Exploratory data analysis of genomic data from breast cancer studies

Urminder Singh

December 8, 2018

Introduction

Our data consists of gene expression data from various samples collected to study breast cancer (Breast Invasive Carcinoma). These samples were collected from a number of different individuals. Then, the samples' RNA were extracted sequenced to get RNA-seq reads from the sample. Using the RNA-seq, the gene expression was estimated. However, we used the data from (Wang et. al. 2018) where the expression data was mapped and normalized again to remove batch effects, which made the dataset comparable across the samples. Apart from the gene expression data for each sample, we also have associated metadata for each sample. This metadata provides information about the individual, tissue, disease etc.associated with each sample.

Our dataset has expression estimates of around 20,000 genes across 1,092 samples. Out of these 1,092 samples, 982 had tumor and 110 were normal tissues without tumor. We also added the gene metadata to our dataset which describes the gene features and functions. This also includes mutation information of the genes. This information can provide crucial information about how different mutations in the genome can lead to cancer.

Overall, our dataset is GxS matrix where each row corresponds to a gene and each column corresponds to a RNA-seq sample. Additionally, we have metadata for all the rows and all the columns.

Data

We used data from TCGA. 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.)

First, we collected data clinical data from TCGA for breast cancer (BRCA) studies. This data contains Then we downloaded the mutation information for the BRCA studies. This data Finally, I collected the gene expression data from (Wang et. al. 2018). This data contains the estimated expression values of genes in the BRCA study. This data was normalized and batch corrected so it is comparable across different TCGA samples.

Data downloading and processing

This section describes how the data was downloaded and pre-processed.

Downloading clinical data from TCGA

We used the TCGABiolinks package to download TCGA data. The function *GDCquery_clinic* is used to download the data. We wrote additional functions to clean the data and arrange it into a dataframe.

The clinical data for BRCA studies is stored in a dataframe *brcaDF*.

Downloading mutation data from TCGA

Next, we downloaded mutation data using TCGABiolinks. 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.

We now have a data frame of dimensions 93612, 145. This data frame contains clinical information and coressponding mutation information for each BRCA patient. We will use this data to explore various…

Downloading gene expression data

Next, we downloaded the gene expression data for the BRCA studies.

We now have two expression datasets for tumor and normal or non-tumor tissues from BRCA studies. 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 expression data is 19738, 982 and for normal expression data is 19738, 110

Finally, our data consist of three data frames

1. *brcaMAF_MD* contains all the mutation information
2. *brcaexp_nontumor* contains the gene expression values over normal samples from TCGA
3. *brcaexp_tumor* contains the gene expression values over tumor samples from TCGA

Analysis

Mutations in 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* 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 mutation can change the amino acid sequence of the protein coded by the mutated gene making an unstable protein. After missense mutations we see a lot of nonsense mutations. A gene with nonsense mutation produces a truncated protein which is usually 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 reported. Out of these genes PIK3CA, TP53, CDH1 and PTEN are already known to be associated with cancer.

Next we looked at how the gene mutations differ for different categories. First, We looked at which genes were mutated for 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.

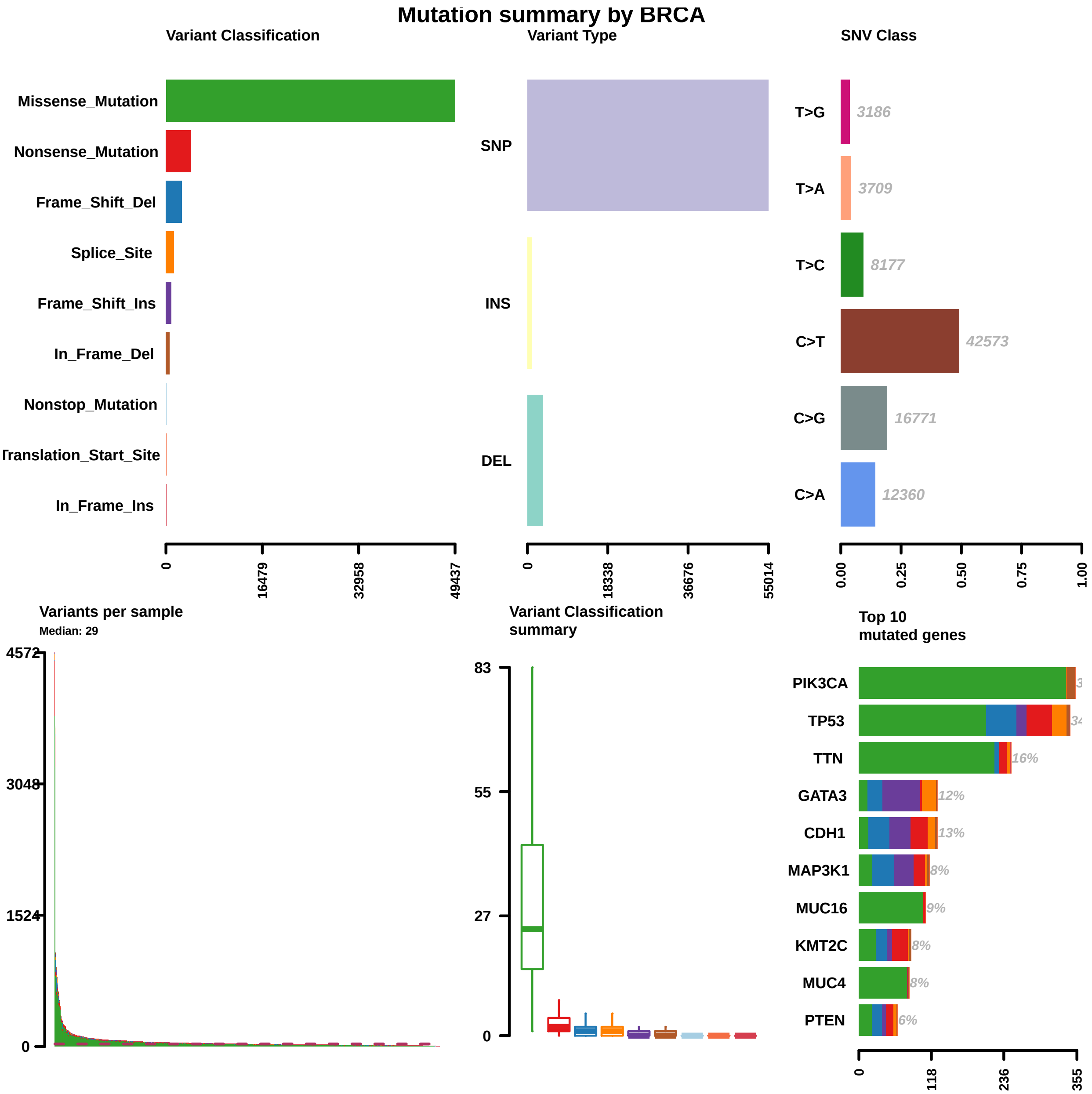

Figure 1: BRCA mutations summary

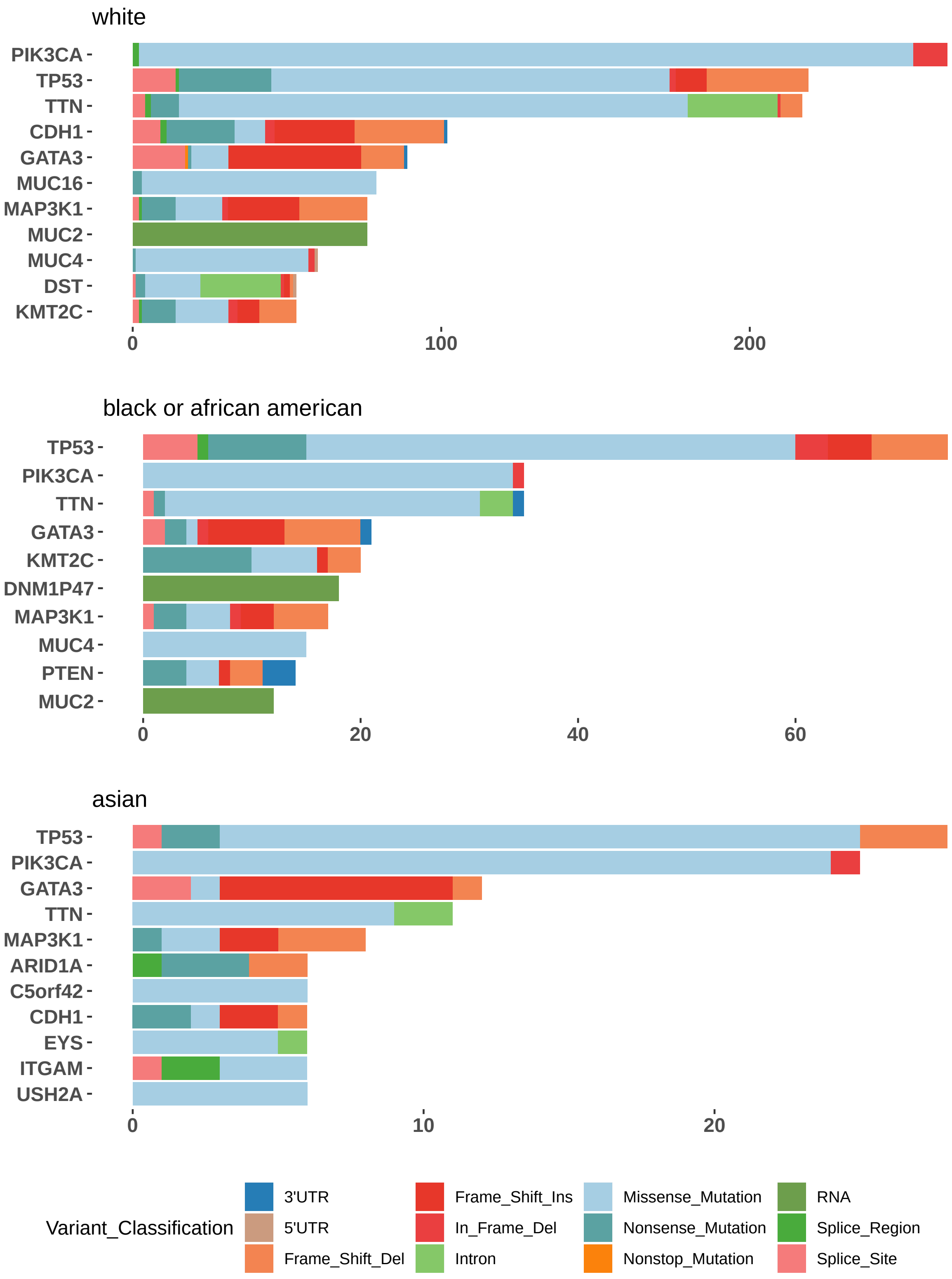

Figure 2: Top mutated genes by race

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. Gene expression is an estimate of the amount of RNA produced by the gene which in turn will be used to create specific proteins. We looked at the distribution of gene expressions for tumor and normal tissues. First, we took the set of highly mutated genes in BRCA and compared their expression for 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. 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. Genes which are differntially expressed could be important biomarkers and give us more insight on which biological pathways are most affected in cancer. Figure 6 shows violin plots for the selected genes. From the plot we can see that genes on the left panel are downregulated and genes in the right panel are all up regulated in BRCA samples as compared to normal samples.

Comparing expression of mutated BRCA genes with other cancer types

Finally, we wanted to compare expression of highly mutated genes in BRCA with other cancers from different tissues. To do this we downloaded more expression data for four tumor from different tissues which were liver,stomach,colon, and lung cancer.

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.

Conclusion

System 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:
```

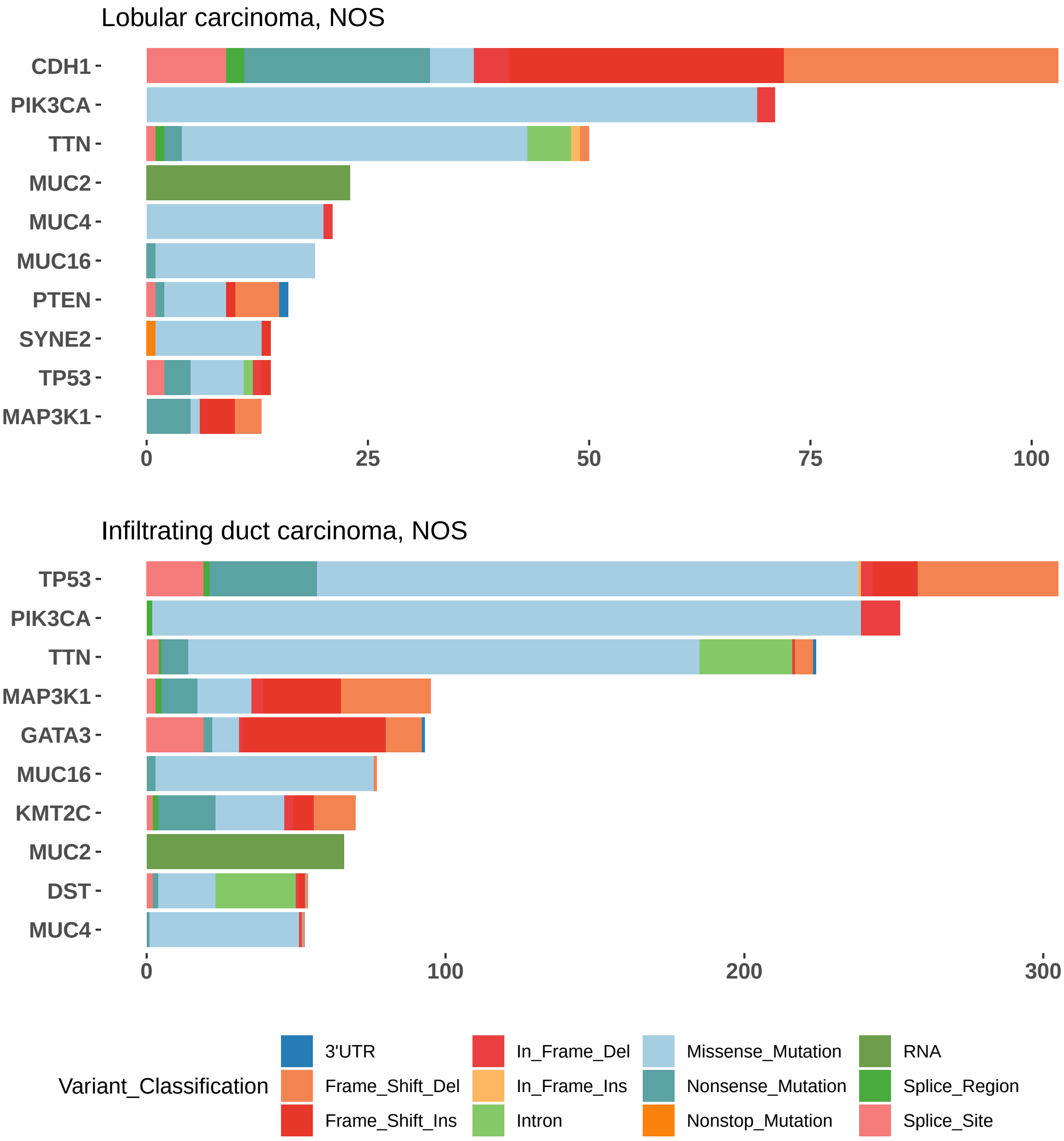

Figure 3: Top mutated genes by cancer type

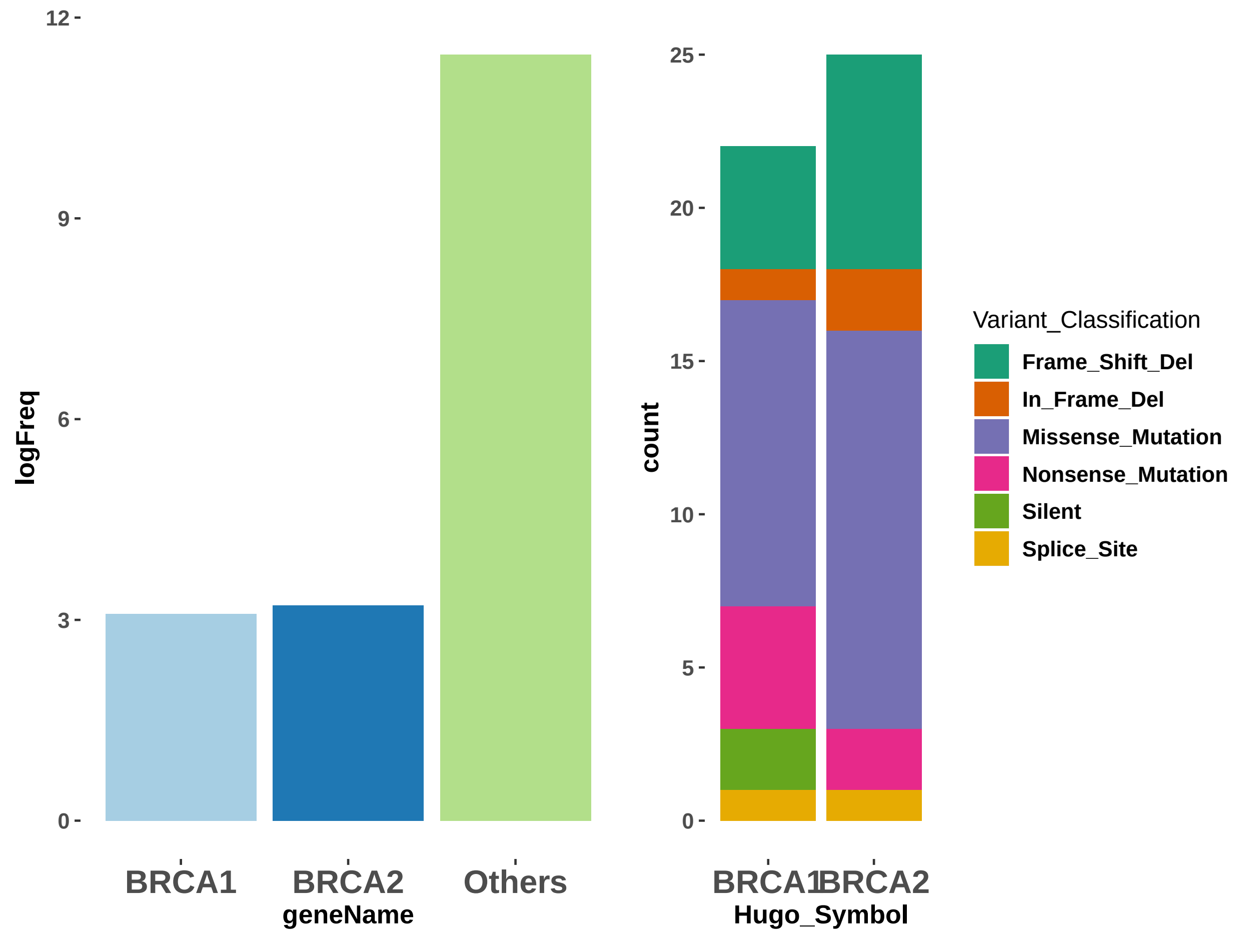

Figure 4: BRCA1/BRCA2 mutation statistics

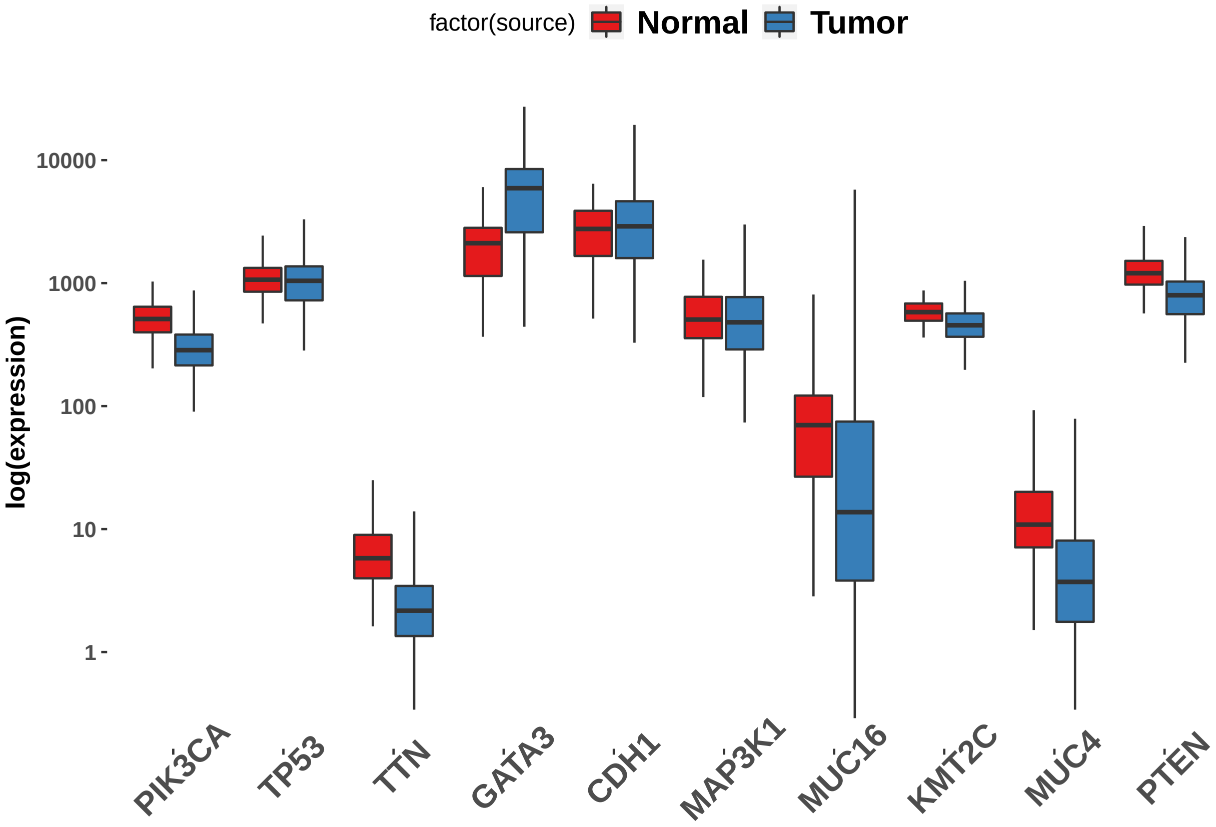

Figure 5: Expression of highly mutated genes in tumor and normal samples

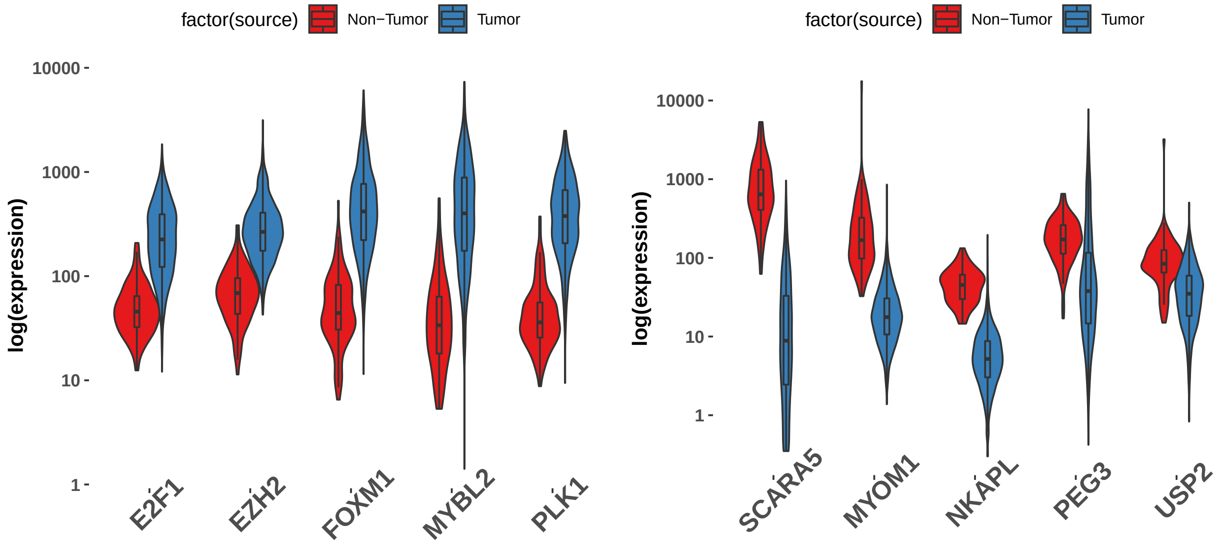

Figure 6: Expression of genes known to be regulated in tumor and normal samples

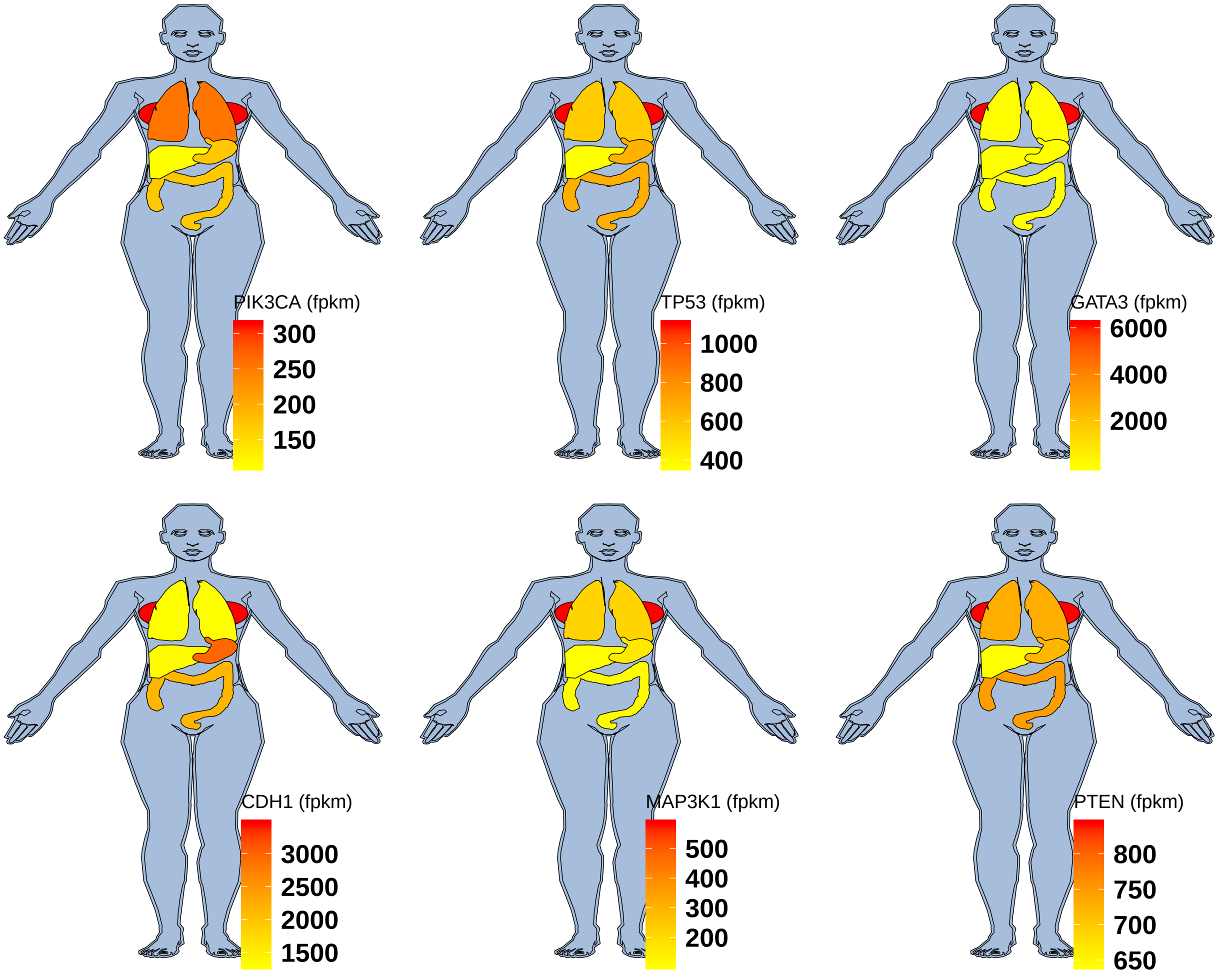

Figure 7: Mean expression of highly mutated BRCA genes in tumors from different tissues

```
## [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] bindrcpp_0.2.2      RColorBrewer_1.1-2  gridExtra_2.3
##  [4] viridis_0.5.1       viridisLite_0.3.0   gganatogram_1.1.1
##  [7] ggpolypath_0.1.0    ggplot2_3.1.0       readr_1.2.1
## [10] maftools_1.8.0      Biobase_2.42.0      BiocGenerics_0.28.0
## [13] plyr_1.8.4          data.table_1.11.8   DT_0.5
## [16] dplyr_0.7.8         TCGAbiolinks_2.10.0
##
## 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              labeling_0.3
##  [73] tidyselect_0.2.5            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
```