Exploratory data analysis of genomic data from breast cancer studies

Carlos Back and Urminder Singh

December 13, 2018

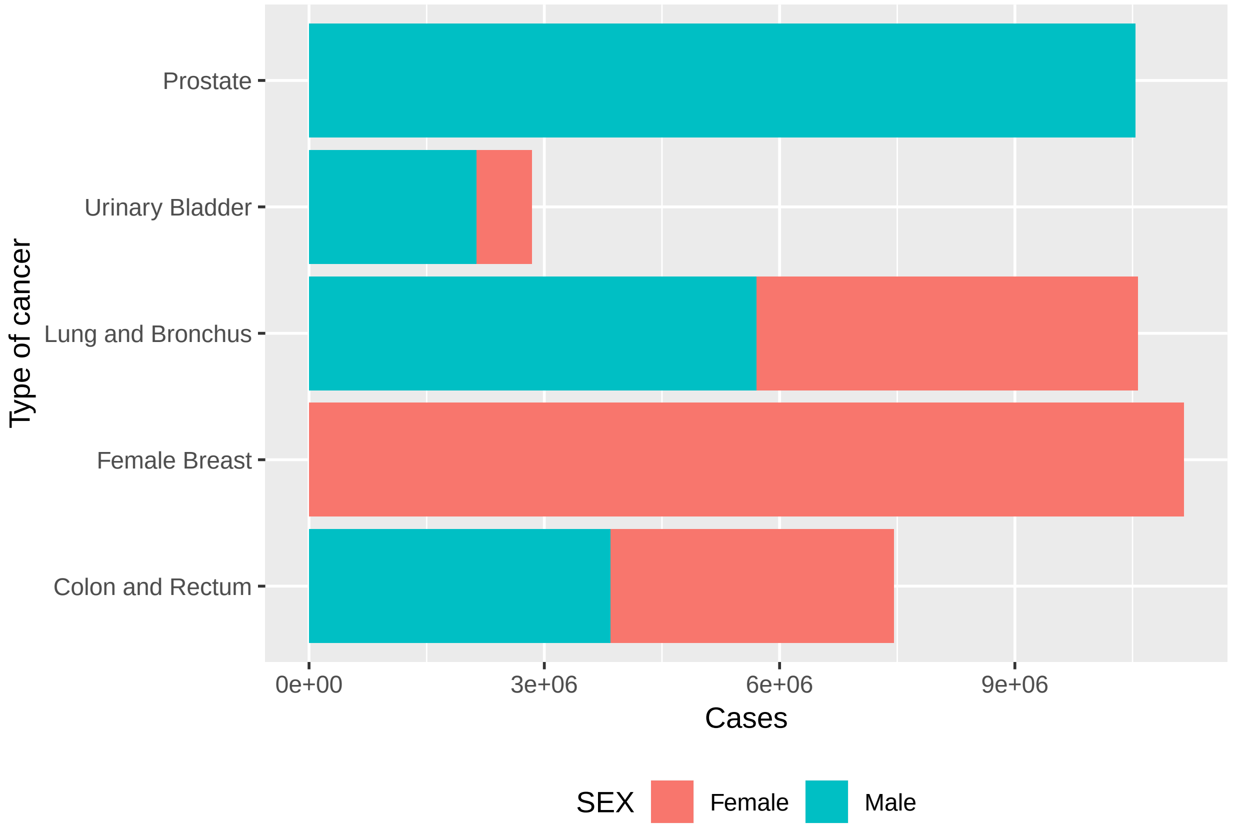
Introduction

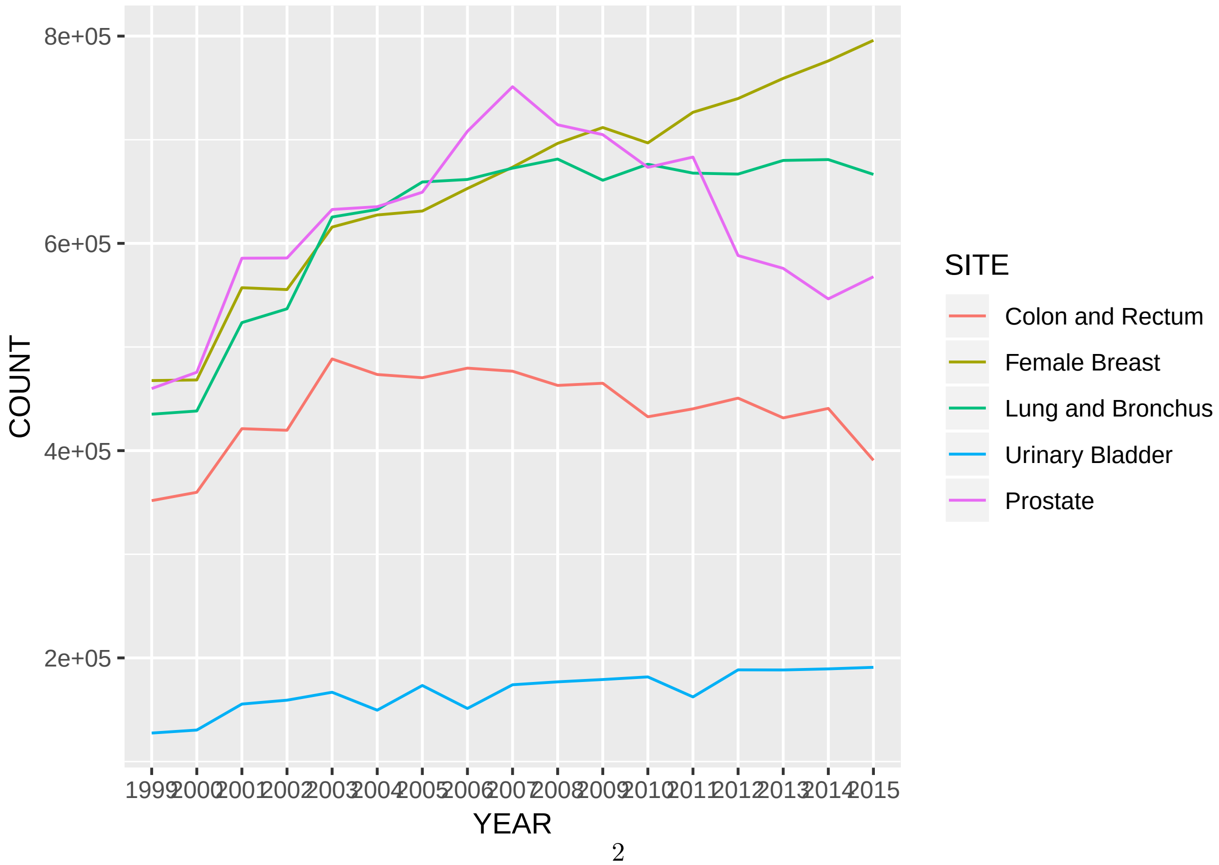
Cancer is a serious and potentially life-threatning disease. Each year the number of people diagnosed with cancer are increasing and this number is expected to increase rapidly in coming years. In 2018, an estimated 1,735,350 new cases of cancer will be diagnosed and 609,640 people will die from the disease in the United States alone ("Cancer Statistics," n.d.). By the year 2030 we expect to see 23.6 million cancer cases per year worldwide. As of 2018, breast cancer is the most common cancer after lung and prostate cancer("Cancer Statistics," n.d.). In order to provide effective treatment for cancer, researchers must fully understand this disease at the genomic level. To do so, The Cancer Genome Atlas (TCGA) project started in 2005 which has a catalogue of genomic data collected from cancer patients (CancerGenomeAtlasNetwork and others 2012; Wang et al. 2018). To further understand the prevalence of cancer among different populations, resources like United States Cancer Statistics (USCS), Centers for Disease Control and Prevention (CDC), National Program of Cancer Registries (NPCR), National Center for Health Statistics(NCHS), and National Cancer Institute's (NCI) keep a comprehensive record of cancer cases and associated attributes.

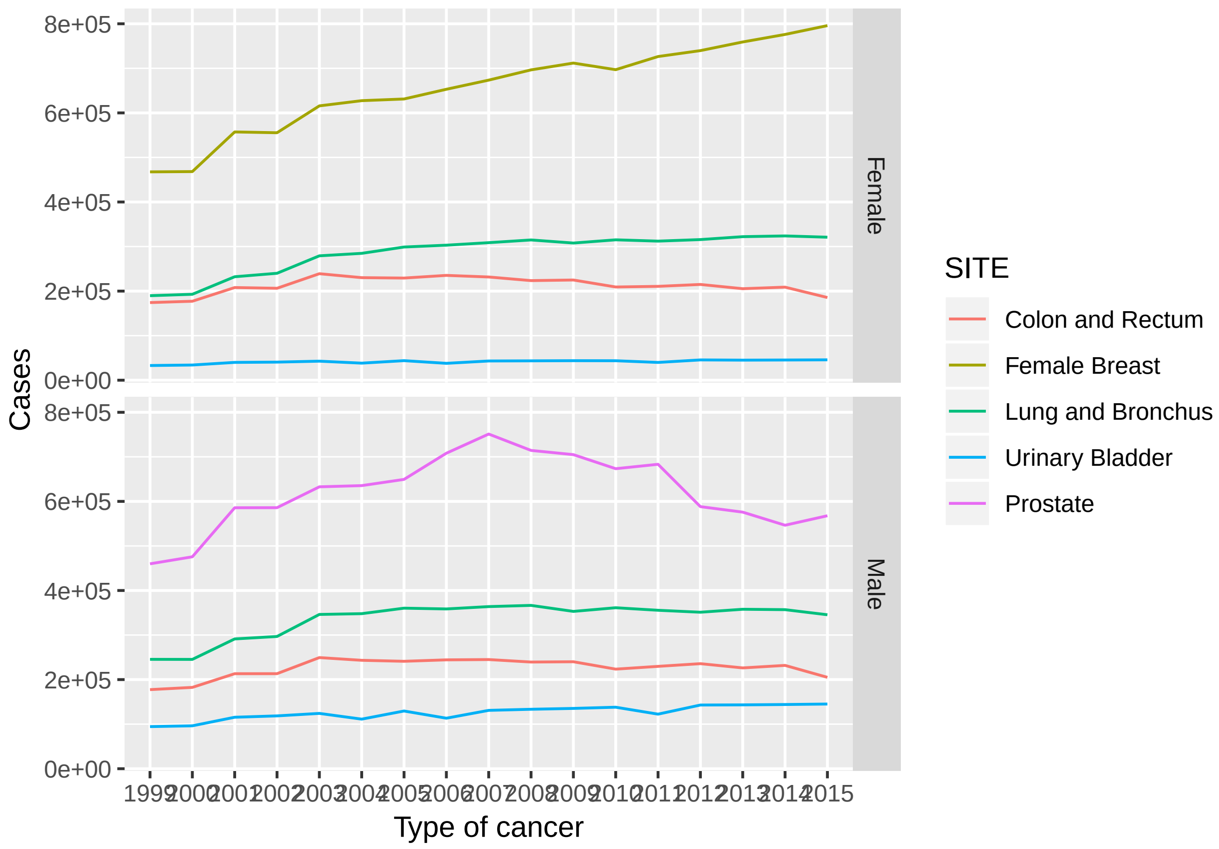
In this project, we analysed data from resources USCS, NPCR, NCHS, NCI and TCGA. In part one, we summarized cancer statistics from the CDC data. This dataset had more than a million cancer cases and cancer deaths for the whole U.S. population from 1999 to 2015, along with information about patients state of residence, states' population, type of cancer, gender and race. We looked at the different trends of various cancer cases to see if they correlate with any other factor like gender, age, geographic location etc. In particular we focused on the incidences of breast cancer accross the United States.

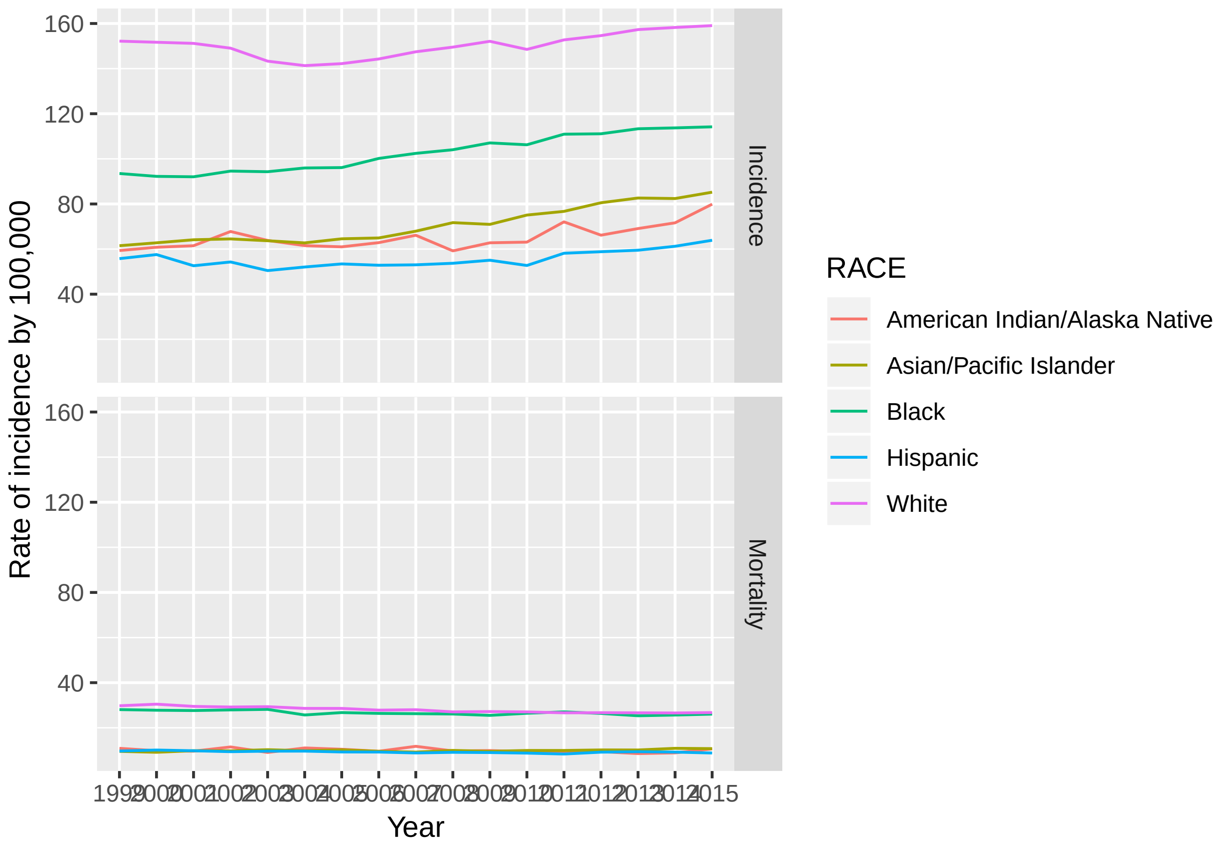
Next, in part two, we collected breast cancer (BRCA) data from TCGA in order to find associations between various clinical attributes, gene mutations, and gene expression. We found that the sets of highly mutated genes in BRCA could be different under different clinical attributes. We observed that the most common type of mutation in all these sets is missense mutation. We also looked at the expression patterns of BRCA associated genes in tumor and normal samples. It appears that some of the higly mutated genes change overall level of expression. Finally, we visually compared mean expression levels of these genes under five different cancers (breast, colon, liver, lung and stomach)

Part I

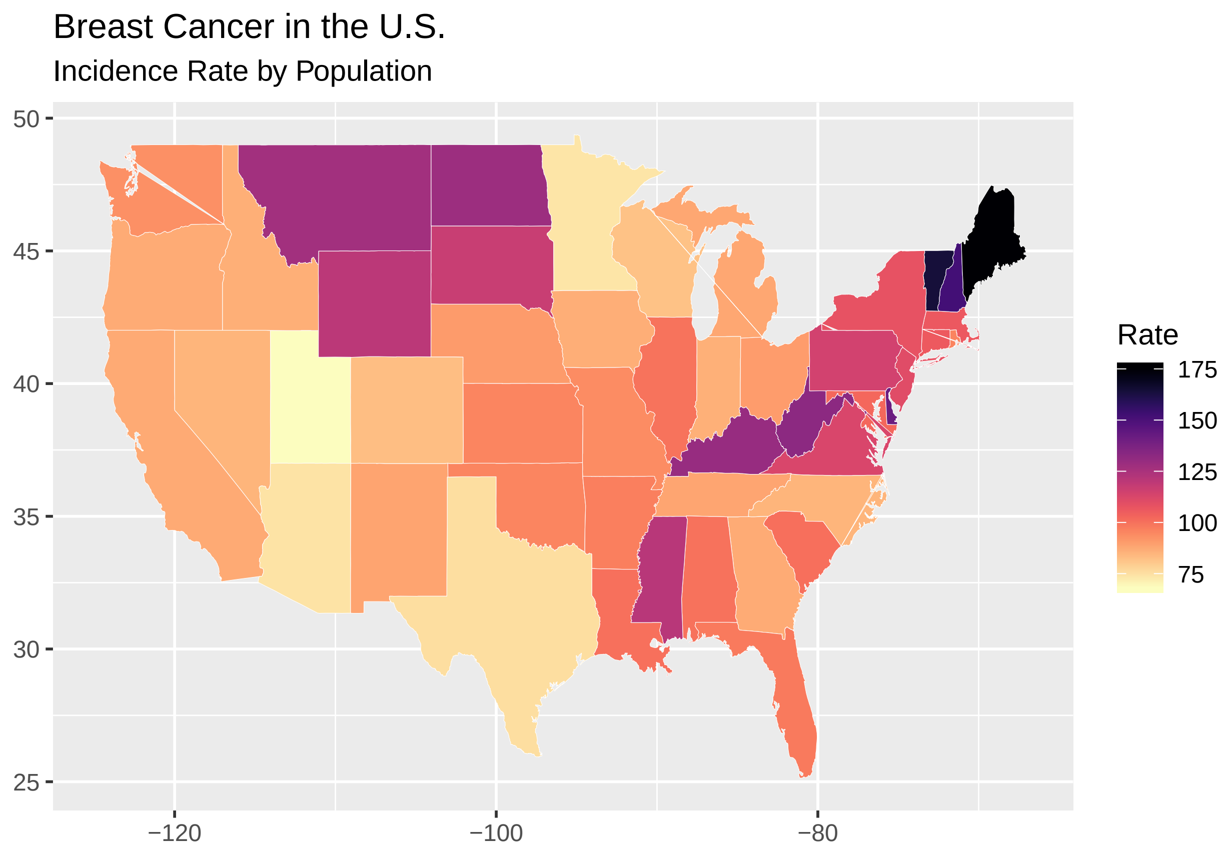

Prostate
Urinary Bladder
Lung and Bronchus
Female Breast
Colon and Rectum
Type of cancer
0e+00
3e+06
6e+06
9e+06
Cases
SEX
🟥 Female
🟦 Male

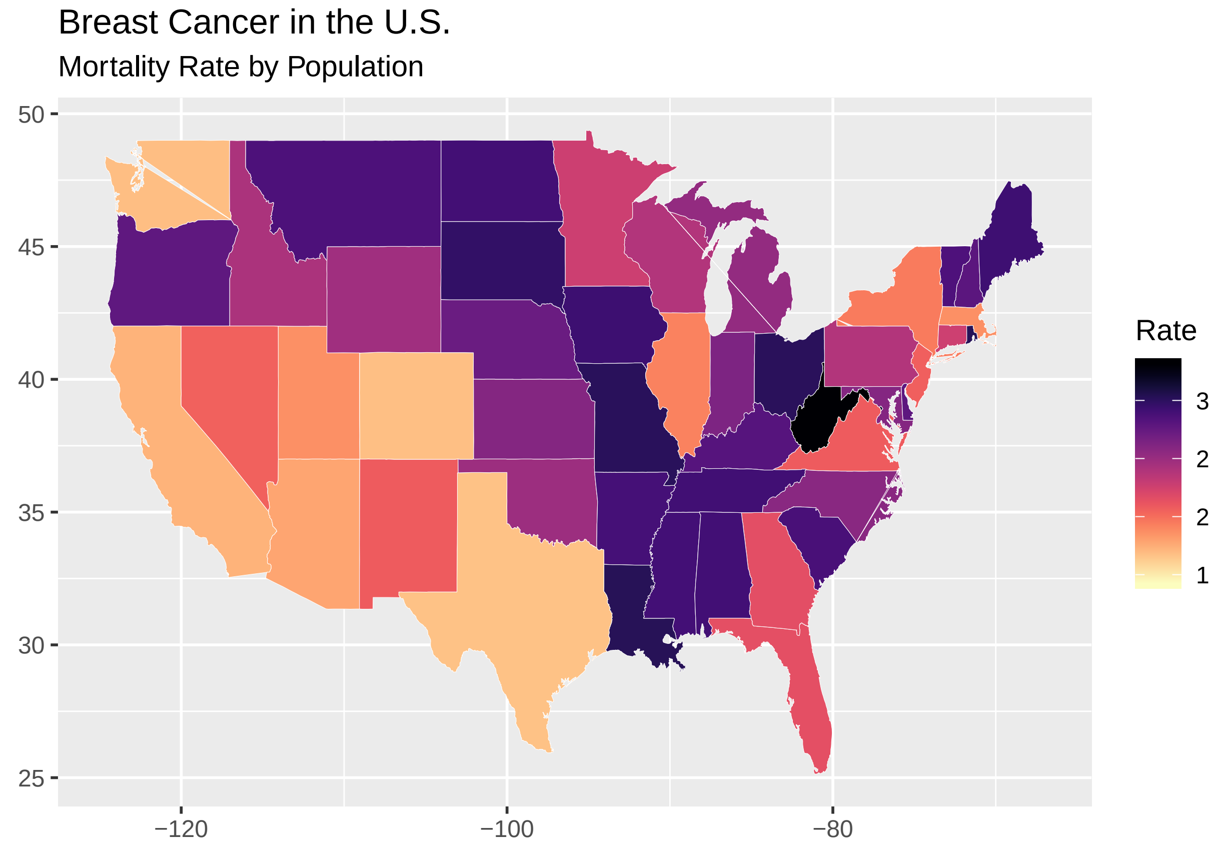

8e+05
6e+05
4e+05
2e+05
COUNT
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
YEAR
SITE
Colon and Rectum
Female Breast
Lung and Bronchus
Urinary Bladder
Prostate


Cases
8e+05
6e+05
4e+05
2e+05
0e+00
Female
Male
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Type of cancer
SITE
Colon and Rectum
Female Breast
Lung and Bronchus
Urinary Bladder
Prostate



The results showed that breast cancer is the main cancer type in women and the main cancer type overall. More than that, the number of cases of breast cancer have been increasing along the years, contrary the any other top 5 type of cancer in the country (urinary bladder, colon and rectum, prostate, lung and bronchus). Prostate cancer cases, the main type of cancer in men, is only the third type of cancer in general, behind lung and bronchus. Incidences of breast cancer are higher in white women, followed by black and asian women. In this data, hispanic women have the lower incidences of cancer. Regarding mortality, white and black women have similar rates, about 30 per 100,000 people. All the other races have similar mortality of around 10 per 100,000 people.


Breast Cancer in the U.S.
Incidence Rate by Population
50
45
40
35
30
25
−120
−100
−80
Rate
175
150
125
100
75



The distribution of cases among US states show that the incidence of breast cancer is higher in northern states such as Maine, Vermont, New Hampshire, Montana and North Dakota. Mississipi, Kentucky and West Virginia have a slightly fewer cases. The mortality by breast cancer is also high in northern states, but is even higher in Mississipi, West Virginia and Missouri.

All these findings shows that we should investigate breast cancer further to find better ways to diagnose and combat the disease.

Part II

Data and Methods

We used data from TCGA (CancerGenomeAtlasNetwork and others 2012). TCGA is a comprehensive repository of human cancer molecular and clinical data, TCGA database has collected clinical and molecular phenotypes of thousands of tumor patients across different tumor types. The TCGA dataset, contains:

- Clinical information about participants
- Metadata about the samples (e.g. the weight of a sample portion, etc.)
- Histopathology slide images from sample portions
- Molecular information derived from the samples (e.g. mRNA/miRNA expression, protein expression, copy number, etc.)

We downloaded three types of data i.e. clinical data, mutation data and gene expression data. The clinical data contains information about each patient who participated in the study. It includes variables like gender, age, race, cancer type etc. The mutation data contains the molecular information about genes for each cancer patient. It describes various mutations found in the genes of the cancer patients. This includes name of the

genes mutated, mutation classification, mutation type etc.
Finally, we downloaded the gene expression data which contains the level of gene expression estimated from RNA samples of each patient. TCGA has tumor samples for each patient and where possible, non-tumor/normal samples are also present. Thus, we have gene expression data for both tumor and normal tissues, although the number of normal samples are much less. Instead of directly using gene expression data from TCGA, we downloaded data from Wang et. al. 2018 (Wang et al. 2018). This data was created by processing the TCGA data. This data was normalized and corrected for batch effects and thus is comparable accross different samples.

Downloading clinical data from TCGA

We used the TCGABiolinks (Colaprico et al. 2015) package to download the TCGA clinical data for BRCA studies. The function *GDCquery_clinic* was used to download the data. We wrote additional functions to clean the data and arrange it into a dataframe.

Downloading mutation data from TCGA

Next, we downloaded mutation data from BRCA studies using TCGABiolinks' *GDCquery_Maf* function (Colaprico et al. 2015). After getting the mutation data, we joined the clinical table with the mutation table to have clinical information for each mutation in the mutation table.

This resulted in a data frame of dimensions 93612, 145. This data frame contains clinical information and coressponding mutation information for each BRCA patient.

Downloading gene expression data

Next, we downloaded the gene expression data for the BRCA studies from (Wang et al. 2018).

These datasets contains the gene expression patterns over the tumor samples and normal samples. Each row corresponds to a gene and each column corresponds to a sample. The data dimentions for tumor samples was 19738, 982 and for normal samples was 19738, 110.

Analysis

We did some exploratory data analysis on our data in order to find associations between various clinical attributes, gene mutations, and gene expression.

Highly mutated genes

First, we looked at the mutation data to find which genes are highly mutated in BRCA and what types of different mutations are present in the data. We used the package *maftools*(Mayakonda et al. 2018) to plot a summary of the BRCA mutation dataset.

Figure 1 shows the summary plot generated by maftools. From the plot we can see that majority of mutations are of type missense mutation. Missense mutations can change the amino acid sequence of the protein coded by the mutated gene making an unstable protein product. After missense mutations, we see a lot of nonsense mutations. A gene with nonsense mutation produces a truncated protein which is non-functional. We also found that, most variant types are Single Neucleotide Polymorphisms (SNPs). These variants are a result of mutation of only a single neucleotide in the DNA. SNPs can have deletarious effects if they result in missense or nonsense mutations. We also found the top 10 genes with highest number of mutations. Out of these genes PIK3CA, TP53, CDH1 and PTEN are already known to be associated with cancer. TP53, for example,

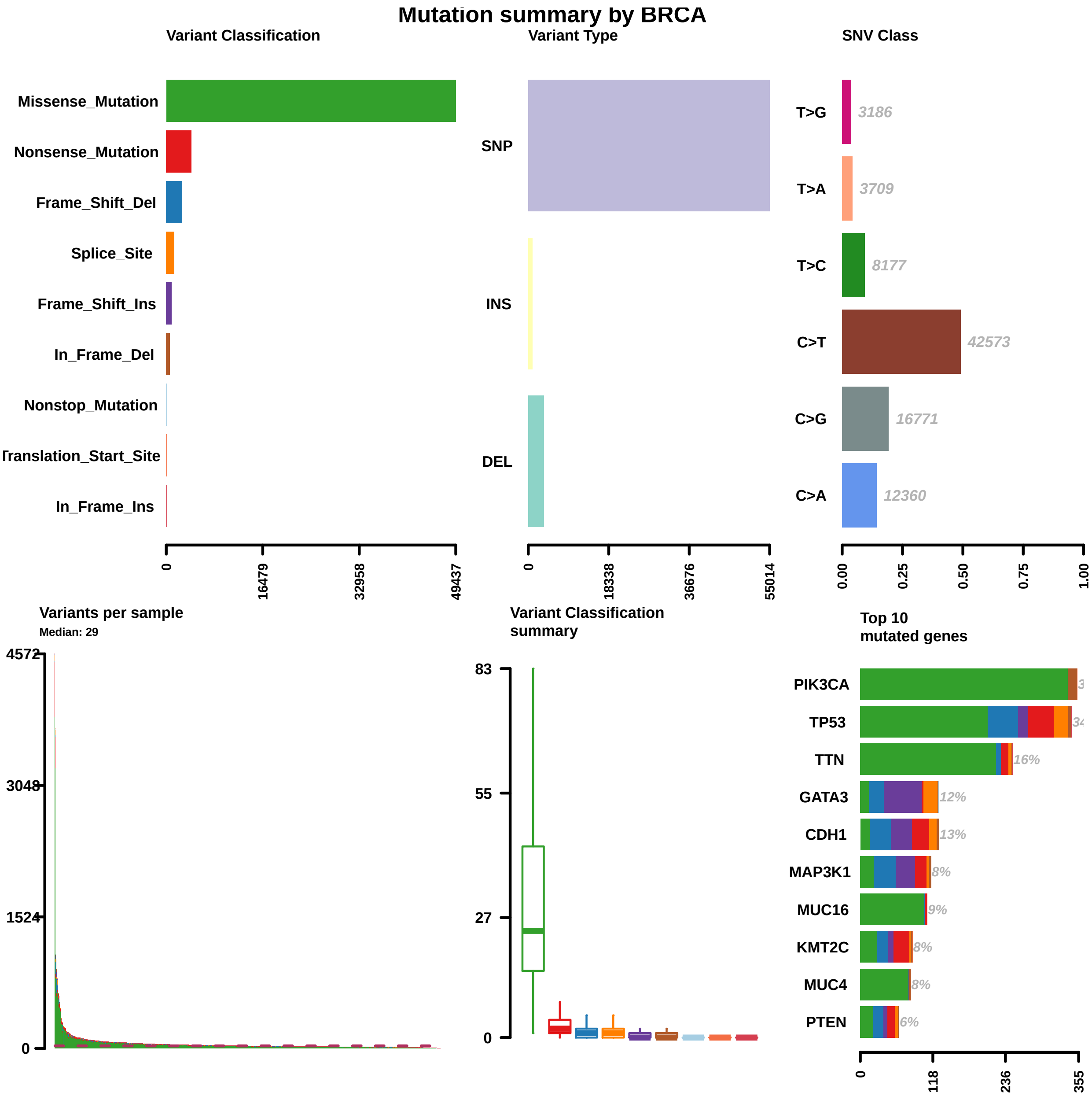


Figure 1: BRCA mutations summary

codes for a tumor suppressor protein which is essential for regulation of cell cycle. Any mutation in this gene can cause changes in the TP53 protein which may not function as a tumor suppressor anymore thus causing tumor.

Next, we looked at how the gene mutations differ for different categories. For this task we wrote our own function to plot summary as barchart. First, we looked at which genes were mutated in samples from different race.

Figure 2 shows the top mutated genes for each race. PIK3CA and TP53 genes are the most mutated for each category. There are unique genes too for each category e.g. MUC16 for white, DNM1p47 for black/african american and C5orf42 for asian populations. We also found majority of mutations are of type missense mutation for each category.

Then, we looked at how genes are mutated for the two most common breast cancer types i.e. ductal carcinoma and lobular carcinoma.

Figure 3 shows the top mutated genes for ductal carcinoma and lobular carcinoma. Interestingly, the gene TP53 has lesser mutations in lobular carcinoma that in ductal carcinoma.

From the analysis above we noticed that BRCA1 and BRCA2 are not in the set of top mutated genes. These genes are known to be associated with cancer and a mutation in these genes can increase the risk of breast and ovarian cancer in women. To find out how many mutations belonged to BRCA1 and BRCA2 in our dataset, we plotted a barchart (Figure 3).

Figure 4 shows the number of mutations (log scale) in BRCA1/2 and all other genes combined and the type of mutation for BRCA1 and BRCA2 genes. We found that mutations in these two genes were very small as compared to other genes. Missense mutation is most common in these two genes too.

Expression patterns of genes

After looking at various mutated genes, we then looked at the expression pattern of these genes. We looked at the distributions of gene expression for tumor and normal tissues. First, we took the set of highly mutated genes in BRCA and compared their expression in tumor and normal samples.

Figure 5 shows the boxplots for the ten genes. We can see visually that expression of genes TP53, CDH1, and MAP3K1 looks very simillar and unchanged in the two categories. While the expression of genes PIK3CA, TTN, GATA3 and MUC4 looks different under the two conditions. This might give some insight on how these genes play a role in cancer. A statistical test is still required to assess the significance of these differences.

After that, we compiled a list of genes are known to be differentially expressed in cancer (Li, Sun, and Wang 2017). Genes which are differntially expressed could be important biomarkers and give us more insights about the most affected biological pathways in cancer.

Figure 6 shows violin plots for the cancer regulated genes. From the plots we can see that all the genes on the left panel appear to be downregulated and all the genes in the right panel seems to be up regulated in BRCA tumor samples as compared to normal samples. This result was expected and also mentioned in Li, Sun, and Wang (2017).

Comparing expression of mutated BRCA genes with other cancer types

Finally, we wanted to compare the expression of highly mutated genes in BRCA with other cancers from different tissues. To do this we downloaded more expression data for four tumors from different tissues which were colon, liver, lung, and stomach. Then we looked at the mean expression of each of these genes accross different tissues. To visualize the mean expression over different tissues we used the package gganatogram (Maag 2018).

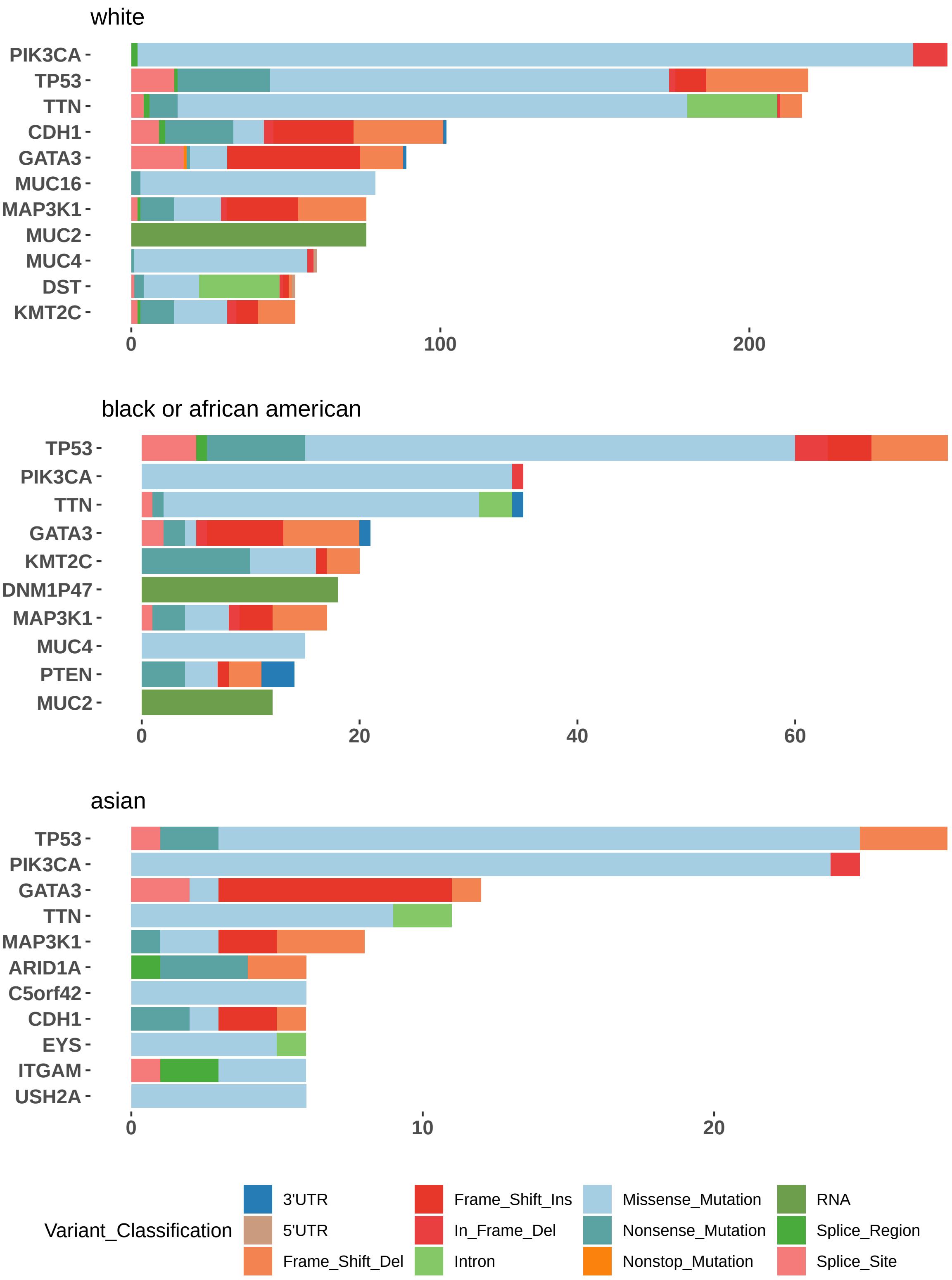


Figure 2: Top mutated genes by race

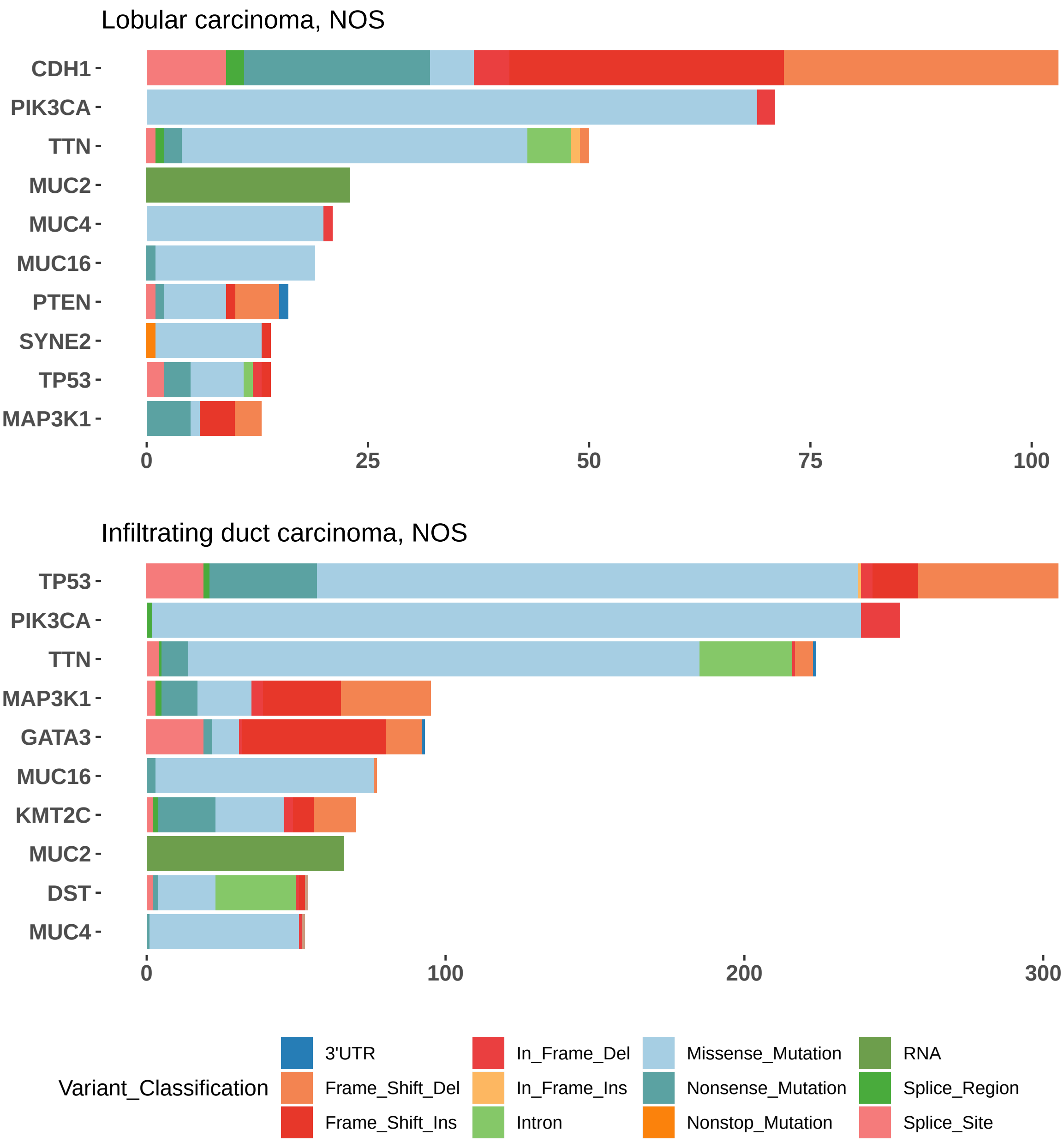


Figure 3: Top mutated genes by cancer type

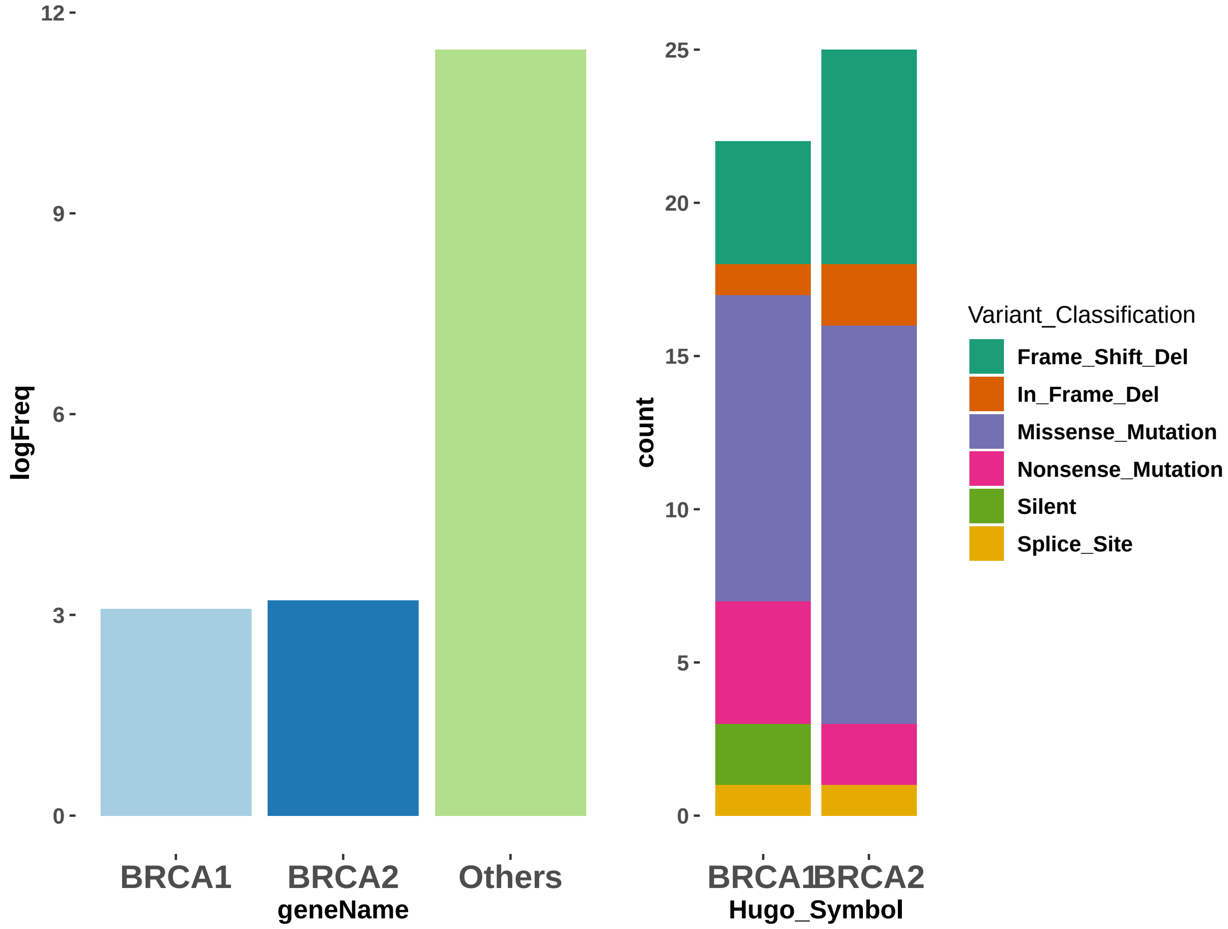


Figure 4: BRCA1/BRCA2 mutation statistics

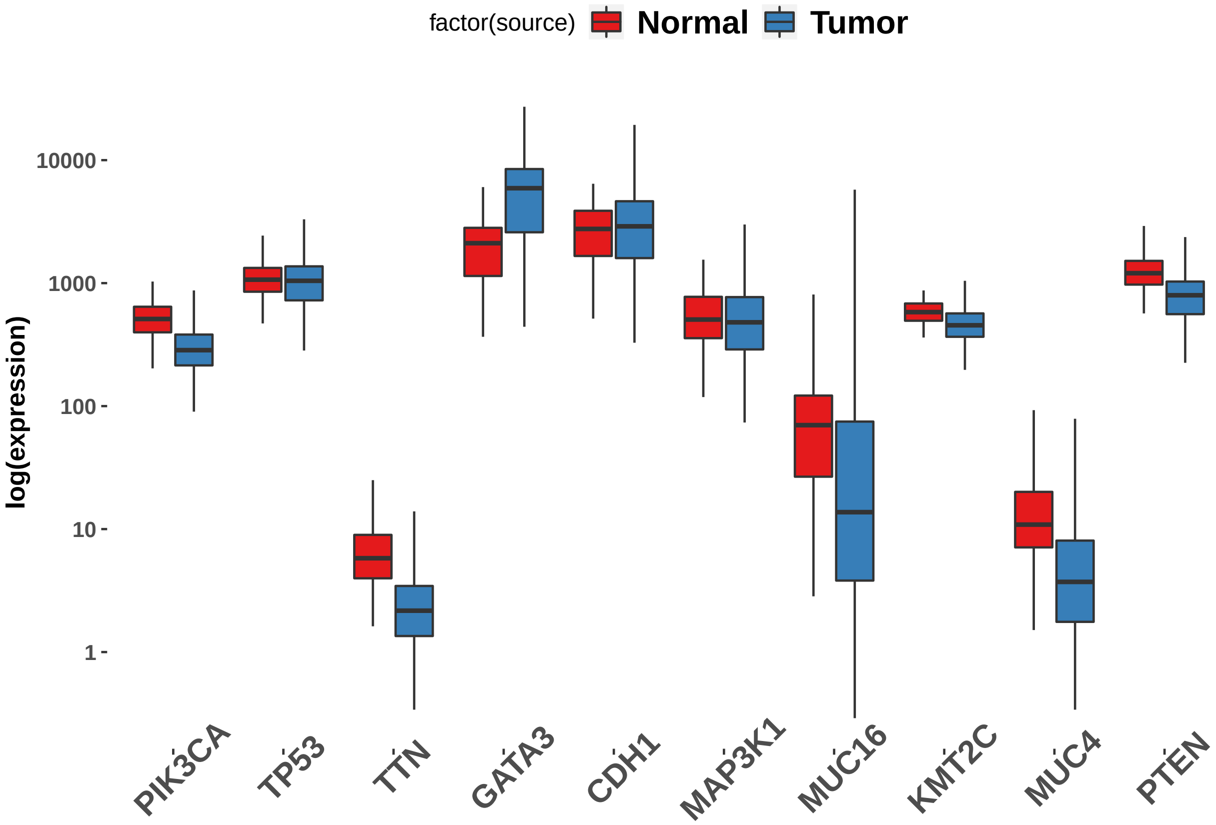


Figure 5: Expression of highly mutated genes in tumor and normal samples

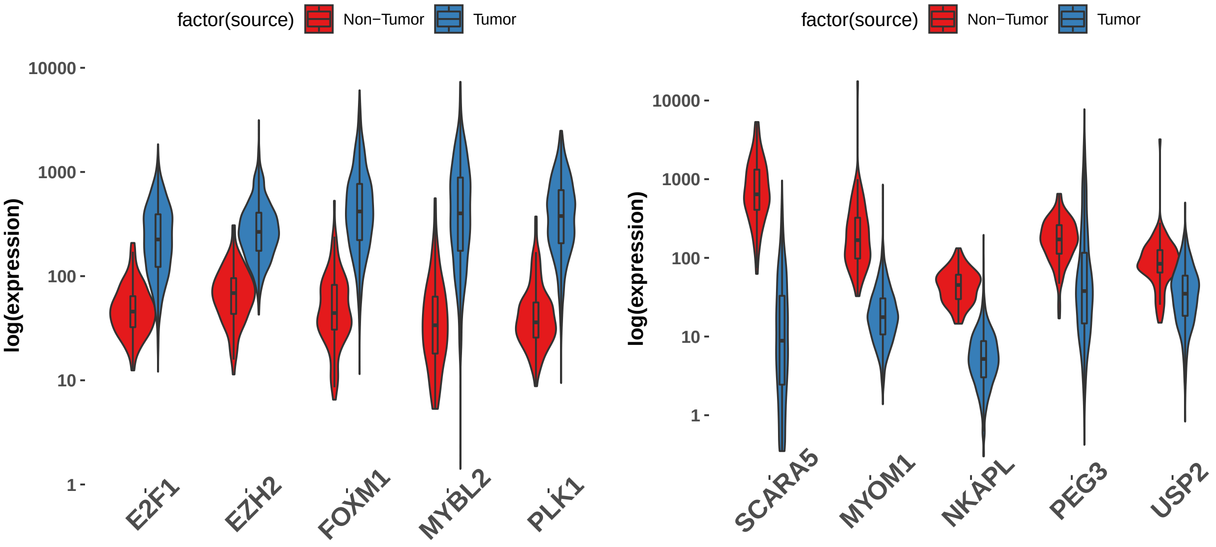


Figure 6: Expression of genes known to be regulated in tumor and normal samples

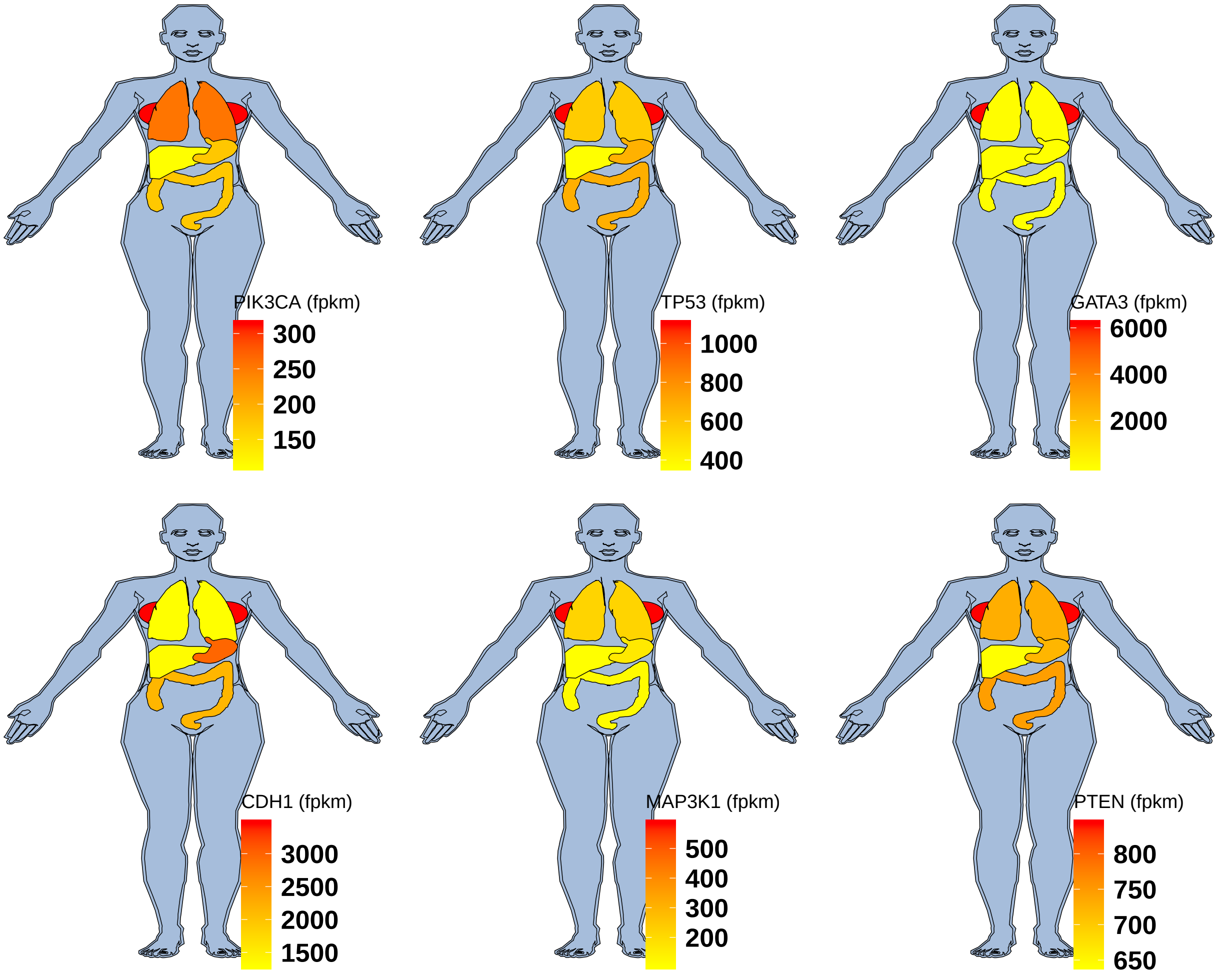


Figure 7: Mean expression of highly mutated BRCA genes in tumors from different tissues

Figure 7 shows the mean expression of the genes in tumors from different tissues. We found that BRCA samples had higher mean expression as compared with other tissues. A definite conclusion requires a statistical test though.

Conclusion

From looking at the cancer statistics we can say that cancer is a growing problem. The statistics may healp researchers, doctors and policy makers

In order to find more effective methods to treat cancer, we must first fully understand the disease itself. A number of mutations in one's DNA can lead to cancer. We saw that in BRCA few genes get mutated most often like PIK3CA and TP53. We already know that mutations in these genes results in an altered protein which is not capable of normal biological function. We also need to better understand how these mutation affects the cellular pathways. Expression patterns of these genes may reveal if they are getting activated or suppressed due to cancer. Genes which are differentially regulated in cancer may prove to be important biomarkers. Studying functions of these genes may lead to new ways of stopping cancer in the body.

Session information

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] grid      parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] RColorBrewer_1.1-2  gridExtra_2.3       gganatogram_1.1.1
##  [4] ggpolypath_0.1.0    maftools_1.8.0      Biobase_2.42.0
##  [7] BiocGenerics_0.28.0 plyr_1.8.4          data.table_1.11.8
## [10] DT_0.5              TCGAbiolinks_2.10.0 bindrcpp_0.2.2
## [13] maps_3.3.0          viridis_0.5.1       viridisLite_0.3.0
## [16] ggplot2_3.1.0       dplyr_0.7.8         readr_1.2.1
##
## loaded via a namespace (and not attached):
##   [1] backports_1.1.2             circlize_0.4.5
##   [3] aroma.light_3.12.0          NMF_0.21.0
##   [5] selectr_0.4-1               ConsensusClusterPlus_1.46.0
##   [7] lazyeval_0.2.1              splines_3.5.1
##   [9] BiocParallel_1.16.2         GenomeInfoDb_1.18.1
##  [11] gridBase_0.4-7              sva_3.30.0
##  [13] digest_0.6.18               foreach_1.4.4
##  [15] htmltools_0.3.6             magrittr_1.5
##  [17] memoise_1.1.0               BSgenome_1.50.0
##  [19] cluster_2.0.7-1             doParallel_1.0.14
##  [21] limma_3.38.2                ComplexHeatmap_1.20.0
##  [23] Biostrings_2.50.1           annotate_1.60.0
##  [25] wordcloud_2.6               matrixStats_0.54.0
##  [27] R.utils_2.7.0               prettyunits_1.0.2
##  [29] colorspace_1.3-2            blob_1.1.1
##  [31] rvest_0.3.2                 ggrepel_0.8.0
##  [33] crayon_1.3.4                RCurl_1.95-4.11
##  [35] jsonlite_1.5                genefilter_1.64.0
##  [37] bindr_0.1.1                 survival_2.42-3
##  [39] zoo_1.8-4                   iterators_1.0.10
##  [41] glue_1.3.0                  survminer_0.4.3
##  [43] registry_0.5                gtable_0.2.0
##  [45] zlibbioc_1.28.0             XVector_0.22.0
##  [47] GetoptLong_0.1.7            DelayedArray_0.8.0
##  [49] shape_1.4.4                 scales_1.0.0
##  [51] DESeq_1.34.0                rngtools_1.3.1
##  [53] DBI_1.0.0                   edgeR_3.24.0
##  [55] bibtex_0.4.2                ggthemes_4.0.1
##  [57] Rcpp_1.0.0                  xtable_1.8-3
##  [59] progress_1.2.0              cmprsk_2.2-7
##  [61] mclust_5.4.2                bit_1.1-14
##  [63] matlab_1.0.2                km.ci_0.5-2
##  [65] stats4_3.5.1                htmlwidgets_1.3
##  [67] httr_1.3.1                  pkgconfig_2.0.2
##  [69] XML_3.98-1.16               R.methodsS3_1.7.1
##  [71] locfit_1.5-9.1              tidyselect_0.2.5
##  [73] labeling_0.3                rlang_0.3.0.1
##  [75] reshape2_1.4.3              AnnotationDbi_1.44.0
##  [77] munsell_0.5.0               tools_3.5.1
##  [79] downloader_0.4              RSQLite_2.1.1
##  [81] broom_0.5.0                 evaluate_0.12
##  [83] stringr_1.3.1               yaml_2.2.0
##  [85] knitr_1.20                  bit64_0.9-7
##  [87] survMisc_0.5.5              purrr_0.2.5
##  [89] EDASeq_2.16.0               nlme_3.1-137
##  [91] R.oo_1.22.0                 xml2_1.2.0
##  [93] biomaRt_2.38.0              compiler_3.5.1
##  [95] curl_3.2                    tibble_1.4.2
##  [97] geneplotter_1.60.0          stringi_1.2.4
##  [99] highr_0.7                   GenomicFeatures_1.34.1
## [101] lattice_0.20-35             Matrix_1.2-14
## [103] KMsurv_0.1-5                pillar_1.3.0
## [105] GlobalOptions_0.1.0         cowplot_0.9.3
## [107] bitops_1.0-6                rtracklayer_1.42.1
## [109] GenomicRanges_1.34.0        R6_2.3.0
## [111] latticeExtra_0.6-28         hwriter_1.3.2
## [113] ShortRead_1.40.0            IRanges_2.16.0
## [115] codetools_0.2-15            assertthat_0.2.0
## [117] SummarizedExperiment_1.12.0 pkgmaker_0.27
## [119] rprojroot_1.3-2             rjson_0.2.20
## [121] withr_2.1.2                  GenomicAlignments_1.18.0
## [123] Rsamtools_1.34.0             S4Vectors_0.20.1
## [125] GenomeInfoDbData_1.2.0       mgcv_1.8-24
## [127] hms_0.4.2                    tidyr_0.8.2
## [129] rmarkdown_1.10               ggpubr_0.2
```

References

"Cancer Statistics." n.d. *National Cancer Institute*. https://www.cancer.gov/about-cancer/understanding/statistics.

CancerGenomeAtlasNetwork, and others. 2012. "Comprehensive Molecular Portraits of Human Breast Tumours." *Nature* 490 (7418). Nature Publishing Group:61.

Colaprico, Antonio, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, et al. 2015. "TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of Tcga Data." *Nucleic Acids Research* 44 (8). Oxford University Press:e71–e71.

Li, Mengyuan, Qingrong Sun, and Xiaosheng Wang. 2017. "Transcriptional Landscape of Human Cancers." *Oncotarget* 8 (21). Impact Journals, LLC:34534.

Maag, Jesper LV. 2018. "Gganatogram: An R Package for Modular Visualisation of Anatograms and Tissues Based on Ggplot2." *F1000Research* 7. Faculty of 1000 Ltd.

Mayakonda, Anand, De-Chen Lin, Yassen Assenov, Christoph Plass, and H Phillip Koeffler. 2018. "Maftools: Efficient and Comprehensive Analysis of Somatic Variants in Cancer." *Genome Research* 28 (11). Cold Spring Harbor Lab:1747–56.

Wang, Qingguo, Joshua Armenia, Chao Zhang, Alexander V Penson, Ed Reznik, Liguo Zhang, Thais Minet, et al. 2018. "Unifying Cancer and Normal Rna Sequencing Data from Different Sources." *Scientific Data* 5. Nature Publishing Group:180061.