TCGA-LIHC Tumor vs. Adjacent

Robet Phavong

2024-06-27

## Set options for printing reports

# this will make sure that the code doesn't run off the page when printing a report  
knitr::opts\_chunk$set(tidy.opts = list(width.cutoff = 50), tidy = TRUE)

# Load required packages

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2)  
library(plotly)

##   
## Attaching package: 'plotly'  
##   
## The following object is masked from 'package:ggplot2':  
##   
## last\_plot  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following object is masked from 'package:graphics':  
##   
## layout

library(TCGAbiolinks)  
library(SummarizedExperiment)

## Loading required package: MatrixGenerics  
## Loading required package: matrixStats  
##   
## Attaching package: 'matrixStats'  
##   
## The following object is masked from 'package:dplyr':  
##   
## count  
##   
##   
## Attaching package: 'MatrixGenerics'  
##   
## The following objects are masked from 'package:matrixStats':  
##   
## colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,  
## colCounts, colCummaxs, colCummins, colCumprods, colCumsums,  
## colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,  
## colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,  
## colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,  
## colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,  
## colWeightedMeans, colWeightedMedians, colWeightedSds,  
## colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,  
## rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,  
## rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,  
## rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,  
## rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,  
## rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,  
## rowWeightedMads, rowWeightedMeans, rowWeightedMedians,  
## rowWeightedSds, rowWeightedVars  
##   
## Loading required package: GenomicRanges  
## Loading required package: stats4  
## Loading required package: BiocGenerics  
##   
## Attaching package: 'BiocGenerics'  
##   
## The following objects are masked from 'package:lubridate':  
##   
## intersect, setdiff, union  
##   
## The following objects are masked from 'package:dplyr':  
##   
## combine, intersect, setdiff, union  
##   
## The following objects are masked from 'package:stats':  
##   
## IQR, mad, sd, var, xtabs  
##   
## The following objects are masked from 'package:base':  
##   
## anyDuplicated, aperm, append, as.data.frame, basename, cbind,  
## colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,  
## get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,  
## match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,  
## Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,  
## table, tapply, union, unique, unsplit, which.max, which.min  
##   
## Loading required package: S4Vectors  
##   
## Attaching package: 'S4Vectors'  
##   
## The following object is masked from 'package:plotly':  
##   
## rename  
##   
## The following objects are masked from 'package:lubridate':  
##   
## second, second<-  
##   
## The following objects are masked from 'package:dplyr':  
##   
## first, rename  
##   
## The following object is masked from 'package:tidyr':  
##   
## expand  
##   
## The following object is masked from 'package:utils':  
##   
## findMatches  
##   
## The following objects are masked from 'package:base':  
##   
## expand.grid, I, unname  
##   
## Loading required package: IRanges  
##   
## Attaching package: 'IRanges'  
##   
## The following object is masked from 'package:plotly':  
##   
## slice  
##   
## The following object is masked from 'package:lubridate':  
##   
## %within%  
##   
## The following objects are masked from 'package:dplyr':  
##   
## collapse, desc, slice  
##   
## The following object is masked from 'package:purrr':  
##   
## reduce  
##   
## Loading required package: GenomeInfoDb  
## Loading required package: Biobase  
## Welcome to Bioconductor  
##   
## Vignettes contain introductory material; view with  
## 'browseVignettes()'. To cite Bioconductor, see  
## 'citation("Biobase")', and for packages 'citation("pkgname")'.  
##   
##   
## Attaching package: 'Biobase'  
##   
## The following object is masked from 'package:MatrixGenerics':  
##   
## rowMedians  
##   
## The following objects are masked from 'package:matrixStats':  
##   
## anyMissing, rowMedians

library(maftools)  
library(dplyr)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(tidyr)  
library(reshape2)

##   
## Attaching package: 'reshape2'  
##   
## The following object is masked from 'package:tidyr':  
##   
## smiths

library(tibble)

# Set directories

working\_path <- "/Users/robertphavongurquidez/Desktop/BIO 593 Applied Project/TCGA\_Tumor\_Adjacent/"  
  
setwd(working\_path)  
  
filepath <- "/Users/robertphavongurquidez/Desktop/BIO 593 Applied Project/TCGA\_Tumor\_Adjacent/"

# Get list of projects within the GDC Portal  
gdcprojects <- getGDCprojects()  
  
getProjectSummary("TCGA-LIHC")

## $file\_count  
## [1] 20820  
##   
## $data\_categories  
## file\_count case\_count data\_category  
## 1 6070 377 Simple Nucleotide Variation  
## 2 3226 377 Sequencing Reads  
## 3 1634 377 Biospecimen  
## 4 803 377 Clinical  
## 5 4197 377 Copy Number Variation  
## 6 1698 376 Transcriptome Profiling  
## 7 1290 377 DNA Methylation  
## 8 184 184 Proteome Profiling  
## 9 22 11 Somatic Structural Variation  
## 10 1696 371 Structural Variation  
##   
## $case\_count  
## [1] 377  
##   
## $file\_size  
## [1] 2.340194e+14

# Building a query for the extraction of project of interest

query\_transcriptome\_LIHC <- GDCquery(project = "TCGA-LIHC",  
 data.category = "Transcriptome Profiling", experimental.strategy = "RNA-Seq",  
 workflow.type = "STAR - Counts", access = "open")

## --------------------------------------

## o GDCquery: Searching in GDC database

## --------------------------------------

## Genome of reference: hg38

## --------------------------------------------

## oo Accessing GDC. This might take a while...

## --------------------------------------------

## ooo Project: TCGA-LIHC

## --------------------

## oo Filtering results

## --------------------

## ooo By access

## ooo By experimental.strategy

## ooo By workflow.type

## ----------------

## oo Checking data

## ----------------

## ooo Checking if there are duplicated cases

## ooo Checking if there are results for the query

## -------------------

## o Preparing output

## -------------------

# Extract the files that are only ‘Transcriptome Profiling’

output\_query\_TCGA\_LIHC\_transcriptome <- getResults(query\_transcriptome\_LIHC)  
  
head(query\_transcriptome\_LIHC)

## results project data.category data.type access  
## 1 c("a7c4c.... TCGA-LIHC Transcriptome Profiling NA open  
## experimental.strategy platform sample.type barcode workflow.type  
## 1 RNA-Seq NA NA NA STAR - Counts

# Build a query to retrieve gene expression data

# Separate the 'Primary Tumor'  
query\_TCGA\_LIHC\_transcriptome\_primary\_tumor <- GDCquery(project = "TCGA-LIHC",  
 data.category = "Transcriptome Profiling", experimental.strategy = "RNA-Seq",  
 data.type = "Gene Expression Quantification", sample.type = "Primary Tumor",  
 workflow.type = "STAR - Counts", access = "open")

## --------------------------------------

## o GDCquery: Searching in GDC database

## --------------------------------------

## Genome of reference: hg38

## --------------------------------------------

## oo Accessing GDC. This might take a while...

## --------------------------------------------

## ooo Project: TCGA-LIHC

## --------------------

## oo Filtering results

## --------------------

## ooo By access

## ooo By experimental.strategy

## ooo By data.type

## ooo By workflow.type

## ooo By sample.type

## ----------------

## oo Checking data

## ----------------

## ooo Checking if there are duplicated cases

## ooo Checking if there are results for the query

## -------------------

## o Preparing output

## -------------------

# Separate the 'Recurrent Tumor'  
query\_TCGA\_LIHC\_transcriptome\_recurrent\_tumor <- GDCquery(project = "TCGA-LIHC",  
 data.category = "Transcriptome Profiling", experimental.strategy = "RNA-Seq",  
 data.type = "Gene Expression Quantification", sample.type = "Recurrent Tumor",  
 workflow.type = "STAR - Counts", access = "open")

## --------------------------------------

## o GDCquery: Searching in GDC database

## --------------------------------------

## Genome of reference: hg38

## --------------------------------------------

## oo Accessing GDC. This might take a while...

## --------------------------------------------

## ooo Project: TCGA-LIHC

## --------------------

## oo Filtering results

## --------------------

## ooo By access

## ooo By experimental.strategy

## ooo By data.type

## ooo By workflow.type

## ooo By sample.type

## ----------------

## oo Checking data

## ----------------

## ooo Checking if there are duplicated cases

## ooo Checking if there are results for the query

## -------------------

## o Preparing output

## -------------------

# Separate the 'Solid Tissue Normal'  
query\_TCGA\_LIHC\_transcriptome\_solid\_tissue\_normal <- GDCquery(project = "TCGA-LIHC",  
 data.category = "Transcriptome Profiling", experimental.strategy = "RNA-Seq",  
 data.type = "Gene Expression Quantification", sample.type = "Solid Tissue Normal",  
 workflow.type = "STAR - Counts", access = "open")

## --------------------------------------

## o GDCquery: Searching in GDC database

## --------------------------------------

## Genome of reference: hg38

## --------------------------------------------

## oo Accessing GDC. This might take a while...

## --------------------------------------------

## ooo Project: TCGA-LIHC

## --------------------

## oo Filtering results

## --------------------

## ooo By access

## ooo By experimental.strategy

## ooo By data.type

## ooo By workflow.type

## ooo By sample.type

## ----------------

## oo Checking data

## ----------------

## ooo Checking if there are duplicated cases

## ooo Checking if there are results for the query

## -------------------

## o Preparing output

## -------------------

# Download data of interest from files filtered for gene expression from Project TCGA-LIHC using GDCdownload

GDCdownload(query\_TCGA\_LIHC\_transcriptome\_primary\_tumor)

## Downloading data for project TCGA-LIHC

## Of the 371 files for download 371 already exist.

## All samples have been already downloaded

GDCdownload(query\_TCGA\_LIHC\_transcriptome\_recurrent\_tumor)

## Downloading data for project TCGA-LIHC

## Of the 3 files for download 3 already exist.

## All samples have been already downloaded

GDCdownload(query\_TCGA\_LIHC\_transcriptome\_solid\_tissue\_normal)

## Downloading data for project TCGA-LIHC

## Of the 50 files for download 50 already exist.

## All samples have been already downloaded

# Prepare data to extract information from the data selected

# Primary Tumor  
tcga\_LIHC\_transcriptome\_primary\_tumor <- GDCprepare(query\_TCGA\_LIHC\_transcriptome\_primary\_tumor,  
 summarizedExperiment = FALSE)

## | | 0% | |0.2695418% ~55 s remaining| |0.5390836% ~44 s remaining| |0.8086253% ~39 s remaining| |1.078167% ~43 s remaining | |1.347709% ~40 s remaining | |1.617251% ~39 s remaining | |1.886792% ~38 s remaining |= |2.156334% ~37 s remaining |= |2.425876% ~55 s remaining |= |2.695418% ~53 s remaining |= |2.96496% ~51 s remaining |= |3.234501% ~49 s remaining |= |3.504043% ~49 s remaining |= |3.773585% ~48 s remaining |== |4.043127% ~47 s remaining |== |4.312668% ~47 s remaining |== |4.58221% ~46 s remaining |== |4.851752% ~46 s remaining |== |5.121294% ~46 s remaining |== |5.390836% ~54 s remaining |== |5.660377% ~53 s remaining |=== |5.929919% ~52 s remaining |=== |6.199461% ~52 s remaining |=== |6.469003% ~51 s remaining |=== |6.738544% ~50 s remaining |=== |7.008086% ~49 s remaining |=== |7.277628% ~49 s remaining |=== |7.54717% ~48 s remaining |==== |7.816712% ~48 s remaining |==== |8.086253% ~47 s remaining |==== |8.355795% ~46 s remaining |==== |8.625337% ~52 s remaining |==== |8.894879% ~51 s remaining |==== |9.16442% ~50 s remaining |==== |9.433962% ~50 s remaining |===== |9.703504% ~49 s remaining |===== |9.973046% ~49 s remaining |===== |10.24259% ~48 s remaining |===== |10.51213% ~48 s remaining |===== |10.78167% ~48 s remaining |===== |11.05121% ~47 s remaining |===== |11.32075% ~47 s remaining |====== |11.5903% ~47 s remaining |====== |11.85984% ~46 s remaining |====== |12.12938% ~46 s remaining |====== |12.39892% ~46 s remaining |====== |12.66846% ~51 s remaining |====== |12.93801% ~50 s remaining |====== |13.20755% ~50 s remaining |======= |13.47709% ~50 s remaining |======= |13.74663% ~50 s remaining |======= |14.01617% ~51 s remaining |======= |14.28571% ~51 s remaining |======= |14.55526% ~51 s remaining |======= |14.8248% ~52 s remaining |======= |15.09434% ~53 s remaining |======= |15.36388% ~53 s remaining |======== |15.63342% ~54 s remaining |======== |15.90296% ~55 s remaining |======== |16.17251% ~56 s remaining |======== |16.44205% ~55 s remaining |======== |16.71159% ~55 s remaining |======== |16.98113% ~55 s remaining |======== |17.25067% ~55 s remaining |========= |17.52022% ~1 m remaining |========= |17.78976% ~1 m remaining |========= |18.0593% ~1 m remaining |========= |18.32884% ~1 m remaining |========= |18.59838% ~1 m remaining |========= |18.86792% ~60 s remaining |========= |19.13747% ~59 s remaining |========== |19.40701% ~59 s remaining |========== |19.67655% ~58 s remaining |========== |19.94609% ~58 s remaining |========== |20.21563% ~57 s remaining |========== |20.48518% ~57 s remaining |========== |20.75472% ~56 s remaining |========== |21.02426% ~56 s remaining |=========== |21.2938% ~55 s remaining |=========== |21.56334% ~55 s remaining |=========== |21.83288% ~54 s remaining |=========== |22.10243% ~54 s remaining |=========== |22.37197% ~54 s remaining |=========== |22.64151% ~53 s remaining |=========== |22.91105% ~55 s remaining |============ |23.18059% ~1 m remaining |============ |23.45013% ~1 m remaining |============ |23.71968% ~60 s remaining |============ |23.98922% ~59 s remaining |============ |24.25876% ~59 s remaining |============ |24.5283% ~58 s remaining |============ |24.79784% ~58 s remaining |============= |25.06739% ~57 s remaining |============= |25.33693% ~57 s remaining |============= |25.60647% ~56 s remaining |============= |25.87601% ~56 s remaining |============= |26.14555% ~55 s remaining |============= |26.41509% ~55 s remaining |============= |26.68464% ~54 s remaining |============== |26.95418% ~54 s remaining |============== |27.22372% ~53 s remaining |============== |27.49326% ~53 s remaining |============== |27.7628% ~53 s remaining |============== |28.03235% ~52 s remaining |============== |28.30189% ~52 s remaining |============== |28.57143% ~52 s remaining |============== |28.84097% ~51 s remaining |=============== |29.11051% ~51 s remaining |=============== |29.38005% ~51 s remaining |=============== |29.6496% ~50 s remaining |=============== |29.91914% ~50 s remaining |=============== |30.18868% ~51 s remaining |=============== |30.45822% ~51 s remaining |=============== |30.72776% ~50 s remaining |================ |30.9973% ~50 s remaining |================ |31.26685% ~50 s remaining |================ |31.53639% ~50 s remaining |================ |31.80593% ~50 s remaining |================ |32.07547% ~49 s remaining |================ |32.34501% ~49 s remaining |================ |32.61456% ~49 s remaining |================= |32.8841% ~48 s remaining |================= |33.15364% ~48 s remaining |================= |33.42318% ~48 s remaining |================= |33.69272% ~47 s remaining |================= |33.96226% ~47 s remaining |================= |34.23181% ~47 s remaining |================= |34.50135% ~46 s remaining |================== |34.77089% ~46 s remaining |================== |35.04043% ~46 s remaining |================== |35.30997% ~46 s remaining |================== |35.57951% ~47 s remaining |================== |35.84906% ~47 s remaining |================== |36.1186% ~47 s remaining |================== |36.38814% ~47 s remaining |=================== |36.65768% ~47 s remaining |=================== |36.92722% ~47 s remaining |=================== |37.19677% ~47 s remaining |=================== |37.46631% ~47 s remaining |=================== |37.73585% ~46 s remaining |=================== |38.00539% ~47 s remaining |=================== |38.27493% ~47 s remaining |==================== |38.54447% ~53 s remaining |==================== |38.81402% ~52 s remaining |==================== |39.08356% ~52 s remaining |==================== |39.3531% ~52 s remaining |==================== |39.62264% ~51 s remaining |==================== |39.89218% ~51 s remaining |==================== |40.16173% ~51 s remaining |===================== |40.43127% ~50 s remaining |===================== |40.70081% ~50 s remaining |===================== |40.97035% ~50 s remaining |===================== |41.23989% ~50 s remaining |===================== |41.50943% ~49 s remaining |===================== |41.77898% ~49 s remaining |===================== |42.04852% ~49 s remaining |====================== |42.31806% ~49 s remaining |====================== |42.5876% ~48 s remaining |====================== |42.85714% ~48 s remaining |====================== |43.12668% ~48 s remaining |====================== |43.39623% ~47 s remaining |====================== |43.66577% ~47 s remaining |====================== |43.93531% ~47 s remaining |====================== |44.20485% ~47 s remaining |======================= |44.47439% ~47 s remaining |======================= |44.74394% ~47 s remaining |======================= |45.01348% ~46 s remaining |======================= |45.28302% ~46 s remaining |======================= |45.55256% ~46 s remaining |======================= |45.8221% ~45 s remaining |======================= |46.09164% ~45 s remaining |======================== |46.36119% ~45 s remaining |======================== |46.63073% ~44 s remaining |======================== |46.90027% ~44 s remaining |======================== |47.16981% ~44 s remaining |======================== |47.43935% ~43 s remaining |======================== |47.70889% ~43 s remaining |======================== |47.97844% ~43 s remaining |========================= |48.24798% ~43 s remaining |========================= |48.51752% ~44 s remaining |========================= |48.78706% ~44 s remaining |========================= |49.0566% ~43 s remaining |========================= |49.32615% ~43 s remaining |========================= |49.59569% ~43 s remaining |========================= |49.86523% ~42 s remaining |========================== |50.13477% ~42 s remaining |========================== |50.40431% ~42 s remaining |========================== |50.67385% ~41 s remaining |========================== |50.9434% ~41 s remaining |========================== |51.21294% ~40 s remaining |========================== |51.48248% ~40 s remaining |========================== |51.75202% ~40 s remaining |=========================== |52.02156% ~40 s remaining |=========================== |52.29111% ~39 s remaining |=========================== |52.56065% ~39 s remaining |=========================== |52.83019% ~39 s remaining |=========================== |53.09973% ~39 s remaining |=========================== |53.36927% ~38 s remaining |=========================== |53.63881% ~38 s remaining |============================ |53.90836% ~38 s remaining |============================ |54.1779% ~38 s remaining |============================ |54.44744% ~37 s remaining |============================ |54.71698% ~37 s remaining |============================ |54.98652% ~37 s remaining |============================ |55.25606% ~36 s remaining |============================ |55.52561% ~36 s remaining |============================= |55.79515% ~36 s remaining |============================= |56.06469% ~36 s remaining |============================= |56.33423% ~35 s remaining |============================= |56.60377% ~35 s remaining |============================= |56.87332% ~35 s remaining |============================= |57.14286% ~34 s remaining |============================= |57.4124% ~34 s remaining |============================= |57.68194% ~34 s remaining |============================== |57.95148% ~34 s remaining |============================== |58.22102% ~33 s remaining |============================== |58.49057% ~33 s remaining |============================== |58.76011% ~33 s remaining |============================== |59.02965% ~32 s remaining |============================== |59.29919% ~32 s remaining |============================== |59.56873% ~32 s remaining |=============================== |59.83827% ~32 s remaining |=============================== |60.10782% ~31 s remaining |=============================== |60.37736% ~32 s remaining |=============================== |60.6469% ~31 s remaining |=============================== |60.91644% ~31 s remaining |=============================== |61.18598% ~31 s remaining |=============================== |61.45553% ~31 s remaining |================================ |61.72507% ~30 s remaining |================================ |61.99461% ~30 s remaining |================================ |62.26415% ~30 s remaining |================================ |62.53369% ~29 s remaining |================================ |62.80323% ~29 s remaining |================================ |63.07278% ~29 s remaining |================================ |63.34232% ~29 s remaining |================================= |63.61186% ~28 s remaining |================================= |63.8814% ~28 s remaining |================================= |64.15094% ~28 s remaining |================================= |64.42049% ~28 s remaining |================================= |64.69003% ~27 s remaining |================================= |64.95957% ~27 s remaining |================================= |65.22911% ~27 s remaining |================================== |65.49865% ~26 s remaining |================================== |65.76819% ~26 s remaining |================================== |66.03774% ~26 s remaining |================================== |66.30728% ~26 s remaining |================================== |66.57682% ~25 s remaining |================================== |66.84636% ~25 s remaining |================================== |67.1159% ~25 s remaining |=================================== |67.38544% ~25 s remaining |=================================== |67.65499% ~25 s remaining |=================================== |67.92453% ~24 s remaining |=================================== |68.19407% ~24 s remaining |=================================== |68.46361% ~24 s remaining |=================================== |68.73315% ~24 s remaining |=================================== |69.0027% ~23 s remaining |==================================== |69.27224% ~23 s remaining |==================================== |69.54178% ~23 s remaining |==================================== |69.81132% ~23 s remaining |==================================== |70.08086% ~22 s remaining |==================================== |70.3504% ~22 s remaining |==================================== |70.61995% ~22 s remaining |==================================== |70.88949% ~22 s remaining |===================================== |71.15903% ~21 s remaining |===================================== |71.42857% ~21 s remaining |===================================== |71.69811% ~21 s remaining |===================================== |71.96765% ~21 s remaining |===================================== |72.2372% ~20 s remaining |===================================== |72.50674% ~20 s remaining |===================================== |72.77628% ~20 s remaining |===================================== |73.04582% ~20 s remaining |====================================== |73.31536% ~19 s remaining |====================================== |73.58491% ~19 s remaining |====================================== |73.85445% ~19 s remaining |====================================== |74.12399% ~19 s remaining |====================================== |74.39353% ~19 s remaining |====================================== |74.66307% ~18 s remaining |====================================== |74.93261% ~18 s remaining |======================================= |75.20216% ~18 s remaining |======================================= |75.4717% ~18 s remaining |======================================= |75.74124% ~18 s remaining |======================================= |76.01078% ~18 s remaining |======================================= |76.28032% ~17 s remaining |======================================= |76.54987% ~17 s remaining |======================================= |76.81941% ~17 s remaining |======================================== |77.08895% ~17 s remaining |======================================== |77.35849% ~16 s remaining |======================================== |77.62803% ~16 s remaining |======================================== |77.89757% ~16 s remaining |======================================== |78.16712% ~16 s remaining |======================================== |78.43666% ~15 s remaining |======================================== |78.7062% ~15 s remaining |========================================= |78.97574% ~15 s remaining |========================================= |79.24528% ~15 s remaining |========================================= |79.51482% ~15 s remaining |========================================= |79.78437% ~14 s remaining |========================================= |80.05391% ~14 s remaining |========================================= |80.32345% ~14 s remaining |========================================= |80.59299% ~14 s remaining |========================================== |80.86253% ~14 s remaining |========================================== |81.13208% ~13 s remaining |========================================== |81.40162% ~13 s remaining |========================================== |81.67116% ~13 s remaining |========================================== |81.9407% ~13 s remaining |========================================== |82.21024% ~12 s remaining |========================================== |82.47978% ~12 s remaining |=========================================== |82.74933% ~12 s remaining |=========================================== |83.01887% ~12 s remaining |=========================================== |83.28841% ~12 s remaining |=========================================== |83.55795% ~11 s remaining |=========================================== |83.82749% ~11 s remaining |=========================================== |84.09704% ~11 s remaining |=========================================== |84.36658% ~11 s remaining |============================================ |84.63612% ~11 s remaining |============================================ |84.90566% ~10 s remaining |============================================ |85.1752% ~10 s remaining |============================================ |85.44474% ~10 s remaining |============================================ |85.71429% ~10 s remaining |============================================ |85.98383% ~10 s remaining |============================================ |86.25337% ~10 s remaining |============================================ |86.52291% ~9 s remaining |============================================= |86.79245% ~9 s remaining |============================================= |87.06199% ~9 s remaining |============================================= |87.33154% ~9 s remaining |============================================= |87.60108% ~9 s remaining |============================================= |87.87062% ~8 s remaining |============================================= |88.14016% ~8 s remaining |============================================= |88.4097% ~8 s remaining |============================================== |88.67925% ~8 s remaining |============================================== |88.94879% ~8 s remaining |============================================== |89.21833% ~7 s remaining |============================================== |89.48787% ~7 s remaining |============================================== |89.75741% ~7 s remaining |============================================== |90.02695% ~7 s remaining |============================================== |90.2965% ~7 s remaining |=============================================== |90.56604% ~6 s remaining |=============================================== |90.83558% ~6 s remaining |=============================================== |91.10512% ~6 s remaining |=============================================== |91.37466% ~6 s remaining |=============================================== |91.6442% ~6 s remaining |=============================================== |91.91375% ~5 s remaining |=============================================== |92.18329% ~5 s remaining |================================================ |92.45283% ~5 s remaining |================================================ |92.72237% ~5 s remaining |================================================ |92.99191% ~5 s remaining |================================================ |93.26146% ~5 s remaining |================================================ |93.531% ~4 s remaining |================================================ |93.80054% ~4 s remaining |================================================ |94.07008% ~4 s remaining |================================================= |94.33962% ~4 s remaining |================================================= |94.60916% ~4 s remaining |================================================= |94.87871% ~4 s remaining |================================================= |95.14825% ~3 s remaining |================================================= |95.41779% ~3 s remaining |================================================= |95.68733% ~3 s remaining |================================================= |95.95687% ~3 s remaining |================================================== |96.22642% ~3 s remaining |================================================== |96.49596% ~2 s remaining |================================================== |96.7655% ~2 s remaining |================================================== |97.03504% ~2 s remaining |================================================== |97.30458% ~2 s remaining |================================================== |97.57412% ~2 s remaining |================================================== |97.84367% ~1 s remaining |=================================================== |98.11321% ~1 s remaining |=================================================== |98.38275% ~1 s remaining |=================================================== |98.65229% ~1 s remaining |=================================================== |98.92183% ~1 s remaining |=================================================== |99.19137% ~1 s remaining |=================================================== |99.46092% ~0 s remaining |=================================================== |99.73046% ~0 s remaining |====================================================|100% ~0 s remaining |====================================================|100% Completed after 1 m

# Recurrent Tumor  
tcga\_LIHC\_transcriptome\_recurrent\_tumor <- GDCprepare(query\_TCGA\_LIHC\_transcriptome\_recurrent\_tumor,  
 summarizedExperiment = FALSE)

## | | 0% |================= |33.33333% ~1 s remaining |================================== |66.66667% ~0 s remaining |====================================================|100% ~0 s remaining |====================================================|100% Completed after 1 s

# Solid Normal Tissue  
tcga\_LIHC\_transcriptome\_solid\_tissue\_normal <- GDCprepare(query\_TCGA\_LIHC\_transcriptome\_solid\_tissue\_normal,  
 summarizedExperiment = FALSE)

## | | 0% |= | 2% ~10 s remaining |== | 4% ~14 s remaining |=== | 6% ~12 s remaining |==== | 8% ~10 s remaining |===== | 10% ~9 s remaining |====== | 12% ~9 s remaining |======= | 14% ~8 s remaining |======== | 16% ~7 s remaining |========= | 18% ~7 s remaining |========== | 20% ~6 s remaining |=========== | 22% ~6 s remaining |============ | 24% ~6 s remaining |============= | 26% ~6 s remaining |============== | 28% ~6 s remaining |=============== | 30% ~5 s remaining |================ | 32% ~5 s remaining |================= | 34% ~5 s remaining |================== | 36% ~5 s remaining |=================== | 38% ~5 s remaining |==================== | 40% ~4 s remaining |===================== | 42% ~4 s remaining |====================== | 44% ~4 s remaining |======================= | 46% ~4 s remaining |======================== | 48% ~4 s remaining |========================== | 50% ~4 s remaining |=========================== | 52% ~4 s remaining |============================ | 54% ~4 s remaining |============================= | 56% ~3 s remaining |============================== | 58% ~3 s remaining |=============================== | 60% ~3 s remaining |================================ | 62% ~3 s remaining |================================= | 64% ~3 s remaining |================================== | 66% ~3 s remaining |=================================== | 68% ~2 s remaining |==================================== | 70% ~2 s remaining |===================================== | 72% ~2 s remaining |====================================== | 74% ~2 s remaining |======================================= | 76% ~2 s remaining |======================================== | 78% ~2 s remaining |========================================= | 80% ~2 s remaining |========================================== | 82% ~1 s remaining |=========================================== | 84% ~1 s remaining |============================================ | 86% ~1 s remaining |============================================= | 88% ~1 s remaining |============================================== | 90% ~1 s remaining |=============================================== | 92% ~1 s remaining |================================================ | 94% ~0 s remaining |================================================= | 96% ~0 s remaining |================================================== | 98% ~0 s remaining |====================================================|100% ~0 s remaining |====================================================|100% Completed after 8 s

# Extract the Genes of Interest and their TPM counts for Primary Tumor

## Primary Tumor Define the gene names of  
## interest  
genes\_of\_interest <- c("XIST", "RPS4Y1", "ZFY", "USP9Y",  
 "DDX3Y", "UTY", "TMSB4Y", "EIF1AY", "NLGN4Y", "KDM5D")  
  
# Filter rows to keep only the genes of interest  
filtered\_genes\_primary\_tumor <- tcga\_LIHC\_transcriptome\_primary\_tumor %>%  
 filter(gene\_name %in% genes\_of\_interest)  
  
# Select columns that start with  
# 'tpm\_unstranded\_' and the 'gene\_name' column  
genes\_tpm\_primary\_tumor <- filtered\_genes\_primary\_tumor %>%  
 select(gene\_name, starts\_with("tpm\_unstranded\_"))  
  
# Transform data, so that cases are on the 'rows'  
# and the genes are as columns  
genes\_tpm\_primary\_tumor <- t(genes\_tpm\_primary\_tumor)  
  
# Set the first row as column names  
colnames(genes\_tpm\_primary\_tumor) <- as.character(genes\_tpm\_primary\_tumor[1,  
 ])  
  
# Remove the first row  
genes\_tpm\_primary\_tumor <- genes\_tpm\_primary\_tumor[-1,  
 ]  
  
# Convert the data to a data frame  
genes\_tpm\_primary\_tumor <- as.data.frame(genes\_tpm\_primary\_tumor)

# Extract the Genes of Interest and their TPM counts for Recurrent Tumor

## Recurrent Tumor Define the gene names of  
## interest  
genes\_of\_interest <- c("XIST", "RPS4Y1", "ZFY", "USP9Y",  
 "DDX3Y", "UTY", "TMSB4Y", "EIF1AY", "NLGN4Y", "KDM5D")  
  
# Filter rows to keep only the genes of interest  
filtered\_genes\_recurrent\_tumor <- tcga\_LIHC\_transcriptome\_recurrent\_tumor %>%  
 filter(gene\_name %in% genes\_of\_interest)  
  
# Select columns that start with  
# 'tpm\_unstranded\_' and the 'gene\_name' column  
genes\_tpm\_recurrent\_tumor <- filtered\_genes\_recurrent\_tumor %>%  
 select(gene\_name, starts\_with("tpm\_unstranded\_"))  
  
# Transform data, so that cases are on the 'rows'  
# and the genes are as columns  
genes\_tpm\_recurrent\_tumor <- t(genes\_tpm\_recurrent\_tumor)  
  
# Set the first row as column names  
colnames(genes\_tpm\_recurrent\_tumor) <- as.character(genes\_tpm\_recurrent\_tumor[1,  
 ])  
  
# Remove the first row  
genes\_tpm\_recurrent\_tumor <- genes\_tpm\_recurrent\_tumor[-1,  
 ]  
  
# Convert the data to a data frame  
genes\_tpm\_recurrent\_tumor <- as.data.frame(genes\_tpm\_recurrent\_tumor)

# Extract the Genes of Interest and their TPM counts for Solid Tissue Normal

## Recurrent Tumor Define the gene names of  
## interest  
genes\_of\_interest <- c("XIST", "RPS4Y1", "ZFY", "USP9Y",  
 "DDX3Y", "UTY", "TMSB4Y", "EIF1AY", "NLGN4Y", "KDM5D")  
  
# Filter rows to keep only the genes of interest  
filtered\_genes\_solid\_tissue\_normal <- tcga\_LIHC\_transcriptome\_solid\_tissue\_normal %>%  
 filter(gene\_name %in% genes\_of\_interest)  
  
# Select columns that start with  
# 'tpm\_unstranded\_' and the 'gene\_name' column  
genes\_tpm\_solid\_tissue\_normal <- filtered\_genes\_solid\_tissue\_normal %>%  
 select(gene\_name, starts\_with("tpm\_unstranded\_"))  
  
# Transform data, so that cases are on the 'rows'  
# and the genes are as columns  
genes\_tpm\_solid\_tissue\_normal <- t(genes\_tpm\_solid\_tissue\_normal)  
  
# Set the first row as column names  
colnames(genes\_tpm\_solid\_tissue\_normal) <- as.character(genes\_tpm\_solid\_tissue\_normal[1,  
 ])  
  
# Remove the first row  
genes\_tpm\_solid\_tissue\_normal <- genes\_tpm\_solid\_tissue\_normal[-1,  
 ]  
  
# Convert the data to a data frame  
genes\_tpm\_solid\_tissue\_normal <- as.data.frame(genes\_tpm\_solid\_tissue\_normal)

# Remove ‘tpm\_unstranded\_’ from all the cases

# Remove 'tpm\_unstranded\_' from the row names of  
# genes\_tpm\_primary\_tumor  
rownames(genes\_tpm\_primary\_tumor) <- sub("^tpm\_unstranded\_",  
 "", rownames(genes\_tpm\_primary\_tumor))  
  
# Remove 'tpm\_unstranded\_' from the row names of  
# genes\_tpm\_recurrent\_tumor  
rownames(genes\_tpm\_recurrent\_tumor) <- sub("^tpm\_unstranded\_",  
 "", rownames(genes\_tpm\_recurrent\_tumor))  
  
# Remove 'tpm\_unstranded\_' from the row names of  
# genes\_tpm\_solid\_tissue\_normal  
rownames(genes\_tpm\_solid\_tissue\_normal) <- sub("^tpm\_unstranded\_",  
 "", rownames(genes\_tpm\_solid\_tissue\_normal))  
  
# Verify the changes  
head(rownames(genes\_tpm\_primary\_tumor))

## [1] "TCGA-5R-AA1C-01A-11R-A41C-07" "TCGA-ED-A66Y-01A-11R-A311-07"  
## [3] "TCGA-DD-AAEI-01A-11R-A41C-07" "TCGA-ED-A8O5-01A-11R-A36F-07"  
## [5] "TCGA-DD-AACH-01A-11R-A41C-07" "TCGA-FV-A3R3-01A-11R-A22L-07"

head(rownames(genes\_tpm\_recurrent\_tumor))

## [1] "TCGA-DD-AACA-02A-11R-A41C-07" "TCGA-DD-AACA-02B-11R-A41C-07"  
## [3] "TCGA-ZS-A9CF-02A-11R-A38B-07"

head(rownames(genes\_tpm\_solid\_tissue\_normal))

## [1] "TCGA-DD-A3A2-11A-11R-A213-07" "TCGA-DD-A39Z-11A-21R-A213-07"  
## [3] "TCGA-FV-A3I0-11A-11R-A22L-07" "TCGA-DD-A39X-11A-11R-A213-07"  
## [5] "TCGA-BC-A10U-11A-11R-A131-07" "TCGA-DD-A114-11A-12R-A131-07"

# Combine ‘cases’, ‘cases.sumbitter\_id’, and ‘sample\_type’

## Note, cases.submitter\_id: The unique key that was used for the case that links the demographic entity to the case. (<https://docs.gdc.cancer.gov/Data_Submission_Portal/Users_Guide/Data_Submission_Walkthrough/> )

# From the  
# 'output\_query\_TCGA\_LIHC\_transcriptome'extract  
# 'cases' and 'cases.sumbitter\_id so that we can  
# combine to the data below to correlate cases  
# between primary, solid tissue normal, and  
# recurrent  
cases\_column <- output\_query\_TCGA\_LIHC\_transcriptome %>%  
 select(cases, cases.submitter\_id, sample\_type)  
  
# Print the cases\_column  
head(cases\_column)

## cases cases.submitter\_id sample\_type  
## 1 TCGA-5R-AA1C-01A-11R-A41C-07 TCGA-5R-AA1C Primary Tumor  
## 2 TCGA-ED-A66Y-01A-11R-A311-07 TCGA-ED-A66Y Primary Tumor  
## 3 TCGA-DD-AAEI-01A-11R-A41C-07 TCGA-DD-AAEI Primary Tumor  
## 4 TCGA-ED-A8O5-01A-11R-A36F-07 TCGA-ED-A8O5 Primary Tumor  
## 5 TCGA-DD-A3A2-11A-11R-A213-07 TCGA-DD-A3A2 Solid Tissue Normal  
## 6 TCGA-DD-AACH-01A-11R-A41C-07 TCGA-DD-AACH Primary Tumor

# Separate the cases based on 'sample\_type'  
# Subset for 'Primary Tumor'  
primary\_tumor\_cases <- cases\_column[cases\_column$sample\_type ==  
 "Primary Tumor", ]  
  
# Subset for 'Solid Tissue Normal'  
solid\_tissue\_normal\_cases <- cases\_column[cases\_column$sample\_type ==  
 "Solid Tissue Normal", ]  
  
# Subset for 'Recurrent Tumor'  
recurrent\_tumor\_cases <- cases\_column[cases\_column$sample\_type ==  
 "Recurrent Tumor", ]  
  
  
  
  
#### Add 'cases.submitter\_id' to gene\_tpm  
#### dataframes Primary Tumor Convert the row  
#### names of genes\_tpm\_primary\_tumor to a column  
genes\_tpm\_primary\_tumor$cases <- rownames(genes\_tpm\_primary\_tumor)  
  
# Perform the merge operation to add  
# 'cases.submitter\_id'  
genes\_tpm\_primary\_tumor\_merged <- merge(genes\_tpm\_primary\_tumor,  
 primary\_tumor\_cases[, c("cases", "cases.submitter\_id")],  
 by = "cases")  
  
# Reorder columns to make 'cases.submitter\_id'  
# the second column Get the current column names  
col\_names\_primary\_tumor <- colnames(genes\_tpm\_primary\_tumor\_merged)  
  
# Create a new order for the columns  
new\_order\_primary\_tumor <- c(col\_names\_primary\_tumor[1],  
 "cases.submitter\_id", col\_names\_primary\_tumor[2:(ncol(genes\_tpm\_primary\_tumor\_merged) -  
 1)])  
  
# Reorder the columns  
genes\_tpm\_primary\_tumor\_merged <- genes\_tpm\_primary\_tumor\_merged[,  
 new\_order\_primary\_tumor]  
  
# Remove the first column  
genes\_tpm\_primary\_tumor\_merged <- genes\_tpm\_primary\_tumor\_merged[,  
 -1]  
  
# Set 'cases.submitter\_id' as row names  
rownames(genes\_tpm\_primary\_tumor\_merged) <- genes\_tpm\_primary\_tumor\_merged$cases.submitter\_id  
  
# Remove the 'cases.submitter\_id' column  
# genes\_tpm\_primary\_tumor\_merged$cases.submitter\_id  
# <- NULL  
  
# Print the first few rows to verify the result  
head(genes\_tpm\_primary\_tumor\_merged)

## cases.submitter\_id KDM5D DDX3Y ZFY USP9Y RPS4Y1  
## TCGA-2V-A95S TCGA-2V-A95S 9.7150 13.7261 2.7309 1.3592 136.3996  
## TCGA-2Y-A9GS TCGA-2Y-A9GS 14.9729 24.2846 5.4247 0.8288 233.6748  
## TCGA-2Y-A9GT TCGA-2Y-A9GT 7.4837 19.2319 3.4732 0.7611 68.2538  
## TCGA-2Y-A9GU TCGA-2Y-A9GU 0.0268 0.0222 0.0114 0.0035 0.2648  
## TCGA-2Y-A9GV TCGA-2Y-A9GV 0.0000 0.0000 0.0000 0.0000 0.0000  
## TCGA-2Y-A9GW TCGA-2Y-A9GW 6.1405 10.7674 2.3214 0.4667 139.2509  
## TMSB4Y NLGN4Y UTY EIF1AY XIST  
## TCGA-2V-A95S 0.5628 1.2850 2.5040 26.3903 0.0174  
## TCGA-2Y-A9GS 2.2689 0.5371 6.4703 22.1026 0.0000  
## TCGA-2Y-A9GT 1.3173 1.4525 3.2338 15.1041 0.0029  
## TCGA-2Y-A9GU 0.0000 0.0148 0.0000 0.0158 6.8480  
## TCGA-2Y-A9GV 0.0000 0.0000 0.0000 0.0000 10.4952  
## TCGA-2Y-A9GW 0.3415 1.5551 2.1054 10.6414 0.0000

#### Add 'cases.submitter\_id' to gene\_tpm  
#### dataframes Recurrent Tumor Convert the row  
#### names of genes\_tpm\_recurrent\_tumor to a  
#### column  
genes\_tpm\_recurrent\_tumor$cases <- rownames(genes\_tpm\_recurrent\_tumor)  
  
# Perform the merge operation to add  
# 'cases.submitter\_id'  
genes\_tpm\_recurrent\_tumor\_merged <- merge(genes\_tpm\_recurrent\_tumor,  
 recurrent\_tumor\_cases[, c("cases", "cases.submitter\_id")],  
 by = "cases")  
  
# Reorder columns to make 'cases.submitter\_id'  
# the second column Get the current column names  
col\_names\_recurrent\_tumor <- colnames(genes\_tpm\_recurrent\_tumor\_merged)  
  
# Create a new order for the columns  
new\_order\_recurrent\_tumor <- c(col\_names\_recurrent\_tumor[1],  
 "cases.submitter\_id", col\_names\_recurrent\_tumor[2:(ncol(genes\_tpm\_recurrent\_tumor\_merged) -  
 1)])  
  
# Reorder the columns  
genes\_tpm\_recurrent\_tumor\_merged <- genes\_tpm\_recurrent\_tumor\_merged[,  
 new\_order\_recurrent\_tumor]  
  
# Remove the first column  
genes\_tpm\_recurrent\_tumor\_merged <- genes\_tpm\_recurrent\_tumor\_merged[,  
 -1]  
  
################################################# Issues  
################################################# with  
################################################# cases.submitter\_id  
################################################# that  
################################################# repeats  
################################################# ##  
################################################# #  
################################################# Set  
################################################# 'cases.submitter\_id'  
################################################# as  
################################################# row  
################################################# names  
################################################# rownames(genes\_tpm\_recurrent\_tumor\_merged)  
################################################# <-  
################################################# genes\_tpm\_recurrent\_tumor\_merged$cases.submitter\_id  
################################################# #  
################################################# Remove  
################################################# the  
################################################# 'cases.submitter\_id'  
################################################# column  
################################################# genes\_tpm\_recurrent\_tumor\_merged$cases.submitter\_id  
################################################# <-  
################################################# NULL  
################################################# #  
################################################# Print  
################################################# the  
################################################# first  
################################################# few  
################################################# rows  
################################################# to  
################################################# verify  
################################################# the  
################################################# result  
################################################# head(genes\_tpm\_recurrent\_tumor\_merged)  
  
  
  
  
#### Add 'cases.submitter\_id' to gene\_tpm  
#### dataframes Solid Tissue Normal Convert the  
#### row names of genes\_tpm\_solid\_tissue\_normal  
#### to a column  
genes\_tpm\_solid\_tissue\_normal$cases <- rownames(genes\_tpm\_solid\_tissue\_normal)  
  
# Perform the merge operation to add  
# 'cases.submitter\_id'  
genes\_tpm\_solid\_tissue\_normal\_merged <- merge(genes\_tpm\_solid\_tissue\_normal,  
 solid\_tissue\_normal\_cases[, c("cases", "cases.submitter\_id")],  
 by = "cases")  
  
# Reorder columns to make 'cases.submitter\_id'  
# the second column Get the current column names  
col\_names\_solid\_tissue\_normal <- colnames(genes\_tpm\_solid\_tissue\_normal\_merged)  
  
# Create a new order for the columns  
new\_order\_solid\_tissue\_normal <- c(col\_names\_solid\_tissue\_normal[1],  
 "cases.submitter\_id", col\_names\_solid\_tissue\_normal[2:(ncol(genes\_tpm\_solid\_tissue\_normal\_merged) -  
 1)])  
  
# Reorder the columns  
genes\_tpm\_solid\_tissue\_normal\_merged <- genes\_tpm\_solid\_tissue\_normal\_merged[,  
 new\_order\_solid\_tissue\_normal]  
  
# Remove the first column  
genes\_tpm\_solid\_tissue\_normal\_merged <- genes\_tpm\_solid\_tissue\_normal\_merged[,  
 -1]  
  
# Set 'cases.submitter\_id' as row names  
rownames(genes\_tpm\_solid\_tissue\_normal\_merged) <- genes\_tpm\_solid\_tissue\_normal\_merged$cases.submitter\_id  
  
# Remove the 'cases.submitter\_id' column  
# genes\_tpm\_solid\_tissue\_normal\_merged$cases.submitter\_id  
# <- NULL  
  
# Print the first few rows to verify the result  
head(genes\_tpm\_solid\_tissue\_normal\_merged)

## cases.submitter\_id KDM5D DDX3Y ZFY USP9Y RPS4Y1  
## TCGA-BC-A10Q TCGA-BC-A10Q 0.0261 0.0691 0.0267 0.0081 0.8233  
## TCGA-BC-A10R TCGA-BC-A10R 0.0000 0.0000 0.0000 0.0000 0.0394  
## TCGA-BC-A10T TCGA-BC-A10T 2.1161 13.4505 3.1068 0.7133 81.6385  
## TCGA-BC-A10U TCGA-BC-A10U 3.2578 15.5952 2.9366 0.8977 96.0961  
## TCGA-BC-A10W TCGA-BC-A10W 3.1482 14.7835 2.0547 0.8642 117.4401  
## TCGA-BC-A10X TCGA-BC-A10X 0.0000 0.0000 0.0000 0.0000 0.0000  
## TMSB4Y NLGN4Y UTY EIF1AY XIST  
## TCGA-BC-A10Q 0.0000 0.0000 0.0215 0.0245 4.3383  
## TCGA-BC-A10R 0.0000 0.0000 0.0000 0.0000 4.8065  
## TCGA-BC-A10T 0.8258 0.7771 1.3324 8.1587 0.0397  
## TCGA-BC-A10U 0.7235 0.7639 1.5558 10.4544 0.0533  
## TCGA-BC-A10W 0.7526 0.8760 1.5136 10.5587 0.0502  
## TCGA-BC-A10X 0.0000 0.0079 0.0000 0.0000 3.6791

# Combine Solid Tissue Normal and Primary Tumor

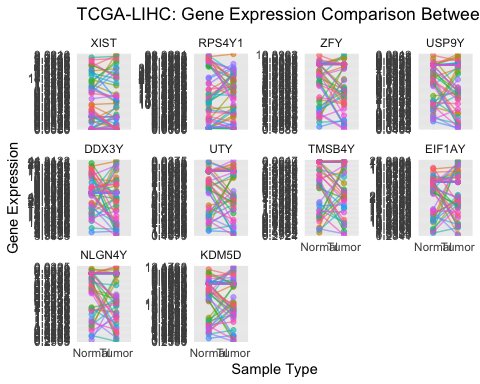
# Find common cases (rows) across all data frames  
common\_cases\_solid\_primary <- intersect(rownames(genes\_tpm\_solid\_tissue\_normal\_merged),  
 rownames(genes\_tpm\_primary\_tumor\_merged))  
  
# Subset the data frames to keep only the common  
# cases, preserving row names  
genes\_tpm\_solid\_tissue\_normal\_common <- genes\_tpm\_solid\_tissue\_normal\_merged[common\_cases\_solid\_primary,  
 , drop = FALSE]  
genes\_tpm\_primary\_tumor\_common <- genes\_tpm\_primary\_tumor\_merged[common\_cases\_solid\_primary,  
 , drop = FALSE]  
  
# Ensure row names are unique within each  
# subsetted data frame  
rownames(genes\_tpm\_solid\_tissue\_normal\_common) <- make.unique(rownames(genes\_tpm\_solid\_tissue\_normal\_common))  
rownames(genes\_tpm\_primary\_tumor\_common) <- make.unique(rownames(genes\_tpm\_primary\_tumor\_common))  
  
# Add a column to each data frame to indicate the  
# sample type  
genes\_tpm\_solid\_tissue\_normal\_common$SampleType <- "Solid Tissue Normal"  
genes\_tpm\_primary\_tumor\_common$SampleType <- "Primary Tumor"  
  
# Combine the data frames while preserving row  
# names  
combined\_common\_normal\_primary <- rbind(genes\_tpm\_solid\_tissue\_normal\_common,  
 genes\_tpm\_primary\_tumor\_common)  
  
# Ensure that SampleType is a factor  
combined\_common\_normal\_primary$SampleType <- factor(combined\_common\_normal\_primary$SampleType,  
 levels = c("Solid Tissue Normal", "Primary Tumor"))

# Make data frame as a long data frame in prep for ggplot

# Separate the data frames  
solid\_normal\_data <- combined\_common\_normal\_primary %>%  
 filter(SampleType == "Solid Tissue Normal") %>%  
 select(cases.submitter\_id, XIST, RPS4Y1, ZFY, USP9Y,  
 DDX3Y, UTY, TMSB4Y, EIF1AY, NLGN4Y, KDM5D)  
  
primary\_tumor\_data <- combined\_common\_normal\_primary %>%  
 filter(SampleType == "Primary Tumor") %>%  
 select(cases.submitter\_id, XIST, RPS4Y1, ZFY, USP9Y,  
 DDX3Y, UTY, TMSB4Y, EIF1AY, NLGN4Y, KDM5D)  
  
# Merge the data frames on 'cases.submitter\_id'  
merged\_data <- merge(solid\_normal\_data, primary\_tumor\_data,  
 by = "cases.submitter\_id", suffixes = c(".Normal",  
 ".Tumor"))  
  
# Reshape data for ggpairs  
long\_data <- melt(merged\_data, id.vars = "cases.submitter\_id")  
long\_data$Gene <- sub("\\..\*", "", long\_data$variable)  
long\_data$SampleType <- sub(".\*\\.", "", long\_data$variable)  
long\_data <- dcast(long\_data, cases.submitter\_id +  
 Gene ~ SampleType, value.var = "value")

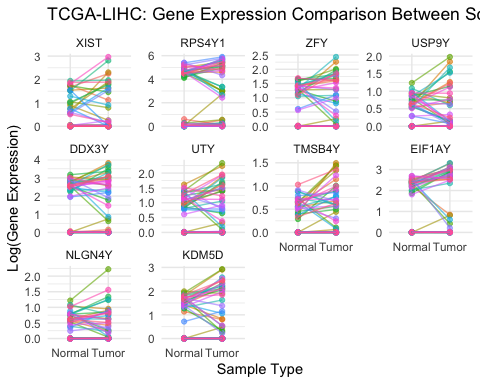
# Plot using ggplot

# Reshape the data from wide to long format  
long\_data\_melted <- melt(long\_data, id.vars = c("cases.submitter\_id",  
 "Gene"), variable.name = "SampleType", value.name = "Expression")  
  
# Define the order of genes  
gene\_order <- c("XIST", "RPS4Y1", "ZFY", "USP9Y", "DDX3Y",  
 "UTY", "TMSB4Y", "EIF1AY", "NLGN4Y", "KDM5D")  
  
# Convert gene column to factor with specified  
# levels  
long\_data\_melted$Gene <- factor(long\_data\_melted$Gene,  
 levels = gene\_order)  
  
# Plot the data using ggplot2  
ggplot(long\_data\_melted, aes(x = SampleType, y = Expression,  
 group = cases.submitter\_id, color = cases.submitter\_id)) +  
 geom\_point(alpha = 0.6) + geom\_line(alpha = 0.6) +  
 facet\_wrap(~Gene, scales = "free\_y") + labs(title = "TCGA-LIHC: Gene Expression Comparison Between Solid Tissue Normal and Primary Tumor",  
 x = "Sample Type", y = "Gene Expression") + theme\_minimal() +  
 theme(legend.position = "none") # Hide legend for case.submitter\_id



## Plotting the above with log transformed gene expressions

# Define the order of genes  
gene\_order <- c("XIST", "RPS4Y1", "ZFY", "USP9Y", "DDX3Y",  
 "UTY", "TMSB4Y", "EIF1AY", "NLGN4Y", "KDM5D")  
  
# Convert gene column to factor with specified  
# levels  
long\_data\_melted$Gene <- factor(long\_data\_melted$Gene,  
 levels = gene\_order)  
  
# Ensure the Expression column is numeric  
long\_data\_melted$Expression <- as.numeric(long\_data\_melted$Expression)  
  
# Log-transform the Expression values  
long\_data\_melted <- long\_data\_melted %>%  
 mutate(LogExpression = log(Expression + 1)) # Adding 1 to avoid log(0)  
  
# Plot the data using ggplot2  
ggplot(long\_data\_melted, aes(x = SampleType, y = LogExpression,  
 group = cases.submitter\_id, color = cases.submitter\_id)) +  
 geom\_point(alpha = 0.6) + geom\_line(alpha = 0.6) +  
 facet\_wrap(~Gene, scales = "free\_y") + labs(title = "TCGA-LIHC: Gene Expression Comparison Between Solid Tissue Normal and Primary Tumor",  
 x = "Sample Type", y = "Log(Gene Expression)") +  
 theme\_minimal() + theme(legend.position = "none")



# Explore Clinical from TCGA to explore contents

query\_clinical\_LIHC <- GDCquery(project = "TCGA-LIHC",  
 data.category = "Clinical")

## --------------------------------------

## o GDCquery: Searching in GDC database

## --------------------------------------

## Genome of reference: hg38

## --------------------------------------------

## oo Accessing GDC. This might take a while...

## --------------------------------------------

## ooo Project: TCGA-LIHC

## --------------------

## oo Filtering results

## --------------------

## ----------------

## oo Checking data

## ----------------

## ooo Checking if there are duplicated cases

## Warning: There are more than one file for the same case. Please verify query results. You can use the command View(getResults(query)) in rstudio

## ooo Checking if there are results for the query

## -------------------

## o Preparing output

## -------------------

# Download the Clinical data for Clinical Supplement only

query\_clinical\_LIHC\_cs <- GDCquery(project = "TCGA-LIHC",  
 data.category = "Clinical", data.type = "Clinical Supplement",  
 data.format = "BCR XML")

## --------------------------------------

## o GDCquery: Searching in GDC database

## --------------------------------------

## Genome of reference: hg38

## --------------------------------------------

## oo Accessing GDC. This might take a while...

## --------------------------------------------

## ooo Project: TCGA-LIHC

## --------------------

## oo Filtering results

## --------------------

## ooo By data.format

## ooo By data.type

## ----------------

## oo Checking data

## ----------------

## ooo Checking if there are duplicated cases

## ooo Checking if there are results for the query

## -------------------

## o Preparing output

## -------------------

GDCdownload(query\_clinical\_LIHC\_cs)

## Downloading data for project TCGA-LIHC

## Of the 377 files for download 377 already exist.

## All samples have been already downloaded

clinical\_LIHC <- GDCprepare\_clinic(query\_clinical\_LIHC\_cs,  
 clinical.info = "patient")

## | | | 0% | | | 1% | |= | 1% | |= | 2% | |== | 2% | |== | 3% | |=== | 4% | |=== | 5% | |==== | 5% | |==== | 6% | |===== | 7% | |===== | 8% | |====== | 8% | |====== | 9% | |======= | 10% | |======= | 11% | |======== | 11% | |======== | 12% | |========= | 12% | |========= | 13% | |========= | 14% | |========== | 14% | |========== | 15% | |=========== | 15% | |=========== | 16% | |============ | 16% | |============ | 17% | |============ | 18% | |============= | 18% | |============= | 19% | |============== | 19% | |============== | 20% | |============== | 21% | |=============== | 21% | |=============== | 22% | |================ | 22% | |================ | 23% | |================= | 24% | |================= | 25% | |================== | 25% | |================== | 26% | |=================== | 27% | |=================== | 28% | |==================== | 28% | |==================== | 29% | |===================== | 29% | |===================== | 30% | |===================== | 31% | |====================== | 31% | |====================== | 32% | |======================= | 32% | |======================= | 33% | |======================== | 34% | |======================== | 35% | |========================= | 35% | |========================= | 36% | |========================== | 37% | |========================== | 38% | |=========================== | 38% | |=========================== | 39% | |============================ | 40% | |============================ | 41% | |============================= | 41% | |============================= | 42% | |============================== | 42% | |============================== | 43% | |============================== | 44% | |=============================== | 44% | |=============================== | 45% | |================================ | 45% | |================================ | 46% | |================================= | 47% | |================================= | 48% | |================================== | 48% | |================================== | 49% | |=================================== | 49% | |=================================== | 50% | |=================================== | 51% | |==================================== | 51% | |==================================== | 52% | |===================================== | 52% | |===================================== | 53% | |====================================== | 54% | |====================================== | 55% | |======================================= | 55% | |======================================= | 56% | |======================================== | 56% | |======================================== | 57% | |======================================== | 58% | |========================================= | 58% | |========================================= | 59% | |========================================== | 59% | |========================================== | 60% | |=========================================== | 61% | |=========================================== | 62% | |============================================ | 62% | |============================================ | 63% | |============================================= | 64% | |============================================= | 65% | |============================================== | 65% | |============================================== | 66% | |=============================================== | 67% | |=============================================== | 68% | |================================================ | 68% | |================================================ | 69% | |================================================= | 69% | |================================================= | 70% | |================================================= | 71% | |================================================== | 71% | |================================================== | 72% | |=================================================== | 72% | |=================================================== | 73% | |==================================================== | 74% | |==================================================== | 75% | |===================================================== | 75% | |===================================================== | 76% | |====================================================== | 77% | |====================================================== | 78% | |======================================================= | 78% | |======================================================= | 79% | |======================================================== | 79% | |======================================================== | 80% | |======================================================== | 81% | |========================================================= | 81% | |========================================================= | 82% | |========================================================== | 82% | |========================================================== | 83% | |========================================================== | 84% | |=========================================================== | 84% | |=========================================================== | 85% | |============================================================ | 85% | |============================================================ | 86% | |============================================================= | 86% | |============================================================= | 87% | |============================================================= | 88% | |============================================================== | 88% | |============================================================== | 89% | |=============================================================== | 89% | |=============================================================== | 90% | |================================================================ | 91% | |================================================================ | 92% | |================================================================= | 92% | |================================================================= | 93% | |================================================================== | 94% | |================================================================== | 95% | |=================================================================== | 95% | |=================================================================== | 96% | |==================================================================== | 97% | |==================================================================== | 98% | |===================================================================== | 98% | |===================================================================== | 99% | |======================================================================| 99% | |======================================================================| 100%

## To get the following information please change the clinical.info argument

## => new\_tumor\_events: new\_tumor\_event   
## => drugs: drug   
## => follow\_ups: follow\_up   
## => radiations: radiation

## Parsing follow up version: follow\_up\_v4.0

## | | | 0% | | | 1% | |= | 1% | |= | 2% | |== | 3% | |=== | 4% | |=== | 5% | |==== | 5% | |==== | 6% | |===== | 7% | |===== | 8% | |====== | 8% | |====== | 9% | |======= | 10% | |======= | 11% | |======== | 11% | |======== | 12% | |========= | 12% | |========= | 13% | |========= | 14% | |========== | 14% | |========== | 15% | |=========== | 15% | |=========== | 16% | |============ | 17% | |============ | 18% | |============= | 18% | |============= | 19% | |============== | 19% | |============== | 20% | |============== | 21% | |=============== | 21% | |=============== | 22% | |================ | 22% | |================ | 23% | |================= | 24% | |================= | 25% | |================== | 25% | |================== | 26% | |=================== | 27% | |=================== | 28% | |==================== | 28% | |==================== | 29% | |===================== | 29% | |===================== | 30% | |===================== | 31% | |====================== | 31% | |====================== | 32% | |======================= | 32% | |======================= | 33% | |======================== | 34% | |======================== | 35% | |========================= | 35% | |========================= | 36% | |========================== | 37% | |========================== | 38% | |=========================== | 38% | |=========================== | 39% | |============================ | 40% | |============================ | 41% | |============================= | 41% | |============================= | 42% | |============================== | 42% | |============================== | 43% | |============================== | 44% | |=============================== | 44% | |=============================== | 45% | |================================ | 45% | |================================ | 46% | |================================= | 47% | |================================= | 48% | |================================== | 48% | |================================== | 49% | |=================================== | 49% | |=================================== | 50% | |=================================== | 51% | |==================================== | 51% | |==================================== | 52% | |===================================== | 52% | |===================================== | 53% | |====================================== | 54% | |====================================== | 55% | |======================================= | 55% | |======================================= | 56% | |======================================== | 56% | |======================================== | 57% | |======================================== | 58% | |========================================= | 58% | |========================================= | 59% | |========================================== | 59% | |========================================== | 60% | |=========================================== | 61% | |=========================================== | 62% | |============================================ | 62% | |============================================ | 63% | |============================================= | 64% | |============================================== | 66% | |=============================================== | 67% | |=============================================== | 68% | |================================================ | 68% | |================================================ | 69% | |================================================= | 69% | |================================================= | 70% | |================================================= | 71% | |================================================== | 71% | |================================================== | 72% | |=================================================== | 72% | |=================================================== | 73% | |==================================================== | 74% | |==================================================== | 75% | |===================================================== | 75% | |===================================================== | 76% | |====================================================== | 77% | |====================================================== | 78% | |======================================================= | 78% | |======================================================= | 79% | |======================================================== | 79% | |======================================================== | 80% | |======================================================== | 81% | |========================================================= | 81% | |========================================================= | 82% | |========================================================== | 82% | |========================================================== | 83% | |========================================================== | 84% | |=========================================================== | 84% | |=========================================================== | 85% | |============================================================ | 85% | |============================================================ | 86% | |============================================================= | 87% | |============================================================= | 88% | |============================================================== | 88% | |============================================================== | 89% | |=============================================================== | 89% | |=============================================================== | 90% | |================================================================ | 91% | |================================================================ | 92% | |================================================================= | 92% | |================================================================= | 93% | |================================================================== | 94% | |================================================================== | 95% | |=================================================================== | 95% | |=================================================================== | 96% | |==================================================================== | 97% | |==================================================================== | 98% | |===================================================================== | 98% | |===================================================================== | 99% | |======================================================================| 99% | |======================================================================| 100%

## Adding stage event information

## | | | 0% | | | 1% | |= | 1% | |= | 2% | |== | 2% | |== | 3% | |=== | 4% | |=== | 5% | |==== | 5% | |==== | 6% | |===== | 7% | |===== | 8% | |====== | 8% | |====== | 9% | |======= | 10% | |======= | 11% | |======== | 11% | |======== | 12% | |========= | 12% | |========= | 13% | |========= | 14% | |========== | 14% | |========== | 15% | |=========== | 15% | |=========== | 16% | |============ | 16% | |============ | 17% | |============ | 18% | |============= | 18% | |============= | 19% | |============== | 19% | |============== | 20% | |============== | 21% | |=============== | 21% | |=============== | 22% | |================ | 22% | |================ | 23% | |================= | 24% | |================= | 25% | |================== | 25% | |================== | 26% | |=================== | 27% | |=================== | 28% | |==================== | 28% | |==================== | 29% | |===================== | 29% | |===================== | 30% | |===================== | 31% | |====================== | 31% | |====================== | 32% | |======================= | 32% | |======================= | 33% | |======================== | 34% | |======================== | 35% | |========================= | 35% | |========================= | 36% | |========================== | 37% | |========================== | 38% | |=========================== | 38% | |=========================== | 39% | |============================ | 40% | |============================ | 41% | |============================= | 41% | |============================= | 42% | |============================== | 42% | |============================== | 43% | |============================== | 44% | |=============================== | 44% | |=============================== | 45% | |================================ | 45% | |================================ | 46% | |================================= | 47% | |================================= | 48% | |================================== | 48% | |================================== | 49% | |=================================== | 49% | |=================================== | 50% | |=================================== | 51% | |==================================== | 51% | |==================================== | 52% | |===================================== | 52% | |===================================== | 53% | |====================================== | 54% | |====================================== | 55% | |======================================= | 55% | |======================================= | 56% | |======================================== | 56% | |======================================== | 57% | |======================================== | 58% | |========================================= | 58% | |========================================= | 59% | |========================================== | 59% | |========================================== | 60% | |=========================================== | 61% | |=========================================== | 62% | |============================================ | 62% | |============================================ | 63% | |============================================= | 64% | |============================================= | 65% | |============================================== | 65% | |============================================== | 66% | |=============================================== | 67% | |=============================================== | 68% | |================================================ | 68% | |================================================ | 69% | |================================================= | 69% | |================================================= | 70% | |================================================= | 71% | |================================================== | 71% | |================================================== | 72% | |=================================================== | 72% | |=================================================== | 73% | |==================================================== | 74% | |==================================================== | 75% | |===================================================== | 75% | |===================================================== | 76% | |====================================================== | 77% | |====================================================== | 78% | |======================================================= | 78% | |======================================================= | 79% | |======================================================== | 79% | |======================================================== | 80% | |======================================================== | 81% | |========================================================= | 81% | |========================================================= | 82% | |========================================================== | 82% | |========================================================== | 83% | |========================================================== | 84% | |=========================================================== | 84% | |=========================================================== | 85% | |============================================================ | 85% | |============================================================ | 86% | |============================================================= | 86% | |============================================================= | 87% | |============================================================= | 88% | |============================================================== | 88% | |============================================================== | 89% | |=============================================================== | 89% | |=============================================================== | 90% | |================================================================ | 91% | |================================================================ | 92% | |================================================================= | 92% | |================================================================= | 93% | |================================================================== | 94% | |================================================================== | 95% | |=================================================================== | 95% | |=================================================================== | 96% | |==================================================================== | 97% | |==================================================================== | 98% | |===================================================================== | 98% | |===================================================================== | 99% | |======================================================================| 99% | |======================================================================| 100%

## Updating days\_to\_last\_followup and vital\_status from follow\_up information using last entry  
## Parsing follow up version: follow\_up\_v4.0

## | | | 0% | | | 1% | |= | 1% | |= | 2% | |== | 3% | |=== | 4% | |=== | 5% | |==== | 5% | |==== | 6% | |===== | 7% | |===== | 8% | |====== | 8% | |====== | 9% | |======= | 10% | |======= | 11% | |======== | 11% | |======== | 12% | |========= | 12% | |========= | 13% | |========= | 14% | |========== | 14% | |========== | 15% | |=========== | 15% | |=========== | 16% | |============ | 17% | |============ | 18% | |============= | 18% | |============= | 19% | |============== | 19% | |============== | 20% | |============== | 21% | |=============== | 21% | |=============== | 22% | |================ | 22% | |================ | 23% | |================= | 24% | |================= | 25% | |================== | 25% | |================== | 26% | |=================== | 27% | |=================== | 28% | |==================== | 28% | |==================== | 29% | |===================== | 29% | |===================== | 30% | |===================== | 31% | |====================== | 31% | |====================== | 32% | |======================= | 32% | |======================= | 33% | |======================== | 34% | |======================== | 35% | |========================= | 35% | |========================= | 36% | |========================== | 37% | |========================== | 38% | |=========================== | 38% | |=========================== | 39% | |============================ | 40% | |============================ | 41% | |============================= | 41% | |============================= | 42% | |============================== | 42% | |============================== | 43% | |============================== | 44% | |=============================== | 44% | |=============================== | 45% | |================================ | 45% | |================================ | 46% | |================================= | 47% | |================================= | 48% | |================================== | 48% | |================================== | 49% | |=================================== | 49% | |=================================== | 50% | |=================================== | 51% | |==================================== | 51% | |==================================== | 52% | |===================================== | 52% | |===================================== | 53% | |====================================== | 54% | |====================================== | 55% | |======================================= | 55% | |======================================= | 56% | |======================================== | 56% | |======================================== | 57% | |======================================== | 58% | |========================================= | 58% | |========================================= | 59% | |========================================== | 59% | |========================================== | 60% | |=========================================== | 61% | |=========================================== | 62% | |============================================ | 62% | |============================================ | 63% | |============================================= | 64% | |============================================== | 66% | |=============================================== | 67% | |=============================================== | 68% | |================================================ | 68% | |================================================ | 69% | |================================================= | 69% | |================================================= | 70% | |================================================= | 71% | |================================================== | 71% | |================================================== | 72% | |=================================================== | 72% | |=================================================== | 73% | |==================================================== | 74% | |==================================================== | 75% | |===================================================== | 75% | |===================================================== | 76% | |====================================================== | 77% | |====================================================== | 78% | |======================================================= | 78% | |======================================================= | 79% | |======================================================== | 79% | |======================================================== | 80% | |======================================================== | 81% | |========================================================= | 81% | |========================================================= | 82% | |========================================================== | 82% | |========================================================== | 83% | |========================================================== | 84% | |=========================================================== | 84% | |=========================================================== | 85% | |============================================================ | 85% | |============================================================ | 86% | |============================================================= | 87% | |============================================================= | 88% | |============================================================== | 88% | |============================================================== | 89% | |=============================================================== | 89% | |=============================================================== | 90% | |================================================================ | 91% | |================================================================ | 92% | |================================================================= | 92% | |================================================================= | 93% | |================================================================== | 94% | |================================================================== | 95% | |=================================================================== | 95% | |=================================================================== | 96% | |==================================================================== | 97% | |==================================================================== | 98% | |===================================================================== | 98% | |===================================================================== | 99% | |======================================================================| 99% | |======================================================================| 100%

# Add ‘gender’ from ‘clinical\_LIHC’ to ‘combined\_common\_normal\_primary’

# Make sure the column names and types are  
# appropriate for merging between data sets  
combined\_common\_normal\_primary$cases.submitter\_id <- as.character(combined\_common\_normal\_primary$cases.submitter\_id)  
clinical\_LIHC$bcr\_patient\_barcode <- as.character(clinical\_LIHC$bcr\_patient\_barcode)  
  
# Merge the data frames  
combined\_common\_normal\_primary\_gender <- merge(combined\_common\_normal\_primary,  
 clinical\_LIHC[, c("bcr\_patient\_barcode", "gender")],  
 by.x = "cases.submitter\_id", by.y = "bcr\_patient\_barcode",  
 all.x = TRUE)

# Determine thresholds for each gene as low, intermediate, high expression

## XIST

# Ensure that XIST column is numeric  
combined\_common\_normal\_primary\_gender$XIST <- as.numeric(as.character(combined\_common\_normal\_primary\_gender$XIST))  
  
# Visualize gene expressions between males and  
# females using violin plot  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), XIST, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("XIST Expression (log counts)") +  
 theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation introduced infinite values.  
## log-10 transformation introduced infinite values.

## Warning: Removed 32 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Determine mean of gene expression for males for  
# XIST  
mean\_males\_XIST <- mean(combined\_common\_normal\_primary\_gender$XIST[combined\_common\_normal\_primary\_gender$gender ==  
 "MALE"])  
  
mean\_males\_XIST

## [1] 0.008796429

# Determine high and low threshold of the mean  
high\_threshold\_XIST <- mean(combined\_common\_normal\_primary\_gender$XIST[combined\_common\_normal\_primary\_gender$gender ==  
 "MALE"]) + 2.5 \* sd(combined\_common\_normal\_primary\_gender$XIST[combined\_common\_normal\_primary\_gender$gender ==  
 "MALE"])  
  
high\_threshold\_XIST

## [1] 0.05287938

low\_threshold\_XIST <- mean(combined\_common\_normal\_primary\_gender$XIST[combined\_common\_normal\_primary\_gender$gender ==  
 "MALE"]) + 1 \* sd(combined\_common\_normal\_primary\_gender$XIST[combined\_common\_normal\_primary\_gender$gender ==  
 "MALE"])  
  
low\_threshold\_XIST

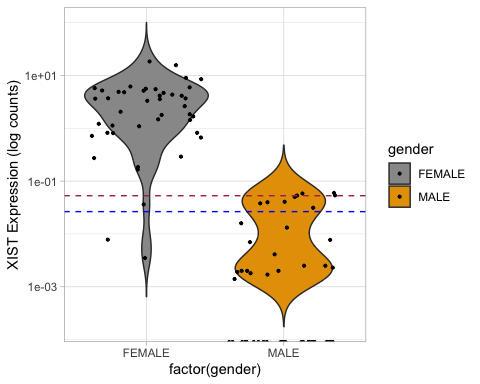
## [1] 0.02642961

# Plot Violin plot with low and high thresholds  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), XIST, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("XIST Expression (log counts)") +  
 geom\_hline(yintercept = high\_threshold\_XIST, linetype = "dashed",  
 color = "maroon") + geom\_hline(yintercept = low\_threshold\_XIST,  
 linetype = "dashed", color = "blue") + theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning: Removed 32 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).

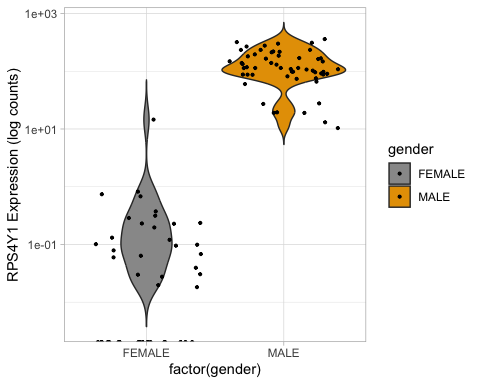


## RPS4Y1

# Ensure that RPS4Y1 column is numeric  
combined\_common\_normal\_primary\_gender$RPS4Y1 <- as.numeric(as.character(combined\_common\_normal\_primary\_gender$RPS4Y1))  
  
# Visualize gene expressions between males and  
# females using violin plot  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), RPS4Y1, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("RPS4Y1 Expression (log counts)") +  
 theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation introduced infinite values.  
## log-10 transformation introduced infinite values.

## Warning: Removed 18 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Determine mean of gene expression for males for  
# RPS4Y1  
mean\_females\_RPS4Y1 <- mean(combined\_common\_normal\_primary\_gender$RPS4Y1[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
mean\_females\_RPS4Y1

## [1] 0.4467068

# Determine high and low threshold of the mean  
high\_threshold\_RPS4Y1 <- mean(combined\_common\_normal\_primary\_gender$RPS4Y1[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 2.5 \* sd(combined\_common\_normal\_primary\_gender$RPS4Y1[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
high\_threshold\_RPS4Y1

## [1] 5.913119

low\_threshold\_RPS4Y1 <- mean(combined\_common\_normal\_primary\_gender$RPS4Y1[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 1 \* sd(combined\_common\_normal\_primary\_gender$RPS4Y1[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
low\_threshold\_RPS4Y1

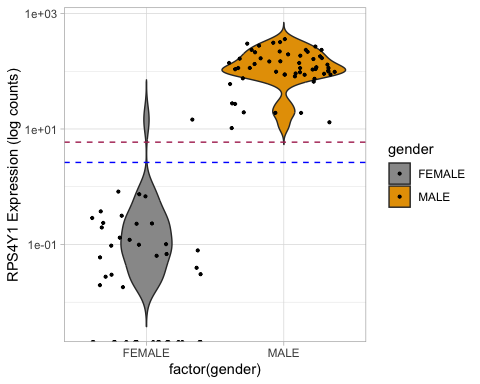
## [1] 2.633272

# Plot Violin plot with low and high thresholds  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), RPS4Y1, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("RPS4Y1 Expression (log counts)") +  
 geom\_hline(yintercept = high\_threshold\_RPS4Y1,  
 linetype = "dashed", color = "maroon") + geom\_hline(yintercept = low\_threshold\_RPS4Y1,  
 linetype = "dashed", color = "blue") + theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning: Removed 18 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# ZFY

# Ensure that ZFY column is numeric  
combined\_common\_normal\_primary\_gender$ZFY <- as.numeric(as.character(combined\_common\_normal\_primary\_gender$ZFY))  
  
# Visualize gene expressions between males and  
# females using violin plot  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), ZFY, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("ZFY Expression (log counts)") +  
 theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation introduced infinite values.  
## log-10 transformation introduced infinite values.

## Warning: Removed 35 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Determine mean of gene expression for males for  
# ZFY  
mean\_females\_ZFY <- mean(combined\_common\_normal\_primary\_gender$ZFY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
mean\_females\_ZFY

## [1] 0.009181818

# Determine high and low threshold of the mean  
high\_threshold\_ZFY <- mean(combined\_common\_normal\_primary\_gender$ZFY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 2.5 \* sd(combined\_common\_normal\_primary\_gender$ZFY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
high\_threshold\_ZFY

## [1] 0.1228894

low\_threshold\_ZFY <- mean(combined\_common\_normal\_primary\_gender$ZFY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 1 \* sd(combined\_common\_normal\_primary\_gender$ZFY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
low\_threshold\_ZFY

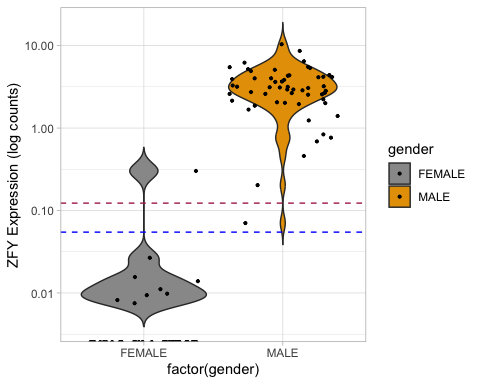
## [1] 0.05466484

# Plot Violin plot with low and high thresholds  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), ZFY, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("ZFY Expression (log counts)") +  
 geom\_hline(yintercept = high\_threshold\_ZFY, linetype = "dashed",  
 color = "maroon") + geom\_hline(yintercept = low\_threshold\_ZFY,  
 linetype = "dashed", color = "blue") + theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning: Removed 35 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# USP9Y

# Ensure that USP9Y column is numeric  
combined\_common\_normal\_primary\_gender$USP9Y <- as.numeric(as.character(combined\_common\_normal\_primary\_gender$USP9Y))  
  
# Visualize gene expressions between males and  
# females using violin plot  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), USP9Y, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("USP9Y Expression (log counts)") +  
 theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation introduced infinite values.  
## log-10 transformation introduced infinite values.

## Warning: Removed 31 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Determine mean of gene expression for males for  
# USP9Y  
mean\_females\_USP9Y <- mean(combined\_common\_normal\_primary\_gender$USP9Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
mean\_females\_USP9Y

## [1] 0.005056818

# Determine high and low threshold of the mean  
high\_threshold\_USP9Y <- mean(combined\_common\_normal\_primary\_gender$USP9Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 2.5 \* sd(combined\_common\_normal\_primary\_gender$USP9Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
high\_threshold\_USP9Y

## [1] 0.06491424

low\_threshold\_USP9Y <- mean(combined\_common\_normal\_primary\_gender$USP9Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 1 \* sd(combined\_common\_normal\_primary\_gender$USP9Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
low\_threshold\_USP9Y

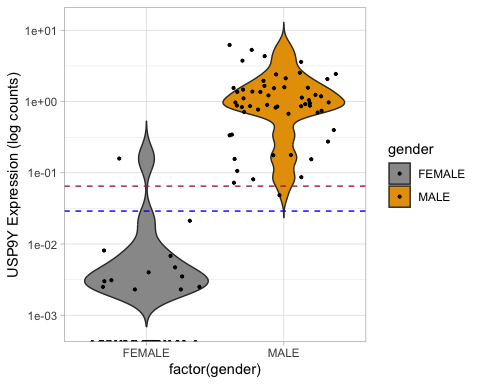
## [1] 0.02899979

# Plot Violin plot with low and high thresholds  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), USP9Y, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("USP9Y Expression (log counts)") +  
 geom\_hline(yintercept = high\_threshold\_USP9Y, linetype = "dashed",  
 color = "maroon") + geom\_hline(yintercept = low\_threshold\_USP9Y,  
 linetype = "dashed", color = "blue") + theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning: Removed 31 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# DDX3Y

# Ensure that DDX3Y column is numeric  
combined\_common\_normal\_primary\_gender$DDX3Y <- as.numeric(as.character(combined\_common\_normal\_primary\_gender$DDX3Y))  
  
# Visualize gene expressions between males and  
# females using violin plot  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), DDX3Y, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("DDX3Y Expression (log counts)") +  
 theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation introduced infinite values.  
## log-10 transformation introduced infinite values.

## Warning: Removed 27 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Determine mean of gene expression for males for  
# DDX3Y  
mean\_females\_DDX3Y <- mean(combined\_common\_normal\_primary\_gender$DDX3Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
mean\_females\_DDX3Y

## [1] 0.03525909

# Determine high and low threshold of the mean  
high\_threshold\_DDX3Y <- mean(combined\_common\_normal\_primary\_gender$DDX3Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 2.5 \* sd(combined\_common\_normal\_primary\_gender$DDX3Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
high\_threshold\_DDX3Y

## [1] 0.4542457

low\_threshold\_DDX3Y <- mean(combined\_common\_normal\_primary\_gender$DDX3Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 1 \* sd(combined\_common\_normal\_primary\_gender$DDX3Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
low\_threshold\_DDX3Y

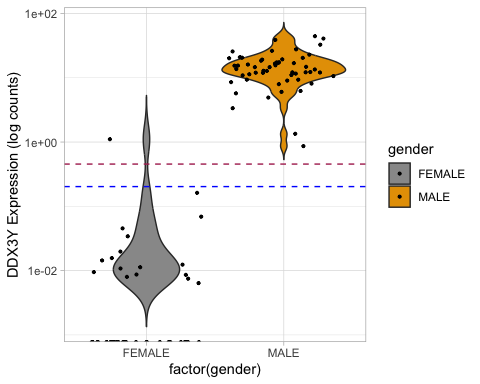
## [1] 0.2028537

# Plot Violin plot with low and high thresholds  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), DDX3Y, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("DDX3Y Expression (log counts)") +  
 geom\_hline(yintercept = high\_threshold\_DDX3Y, linetype = "dashed",  
 color = "maroon") + geom\_hline(yintercept = low\_threshold\_DDX3Y,  
 linetype = "dashed", color = "blue") + theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning: Removed 27 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).

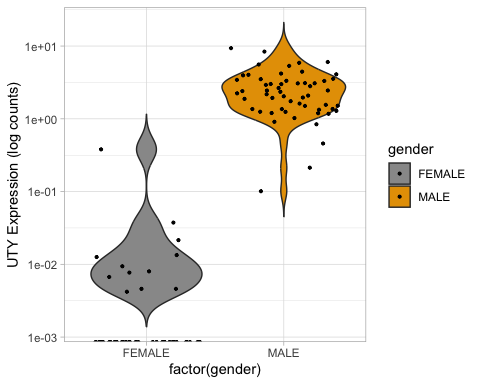


# UTY

# Ensure that UTY column is numeric  
combined\_common\_normal\_primary\_gender$UTY <- as.numeric(as.character(combined\_common\_normal\_primary\_gender$UTY))  
  
# Visualize gene expressions between males and  
# females using violin plot  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), UTY, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("UTY Expression (log counts)") +  
 theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation introduced infinite values.  
## log-10 transformation introduced infinite values.

## Warning: Removed 32 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Determine mean of gene expression for males for  
# UTY  
mean\_females\_UTY <- mean(combined\_common\_normal\_primary\_gender$UTY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
mean\_females\_UTY

## [1] 0.01159773

# Determine high and low threshold of the mean  
high\_threshold\_UTY <- mean(combined\_common\_normal\_primary\_gender$UTY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 2.5 \* sd(combined\_common\_normal\_primary\_gender$UTY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
high\_threshold\_UTY

## [1] 0.1547968

low\_threshold\_UTY <- mean(combined\_common\_normal\_primary\_gender$UTY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 1 \* sd(combined\_common\_normal\_primary\_gender$UTY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
low\_threshold\_UTY

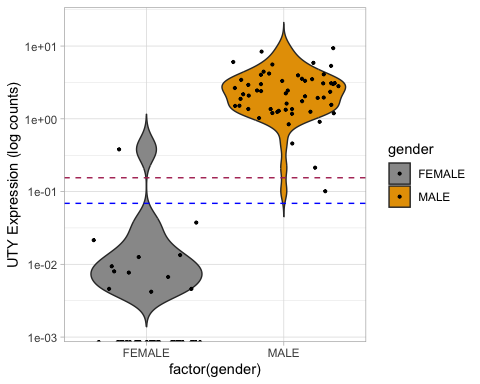
## [1] 0.06887734

# Plot Violin plot with low and high thresholds  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), UTY, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("UTY Expression (log counts)") +  
 geom\_hline(yintercept = high\_threshold\_UTY, linetype = "dashed",  
 color = "maroon") + geom\_hline(yintercept = low\_threshold\_UTY,  
 linetype = "dashed", color = "blue") + theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning: Removed 32 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# TMSB4Y

# Ensure that TMSB4Y column is numeric  
combined\_common\_normal\_primary\_gender$TMSB4Y <- as.numeric(as.character(combined\_common\_normal\_primary\_gender$TMSB4Y))  
  
# Visualize gene expressions between males and  
# females using violin plot  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), TMSB4Y, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("TMSB4Y Expression (log counts)") +  
 theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation introduced infinite values.  
## log-10 transformation introduced infinite values.

## Warning: Removed 42 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Determine mean of gene expression for males for  
# TMSB4Y  
mean\_females\_TMSB4Y <- mean(combined\_common\_normal\_primary\_gender$TMSB4Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
mean\_females\_TMSB4Y

## [1] 0.01413864

# Determine high and low threshold of the mean  
high\_threshold\_TMSB4Y <- mean(combined\_common\_normal\_primary\_gender$TMSB4Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 2.5 \* sd(combined\_common\_normal\_primary\_gender$TMSB4Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
high\_threshold\_TMSB4Y

## [1] 0.2250631

low\_threshold\_TMSB4Y <- mean(combined\_common\_normal\_primary\_gender$TMSB4Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 1 \* sd(combined\_common\_normal\_primary\_gender$TMSB4Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
low\_threshold\_TMSB4Y

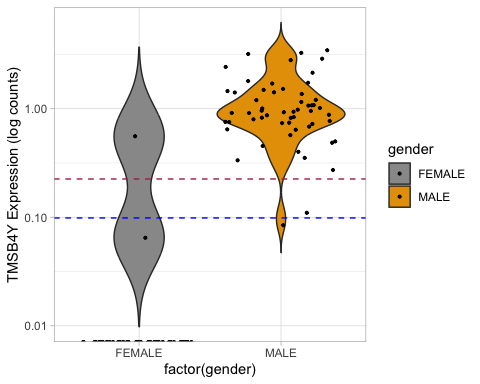
## [1] 0.09850842

# Plot Violin plot with low and high thresholds  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), TMSB4Y, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("TMSB4Y Expression (log counts)") +  
 geom\_hline(yintercept = high\_threshold\_TMSB4Y,  
 linetype = "dashed", color = "maroon") + geom\_hline(yintercept = low\_threshold\_TMSB4Y,  
 linetype = "dashed", color = "blue") + theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning: Removed 42 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).

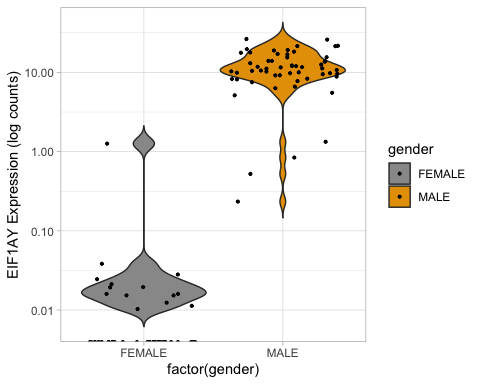


# EIF1AY

# Ensure that EIF1AY column is numeric  
combined\_common\_normal\_primary\_gender$EIF1AY <- as.numeric(as.character(combined\_common\_normal\_primary\_gender$EIF1AY))  
  
# Visualize gene expressions between males and  
# females using violin plot  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), EIF1AY, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("EIF1AY Expression (log counts)") +  
 theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation introduced infinite values.  
## log-10 transformation introduced infinite values.

## Warning: Removed 30 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Determine mean of gene expression for males for  
# EIF1AY  
mean\_females\_EIF1AY <- mean(combined\_common\_normal\_primary\_gender$EIF1AY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
mean\_females\_EIF1AY

## [1] 0.03430909

# Determine high and low threshold of the mean  
high\_threshold\_EIF1AY <- mean(combined\_common\_normal\_primary\_gender$EIF1AY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 2.5 \* sd(combined\_common\_normal\_primary\_gender$EIF1AY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
high\_threshold\_EIF1AY

## [1] 0.508388

low\_threshold\_EIF1AY <- mean(combined\_common\_normal\_primary\_gender$EIF1AY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 1 \* sd(combined\_common\_normal\_primary\_gender$EIF1AY[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
low\_threshold\_EIF1AY

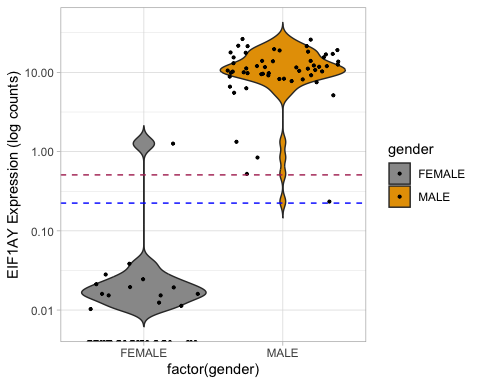
## [1] 0.2239406

# Plot Violin plot with low and high thresholds  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), EIF1AY, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("EIF1AY Expression (log counts)") +  
 geom\_hline(yintercept = high\_threshold\_EIF1AY,  
 linetype = "dashed", color = "maroon") + geom\_hline(yintercept = low\_threshold\_EIF1AY,  
 linetype = "dashed", color = "blue") + theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning: Removed 30 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# NLGN4Y

# Ensure that NLGN4Y column is numeric  
combined\_common\_normal\_primary\_gender$NLGN4Y <- as.numeric(as.character(combined\_common\_normal\_primary\_gender$NLGN4Y))  
  
# Visualize gene expressions between males and  
# females using violin plot  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), NLGN4Y, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("NLGN4Y Expression (log counts)") +  
 theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation introduced infinite values.  
## log-10 transformation introduced infinite values.

## Warning: Removed 36 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Determine mean of gene expression for males for  
# NLGN4Y  
mean\_females\_NLGN4Y <- mean(combined\_common\_normal\_primary\_gender$NLGN4Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
mean\_females\_NLGN4Y

## [1] 0.001727273

# Determine high and low threshold of the mean  
high\_threshold\_NLGN4Y <- mean(combined\_common\_normal\_primary\_gender$NLGN4Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 2.5 \* sd(combined\_common\_normal\_primary\_gender$NLGN4Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
high\_threshold\_NLGN4Y

## [1] 0.01259652

low\_threshold\_NLGN4Y <- mean(combined\_common\_normal\_primary\_gender$NLGN4Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 1 \* sd(combined\_common\_normal\_primary\_gender$NLGN4Y[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
low\_threshold\_NLGN4Y

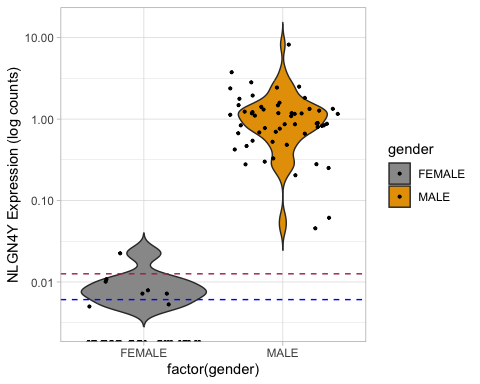
## [1] 0.006074973

# Plot Violin plot with low and high thresholds  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), NLGN4Y, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("NLGN4Y Expression (log counts)") +  
 geom\_hline(yintercept = high\_threshold\_NLGN4Y,  
 linetype = "dashed", color = "maroon") + geom\_hline(yintercept = low\_threshold\_NLGN4Y,  
 linetype = "dashed", color = "blue") + theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning: Removed 36 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).

 ### Only one that seems a bit off based on the thresholds is this one (NLGN4Y) where the thresholds appear to be much within the violin plot of females, where all others appear in between.

# KDM5D

# Ensure that KDM5D column is numeric  
combined\_common\_normal\_primary\_gender$KDM5D <- as.numeric(as.character(combined\_common\_normal\_primary\_gender$KDM5D))  
  
# Visualize gene expressions between males and  
# females using violin plot  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), KDM5D, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("KDM5D Expression (log counts)") +  
 theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation introduced infinite values.  
## log-10 transformation introduced infinite values.

## Warning: Removed 27 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Determine mean of gene expression for males for  
# KDM5D  
mean\_females\_KDM5D <- mean(combined\_common\_normal\_primary\_gender$KDM5D[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
mean\_females\_KDM5D

## [1] 0.017175

# Determine high and low threshold of the mean  
high\_threshold\_KDM5D <- mean(combined\_common\_normal\_primary\_gender$KDM5D[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 2.5 \* sd(combined\_common\_normal\_primary\_gender$KDM5D[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
high\_threshold\_KDM5D

## [1] 0.2483586

low\_threshold\_KDM5D <- mean(combined\_common\_normal\_primary\_gender$KDM5D[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"]) + 1 \* sd(combined\_common\_normal\_primary\_gender$KDM5D[combined\_common\_normal\_primary\_gender$gender ==  
 "FEMALE"])  
  
low\_threshold\_KDM5D

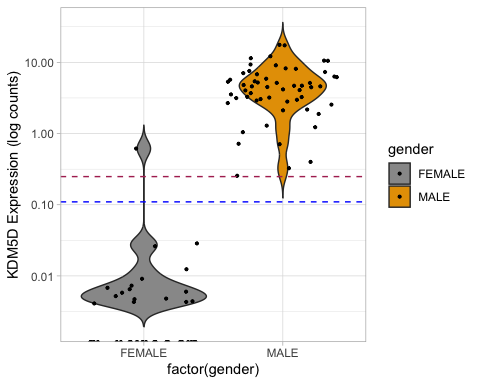
## [1] 0.1096484

# Plot Violin plot with low and high thresholds  
ggplot(data = combined\_common\_normal\_primary\_gender,  
 aes(factor(gender), KDM5D, fill = gender)) + geom\_violin(trim = FALSE) +  
 scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#999999",  
 "#E69F00")) + geom\_jitter(size = 0.75) + ylab("KDM5D Expression (log counts)") +  
 geom\_hline(yintercept = high\_threshold\_KDM5D, linetype = "dashed",  
 color = "maroon") + geom\_hline(yintercept = low\_threshold\_KDM5D,  
 linetype = "dashed", color = "blue") + theme\_light()

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning in scale\_y\_continuous(trans = "log10"): log-10 transformation  
## introduced infinite values.

## Warning: Removed 27 rows containing non-finite outside the scale range  
## (`stat\_ydensity()`).



# Now based on the thresholds determined for each gene of interest above, will add low, intermediate, and high expressions for each case for each gene

# Note to self, Gene order: 'XIST', 'RPS4Y1',  
# 'ZFY', 'USP9Y', 'DDX3Y', 'UTY', 'TMSB4Y',  
# 'EIF1AY', 'NLGN4Y', 'KDM5D'  
  
# Add gene expression category into the data for  
# low, intermediate and high expressions for each  
# gene  
counts\_common\_normal\_primary\_gender <- combined\_common\_normal\_primary\_gender %>%  
 mutate(XIST\_expression\_category = case\_when(XIST >=  
 high\_threshold\_XIST ~ "high\_expression", XIST <=  
 low\_threshold\_XIST ~ "low\_expression", TRUE ~  
 "intermediate\_expression"), RPS4Y1\_expression\_category = case\_when(RPS4Y1 >=  
 high\_threshold\_RPS4Y1 ~ "high\_expression",  
 RPS4Y1 <= low\_threshold\_RPS4Y1 ~ "low\_expression",  
 TRUE ~ "intermediate\_expression"), ZFY\_expression\_category = case\_when(ZFY >=  
 high\_threshold\_ZFY ~ "high\_expression", ZFY <=  
 low\_threshold\_ZFY ~ "low\_expression", TRUE ~  
 "intermediate\_expression"), USP9Y\_expression\_category = case\_when(USP9Y >=  
 high\_threshold\_USP9Y ~ "high\_expression", USP9Y <=  
 low\_threshold\_USP9Y ~ "low\_expression", TRUE ~  
 "intermediate\_expression"), DDX3Y\_expression\_category = case\_when(DDX3Y >=  
 high\_threshold\_USP9Y ~ "high\_expression", DDX3Y <=  
 low\_threshold\_USP9Y ~ "low\_expression", TRUE ~  
 "intermediate\_expression"), UTY\_expression\_category = case\_when(UTY >=  
 high\_threshold\_UTY ~ "high\_expression", UTY <=  
 low\_threshold\_UTY ~ "low\_expression", TRUE ~  
 "intermediate\_expression"), TMSB4Y\_expression\_category = case\_when(TMSB4Y >=  
 high\_threshold\_TMSB4Y ~ "high\_expression",  
 TMSB4Y <= low\_threshold\_TMSB4Y ~ "low\_expression",  
 TRUE ~ "intermediate\_expression"), EIF1AY\_expression\_category = case\_when(EIF1AY >=  
 high\_threshold\_EIF1AY ~ "high\_expression",  
 EIF1AY <= low\_threshold\_EIF1AY ~ "low\_expression",  
 TRUE ~ "intermediate\_expression"), NLGN4Y\_expression\_category = case\_when(NLGN4Y >=  
 high\_threshold\_NLGN4Y ~ "high\_expression",  
 NLGN4Y <= low\_threshold\_NLGN4Y ~ "low\_expression",  
 TRUE ~ "intermediate\_expression"), KDM5D\_expression\_category = case\_when(KDM5D >=  
 high\_threshold\_KDM5D ~ "high\_expression", KDM5D <=  
 low\_threshold\_KDM5D ~ "low\_expression", TRUE ~  
 "intermediate\_expression"))

# Adding inferred sex based on threshold expressions

# Create a new column based on values in  
# XIST\_expression\_category for X chromosome  
# complement  
inferred\_sex\_chromosome\_complement <- counts\_common\_normal\_primary\_gender %>%  
 mutate(X\_inferred\_complement = case\_when(XIST\_expression\_category ==  
 "intermediate\_expression" ~ "X", XIST\_expression\_category ==  
 "high\_expression" ~ "X", XIST\_expression\_category ==  
 "low\_expression" ~ "O"))

# Since there are multiple Y-Linked genes, will take a count for all cases that show differing expressions, then if more than half are low expressed will assign ‘O’

# Define the columns of interest  
Y\_category\_columns <- c("RPS4Y1\_expression\_category",  
 "ZFY\_expression\_category", "USP9Y\_expression\_category",  
 "DDX3Y\_expression\_category", "UTY\_expression\_category",  
 "TMSB4Y\_expression\_category", "EIF1AY\_expression\_category",  
 "NLGN4Y\_expression\_category", "KDM5D\_expression\_category")  
  
# Convert relevant columns to a logical matrix  
# where 'low\_expression' is TRUE  
logical\_matrix\_Y <- inferred\_sex\_chromosome\_complement[Y\_category\_columns] ==  
 "low\_expression"  
  
# Count the number of TRUE values in each row and  
# add it as a new column  
inferred\_sex\_chromosome\_complement <- inferred\_sex\_chromosome\_complement %>%  
 mutate(Y\_low\_expression\_count = rowSums(logical\_matrix\_Y,  
 na.rm = TRUE))  
  
# For samples that have <4 genes expressing  
# 'low\_expression' assign a 'Y' chromosome  
# complement  
inferred\_sex\_chromosome\_complement <- inferred\_sex\_chromosome\_complement %>%  
 mutate(Y\_inferred\_complement = if\_else(Y\_low\_expression\_count <  
 4, "Y", NA\_character\_))

# Adding another column with inferred sex chromosome complement pairs

inferred\_sex\_chromosome\_complement <- inferred\_sex\_chromosome\_complement %>%  
 mutate(sex\_chromosome\_pairs = if\_else(is.na(Y\_inferred\_complement),  
 paste0("X", X\_inferred\_complement), paste0("X",  
 X\_inferred\_complement, Y\_inferred\_complement))) %>%  
 mutate(sex\_chromosome\_pairs = if\_else(gender ==  
 "MALE", gsub("O", "", sex\_chromosome\_pairs),  
 sex\_chromosome\_pairs))

# Make new data frame with case, gender, and sex chromosome pairs

# Extract the desired columns into a new data  
# frame  
sex\_complements <- subset(inferred\_sex\_chromosome\_complement,  
 select = c(cases.submitter\_id, gender, SampleType,  
 sex\_chromosome\_pairs))

# From the sex\_complements data created above, table out the different combinations of

# Change 'Solid Tissue Normal' to  
# 'Normal/Adjacent'  
sex\_complements <- sex\_complements %>%  
 mutate(SampleType = if\_else(SampleType == "Solid Tissue Normal",  
 "Normal/Adjacent", SampleType))  
  
  
# # Note to self of all the potential  
# combinations for sex chromosome pairs FEMALE  
# 'Normal/Adjacent' with XX and Primary Tumor XX  
# FEMALE 'Normal/Adjacent' with XX and Primary  
# Tumor XO FEMALE 'Normal/Adjacent' with XO and  
# Primary Tumor XO FEMALE 'Normal/Adjacent' with  
# XO and Primary Tumor XX FEMALE  
# 'Normal/Adjacent' with XXX and Primary Tumor  
# XXX FEMALE 'Normal/Adjacent' with XO and  
# Primary Tumor XXX FEMALE 'Normal/Adjacent' with  
# XXX and Primary Tumor XO FEMALE  
# 'Normal/Adjacent' with XX and Primary Tumor XXX  
# FEMALE 'Normal/Adjacent' with XXX and Primary  
# Tumor XX FEMALE 'Normal/Adjacent' with XX and  
# Primary Tumor XXY MALE 'Normal/Adjacent' with  
# XY and Primary Tumor XY MALE 'Normal/Adjacent'  
# with XY and Primary Tumor XO MALE  
# 'Normal/Adjacent' with XO and Primary Tumor XO  
# MALE 'Normal/Adjacent' with XO and Primary  
# Tumor XY MALE 'Normal/Adjacent' with XO and  
# Primary Tumor XXY MALE 'Normal/Adjacent' with  
# XXY and Primary Tumor XY MALE 'Normal/Adjacent'  
# with XXY and Primary Tumor XO  
  
  
# Create two subsets: one for 'Normal/Adjacent'  
# and one for 'Primary Tumor'  
normal\_adjacent <- sex\_complements %>%  
 dplyr::filter(SampleType == "Normal/Adjacent") %>%  
 dplyr::select(cases.submitter\_id, gender, sex\_chromosome\_pairs) %>%  
 dplyr::rename(Normal\_adjacent\_sex\_chromosome\_pairs = sex\_chromosome\_pairs)  
  
primary\_tumor <- sex\_complements %>%  
 dplyr::filter(SampleType == "Primary Tumor") %>%  
 dplyr::select(cases.submitter\_id, sex\_chromosome\_pairs) %>%  
 dplyr::rename(Primary\_tumor\_sex\_chromosome\_pairs = sex\_chromosome\_pairs)  
  
  
# Merge the two subsets on 'cases.submitter\_id'  
combined\_data <- dplyr::inner\_join(normal\_adjacent,  
 primary\_tumor, by = "cases.submitter\_id")  
  
  
# Define the combinations of interest  
combinations <- c("FEMALE Normal/Adjacent XX Primary Tumor XX",  
 "FEMALE Normal/Adjacent XX Primary Tumor XO", "FEMALE Normal/Adjacent XO Primary Tumor XO",  
 "FEMALE Normal/Adjacent XO Primary Tumor XX", "FEMALE Normal/Adjacent XXX Primary Tumor XXX",  
 "FEMALE Normal/Adjacent XO Primary Tumor XXX",  
 "FEMALE Normal/Adjacent XXX Primary Tumor XO",  
 "FEMALE Normal/Adjacent XX Primary Tumor XXX",  
 "FEMALE Normal/Adjacent XXX Primary Tumor XX",  
 "FEMALE Normal/Adjacent XX Primary Tumor XXY",  
 "MALE Normal/Adjacent XY Primary Tumor XY", "MALE Normal/Adjacent XY Primary Tumor XO",  
 "MALE Normal/Adjacent XO Primary Tumor XO", "MALE Normal/Adjacent XO Primary Tumor XY",  
 "MALE Normal/Adjacent XO Primary Tumor XXY", "MALE Normal/Adjacent XXY Primary Tumor XY",  
 "MALE Normal/Adjacent XXY Primary Tumor XO", "MALE Normal/Adjacent XXY Primary Tumor XXY",  
 "MALE Normal/Adjacent XY Primary Tumor XXY")  
  
# Create a column to identify the combinations  
combined\_data <- combined\_data %>%  
 dplyr::mutate(combination = paste(gender, "Normal/Adjacent",  
 Normal\_adjacent\_sex\_chromosome\_pairs, "Primary Tumor",  
 Primary\_tumor\_sex\_chromosome\_pairs))  
  
# Filter and count the combinations of interest  
result <- combined\_data %>%  
 dplyr::filter(combination %in% combinations) %>%  
 dplyr::group\_by(combination) %>%  
 dplyr::summarize(count = n())  
  
result

## # A tibble: 7 × 2  
## combination count  
## <chr> <int>  
## 1 FEMALE Normal/Adjacent XX Primary Tumor XO 2  
## 2 FEMALE Normal/Adjacent XX Primary Tumor XX 19  
## 3 FEMALE Normal/Adjacent XX Primary Tumor XXY 1  
## 4 MALE Normal/Adjacent XXY Primary Tumor XXY 3  
## 5 MALE Normal/Adjacent XXY Primary Tumor XY 2  
## 6 MALE Normal/Adjacent XY Primary Tumor XXY 1  
## 7 MALE Normal/Adjacent XY Primary Tumor XY 22