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 for further analysis.

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.

```
#required packages
library(TCGAbiolinks)
library(dplyr)
library(DT)
library(data.table)
library(plyr)

colsToKeep<-c("clinical.submitter_id","clinical.classification_of_tumor","clinical.primary_diagnosis","
#Function takes a df and expands it by unlisting elements at a column
expand<-function(df,colName){
  res<-data.frame()
  #for each row
  for(i in 1: dim(df)[1]){
    thisRow<-df[i, ! (colnames(df) %in% c(colName))]
    tempdf<-as.data.frame(df[i, c(colName)])
    #if list is empty skip that row
    if(dim(tempdf)[1]<1){
      next
    }
    #change colnames so they are unique
    colnames(tempdf)<-paste(paste(colName,".",sep = ""),colnames(tempdf),sep = "")
    #print(paste(i,colnames(tempdf)))
    newRow<-cbind(thisRow,tempdf,row.names = NULL)
    res<-bind_rows(res,newRow)

  }
  #print(res)
  return(res)
}

getjoinedBiospcClinc<-function(projName){
  print(paste("Downloading",projName))
  clinicalBRCA <- GDCquery_clinic(project = projName, type = "clinical")
  biospecimenBRCA <- GDCquery_clinic(project = projName, type = "Biospecimen")

  #rename all cols from clinical table with suffix clinical
  colnames(clinicalBRCA)<- paste0("clinical.",colnames(clinicalBRCA))

  #expand biospecimen data in the order portions, portions.analytes, portions.analytes.aliquots
  toUnpack<-c("portions", "portions.analytes", "portions.analytes.aliquots")
  for(s in toUnpack){
    biospecimenBRCA<-expand(biospecimenBRCA,s)
  }
  #add patient barcode to biospecimen data
  biospecimenBRCA<- biospecimenBRCA %>% mutate(clinical.bcr_patient_barcode=substr(submitter_id,1,nchar
  #join clinical and biospecimen
  finalJoined<-join(clinicalBRCA,biospecimenBRCA,by="clinical.bcr_patient_barcode")
  return(finalJoined)
}
```

```
########################################
##Download only BRCA metadata
brcaDF<-getjoinedBiospcClinc("TCGA-CHOL")
```

```
## [1] "Downloading TCGA-CHOL"
```

```
brcaDF<-brcaDF[,colsToKeep]
#remove cols with all NA values
naCols<-colnames(brcaDF)[sapply(brcaDF, function(x)all(is.na(x)))]
brcaDF<-brcaDF[,!(colnames(brcaDF) %in% naCols)]

head(brcaDF)
```

```
##   clinical.submitter_id clinical.classification_of_tumor
## 1          TCGA-3X-AAV9                     not reported
## 2          TCGA-3X-AAV9                     not reported
## 3          TCGA-3X-AAV9                     not reported
## 4          TCGA-3X-AAV9                     not reported
## 5          TCGA-3X-AAV9                     not reported
## 6          TCGA-3X-AAV9                     not reported
##   clinical.primary_diagnosis clinical.tumor_stage
## 1         Cholangiocarcinoma              stage i
## 2         Cholangiocarcinoma              stage i
## 3         Cholangiocarcinoma              stage i
## 4         Cholangiocarcinoma              stage i
## 5         Cholangiocarcinoma              stage i
## 6         Cholangiocarcinoma              stage i
##   clinical.age_at_diagnosis clinical.vital_status clinical.days_to_death
## 1                     26349                 dead                    339
## 2                     26349                 dead                    339
## 3                     26349                 dead                    339
## 4                     26349                 dead                    339
## 5                     26349                 dead                    339
## 6                     26349                 dead                    339
##   clinical.tissue_or_organ_of_origin clinical.days_to_birth
## 1             Intrahepatic bile duct                -26349
## 2             Intrahepatic bile duct                -26349
## 3             Intrahepatic bile duct                -26349
## 4             Intrahepatic bile duct                -26349
## 5             Intrahepatic bile duct                -26349
## 6             Intrahepatic bile duct                -26349
##   clinical.site_of_resection_or_biopsy clinical.days_to_last_follow_up
## 1              Intrahepatic bile duct                              NA
## 2              Intrahepatic bile duct                              NA
## 3              Intrahepatic bile duct                              NA
## 4              Intrahepatic bile duct                              NA
## 5              Intrahepatic bile duct                              NA
## 6              Intrahepatic bile duct                              NA
##   clinical.weight clinical.bmi clinical.height clinical.gender
## 1              52     18.20665             169            male
## 2              52     18.20665             169            male
## 3              52     18.20665             169            male
## 4              52     18.20665             169            male
## 5              52     18.20665             169            male
## 6              52     18.20665             169            male
```

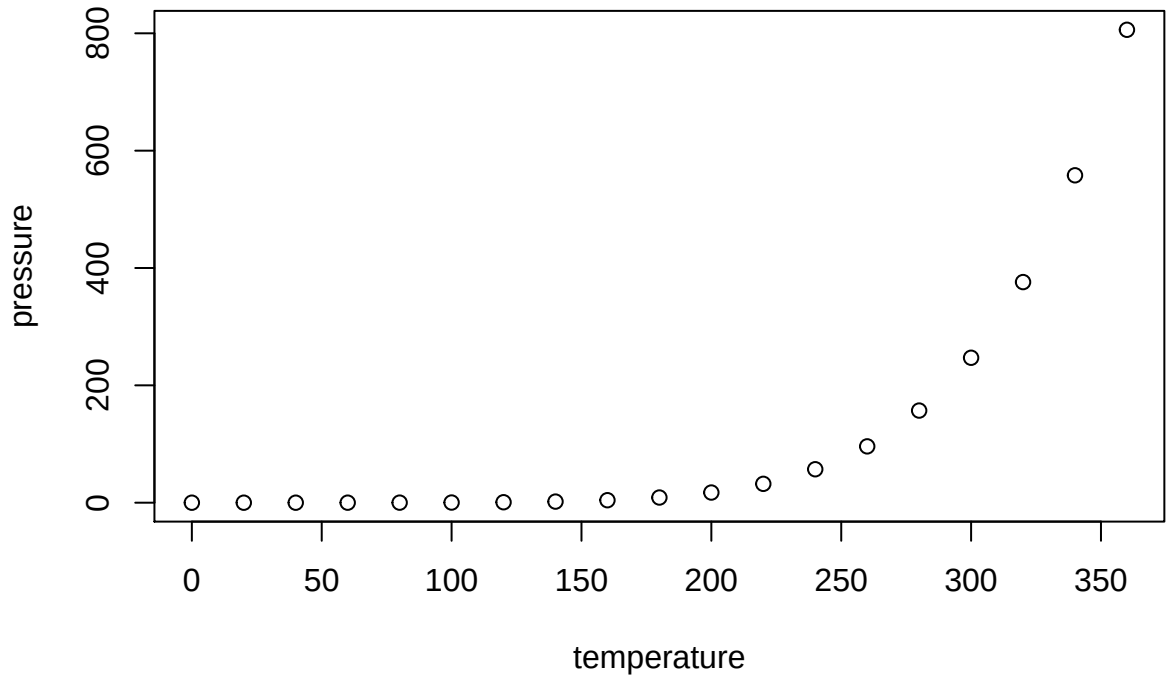
```
##   clinical.year_of_birth clinical.race     clinical.ethnicity
## 1                   1938         asian not hispanic or latino
## 2                   1938         asian not hispanic or latino
## 3                   1938         asian not hispanic or latino
## 4                   1938         asian not hispanic or latino
## 5                   1938         asian not hispanic or latino
## 6                   1938         asian not hispanic or latino
##   clinical.year_of_death clinical.bcr_patient_barcode clinical.disease
## 1                   2010                 TCGA-3X-AAV9             CHOL
## 2                   2010                 TCGA-3X-AAV9             CHOL
## 3                   2010                 TCGA-3X-AAV9             CHOL
## 4                   2010                 TCGA-3X-AAV9             CHOL
## 5                   2010                 TCGA-3X-AAV9             CHOL
## 6                   2010                 TCGA-3X-AAV9             CHOL
##       submitter_id          sample_type portions.submitter_id
## 1 TCGA-3X-AAV9-10A Blood Derived Normal   TCGA-3X-AAV9-10A-01
## 2 TCGA-3X-AAV9-10A Blood Derived Normal   TCGA-3X-AAV9-10A-01
## 3 TCGA-3X-AAV9-01A        Primary Tumor   TCGA-3X-AAV9-01A-72
## 4 TCGA-3X-AAV9-01A        Primary Tumor   TCGA-3X-AAV9-01A-72
## 5 TCGA-3X-AAV9-01A        Primary Tumor   TCGA-3X-AAV9-01A-72
## 6 TCGA-3X-AAV9-01A        Primary Tumor   TCGA-3X-AAV9-01A-72
##   portions.analytes.analyte_type portions.analytes.submitter_id
## 1                            DNA           TCGA-3X-AAV9-10A-01D
## 2                            DNA           TCGA-3X-AAV9-10A-01D
## 3                            DNA           TCGA-3X-AAV9-01A-72D
## 4                            DNA           TCGA-3X-AAV9-01A-72D
## 5                            DNA           TCGA-3X-AAV9-01A-72D
## 6                            RNA           TCGA-3X-AAV9-01A-72R
##   portions.analytes.analyte_type_id
## 1                                 D
## 2                                 D
## 3                                 D
## 4                                 D
## 5                                 D
## 6                                 R
##   portions.analytes.aliquots.analyte_type
## 1                                    <NA>
## 2                                    <NA>
## 3                                    <NA>
## 4                                    <NA>
## 5                                    <NA>
## 6                                    <NA>
##   portions.analytes.aliquots.submitter_id
## 1            TCGA-3X-AAV9-10A-01D-A41A-09
## 2            TCGA-3X-AAV9-10A-01D-A419-01
## 3            TCGA-3X-AAV9-01A-72D-A418-05
## 4            TCGA-3X-AAV9-01A-72D-A417-09
## 5            TCGA-3X-AAV9-01A-72D-A416-01
## 6            TCGA-3X-AAV9-01A-72R-A41D-13
```

Processing

Analysis

You can also embed plots, for example:

```
plot(pressure)
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Conclusion

System information

```
sessionInfo()
```

```
## 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] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] plyr_1.8.4          data.table_1.11.8   DT_0.5
## [4] dplyr_0.7.8         TCGAbiolinks_2.10.0
##
## loaded via a namespace (and not attached):
##   [1] colorspace_1.3-2            selectr_0.4-1
##   [3] rjson_0.2.20                hwriter_1.3.2
##   [5] rprojroot_1.3-2             circlize_0.4.5
##   [7] XVector_0.22.0              GenomicRanges_1.34.0
##   [9] GlobalOptions_0.1.0         rstudioapi_0.8
##  [11] ggpubr_0.2                  matlab_1.0.2
##  [13] ggrepel_0.8.0               bit64_0.9-7
##  [15] AnnotationDbi_1.44.0        xml2_1.2.0
##  [17] codetools_0.2-15            splines_3.5.1
##  [19] R.methodsS3_1.7.1           doParallel_1.0.14
##  [21] DESeq_1.34.0                geneplotter_1.60.0
##  [23] knitr_1.20                  jsonlite_1.5
##  [25] Rsamtools_1.34.0            km.ci_0.5-2
##  [27] broom_0.5.0                 annotate_1.60.0
##  [29] cluster_2.0.7-1             R.oo_1.22.0
##  [31] readr_1.2.1                 compiler_3.5.1
##  [33] httr_1.3.1                  backports_1.1.2
##  [35] assertthat_0.2.0            Matrix_1.2-14
##  [37] lazyeval_0.2.1              limma_3.38.2
##  [39] formatR_1.5                 htmltools_0.3.6
##  [41] prettyunits_1.0.2           tools_3.5.1
##  [43] bindrcpp_0.2.2              gtable_0.2.0
##  [45] glue_1.3.0                  GenomeInfoDbData_1.2.0
##  [47] ggthemes_4.0.1              ShortRead_1.40.0
##  [49] Rcpp_1.0.0                  Biobase_2.42.0
##  [51] Biostrings_2.50.1           nlme_3.1-137
##  [53] rtracklayer_1.42.1          iterators_1.0.10
##  [55] stringr_1.3.1               rvest_0.3.2
##  [57] XML_3.98-1.16               edgeR_3.24.0
##  [59] zoo_1.8-4                   zlibbioc_1.28.0
##  [61] scales_1.0.0                aroma.light_3.12.0
##  [63] hms_0.4.2                   parallel_3.5.1
##  [65] SummarizedExperiment_1.12.0 RColorBrewer_1.1-2
##  [67] curl_3.2                    ComplexHeatmap_1.20.0
##  [69] yaml_2.2.0                  memoise_1.1.0
##  [71] gridExtra_2.3               KMsurv_0.1-5
##  [73] ggplot2_3.1.0               downloader_0.4
##  [75] biomaRt_2.38.0              latticeExtra_0.6-28
##  [77] stringi_1.2.4               RSQLite_2.1.1
##  [79] genefilter_1.64.0           S4Vectors_0.20.1
```

```
##  [81] foreach_1.4.4               GenomicFeatures_1.34.1
##  [83] BiocGenerics_0.28.0         BiocParallel_1.16.2
##  [85] shape_1.4.4                 GenomeInfoDb_1.18.1
##  [87] rlang_0.3.0.1               pkgconfig_2.0.2
##  [89] matrixStats_0.54.0          bitops_1.0-6
##  [91] evaluate_0.12               lattice_0.20-35
##  [93] purrr_0.2.5                 bindr_0.1.1
##  [95] htmlwidgets_1.3             cmprsk_2.2-7
##  [97] GenomicAlignments_1.18.0    bit_1.1-14
##  [99] tidyselect_0.2.5            magrittr_1.5
## [101] R6_2.3.0                    IRanges_2.16.0
## [103] DelayedArray_0.8.0          DBI_1.0.0
## [105] mgcv_1.8-24                 pillar_1.3.0
## [107] survival_2.42-3             RCurl_1.95-4.11
## [109] tibble_1.4.2                EDASeq_2.16.0
## [111] crayon_1.3.4                survMisc_0.5.5
## [113] rmarkdown_1.10              GetoptLong_0.1.7
## [115] progress_1.2.0              locfit_1.5-9.1
## [117] grid_3.5.1                  sva_3.30.0
## [119] blob_1.1.1                  ConsensusClusterPlus_1.46.0
## [121] digest_0.6.18               xtable_1.8-3
## [123] tidyr_0.8.2                 R.utils_2.7.0
## [125] stats4_3.5.1                munsell_0.5.0
## [127] survminer_0.4.3
```