RNA Sequencing Differential Analysis Project

Gavin Fortenberry

August 14th, 2018

Contents

Introduction

What is Myleofibrosis?

- Myleofibrosis is a type of bone marrow cancer in which an rapidly increasing number of blood forming cells form a fibrous like structure that sometimes leads to acute leukemia. Certain genotypes like the JAK2 V617F mutation have been a determining factor of Blast Transformation in Myleofibrosis.

What is differential gene expression?

- Differential gene expression is a way to analyze factors in different groups that may or may not be associated with different gene counts, from RNA sequence data.

What is gene expression?

- Gene expression is the process of which information inside a gene is used to make RNA and proteins. The genotype leads to the phenotype.

Why is differential gene expression important to a biological question?

How to do differential gene expression:

What inputs go in:

24 patients, 12 MF, 12 normal ### What is the analysis going to do

What are the outputs/what is the meaning of the outputs

(My) Experimental Design

What groups am I going to compare? (Bio factors vs. tech factors)

- The groups of phenotype data being used can be sorted into two overarching categories: biological factors and technical factors. Biological factors include "Tissue_type", "genotype_jak2", "genotype_calr", and "genotype_mpl". Technical factors include: "collection_type", "time_to_processing", and "extraction_type".
- The reason why I am splitting the data/analysis into two categories is to attempt to see whether certain factors from each category have an impact on specific gene counts or not.
- To be completed: Why im picking each of the columns within those sections
- I will separate demographic groups by phenotype, and compare gene expression counts for each of the genes between the two groups, groups defined by conditions of that factor. For example, the conditions of the factor of age could be different age groups, sex could be male or female, genotype could be present or not present, etc. ultimate goal is to make observations/conclusions about possible differences in gene counts between different groups that may or may not be significant in diagnosis of Myleofibrosis.

Data Import

Install R Packages

Installs packages of functions that can be used to help manipulate variables and data libraries. These packages can be installed through base R code that downloads them and installs them into R studio. Then certain programming phrases can be used to do new functions in connection with base R programming phrases, with an indicator of the package being used mentioned before using that package's functions. For example, if I wanted to use a function of the "dplyr" package, I would write -> "dplyr::(insert function here)(insert arguments of functions here)". Note that functions or phrases that are in base R programming vocabulary do not require a name indicator of the package being used, as it is from the base software and the software knows how to interperet those phrases without an idicator of a package. It is still required to write names of functions from Base R that are wanting to be used though, and if shortcut variables for functions are desired then you can assign new variables to those functions.

Bioconductor packages need to be installed by "biocLite" rather than install.packages which is for CRAN/base R

```
install.packages("dplyr")
install.packages("knitr")
install.packages("rmarkdown")
```

```
### Install Bioconductor, the DESeq2 Package and ggplot2
source("https://bioconductor.org/biocLite.R")
biocLite("DESeq2")
install.packages("ggplot2")
```

Load Additional Packages

```
library(DESeq2); library(ggplot2)
library(plyr); library(dplyr)
library(gghighlight)
```

Read in Project Metadata to R

Reads the two CSV's and combines them into one data frame: The code below creates three data frames: one labeled as "df_pheno" for data of a file called phenotable, which contains demographics and general information from patients and samples taken from them, another labeled as "df_molecular" for data from another file called molecularDataSets, which contains more specific info on samples collected from patients and their diagnoses from these samples, and another labeled as "df_combined" that combines the two data frames according to common columns.

The overall purpose of this code chunk is to create the combined data frame.

The first two data frames created ("df_pheno" and "df_molecular") were created using base R code "read.csv" function, which reads in data into a data frame from a CSV, a comma seperated values document.

In the process of formatting the df_pheno dataframe in order to be compatible when joining with the df_pheno dataframe, the "rename" function from the dplyr package is used on the df_pheno dataframe to change the name of a mismatching column name.

The function used to actually combine the two dataframes into one is the "full_join" function from the dplyr package which joins two indicated dataframes together by a common column.

"str()" is short for structure, and is a base R function that allows the user to get an idea of the format of the data they are looking at

```
df_pheno <- read.csv(file="../RNASeqData/phenotypeTable.csv",
                     header =TRUE)
df_pheno <- df_pheno %>% rename(assay_material_id = ends_with("assay_material_id"))

df_molecular <- read.csv(file="../RNASeqData/molecularDataSets.csv",
                         header =TRUE, sep=",")

df_combined <- dplyr::full_join(df_pheno, df_molecular,
                                by = "assay_material_id")
head(df_combined, n = 3)
```

```
##   assay_material_id     sex    race age_range diagnosis collection_event
## 1             R0297 unknown unknown   unknown        mf        diagnosis
## 2             R0299    male unknown elderly61        mf        diagnosis
## 3             R0298 unknown unknown   unknown        mf        diagnosis
##   acquisition_date     tissue_type tissue_type_origin collection_type
## 1          8/14/13 peripheralblood                 NA            edta
## 2          8/22/13 peripheralblood                 NA            edta
```

```
## 3          8/19/13 peripheralblood                 NA            edta
##    processing_type   time_to_processing       storage shipping
## 1 mononuclearcells repositoryprocessing cryopreserved    fresh
## 2 mononuclearcells repositoryprocessing cryopreserved    fresh
## 3 mononuclearcells repositoryprocessing cryopreserved    fresh
##   genotype_jak2 genotype_calr  genotype_mpl material_type extraction_type
## 1      positive      negative notdetermined           RNA          column
## 2      negative      positive notdetermined           RNA          column
## 3      negative      positive notdetermined           RNA          column
##   na_260280 na_260230   rin_range jak2_vaf molecular_id genomics_types
## 1      2.05      1.86 highquality       70    M00000298         rnaseq
## 2      2.05      1.81 highquality        0    M00000300         rnaseq
## 3      2.07      1.78 highquality        0    M00000299         rnaseq
##   omics_sample_name omics_contact_id omics_date
## 1        JAK2-6-1-D          jradich    2/27/14
## 2       JAK2-10-1-D          jradich    2/27/14
## 3           MF-D-07          jradich    3/24/14
##                seq_flowcell_id seq_readlength seq_paired seq_libtype
## 1 140227_SN367_0370_AH8JPDADXX             99        yes      truseq
## 2 140227_SN367_0370_AH8JPDADXX             99        yes      truseq
## 3 140324_SN367_0381_BH929TADXX             99        yes      truseq
```

Save Prepared Metadata

Writes a CSV (a comma seperated values document) from the new combined data frame(df_combined) of dataframes df_pheno and df_molecular into a document called "combineddata.csv" stored in the working directory (the main location of the files created from the Rstudio

```
write.csv(df_combined, file = "../ProjectFiles/combineddata.csv",
          row.names = FALSE)
```

The read.csv function is used to re-read the newly written CSV to make sure the data is the same as it was when written. The "str()" and "summary" functions are used to compare the statistics of the new dataframe to the original created one to confirm similarity.

```
test_set_combined_data <- read.csv("combineddata.csv")
str(test_set_combined_data)
```

Read in RNA Sequencing Data from HiSat2/htseqcount

Makes a list of directories(folders) with the R base function "list.dirs" which takes the indicated path of the main directory containing the directories its making a list of in, and the econd argument written "recursive" is set to false becuse the main directory is not desired to be listed in the list of its components.

```
#":/Users/gfortenb/Documents/GitHub/bioDS-bootcamp/RNASeqData"
RNADirectoryList = list.dirs(path = "../RNASeqData", recursive = FALSE)
```

Makes a List of files within each folder of each directory in the main directory. Uses a function to go through the files within each folder and only list files with a certain phrase in the name of the file, "htseq.txt", using the pattern function (only argument usedin function is name of character phrase its looking for).

Binds path's of files(locations of the files in the computer) on the list made to the molecular ID of the files with "as.data.frame"" function"/"cbind" function, created by finding key character sequences in the title of the folders containing the "htseq.txt"" files, using the gsub function (from base R). The first argument used

in the gsub function is a character phrase that indicates where in the string to look for the character phrase of the molecular ID, and the second argument is the list of data of file locations/names to look for the ID in.

The "colnames" function (base R) sets the column names if the newly created dataframe.

```
FileList_htseq1 = sapply(RNADirectoryList,
                  function(x){list.files(path = x,
                                         full.names = TRUE,
                                         pattern = "htseq.txt") })

htseq_FileID_df <- as.data.frame(cbind(FileList_htseq1,
                                 gsub("^.*-","", RNADirectoryList)),
                            stringsAsFactors = F)
colnames(htseq_FileID_df) <- c("Path", "molecular_id")
head(htseq_FileID_df, n = 3)
```

```
##
## ../RNASeqData/JAK2-10-1-D-R0299-M00000300 ../RNASeqData/JAK2-10-1-D-R0299-M00000300/JAK2-10-1-D.htsec
## ../RNASeqData/JAK2-30-D-R0301-M00000302       ../RNASeqData/JAK2-30-D-R0301-M00000302/JAK2-30-D.htsec
## ../RNASeqData/JAK2-36-D-R0303-M00000304       ../RNASeqData/JAK2-36-D-R0303-M00000304/JAK2-36-D.htsec
##                                             molecular_id
## ../RNASeqData/JAK2-10-1-D-R0299-M00000300    M00000300
## ../RNASeqData/JAK2-30-D-R0301-M00000302      M00000302
## ../RNASeqData/JAK2-36-D-R0303-M00000304      M00000304
```

Reads all data from list of selected files and compiles into different data frames/list of different data frames using LApply function.

```
listOf_alldf <- lapply(seq(1:nrow(htseq_FileID_df)),
                  function(i){
                    X <- read.delim(file = htseq_FileID_df$Path[i],
                                    header = FALSE);
                    colnames(X) <- c("Gene", htseq_FileID_df$molecular_id[i]);
                    return(X)
                    } )
```

SALMON HERE<<<—

```
### Reads in data/creates salmon counts dataframe
salmonCounts_df <-read.csv(file="../RNASeqData/2018-08-06_SalmonGeneLevelCounts.csv",
                      header =TRUE)
```

Data Summarizing

Summarizing PHENOTYPES/defining experimental design

Biological factors

```
bio_factors_summary <- df_pheno %>% group_by(diagnosis, genotype_jak2, genotype_calr, age_range, sex) %
bio_factors_summary
```

```
## # A tibble: 6 x 6
## # Groups:   diagnosis, genotype_jak2, genotype_calr, age_range [?]
```

```
##   diagnosis genotype_jak2 genotype_calr age_range   sex     `n()`
##   <fct>     <fct>         <fct>         <fct>       <fct>   <int>
## 1 mf        negative      negative      unknown     unknown     1
## 2 mf        negative      positive      elderly61   male        1
## 3 mf        negative      positive      unknown     unknown     5
## 4 mf        positive      negative      adult18to60 male        1
## 5 mf        positive      negative      unknown     unknown     4
## 6 normal    notdetermined notdetermined adult18to60 unknown    12
```

Technological factors

```
tech_factors_summary <- df_pheno %>% group_by(time_to_processing, collection_type, collection_event, ex
tech_factors_summary
```

```
## # A tibble: 4 x 5
## # Groups:   time_to_processing, collection_type, collection_event [?]
##   time_to_processi~ collection_type collection_event extraction_type `n()`
##   <fct>             <fct>           <fct>            <fct>           <int>
## 1 repositoryproces~ edta            diagnosis        column              4
## 2 repositoryproces~ unknown         diagnosis        column              8
## 3 under3h           acd             normal           column             10
## 4 under3h           acd             normal           trizol              2
```

Bio & tech factors

```
bio_and_tech_summary <- df_pheno %>% group_by(diagnosis, time_to_processing, genotype_jak2, collection_
bio_and_tech_summary
```

```
## # A tibble: 9 x 10
## # Groups:   diagnosis, time_to_processing, genotype_jak2, collection_type,
## #   genotype_calr, collection_event, age_range, extraction_type [?]
##   diagnosis time_to_process~ genotype_jak2 collection_type genotype_calr
##   <fct>     <fct>            <fct>         <fct>           <fct>
## 1 mf        repositoryproce~ negative      edta            positive
## 2 mf        repositoryproce~ negative      edta            positive
## 3 mf        repositoryproce~ negative      unknown         negative
## 4 mf        repositoryproce~ negative      unknown         positive
## 5 mf        repositoryproce~ positive      edta            negative
## 6 mf        repositoryproce~ positive      edta            negative
## 7 mf        repositoryproce~ positive      unknown         negative
## 8 normal    under3h          notdetermined acd             notdetermined
## 9 normal    under3h          notdetermined acd             notdetermined
## # ... with 5 more variables: collection_event <fct>, age_range <fct>,
## #   extraction_type <fct>, sex <fct>, `n()` <int>
```

Create a Summarized Experiment Data set

("Summarized Experiment" - something specific to DESeq2 package) - Normalize RNA Sequencing Counts - in new normalized data frame - (Normalize each sample's counts data based on over all library size for each sample.)

Data Setup For Analysis

Differential Gene Expression Analysis with DESeq2

HERE DOWN->

Make the list of dataframes into one dataframe w/all contents of each dataframe in the dataframe of dataframes as columns using the "join_all" function from the plyr package. This produces a dataframe with the molecular ID and Genes in each sample assigned to a molecular ID columns.

```
###THIS MATTERS- MOLECULAR ID's LINKED TO THE PHENOTABLE
htseq_genecounts_df <- plyr:: join_all(listOf_alldf, by = NULL,
                                type = "full", match = "all")
head(htseq_genecounts_df, n = 3)
```

```
##       Gene M00000300 M00000302 M00000304 M00000305 M00000298 M00000297
## 1     A1BG       226       138       233       438       130       148
## 2 A1BG-AS1         7         7         8         4        10         6
## 3     A1CF         0         0         0         0         0         0
##   M00000299 M00000301 M00000303 M00000306 M00000307 M00000308 M00000019
## 1        88       163        88       132       232       240       119
## 2         0         7         6         0         6        22        10
## 3         0         0         0         0         0         0         0
##   M00000020 M00000021 M00000022 M00000002 M00000001 M00000003 M00000004
## 1       164       230       254       129        63       140       128
## 2        26        39        43        21        13        29        32
## 3         0         0         1         0         0         2         1
##   M00000007 M00000005 M00000006 M00000008
## 1       160       125       133       203
## 2        28        52        25        45
## 3         5         0         3         3
```

Optimizing compatability for creating DataSets for DESEQ2 1(EDITS = DONE)

```
#Creats HTSEQ counts matrix
#Imports everything from HTSEQ Counts dataframe except the first column (Doesn't delete column from ori
htseqCountsMat = as.matrix(htseq_genecounts_df[,-1]); ncol(htseqCountsMat)          #COLS
```

```
## [1] 24
```

```
head(htseqCountsMat)
```

```
##      M00000300 M00000302 M00000304 M00000305 M00000298 M00000297 M00000299
## [1,]       226       138       233       438       130       148        88
## [2,]         7         7         8         4        10         6         0
## [3,]         0         0         0         0         0         0         0
## [4,]        12        15        23        67         7         8         5
## [5,]        17        39         4        28         7        18         1
## [6,]         0         5         4         7        15         0         3
##      M00000301 M00000303 M00000306 M00000307 M00000308 M00000019 M00000020
## [1,]       163        88       132       232       240       119       164
## [2,]         7         6         0         6        22        10        26
## [3,]         0         0         0         0         0         0         0
## [4,]        29        24        32        28        16         9       136
## [5,]        23        46        37         5        48        20        41
## [6,]         5        31         4         6        37         0         3
##      M00000021 M00000022 M00000002 M00000001 M00000003 M00000004 M00000007
```

```
## [1,]       230       254       129        63       140       128       160
## [2,]        39        43        21        13        29        32        28
## [3,]         0         1         0         0         2         1         5
## [4,]       142        47        24        26        15        59       242
## [5,]        44        23        22        28        36        33        87
## [6,]         2         3         4         1         3         4        17
##      M00000005 M00000006 M00000008
## [1,]       125       133       203
## [2,]        52        25        45
## [3,]         0         3         3
## [4,]        34        60       437
## [5,]        43        41        90
## [6,]        10         8         7
```

```
#Sets rownames of new matrix of HTSEQ count data to gene names in 1st column of original dataframe
rownames(htseqCountsMat)<- htseq_genecounts_df[,1]



#Repeated processes for SALMON
```

Repeated processes w/salmon.txt files

```
# Assigns gene names from salmon counts dataframe to variable in order to keep them set as characters/s
salmonCounts_genes <- as.character(salmonCounts_df[ , 25] )
# Creates matrix of salmon counts from salmon count data frame
# Converts numbers with decimals to integers (no decimals)  in salmon counts matrix,
salmonCountsMat <- sapply(salmonCounts_df[ , -25],as.integer)
#head(salmonCounts_df)
# Sets rownames of new salmon counts matrix to gene names of salmon counts dataframe
rownames(salmonCountsMat) <- salmonCounts_genes
head(salmonCountsMat, n = 4)
```

```
##          M00000300 M00000302 M00000304 M00000305 M00000298 M00000297
## A1BG           226       139       230       446       131       154
## A1BG-AS1        36        24        34        29        13        38
## A1CF             0         0         0         0         0         0
## A2M              0         0         0         0         0         0
##          M00000299 M00000301 M00000303 M00000306 M00000307 M00000308
## A1BG            85       174        88       131       244       243
## A1BG-AS1        14        36        34        19        29        69
## A1CF             0         0         0         0         0         0
## A2M              0         0         0         0         0         0
##          M00000019 M00000020 M00000021 M00000022 M00000002 M00000001
## A1BG           115       150       212       247       137        64
## A1BG-AS1        75       145       170       161       102        96
## A1CF             0         0         0         0         0        18
## A2M              0        46        17         0         0        15
##          M00000003 M00000004 M00000007 M00000005 M00000006 M00000008
## A1BG           127       138       164       125       130       194
## A1BG-AS1       189       181       199       216       125       257
## A1CF             0         0        20         0         0        38
```

```
## A2M              0         0        62         0        18       101
```

```
str(salmonCountsMat)
```

```
##  int [1:23537, 1:24] 226 36 0 0 12 0 0 14 1 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:23537] "A1BG" "A1BG-AS1" "A1CF" "A2M" ...
##   ..$ : chr [1:24] "M00000300" "M00000302" "M00000304" "M00000305" ...
```

```
#Creates matrix phenoMat from df_combined dataframe
# Imports all data/columns except molecular ID column (doesnt delete from original dataframe, just does
phenoMat = as.matrix(subset(df_combined, select=-molecular_id)); nrow(phenoMat)
```

```
## [1] 24
```

```
#Sets rownames of phenoMat to data from df_combined  column of molecular ID's
rownames(phenoMat)<- df_combined$molecular_id
```

```
#Link column names AND column values  of htseqCountsMat to phenoMat row names in same order
phenoMat <- phenoMat[match(colnames(htseqCountsMat),row.names(phenoMat)),]
head(phenoMat)
```

```
##           assay_material_id sex       race      age_range     diagnosis
## M00000300 "R0299"           "male"    "unknown" "elderly61"   "mf"
## M00000302 "R0301"           "unknown" "unknown" "unknown"     "mf"
## M00000304 "R0303"           "unknown" "unknown" "unknown"     "mf"
## M00000305 "R0304"           "unknown" "unknown" "unknown"     "mf"
## M00000298 "R0297"           "unknown" "unknown" "unknown"     "mf"
## M00000297 "R0296"           "male"    "unknown" "adult18to60" "mf"
##           collection_event acquisition_date tissue_type
## M00000300 "diagnosis"      "8/22/13"        "peripheralblood"
## M00000302 "diagnosis"      "4/17/13"        "peripheralblood"
## M00000304 "diagnosis"      "6/14/13"        "peripheralblood"
## M00000305 "diagnosis"      "6/20/13"        "peripheralblood"
## M00000298 "diagnosis"      "8/14/13"        "peripheralblood"
## M00000297 "diagnosis"      "7/22/13"        "peripheralblood"
##           tissue_type_origin collection_type processing_type
## M00000300 NA                 "edta"          "mononuclearcells"
## M00000302 NA                 "unknown"       "mononuclearcells"
## M00000304 NA                 "unknown"       "mononuclearcells"
## M00000305 NA                 "unknown"       "mononuclearcells"
## M00000298 NA                 "edta"          "mononuclearcells"
## M00000297 NA                 "edta"          "mononuclearcells"
##           time_to_processing     storage         shipping genotype_jak2
## M00000300 "repositoryprocessing" "cryopreserved" "fresh"  "negative"
## M00000302 "repositoryprocessing" "cryopreserved" "fresh"  "positive"
## M00000304 "repositoryprocessing" "cryopreserved" "fresh"  "positive"
## M00000305 "repositoryprocessing" "cryopreserved" "fresh"  "negative"
## M00000298 "repositoryprocessing" "cryopreserved" "fresh"  "positive"
## M00000297 "repositoryprocessing" "cryopreserved" "fresh"  "positive"
##           genotype_calr genotype_mpl    material_type extraction_type
## M00000300 "positive"    "notdetermined" "RNA"         "column"
## M00000302 "negative"    "notdetermined" "RNA"         "column"
## M00000304 "negative"    "notdetermined" "RNA"         "column"
## M00000305 "positive"    "notdetermined" "RNA"         "column"
## M00000298 "negative"    "notdetermined" "RNA"         "column"
## M00000297 "negative"    "notdetermined" "RNA"         "column"
```

```
##           na_260280 na_260230 rin_range       jak2_vaf genomics_types
## M00000300 "2.05"    "1.81"    "highquality"   " 0.0"   "rnaseq"
## M00000302 "2.09"    "1.67"    "mediumquality" "93.4"   "rnaseq"
## M00000304 "2.11"    "1.74"    "mediumquality" "46.2"   "rnaseq"
## M00000305 "2.07"    "2.06"    "highquality"   " 0.0"   "rnaseq"
## M00000298 "2.05"    "1.86"    "highquality"   "70.0"   "rnaseq"
## M00000297 "2.07"    "1.74"    "mediumquality" "91.8"   "rnaseq"
##           omics_sample_name omics_contact_id omics_date
## M00000300 "JAK2-10-1-D"     "jradich"        "2/27/14"
## M00000302 "JAK2-30-D"       "jradich"        "2/27/14"
## M00000304 "JAK2-36-D"       "jradich"        "2/27/14"
## M00000305 "JAK2-37-D"       "jradich"        "2/27/14"
## M00000298 "JAK2-6-1-D"      "jradich"        "2/27/14"
## M00000297 "MF-D-02"         "jradich"        "3/24/14"
##           seq_flowcell_id                seq_readlength seq_paired
## M00000300 "140227_SN367_0370_AH8JPDADXX" "99"           "yes"
## M00000302 "140227_SN367_0370_AH8JPDADXX" "99"           "yes"
## M00000304 "140227_SN367_0370_AH8JPDADXX" "99"           "yes"
## M00000305 "140227_SN367_0370_AH8JPDADXX" "99"           "yes"
## M00000298 "140227_SN367_0370_AH8JPDADXX" "99"           "yes"
## M00000297 "140324_SN367_0380_AH91T2ADXX" "99"           "yes"
##           seq_libtype
## M00000300 "truseq"
## M00000302 "truseq"
## M00000304 "truseq"
## M00000305 "truseq"
## M00000298 "truseq"
## M00000297 "truseq"
```

Creating original DeseqDatasets

(EDITS = DONE)

```
#DataSet for DESEQ from HTSEQ Matrix created
# variable comparing is diagnosis [levels  = mf (myleofibrosis) and normal (no myleofibrosis) ]
dseq_set_htseq <- DESeqDataSetFromMatrix(htseqCountsMat,phenoMat, design = ~ diagnosis)

# with base level being at "normal" (not myleofibrosis)
dseq_set_htseq$diagnosis <- relevel(dseq_set_htseq$diagnosis, "normal")

#gets rid of rows of gene counts of 1's and 0's
dseq_set_htseq <- dseq_set_htseq[ rowSums(counts(dseq_set_htseq)) > 1, ]
head(dseq_set_htseq)
```

```
## class: DESeqDataSet
## dim: 6 24
## metadata(1): version
## assays(1): counts
## rownames(6): A1BG A1BG-AS1 ... A2M-AS1 A2ML1
## rowData names(0):
## colnames(24): M00000300 M00000302 ... M00000006 M00000008
## colData names(31): assay_material_id sex ... seq_paired
##   seq_libtype
```

```
dseq_set_salmon <- DESeqDataSetFromMatrix(salmonCountsMat,phenoMat, design = ~ diagnosis)
```

```
dseq_set_salmon$diagnosis <- relevel(dseq_set_salmon$diagnosis, "normal")

dseq_set_salmon <- dseq_set_salmon[ rowSums(counts(dseq_set_salmon)) > 1, ] #gets rid of 1's and 0's
head(dseq_set_salmon)
```

```
## class: DESeqDataSet
## dim: 6 24
## metadata(1): version
## assays(1): counts
## rownames(6): A1BG A1BG-AS1 ... A2M-AS1 A2ML1
## rowData names(0):
## colnames(24): M00000300 M00000302 ... M00000006 M00000008
## colData names(31): assay_material_id sex ... seq_paired
##   seq_libtype
```

Creating copies of original datasets for analysis/design formulas

```
###base lvl = diagnosis
#Make copy of htseq data set so original is not modified, multiple formulas can be applied
dseq_set_htseq_copy <- dseq_set_htseq
design(dseq_set_htseq_copy) <- formula(~ diagnosis)
dseq_set_htseq_copy <- DESeq(dseq_set_htseq_copy)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 448 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

```
###base lvl = age range
dseqstuffCopy2 <- dseq_set_htseq
design(dseqstuffCopy2) <- formula(~ age_range)
dseqstuffCopy2 <- DESeq(dseqstuffCopy2)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates
```

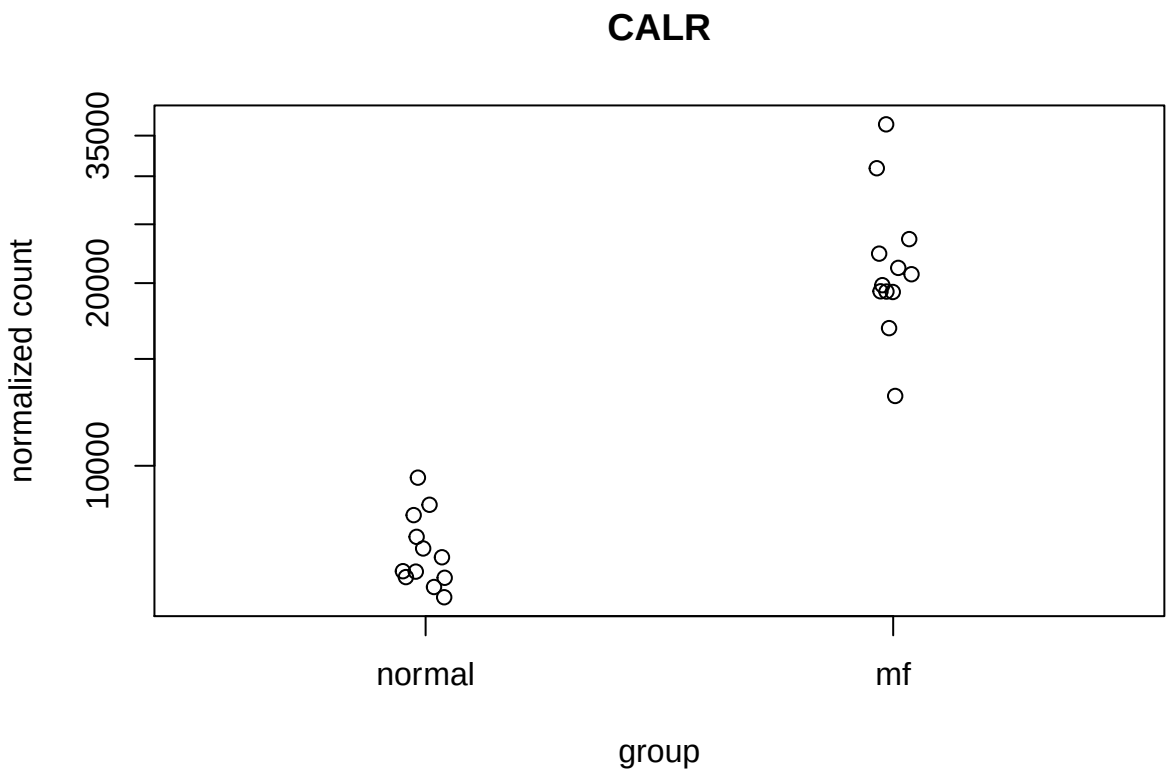
```
## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 5261 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

```
dseq_set_salmon_copy <- dseq_set_salmon
design(dseq_set_salmon_copy ) <- formula(~ diagnosis)
dseq_set_salmon_copy  <- DESeq(dseq_set_salmon_copy )
```

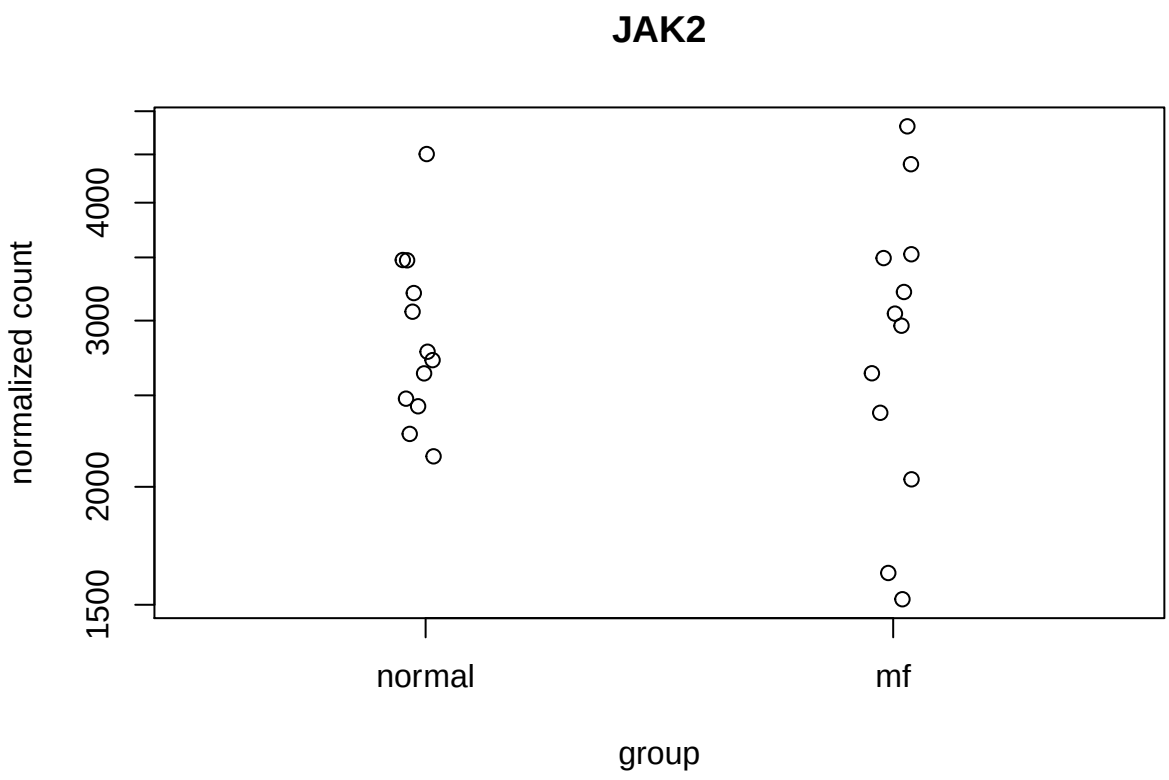
```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 623 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

Analysis

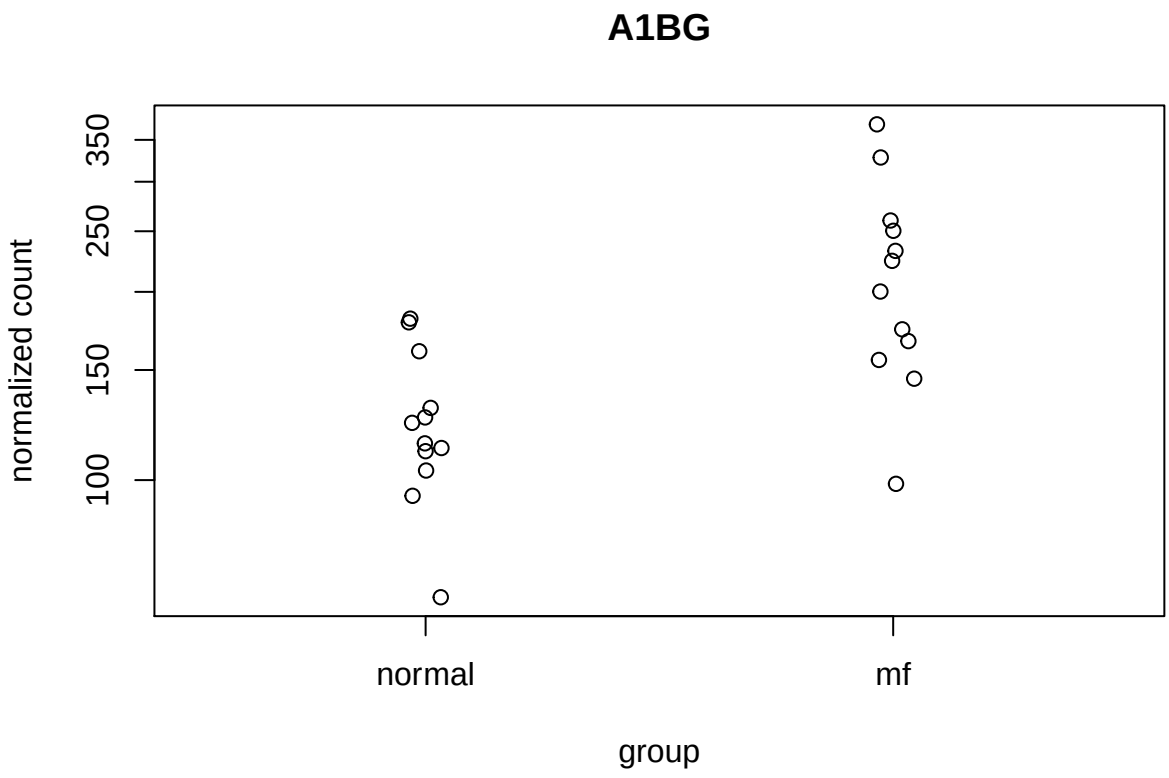
```
DESeq2::plotCounts(dseq_set_htseq, "CALR", intgroup = "diagnosis")
```



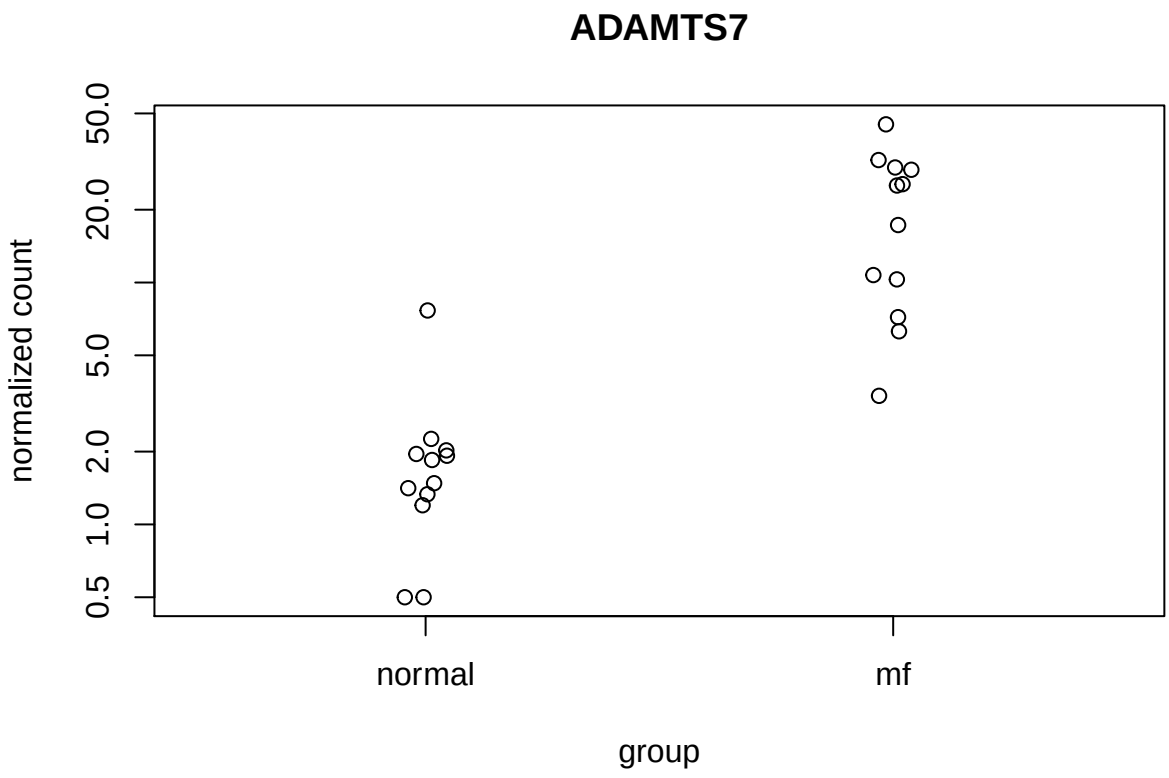

```
DESeq2::plotCounts(dseq_set_htseq, "JAK2", intgroup = "diagnosis")
```




```
DESeq2::plotCounts(dseq_set_htseq, "A1BG", intgroup = "diagnosis")
```




```
DESeq2::plotCounts(dseq_set_htseq, "ADAMTS7", intgroup = "diagnosis")
```



use Deseqdataset to do analysis using deseq2 tools???

Following paragraph/description Copied from https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#differential-expression-analysis . . . IMPORTANT!!!!!!!!!!

A DESeqDataSet object must have an associated design formula. The design formula expresses the variables which will be used in modeling. The formula should be a tilde (~) followed by the variables with plus signs between them (it will be coerced into an formula if it is not already). The design can be changed later, however then all differential analysis steps should be repeated, as the design formula is used to estimate the dispersions and to estimate the log2 fold changes of the model.

Note: In order to benefit from the default settings of the package, you should put the variable of interest at the end of the formula and make sure the control level is the first level.

Gives results of HTSEQ design formula 1:

```
resDseqCopy <- results(dseq_set_htseq_copy)
head(resDseqCopy)
```

```
## log2 fold change (MLE): diagnosis mf vs normal
## Wald test p-value: diagnosis mf vs normal
## DataFrame with 6 rows and 6 columns
##             baseMean log2FoldChange     lfcSE       stat       pvalue
##            <numeric>      <numeric> <numeric>  <numeric>    <numeric>
## A1BG     170.6840521      0.7958968 0.1953430  4.0743564 4.614177e-05
```

```
## A1BG-AS1  16.1712821     -1.5957025 0.3378240 -4.7234733 2.318504e-06
## A1CF       0.4571053     -2.0708240 1.4841992 -1.3952467 1.629415e-01
## A2M       47.3634595     -1.5682386 0.4316494 -3.6331307 2.800031e-04
## A2M-AS1   29.7386172     -0.3417973 0.3672911 -0.9305897 3.520659e-01
## A2ML1      7.2833133      1.4117902 0.5805970  2.4316181 1.503155e-02
##                 padj
##            <numeric>
## A1BG     1.291139e-04
## A1BG-AS1 7.867270e-06
## A1CF     2.238507e-01
## A2M      6.915249e-04
## A2M-AS1  4.314028e-01
## A2ML1    2.700069e-02
```

Gives results of HTSEQ design formula 2:

```
resDseqCopy2 <- results(dseqstuffCopy2)
head(resDseqCopy2)
```

```
## log2 fold change (MLE): age range unknown vs adult18to60
## Wald test p-value: age range unknown vs adult18to60
## DataFrame with 6 rows and 6 columns
##             baseMean log2FoldChange     lfcSE      stat       pvalue
##            <numeric>      <numeric> <numeric> <numeric>    <numeric>
## A1BG     170.6840521      0.4789206 0.2418622  1.980138 4.768799e-02
## A1BG-AS1  16.1712821     -1.6439004 0.3623803 -4.536395 5.722386e-06
## A1CF       0.4571053     -2.0950798 1.6585660 -1.263187 2.065218e-01
## A2M       47.3634595     -1.4064408 0.4623021 -3.042255 2.348129e-03
## A2M-AS1   29.7386172     -0.4274496 0.3834509 -1.114744 2.649601e-01
## A2ML1      7.2833133      1.7346729 0.5394303  3.215750 1.301042e-03
##                 padj
##            <numeric>
## A1BG     8.137994e-02
## A1BG-AS1 2.475171e-05
## A1CF     2.866437e-01
## A2M      5.652116e-03
## A2M-AS1  3.518589e-01
## A2ML1    3.325306e-03
```

```
###p val dif = low, chance there is a sig. dif.. threshhold = 0.05?
```

```
res_dseq_set_salmon_copy <- results(dseq_set_salmon_copy)
head(res_dseq_set_salmon_copy)
```

```
## log2 fold change (MLE): diagnosis mf vs normal
## Wald test p-value: diagnosis mf vs normal
## DataFrame with 6 rows and 6 columns
##             baseMean log2FoldChange     lfcSE       stat       pvalue
##            <numeric>      <numeric> <numeric>  <numeric>    <numeric>
## A1BG     171.549321      0.8942849 0.1972155  4.5345574 5.772439e-06
## A1BG-AS1  83.055858     -1.7277027 0.2393654 -7.2178460 5.281749e-13
## A1CF       2.101638     -4.2269266 2.8041804 -1.5073662 1.317168e-01
## A2M        7.180859     -5.9996433 1.6872217 -3.5559307 3.766434e-04
## A2M-AS1    8.226102      0.7486593 1.1921884  0.6279707 5.300231e-01
## A2ML1      6.030153      0.5870094 1.0072706  0.5827723 5.600466e-01
##                 padj
```

```
##             <numeric>
## A1BG     1.952722e-05
## A1BG-AS1 4.194193e-12
## A1CF     1.901732e-01
## A2M      9.493839e-04
## A2M-AS1  6.127089e-01
## A2ML1    6.391913e-01
```

Prep for plotting visual analysis

```
#boxplot


# htseq then salmon -
# boxplot of gene expression of gene x based on counts
#for mf,
#for normal

#Change data frame/matrix to graph from
#box1 = ggplot(data = gene_l2fc_htseqSalmon1_combo, aes(x= L2FC_SALMON, y = L2FC_HTSEQ)) + geom_boxplot
#box1
```

Log2 fold change

HTSEQ

```
###Extract log2FoldChange columns from htseq and salmon datasets,
###merge into combined DF with genes as rownames, use geom-point to plot in GGplot
### x= LF2C Salmon, Y= LF2C HTSEQ



###resDseqCopy vs. res_dseq_set_salmon_copy =diagnosis base level = normal

###head(resDseqCopy)
###head(res_dseq_set_salmon_copy)


###convert rownames to column name gene in both df's
###full join by gene


###creates two new df's with l2fc data

###create new column with values of rownames
resDseqCopy$Gene <- rownames(resDseqCopy)

head(resDseqCopy)
```

```
## log2 fold change (MLE): diagnosis mf vs normal
## Wald test p-value: diagnosis mf vs normal
```

```
## DataFrame with 6 rows and 7 columns
##              baseMean log2FoldChange     lfcSE       stat       pvalue
##             <numeric>      <numeric> <numeric>  <numeric>    <numeric>
## A1BG     170.6840521      0.7958968 0.1953430  4.0743564 4.614177e-05
## A1BG-AS1  16.1712821     -1.5957025 0.3378240 -4.7234733 2.318504e-06
## A1CF       0.4571053     -2.0708240 1.4841992 -1.3952467 1.629415e-01
## A2M       47.3634595     -1.5682386 0.4316494 -3.6331307 2.800031e-04
## A2M-AS1   29.7386172     -0.3417973 0.3672911 -0.9305897 3.520659e-01
## A2ML1      7.2833133      1.4117902 0.5805970  2.4316181 1.503155e-02
##                   padj        Gene
##              <numeric> <character>
## A1BG     1.291139e-04        A1BG
## A1BG-AS1 7.867270e-06    A1BG-AS1
## A1CF     2.238507e-01        A1CF
## A2M      6.915249e-04         A2M
## A2M-AS1  4.314028e-01     A2M-AS1
## A2ML1    2.700069e-02       A2ML1
```

```
###Subset/put in a copy of the gene column and logfold2 column from the htseq data into new object/df
gene_l2fc_htseq1 <- as.data.frame(resDseqCopy[ ,c(7, 2)])

###Getting rid of rownames (gene list is in new column)
rownames(gene_l2fc_htseq1) = NULL
###Repeating same process for Salmon as HTSEQ
```

SALMON

```
res_dseq_set_salmon_copy$Gene <- rownames(res_dseq_set_salmon_copy) #salmon
gene_l2fc_salmon1 <- as.data.frame(res_dseq_set_salmon_copy[ ,c(7, 2)])
rownames(gene_l2fc_salmon1) = NULL
```

Tests

```
###Tests
head(gene_l2fc_htseq1)
```

```
##       Gene log2FoldChange
## 1     A1BG      0.7958968
## 2 A1BG-AS1     -1.5957025
## 3     A1CF     -2.0708240
## 4      A2M     -1.5682386
## 5  A2M-AS1     -0.3417973
## 6    A2ML1      1.4117902
```

```
str(gene_l2fc_htseq1)
```

```
## 'data.frame':    21920 obs. of  2 variables:
##  $ Gene          : chr  "A1BG" "A1BG-AS1" "A1CF" "A2M" ...
##  $ log2FoldChange: num  0.796 -1.596 -2.071 -1.568 -0.342 ...
```

```
head(gene_l2fc_salmon1)
```

```
##       Gene log2FoldChange
```

```
## 1     A1BG      0.8942849
## 2 A1BG-AS1     -1.7277027
## 3     A1CF     -4.2269266
## 4      A2M     -5.9996433
## 5  A2M-AS1      0.7486593
## 6    A2ML1      0.5870094
```

```
str(gene_l2fc_salmon1)
```

```
## 'data.frame':    19183 obs. of  2 variables:
##  $ Gene          : chr  "A1BG" "A1BG-AS1" "A1CF" "A2M" ...
##  $ log2FoldChange: num  0.894 -1.728 -4.227 -6 0.749 ...
```

HTSEQ & Salmon

```
###Joining two DF's of LF2C by Gene
gene_l2fc_htseqSalmon1_combo <- full_join(gene_l2fc_htseq1, gene_l2fc_salmon1, by = "Gene")

#Setting Column names
colnames(gene_l2fc_htseqSalmon1_combo) <- c("Gene", "L2FC_HTSEQ","L2FC_SALMON")

head(gene_l2fc_htseqSalmon1_combo)
```

```
##       Gene L2FC_HTSEQ L2FC_SALMON
## 1     A1BG  0.7958968   0.8942849
## 2 A1BG-AS1 -1.5957025  -1.7277027
## 3     A1CF -2.0708240  -4.2269266
## 4      A2M -1.5682386  -5.9996433
## 5  A2M-AS1 -0.3417973   0.7486593
## 6    A2ML1  1.4117902   0.5870094
```

P-value

HTSEQ

```
#resDseqCopy #htseq

#gene_l2fc_htseq1_all <- as.data.frame(resDseqCopy)



#res_dseq_set_salmon_copy #Salmon


gene_pvalue_htseq1 <- as.data.frame(resDseqCopy[ ,c(7, 5)])
###Getting rid of rownames (gene list is in new column)
rownames(gene_pvalue_htseq1) = NULL
```

Salmon

```
###Repeating same process for Salmon as HTSEQ
gene_pvalue_salmon1 <- as.data.frame(res_dseq_set_salmon_copy[ ,c(7, 5)])
rownames(gene_pvalue_salmon1) = NULL
```

Tests

```
###Tests
head(gene_pvalue_htseq1)
```

```
##       Gene       pvalue
## 1     A1BG 4.614177e-05
## 2 A1BG-AS1 2.318504e-06
## 3     A1CF 1.629415e-01
## 4      A2M 2.800031e-04
## 5  A2M-AS1 3.520659e-01
## 6    A2ML1 1.503155e-02
```

```
str(gene_pvalue_htseq1)
```

```
## 'data.frame':    21920 obs. of  2 variables:
##  $ Gene  : chr  "A1BG" "A1BG-AS1" "A1CF" "A2M" ...
##  $ pvalue: num  4.61e-05 2.32e-06 1.63e-01 2.80e-04 3.52e-01 ...
```

```
head(gene_pvalue_salmon1)
```

```
##       Gene       pvalue
## 1     A1BG 5.772439e-06
## 2 A1BG-AS1 5.281749e-13
## 3     A1CF 1.317168e-01
## 4      A2M 3.766434e-04
## 5  A2M-AS1 5.300231e-01
## 6    A2ML1 5.600466e-01
```

```
str(gene_pvalue_salmon1)
```

```
## 'data.frame':    19183 obs. of  2 variables:
##  $ Gene  : chr  "A1BG" "A1BG-AS1" "A1CF" "A2M" ...
##  $ pvalue: num  5.77e-06 5.28e-13 1.32e-01 3.77e-04 5.30e-01 ...
```

HTSEQ & Salmon

```
###Joining two DF's of LF2C by Gene
gene_pvalue_htseqSalmon1_combo <- full_join(gene_pvalue_htseq1, gene_pvalue_salmon1, by = "Gene")

#Setting Column names
colnames(gene_pvalue_htseqSalmon1_combo) <- c("Gene", "PVAL_HTSEQ","PVAL_SALMON")

head(gene_pvalue_htseqSalmon1_combo)
```

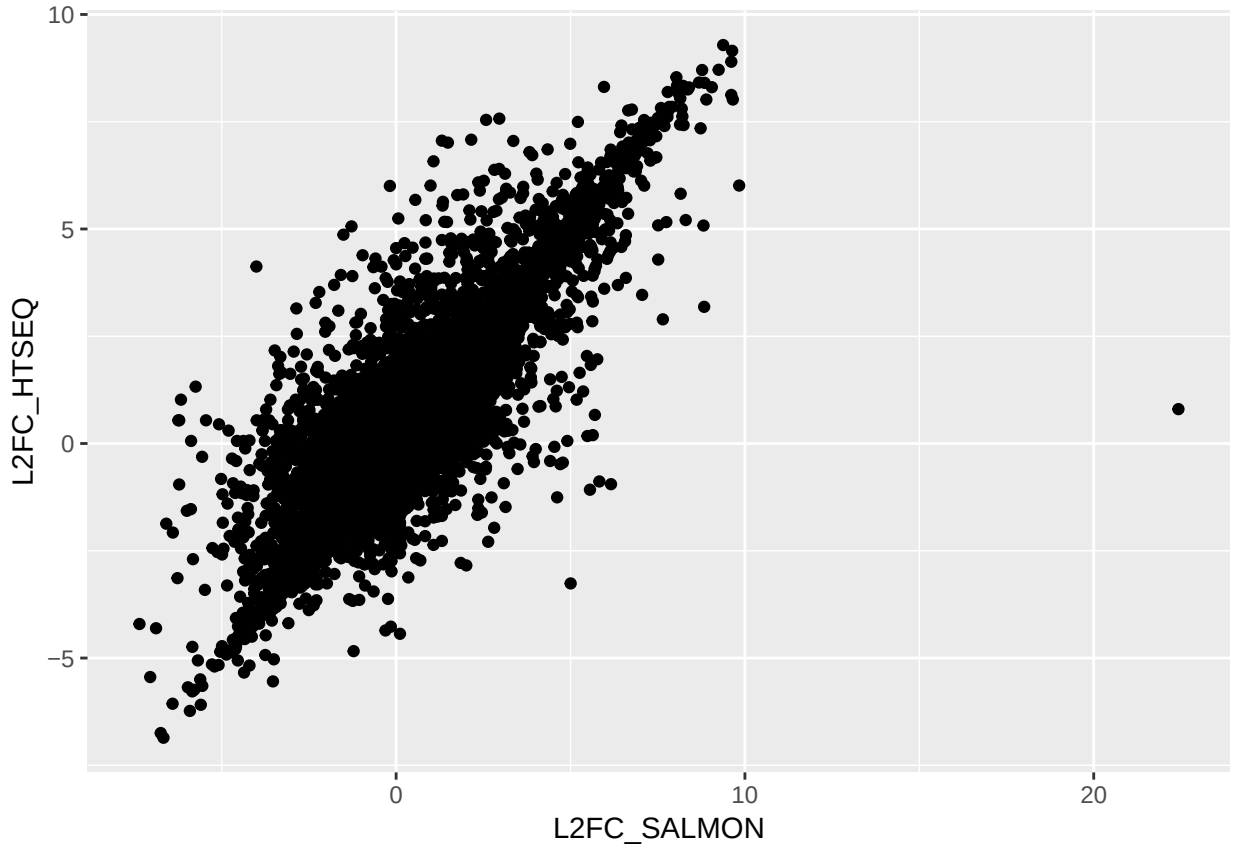
```
##       Gene   PVAL_HTSEQ  PVAL_SALMON
## 1     A1BG 4.614177e-05 5.772439e-06
## 2 A1BG-AS1 2.318504e-06 5.281749e-13
## 3     A1CF 1.629415e-01 1.317168e-01
```

```
## 4      A2M 2.800031e-04 3.766434e-04
## 5  A2M-AS1 3.520659e-01 5.300231e-01
## 6    A2ML1 1.503155e-02 5.600466e-01
```

Visualizing analysis/plotting

```
p1 = ggplot(data = gene_l2fc_htseqSalmon1_combo, aes(x= L2FC_SALMON, y = L2FC_HTSEQ)) + geom_point()
p1
```

```
## Warning: Removed 4933 rows containing missing values (geom_point).
```

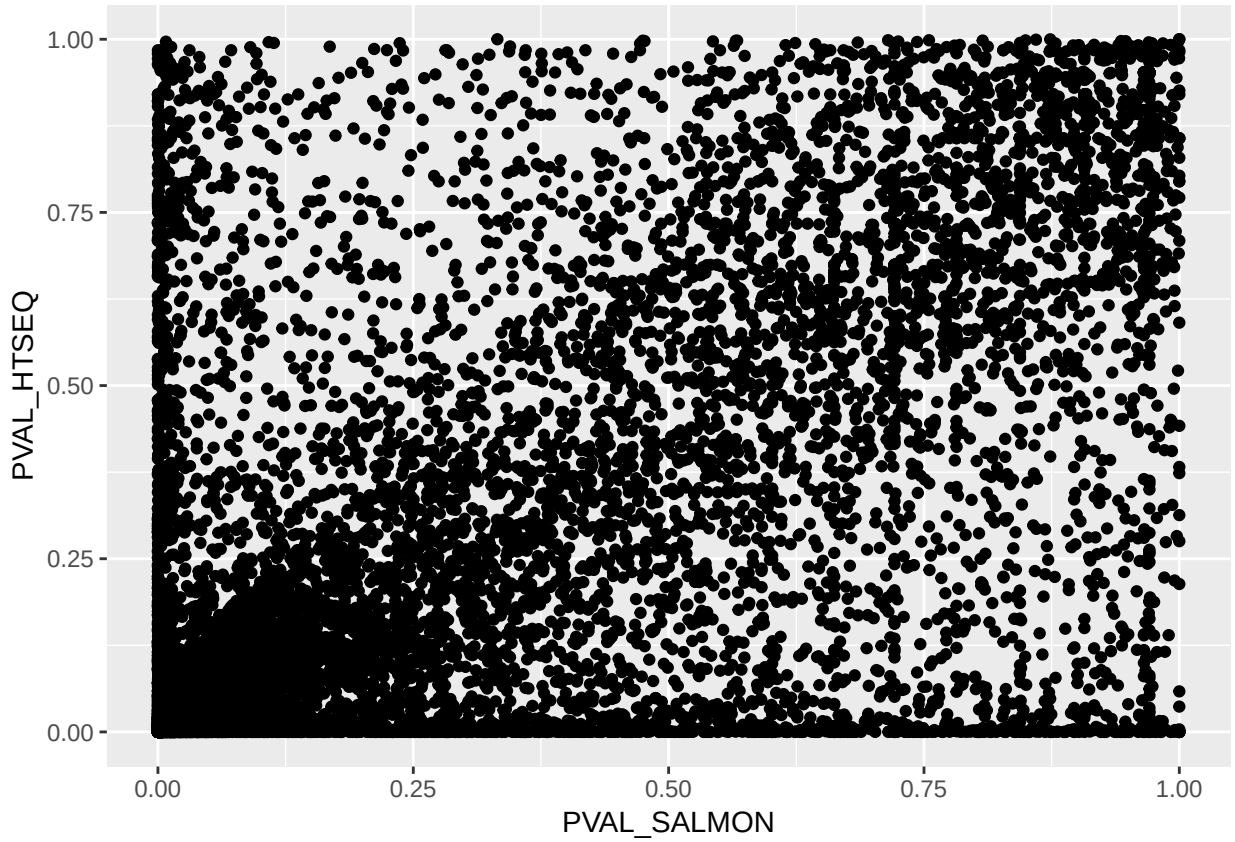



```
#EXPLAIN WHAT L2FC IS
#each dot corresponds to 2 logfoldchanges for one gene
#numbers are l2fc of mf vs normal gene expression
```

PVAL Htseq/salmon for diagnosis mf vs. normal

```
p2 = ggplot(data = gene_pvalue_htseqSalmon1_combo, aes(x = PVAL_SALMON, y = PVAL_HTSEQ)) + geom_point()
p2
```

```
## Warning: Removed 4933 rows containing missing values (geom_point).
```




```
#geom_boxplot
```