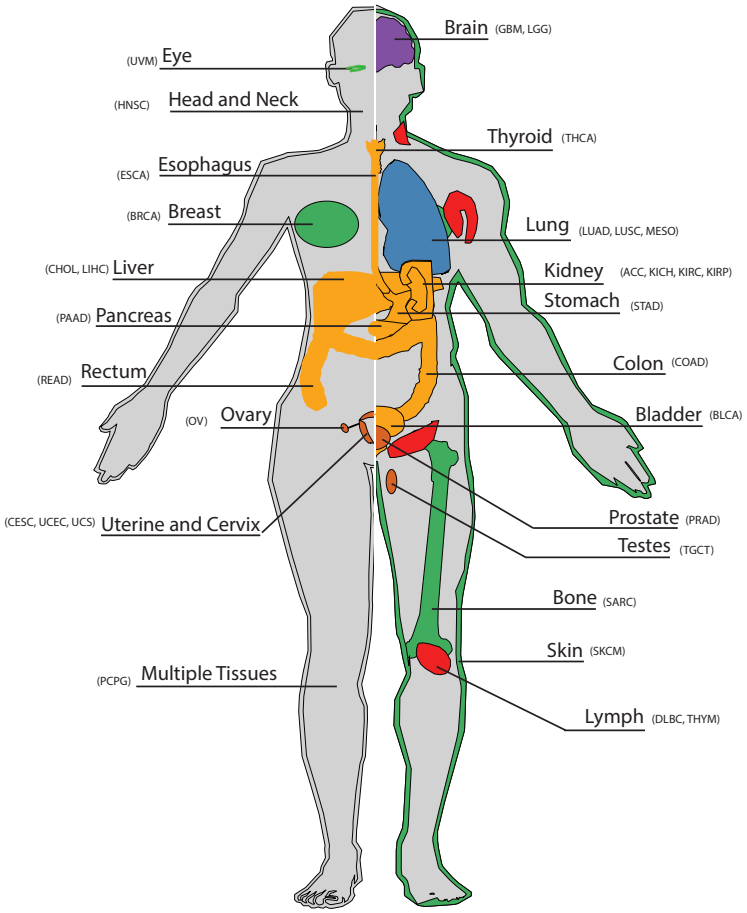

Figure 2: Human tissue map: Organs are labeled with organ name and its matching specific TCGA tissue abbreviated name.

Table 1: Tissue abbreviation and full name are recorded for reference.

Tissue.Abbreviation	Tissue.Full.Name
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UVM	Uveal Melanoma

Materials and Methods

Tissue from TCGA have different number of samples [Figure 3]. BRCA has 1,111 tumor samples and 113 normal, which is the highest number of samples in all tissues in our data. While all tissues have tumor samples, not all tissues have normal samples. In particular, 25% of tissues have no normal samples for gene expression; these tissues are: ACC, BLCA, LGG, MESO, OV, TGCT, UCS, and UVM. Although we are unable to perform gene expression comparisons between tumor and normal samples for these tissues, a comparison of their expression pattern with other tumors is certainly beneficial and informative.

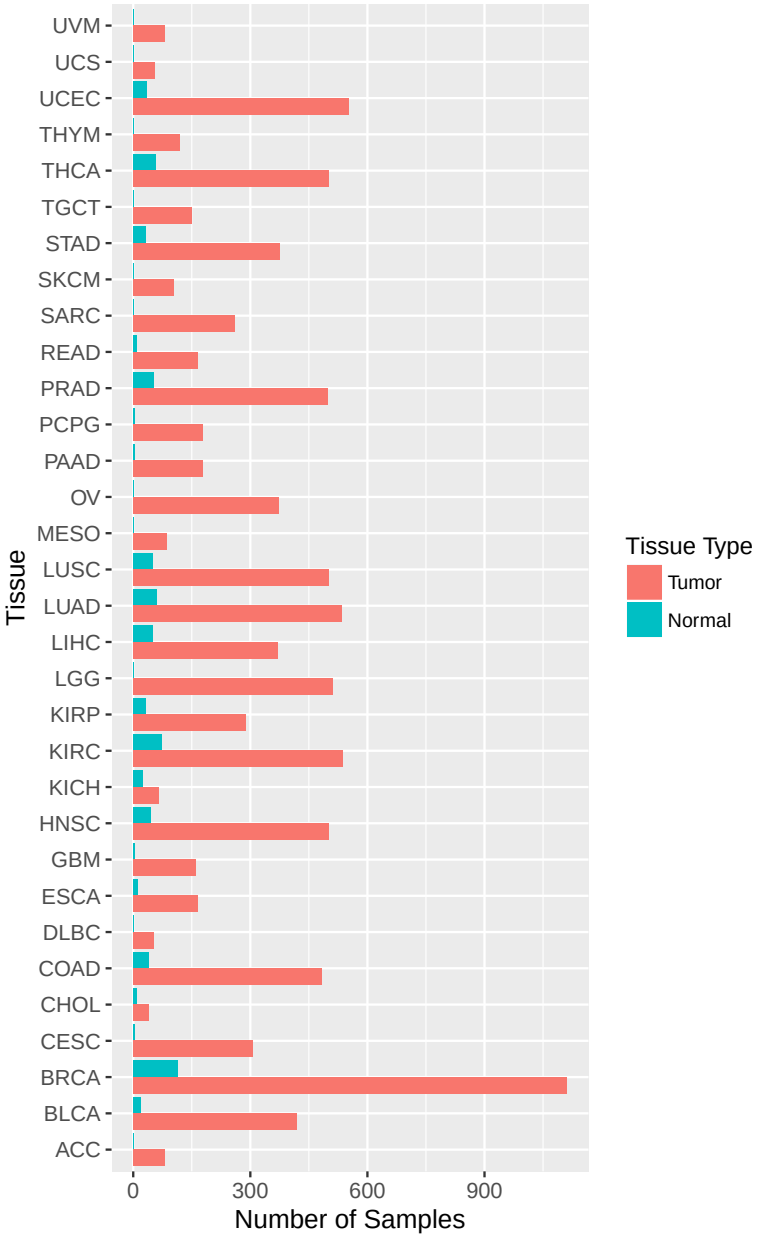

Figure 3: Distribution of samples across 32 tissues. Number of tumor samples is shown in pink and normal samples in teal.

Expression Data

Gene expression data is downloaded from TCGA, which is made publicly available for all tissue types. All files are downloaded through R package, TCGAbiolinks, using a query specifying tissue name, data type, and sample type. In this study, we queried for all 32 tissues, gene expression as data type, and normal or tumor as sample type. TCGA has samples collected from 33 tissues, but gene expression data is only available for 32 tissues, the tissue that lacks gene expression data is acute myeloid leukemia (LAML). Although all tissues have tumor samples, not all tissues have normal samples. A sample file contains two columns, gene ID and gene expression value. All samples have 60,483 gene IDs and corresponding expression value. Samples are downloaded from TCGA as zipped text files. Data preparation is completed with Unix shell scripts and R.

Data Preparation

Gene expression files are prepared using Shell and R. The raw data downloaded is described as below: Raw expression files:

1. A tissue folder is downloaded via TCGAbiolinks in R. The folder contains subfolders, one for each sample.
2. A sample folder has one zipped tab delimited text file (.txt.gz).
3. Normal and tumor samples are downloaded in a separate tissue folder.
4. Data has 24 normal tissues and 32 tumor tissues.
5. Shell: For samples in a tissue, Files are:
    - Unzipped using gunzip command, then extracted from their individual subfolders to one folder. Normal and tumor remain separated.
    - Sorted according to gene ID column using sort command.
    - Joined by gene ID column using paste and column command into a single file. The product, for each tissue there are two files: normal, if any, and tissue. Number of columns depend on number of samples in a tissue, and number or rows is the number of gene ID, 60,483. Some tissues have multiple files per tissue because of the number of samples. An example of this is BRCA that has six files for tumor tissue. More explanation is provided in Challenges section.
6. R:
    - A tissue file is read as data frame using read.table() function from dplyr package.
    - Data frame is converted to long form using melt function from reshape package.
    - The expression is visualized using geom_tile() from ggplot2 package. We will refer to the tile graph as heatmap in this study. Additionally, R is used for:
        1. Producing this report: knitr package[15]
        2. Figure 1: chromoMap package[16]
        3. Figure 2: gganatogram package[17] and added captions to organs using Adobe Illustrator.
        4. Heatmaps and boxplots: used ggplot2 package[19] and Shiny App[20].

Results

Heatmaps

All heatmaps may be viewed on Shiny App: https://ajvanderwal.shinyapps.io/ShinyApp/ Result of selected genes expression analysis show an overexpression of MYC genes in majority of analyzed tissues despite of

[15]Yihui Xie (2018). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.20.
[16]https://cran.r-project.org/web/packages/chromoMap/index.html
[17]Maag JLV. gganatogram: An R package for modular visualisation of anatograms and tissues based on ggplot2[18]. F1000Research 2018, 7:1576 (doi: 10.12688/f1000research.16409.1)
[19]H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
[20]https://cran.r-project.org/web/packages/shiny/index.html

them being normal or tumor. Exception to this pattern is observed in CESC, KICH, and TGCT tumor tissues, which have also lower expression in normal sample in CESC and KICH, whereas TGCT has no normal samples. MYC is higher expression in GBM tumor in comparison to normal tissue.

In contrast to MYC, BRCA1 and BRCA2 have no expression in normal nor tumor across all tissues. Interestingly, BRCA1/2 are genes coding for DNA damage repair function and are well-known to be hereditary mutated in cancer tissues such as breast and ovarian. Although a previous report[21] on ovarian cancer samples shows higher expression of BRCA1, we did not detect any increase of expression in OV tumor tissue and we have no normal tissue samples to compare with. BRCA1 has a slight increased expression in TGCT, whereas BRCA2 has no expression across different tissue types, which confirmed from a study report. Around 40% of normal and/or tumor tissues show a high of of TP53 expression. TP53 has an increased expression in tumor compared to normal tissue in CHOL, GBM, SARC, and THCA. Other tumor tissues show a lower TP53 expression compared to normal in CESC, COAD, KIRP, READ, and THYM. A previous study[22] reports tumor tissues such as ovarian, uterine, and gastrointestinal have low expression of TP53, confirming our observation in READ. Similar to TP53, 43% of tumor tissues have high ABL1 expression but it is uniquely expressed in tumor tissues in two CHOL and KICH. When comparing the overall expression of genes across normal and tumor tissues, ABL1 and KRAS are more expressed in tumor tissue.

To show an example of the heatmaps we created in Shiny App, Figure 4 illustrates gene expression patterns in KICH normal and tumor tissues. In KICH, our selected genes in tumor suppressors and oncogenes show low expression in normal samples in general [Figure 4]. Oncogenes, ABL1 and MYC in specific, have a slightly more expression than tumor suppressors in normal tissue. One sample has a higher expression in MYC, which is opposite to pattern for normal samples. This high expression is unexpected considering considering that samples must have a unified pattern in expression. This case has a great possibility of having an error due biological sample contamination. When compared to tumor tissue, oncogenes are highly expressed. A clear pattern is seen in ABL1 that is regular across tumor samples. MYC has high expression in one of tumor samples. Out of the tumor suppressor genes, TP53 has highest expression in tumor samples. Although a low expression would make more biological sense, since tumor suppressors function as anticancer and we expected to have low expression in a tumor tissue, we explain the relatively higher expression has no effect on tumor because it might be mutated and expresses a dysfunctional product.

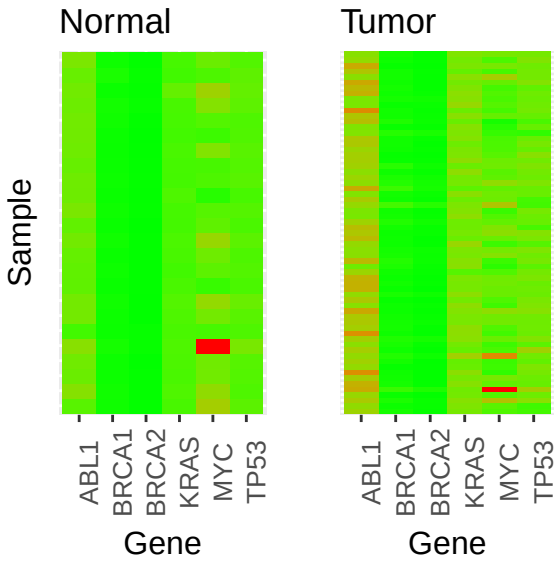

Figure 4: Heatmap tumor suppressors and oncogenes in KICH normal and tumor tissue samples. High expression is shown in red. Low expression is shown in green.

[21] https://www.nature.com/articles/s41416-018-0217-4
[22] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5352183/

Box plots

We created box plots comparing normal and tumor tissues in each selected gene. From Figure 5, we observe an increase number of outliers in tumor tissues in comparison to normal tissues, which shows a feature of tumor tissues of being discrepant. In general, tumor tissues shift expression showing an increase in selected genes. While majority of samples in tumor tissues express the selected genes in an increased level, some samples from tumor display extreme decreased or increased expression.

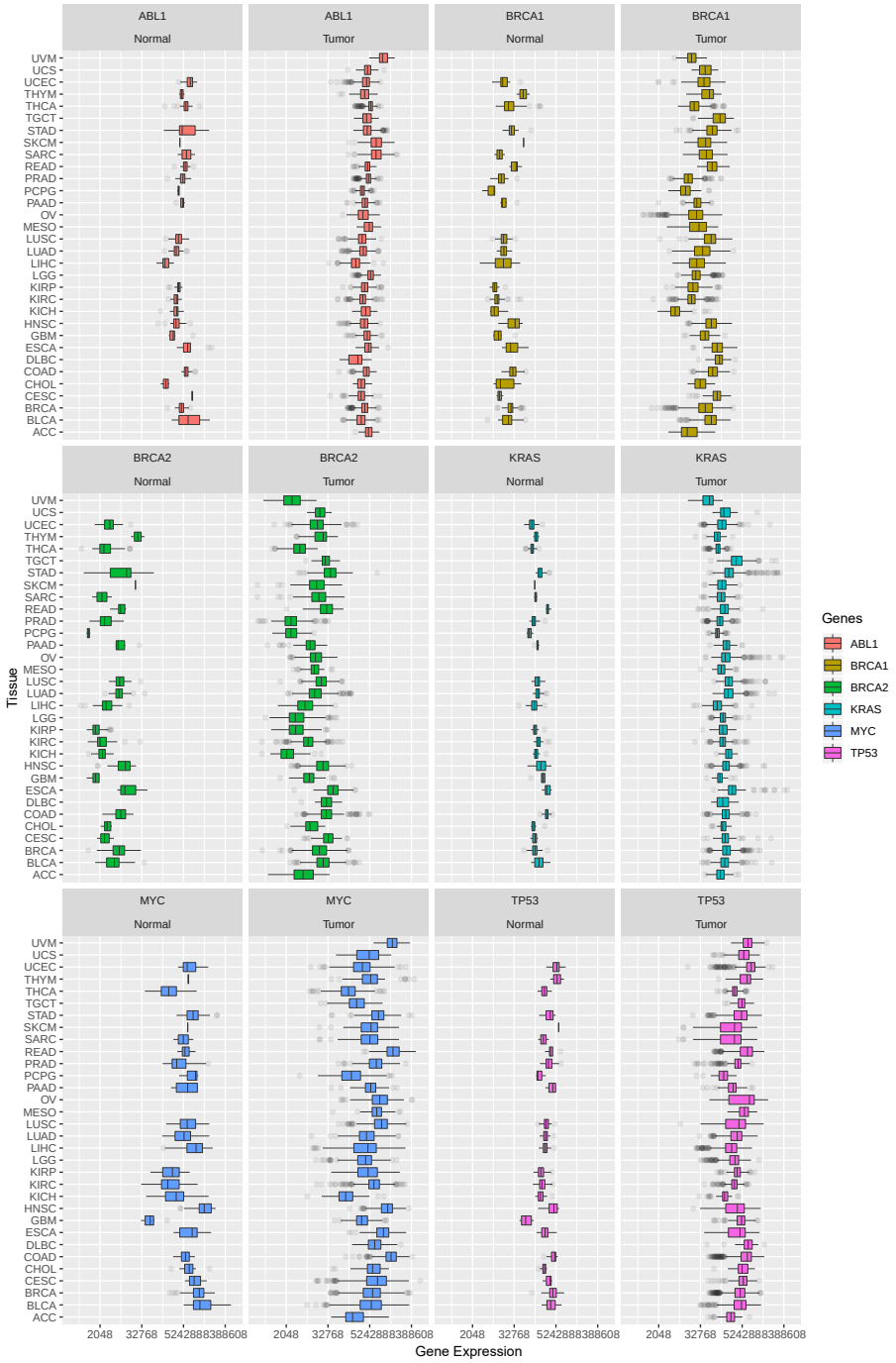

Figure 5: Box plots of selected genes expression in normal and tumor tissues. Appendix A for full size figure.

Paired t Test

Other than specific genes, it is also worthwhile to evaluate each tissue separately given all genes. This would tell us how active tumor suppressor genes and oncogenes are in different tissues. In a more applied sense, it would allow researchers to narrow down which genes (tumor supressors or oncogenes) to focus on when examining a distinct tissue. We can use a paired t-test to calculate significance of the difference between tumor tissue gene expression and normal tissue gene expression. We can also confirm that genes, as a whole, are infact expressed differently amongst tumor and normal tissues.

60,483 different gene IDs were used to evaluate the expressiveness. Since tumor and normal tissue had different sample sizes from patients (e.g. for the BLCA tissue, there were 415 tumor tissue samples and 20 normal tissue samples) we averaged the expressiveness based on each individual gene ID and then compared the average expressiveness from tumor tissue gene ID to the average expressiveness from normal tissue gene ID. Note that the samples for the t-test are the gene IDs and that values are the average expressiveness of that specific gene ID. We conducted a two-tailed paired t-test to see if there is a statistically significant difference, with our null hypothesis (1) stating that the average gene expressiveness from tumor and normal tissues are equal, and our alternative hypothesis (2) stating that the average gene expressiveness from tumor and normal tissues are not equal. We did this by taking the average tumor gene expressiveness minus the average normal gene expressiveness for each gene ID. The results will state which tissues' genes are expressed differently.

$$H_o : \alpha_{Tumor} - \alpha_{Normal} = 0 \tag{1}$$

$$H_a : \alpha_{Tumor} \neq \alpha_{Normal} \tag{2}$$

$$\text{where} \quad \alpha = \text{Average Expressiveness}$$

We performed t-tests on all tissues which had data for both tumor and normal, and there were atleast 20 samples (20 samples allowed use to be confident in our gene ID average expression). Table 2 shows the results from analyzing the activity (expressiveness) between the tumor tissue and normal tissue. The Tissue column shows which tissue was being evaluated, the n column shows the number of gene IDs that were used in the analysis, the p-value column shows the level of significance, the Mean of Difference column shows the average difference between the two, and the .95 Confidence Interval column shows the 95% confidence interval for the difference between the two, meaning if an infinite amount of samples were taken for analysis, 95% of them would contain the true difference between the two. To show that there is significant difference between two, the 95% confidence interval must not contain the value 0.

From Table 2, we can see that 8 of the tissues are statistically significant. Tissues COAD, HNSC, LIHC, and STAD all have a negative mean of difference values. This would suggest that the normal tissue genes are more expressive than the tumor tissue genes. This means genes that are tumor suppressors are more active in these tissues. On the other hand, tissues KICH, KIRP, THCA, and UCEC all have positive mean of difference values. This would suggest that the tumor tissue genes are more expressive than the normal tissue genes. This means genes that are oncogenes are more active in these tissues.

Discussion

Challenges in Data Preparation

Tumor tissues that have high number of samples, such as BRCA (n=1,111) and HNSC (n=500), set a challenge when joining them into a single text file. To overcome this, we group sample, so instead of 1,111 sample files, incase of BRCA, we have only six files. Since we limited this step in preparing the data in shell commands, we could try using R and bind these files into one file using an alternative approach.

Moreover, due to large number of genes per sample (n=60,483), analyzing the data becomes a challenge. We initially attempted to visualize all genes in a single heatmap, but failed because of the data size.

Table 2: Paired Sample t-Test of the Difference Between Tumor and Normal Tissue Gene Expression.

	Tissue	n	t Statistic	p-value	Mean of Difference	.95 Confidence Interval
1	BLCA	60,483	0.3723	0.7097	1260.16	(-5374.02, 7894.34)
2	BRCA	60,483	-0.5383	0.5903	-1663.95	(-7722.14, 4394.24)
3	COAD	60,483	-2.9470	0.0032	-14363.75	(-23916.84, -4810.65)
4	HNSC	60,483	-5.0159	0.0000	-51655.27	(-71839.91, -31470.62)
5	KICH	60,483	3.8394	0.0001	109173.65	(53440.82, 164906.48)
6	KIRC	60,483	0.1603	0.8726	725.73	(-8145.42, 9596.88)
7	KIRP	60,483	4.3660	0.0000	21988.02	(12117.12, 31858.93)
8	LIHC	60,483	-4.6364	0.0000	-138267.41	(-196718.89, -79815.93)
9	LUAD	60,483	0.2368	0.8128	1350.79	(-9831.59, 12533.17)
10	LUSC	60,483	-0.8799	0.3789	-4962.00	(-16015.15, 6091.15)
11	PRAD	60,483	0.4279	0.6687	3242.19	(-11608.54, 18092.92)
12	STAD	60,483	-7.4616	0.0000	-79646.53	(-100567.87, -58725.19)
13	THCA	60,483	4.0427	0.0001	16474.15	(8487.09, 24461.21)
14	UCEC	60,483	6.4985	0.0000	27033.27	(18879.79, 35186.75)

Nature of Normal Tissue Samples

As shown in heatmaps in Results section, some of normal samples have unique expression compared to other normal sample expression. This draws attention to the origin of normal tissue sample, which are defined as adjacent non-tumor tissue samples. This means that these samples are extracted from tissues that have tumor cells, but do not exhibit the characteristic of proliferation as do cancerous cells. This observation is a subject for study in recent years, suggesting that the nature of adjacent non-tumor is different than a real tumor tissue[23].

Conclusion

In this report, we downloaded and analyzed samples from TCGA (32 tissues) and used R packages to visualize gene expression of 6 genes related to cancer. Then, we performed paired t-test to find significant difference between normal and tumor tissue samples for all genes. From these visualizations, we found that MYC is generally highly expressed regardless of nature of tissue, normal or tumor. BRCA1 and BRCA2 have slight to no expression any tissue. ABL1, KRAS, and TP53 have changed expression depending on tissue type. The increase of gene expression is not limited to tumor tissue, in some cases we identified normal tissues that have higher expression than tumor. That draws the attention to the fact that changes in expression between normal and tumor tissues are important,not which tissue has the increased or decreased expression. Additionally, we confirmed the significant difference in expression statistically using paired t-test, which confirmed some tissues having significant difference between normal and tumor tissues.

Future Directions

We explained the concern about the normal tissue samples in TCGA being accurate representation of a normal tissue, thus, as a next step, we would consider using true normal samples from GTEx[24], which is a widely used source providing normal tissue samples, but we have to consider that the data is not publicly available. Moreover, we referred in the background section about the effect mutations have on gene function.

[23] https://www.nature.com/articles/s41467-017-01027-z
[24] https://gtexportal.org/home/

Mutation analysis is often coupled with gene expression analysis, and it is a next step we could perform to further advance this report. Lastly, we could conduct furter research on individual gene expression, however, this would require paired sampling, which is not easily available, and in some cases impossible to retrieve.

Appendix

Appendix A