CountsPlotting

John E Froberg

Sys.Date()

```
knitr::opts_chunk$set(echo=FALSE, message=FALSE)

library('devtools')
```

```
## Loading required package: usethis
```

```
library('tidyverse')
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library('riboWaltz')
library("rmarkdown")
library("patchwork")
library("pheatmap")
library("RColorBrewer")
library("ggplotify")
library("topGO")
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## 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, 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: graph

##
## Attaching package: 'graph'

## The following object is masked from 'package:stringr':
##
##     boundary

## 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")'.

## Loading required package: GO.db

## Loading required package: AnnotationDbi

## Loading required package: stats4

## Loading required package: IRanges

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:dplyr':
##
##     first, rename

## The following object is masked from 'package:tidyr':
##
##     expand

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

##
## Attaching package: 'IRanges'

## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice

## The following object is masked from 'package:purrr':
##
##     reduce

##
## Attaching package: 'AnnotationDbi'

## The following object is masked from 'package:dplyr':
##
##     select

##
```

```
## Loading required package: SparseM
##
## Attaching package: 'SparseM'
## The following object is masked from 'package:base':
##
##     backsolve
##
## groupGOTerms:    GOBPTerm, GOMFTerm, GOCCTerm environments built.
##
## Attaching package: 'topGO'
## The following object is masked from 'package:IRanges':
##
##     members
library("ggstance")
##
## Attaching package: 'ggstance'
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
library("DESeq2")
## Loading required package: GenomicRanges
## Loading required package: GenomeInfoDb
## Loading required package: SummarizedExperiment
## Loading required package: MatrixGenerics
## Loading required package: matrixStats
##
## Attaching package: 'matrixStats'
## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians
## 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

## The following object is masked from 'package:Biobase':
##
##     rowMedians
```

```
library("biomaRt")
library("UpSetR")
library("GGally")
```

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

```
library("clusterProfiler")
```

```
## Registered S3 method overwritten by 'ggtree':
##   method      from
##   identify.gg ggfun

## clusterProfiler v4.2.2  For help: https://yulab-smu.top/biomedical-knowledge-mining-book/
##
## If you use clusterProfiler in published research, please cite:
## T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, L Zhan, X Fu, S Liu, X Bo, and G Yu.

##
## Attaching package: 'clusterProfiler'

## The following object is masked from 'package:biomaRt':
##
##     select

## The following object is masked from 'package:AnnotationDbi':
##
##     select

## The following object is masked from 'package:IRanges':
##
##     slice

## The following object is masked from 'package:S4Vectors':
##
##     rename

## The following object is masked from 'package:purrr':
##
##     simplify

## The following object is masked from 'package:stats':
##
##     filter
```

```
library("org.Mm.eg.db")
```

```
##
```

```
# #variables that are normally passed in from the Snakemake run:
experimentTypeFile="experimentTypesFull_SubFall2021.csv"
lengthCountsFile="dedup_lengthDistroTidy.txt"
CDS="dedupBams/featureCounts_CDS_summary.txt"
three_utr="dedupBams/featureCounts_three_prime_utr_summary.txt"
five_utr="dedupBams/featureCounts_five_prime_utr_summary.txt"
gtf="Mus_musculus.GRCm38.95_chrNamed_headFix.gtf"
bamFiles="dedup_RPbams"
outDir="interactivePlots"
dir.create(outDir)
```

```
## Warning in dir.create(outDir): 'interactivePlots' already exists
```

```
RiboCodeFile="RiboCode_ORFs_out.txt"
```

```
JF_theme <- theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),panel.backgrou
```

###Load in the metadata from the experiment types file. ###Compute "descriptive names" by dropping the "samps" and "date" fields and pasting together. ###Save descriptiveNames as Upper Case "Samples" in experimentTypes.

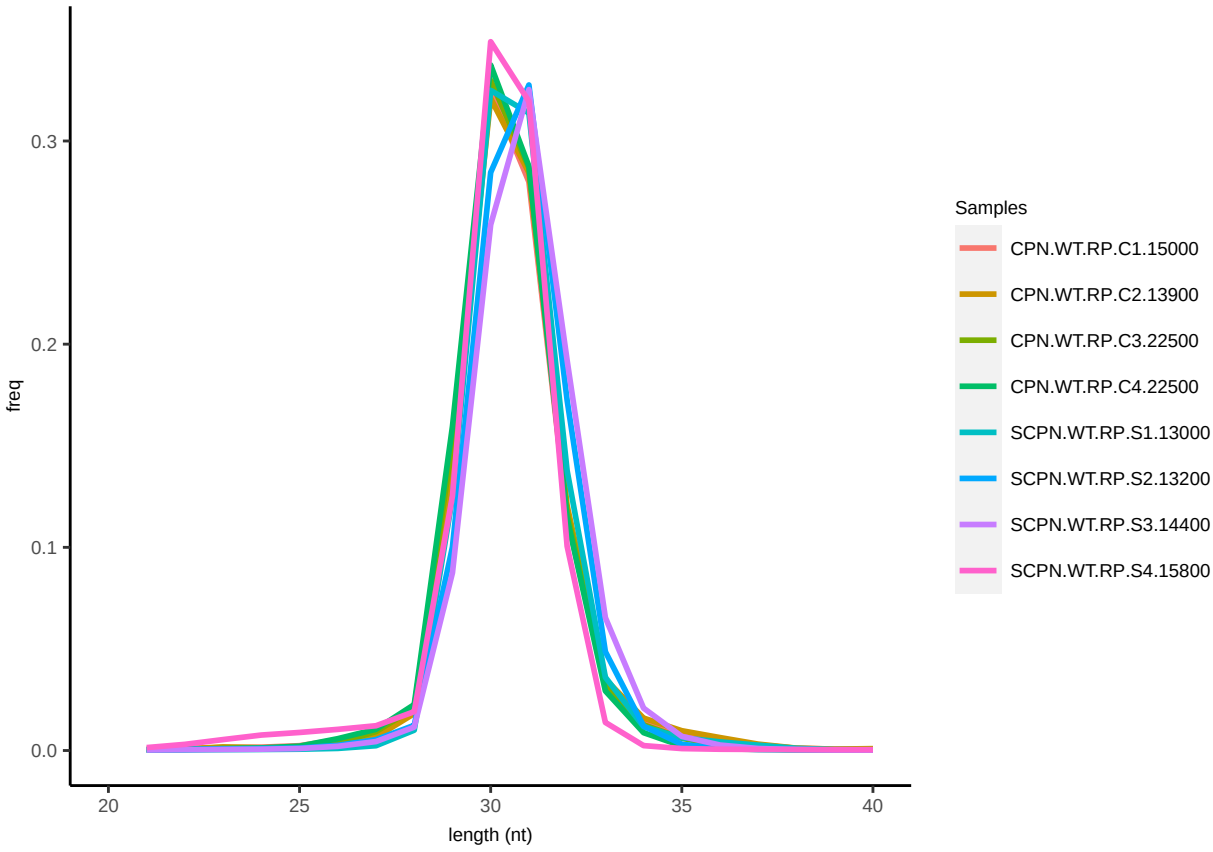
```
##        samps exp subtype genotype   rep    date cells             Samples
## 1   JF134_S1  RP     CPN       WT RP.C1  8/3/21 15000  CPN.WT.RP.C1.15000
## 2   JF135_S2  RP     CPN       WT RP.C2  8/3/21 13900  CPN.WT.RP.C2.13900
## 3   JF136_S3  RP    SCPN       WT RP.S1 8/25/21 13000 SCPN.WT.RP.S1.13000
## 4   JF137_S4  RP    SCPN       WT RP.S2 8/26/21 13200 SCPN.WT.RP.S2.13200
## 5   JF138_S5  AF     CPN       WT AF.C1  8/3/21  5000   CPN.WT.AF.C1.5000
## 6   JF139_S6  AF     CPN       WT AF.C2  8/3/21  4600   CPN.WT.AF.C2.4600
## 7   JF140_S7  AF    SCPN       WT AF.S1 8/25/21  3100  SCPN.WT.AF.S1.3100
## 8   JF141_S8  AF    SCPN       WT AF.S2 8/25/21  3300  SCPN.WT.AF.S2.3300
## 9   JF142_S9  RP     CPN       WT RP.C3  8/3/21 22500  CPN.WT.RP.C3.22500
## 10 JF143_S10  RP     CPN       WT RP.C4  8/3/21 22500  CPN.WT.RP.C4.22500
## 11 JF144_S11  RP    SCPN       WT RP.S3 8/30/21 14400 SCPN.WT.RP.S3.14400
## 12 JF145_S12  RP    SCPN       WT RP.S4  9/8/21 15800 SCPN.WT.RP.S4.15800
## 13 JF146_S13  AF     CPN       WT AF.C3  8/3/21  7500   CPN.WT.AF.C3.7500
## 14 JF147_S14  AF     CPN       WT AF.C4  8/3/21  7500   CPN.WT.AF.C4.7500
## 15 JF148_S15  AF    SCPN       WT AF.S3 8/30/21  3600  SCPN.WT.AF.S3.3600
## 16 JF149_S16  AF    SCPN       WT AF.S4  9/8/21  5200  SCPN.WT.AF.S4.5200
```

###Convert length counts to relative frequencies, add in the metadata

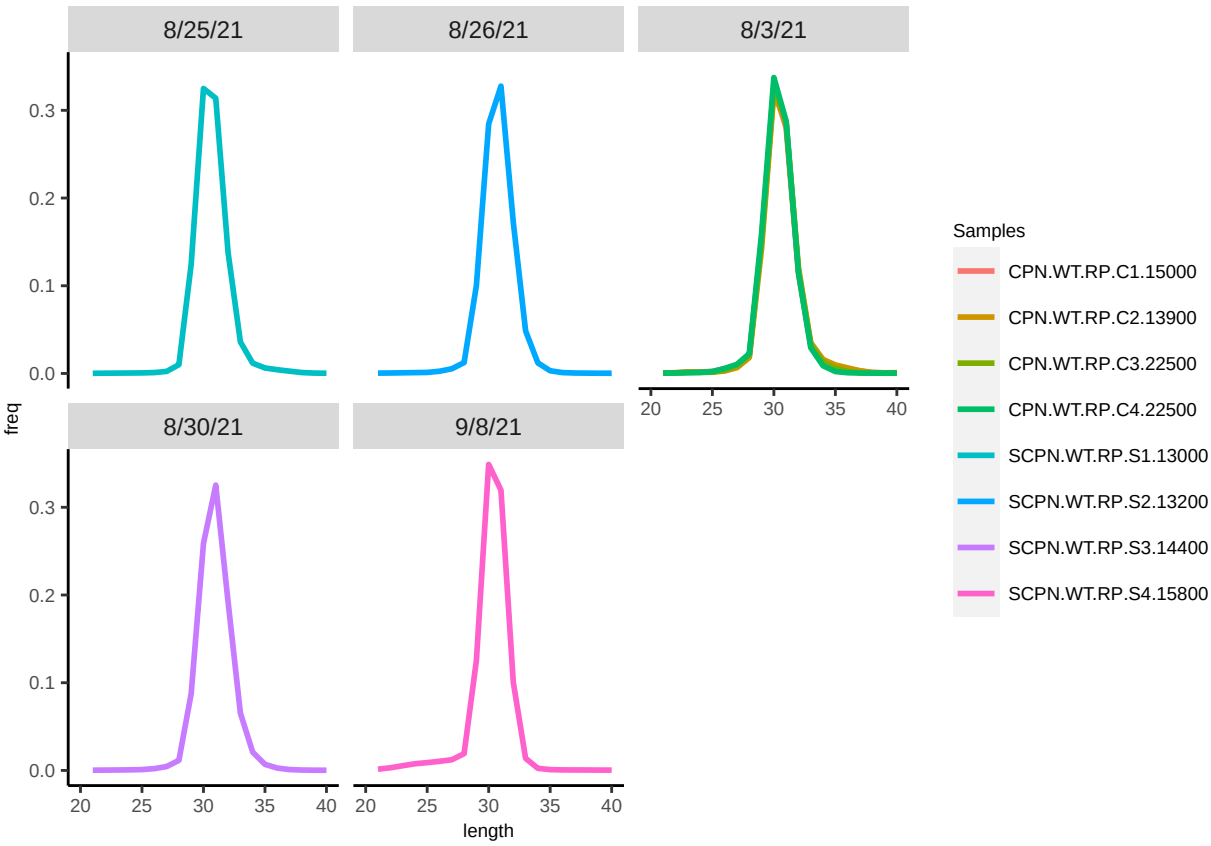
###Find the length with max freq for all samples

###Plot Freq vs length for the CDS with various groupings of the samples

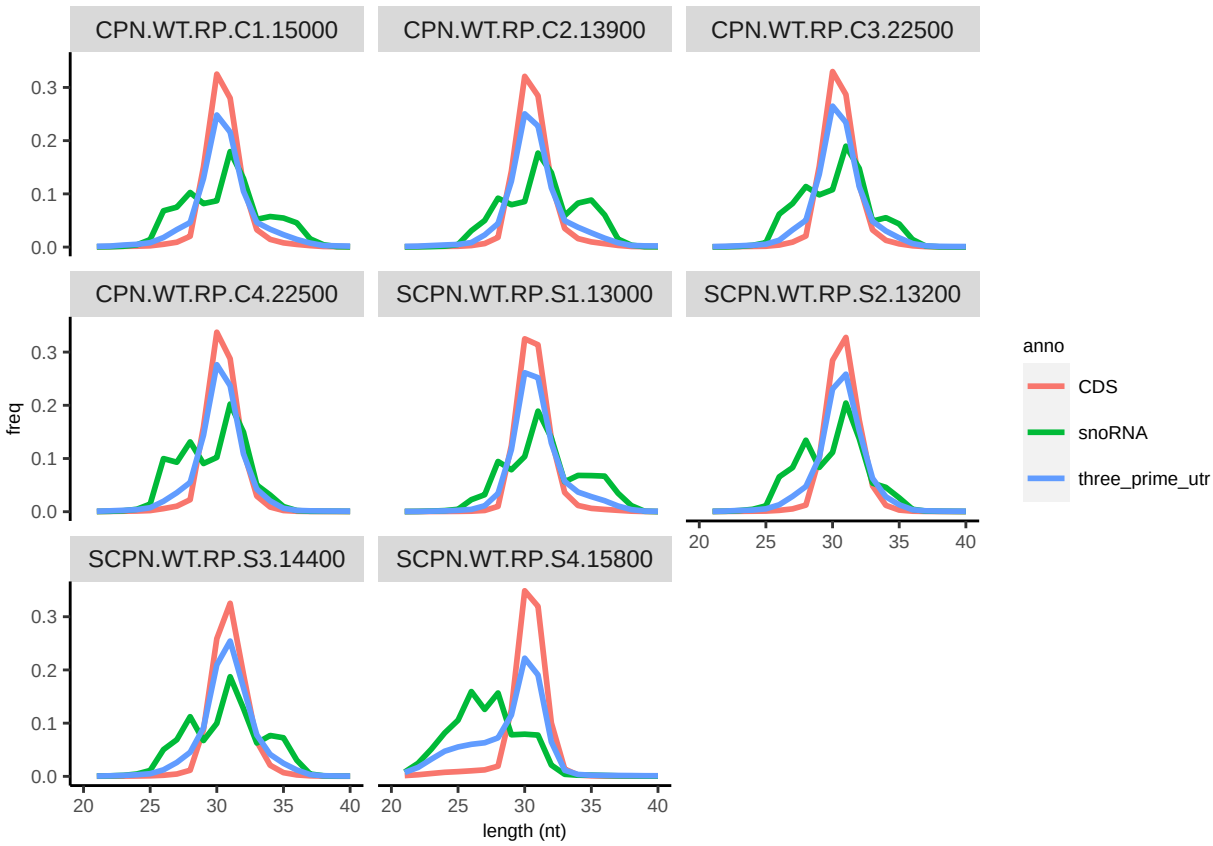
```
## Warning: Removed 72 row(s) containing missing values (geom_path).
```




```
## Warning: Removed 72 row(s) containing missing values (geom_path).
```

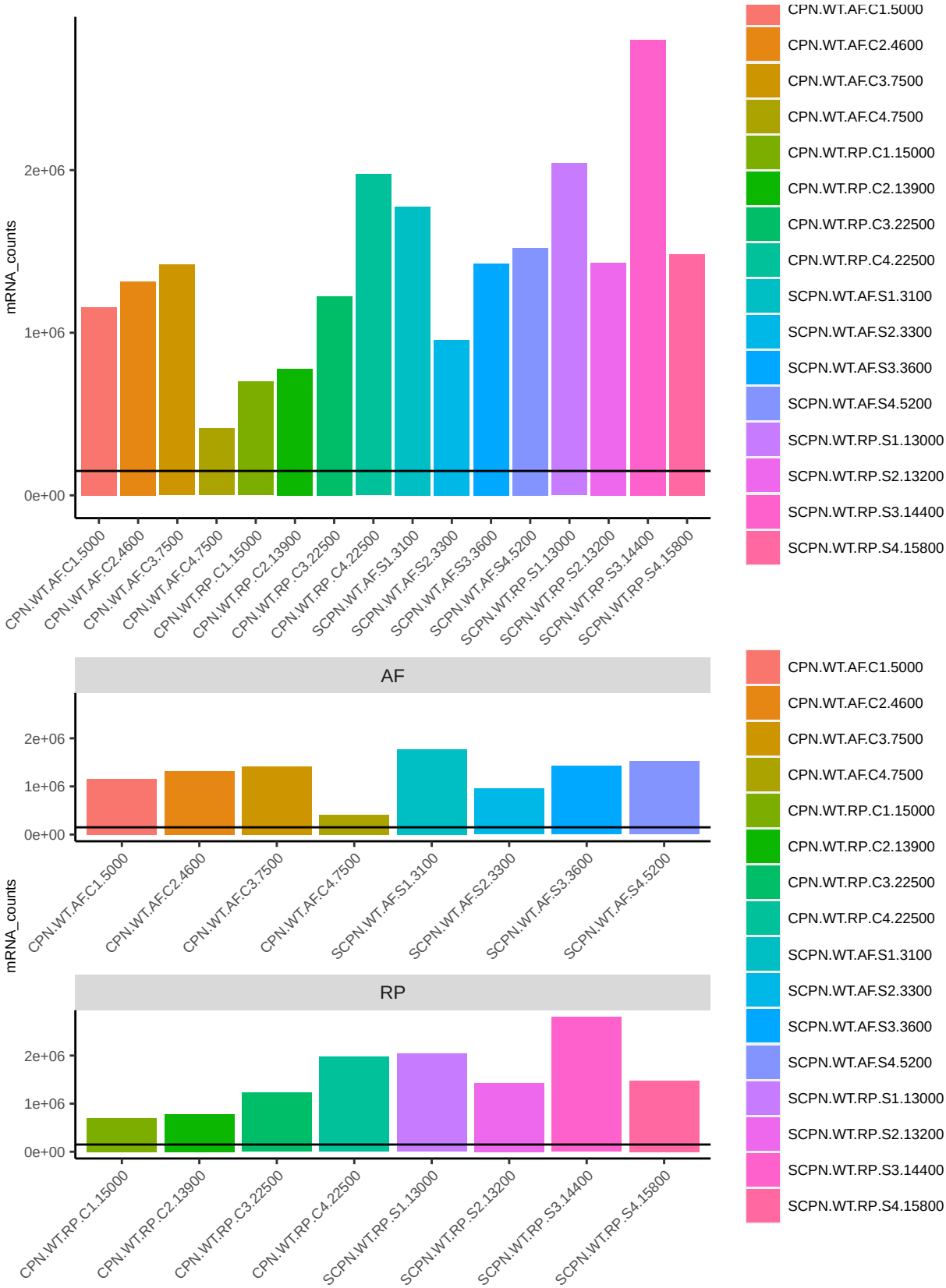



```
## Warning: Removed 27 row(s) containing missing values (geom_path).
```

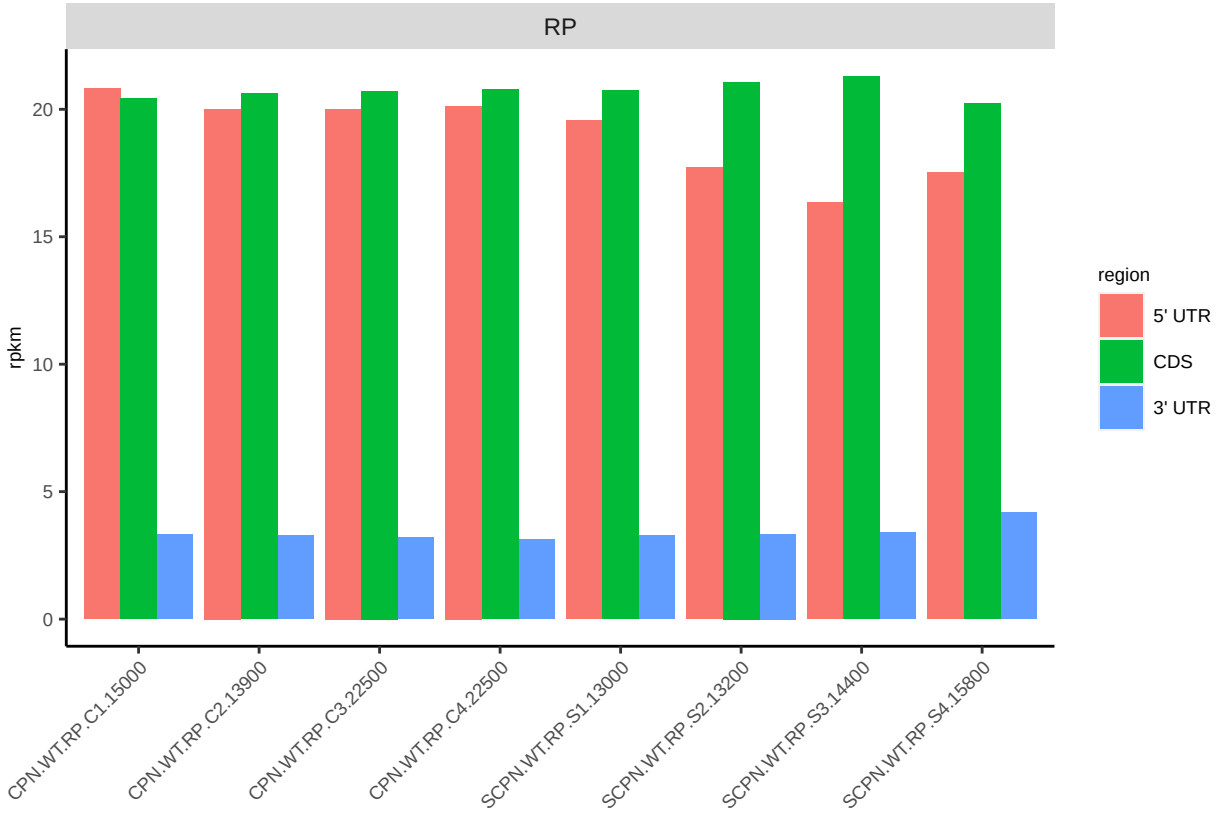


###Load in the featureCounts. Use only CDS reads for RiboProf, 3'+5'UTR+CDS reads for AlkFrag libraries

```
## # A tibble: 16 x 2
##    Samples              mRNA_counts
##    <chr>                      <int>
##  1 CPN.WT.AF.C1.5000        1157569
##  2 CPN.WT.AF.C2.4600        1315772
##  3 CPN.WT.AF.C3.7500        1419054
##  4 CPN.WT.AF.C4.7500         411772
##  5 CPN.WT.RP.C1.15000        698897
##  6 CPN.WT.RP.C2.13900        778664
##  7 CPN.WT.RP.C3.22500       1225160
##  8 CPN.WT.RP.C4.22500       1974973
##  9 SCPN.WT.AF.S1.3100       1775011
## 10 SCPN.WT.AF.S2.3300        955117
## 11 SCPN.WT.AF.S3.3600       1425382
## 12 SCPN.WT.AF.S4.5200       1522812
## 13 SCPN.WT.RP.S1.13000      2041454
## 14 SCPN.WT.RP.S2.13200      1430694
## 15 SCPN.WT.RP.S3.14400      2802151
## 16 SCPN.WT.RP.S4.15800      1481631
```

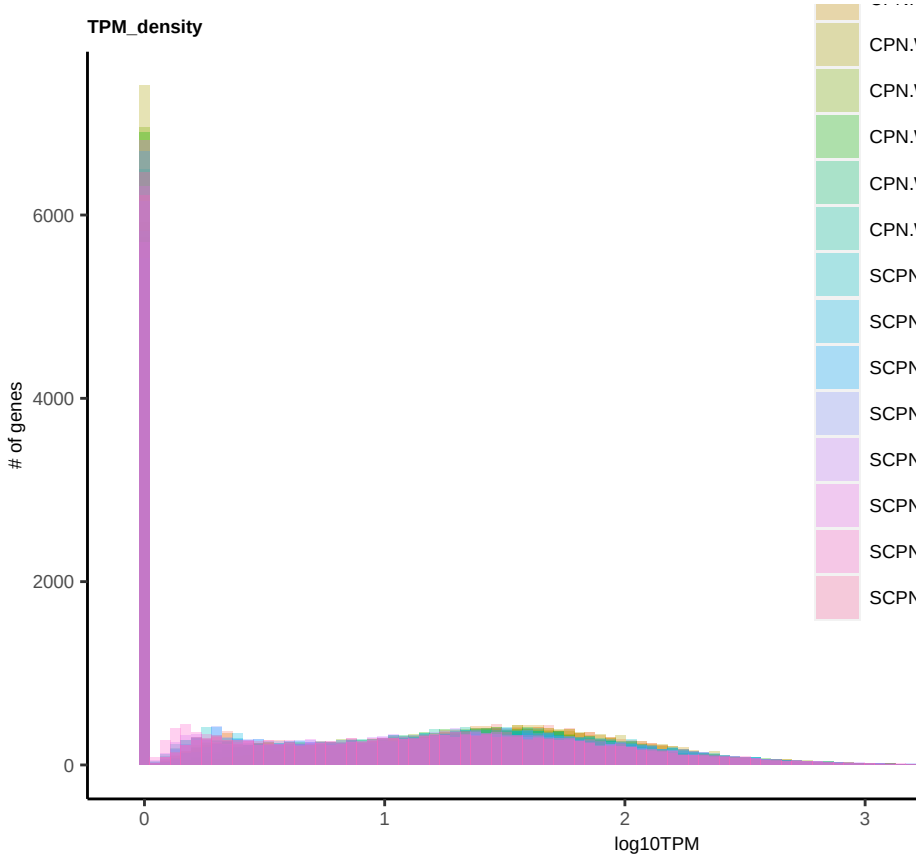


###Calculate the and plot reads per kilobase per million mapped (to mRNA) for the ribosome profiling experi-

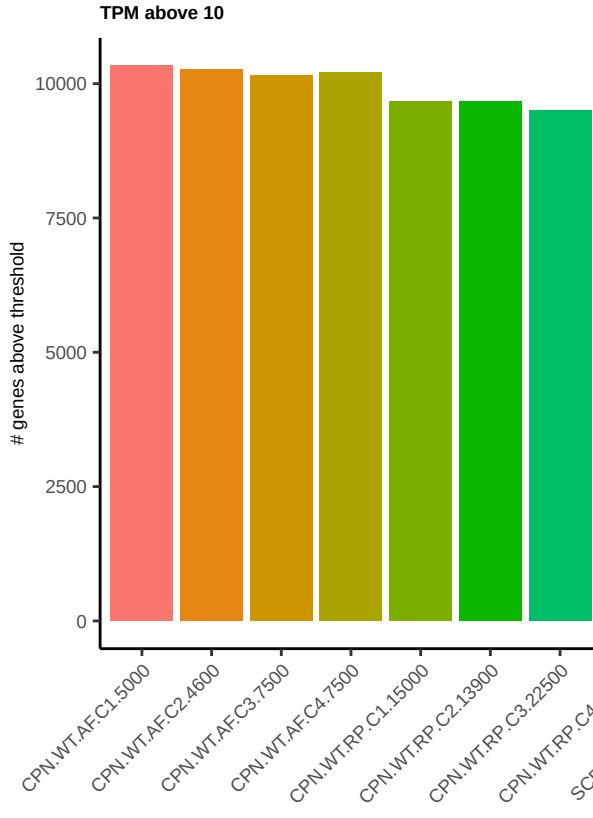


ments

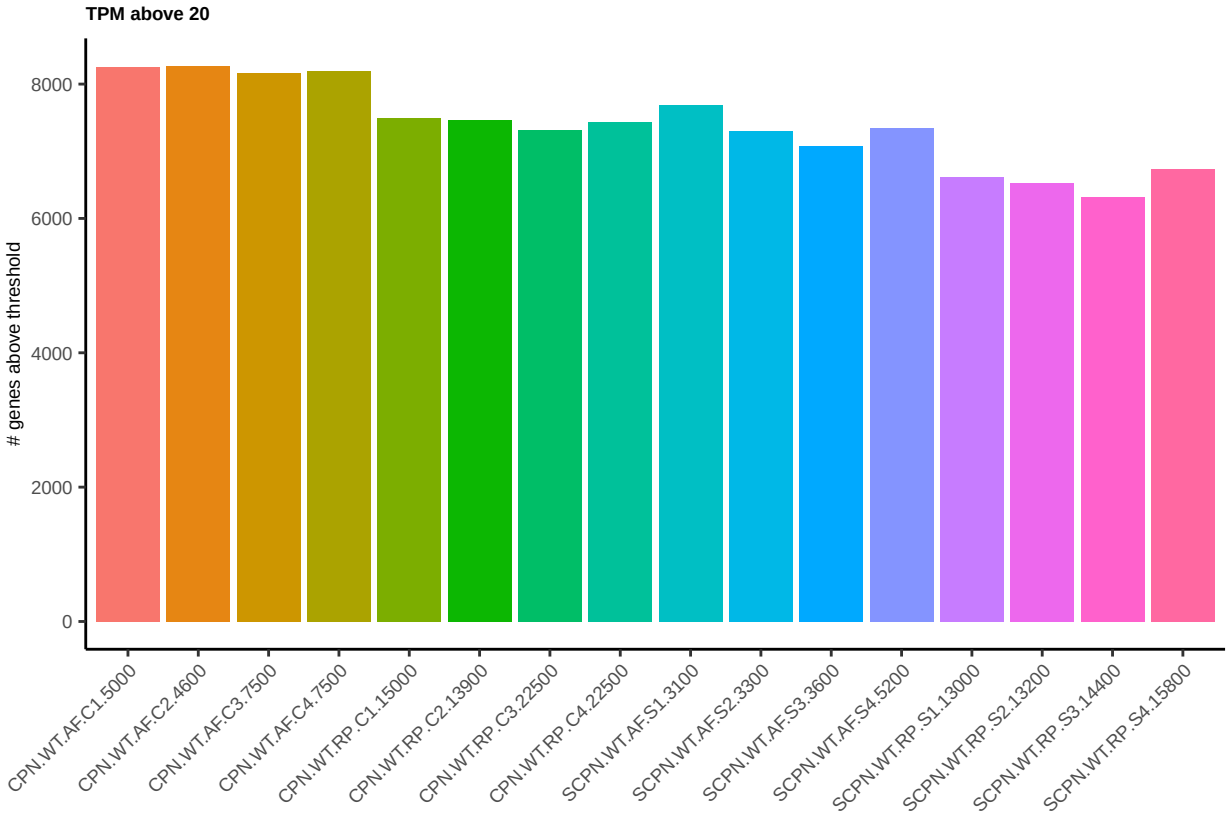
###Calculate TPMs from the raw counts

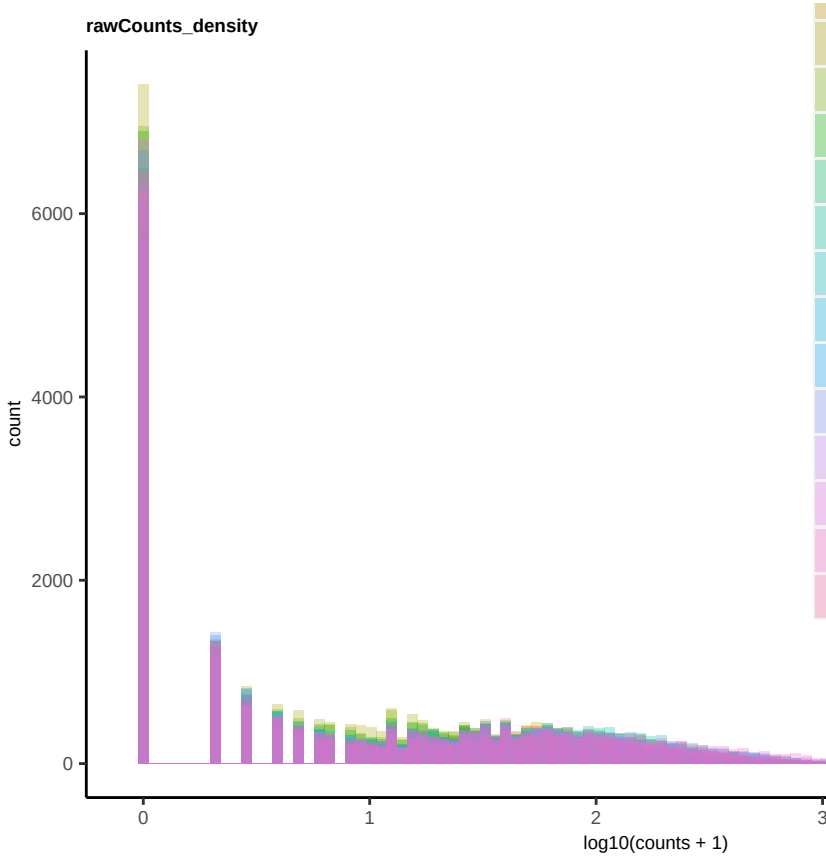


###Now plot the log10TPM+1 distributions

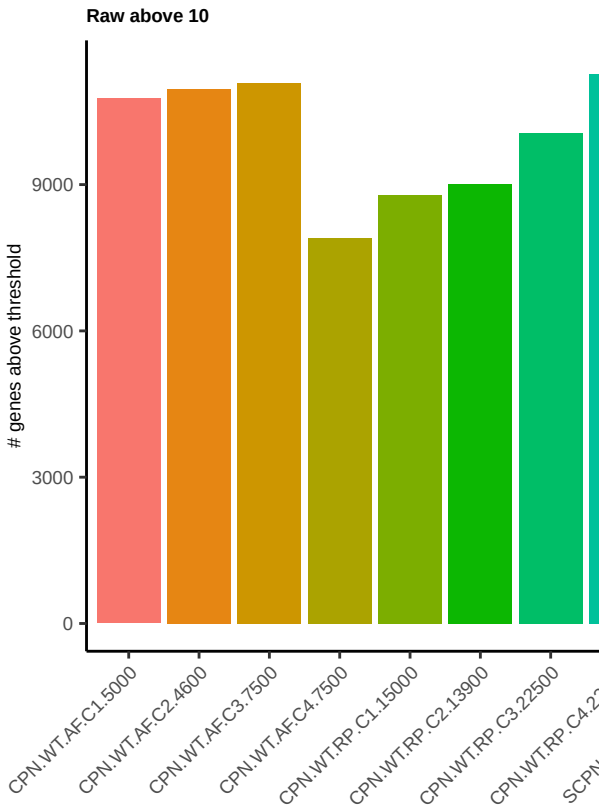


###Plot the number of genes above TPM threshold=10 and threshold=20

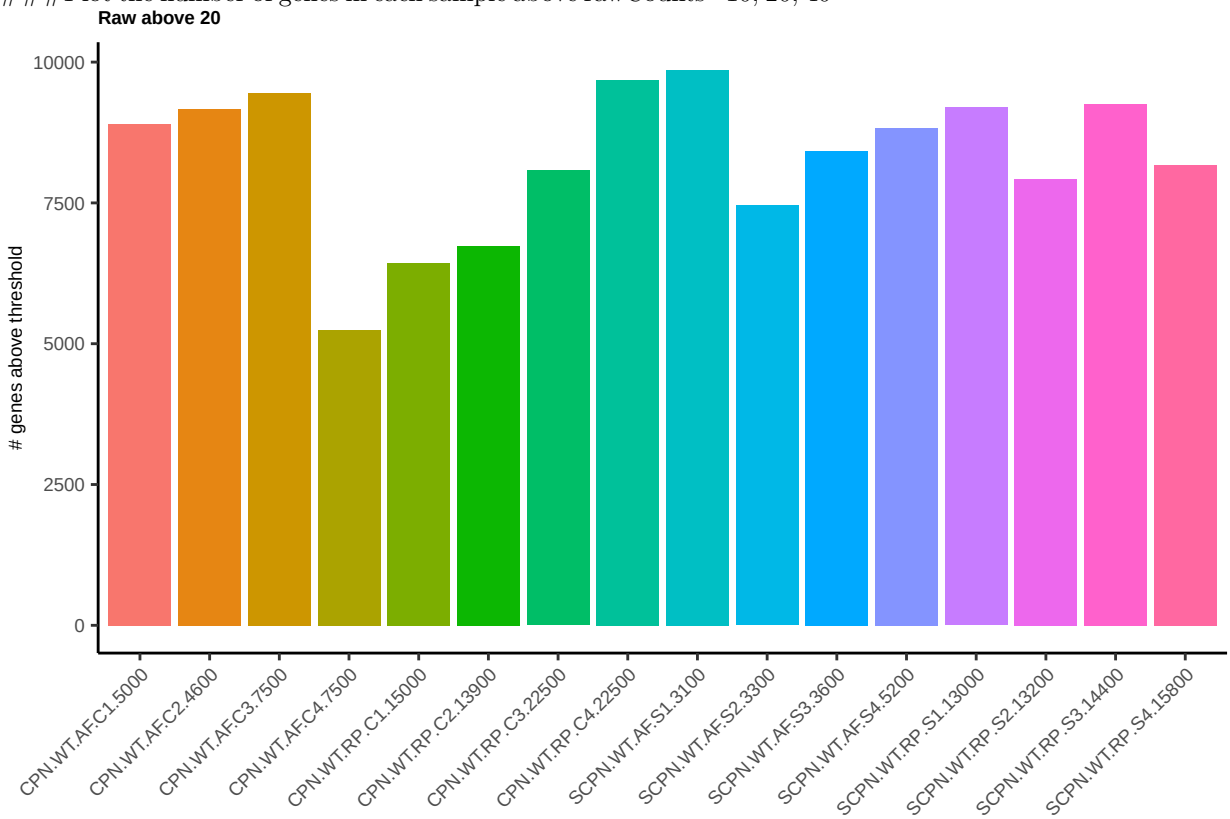


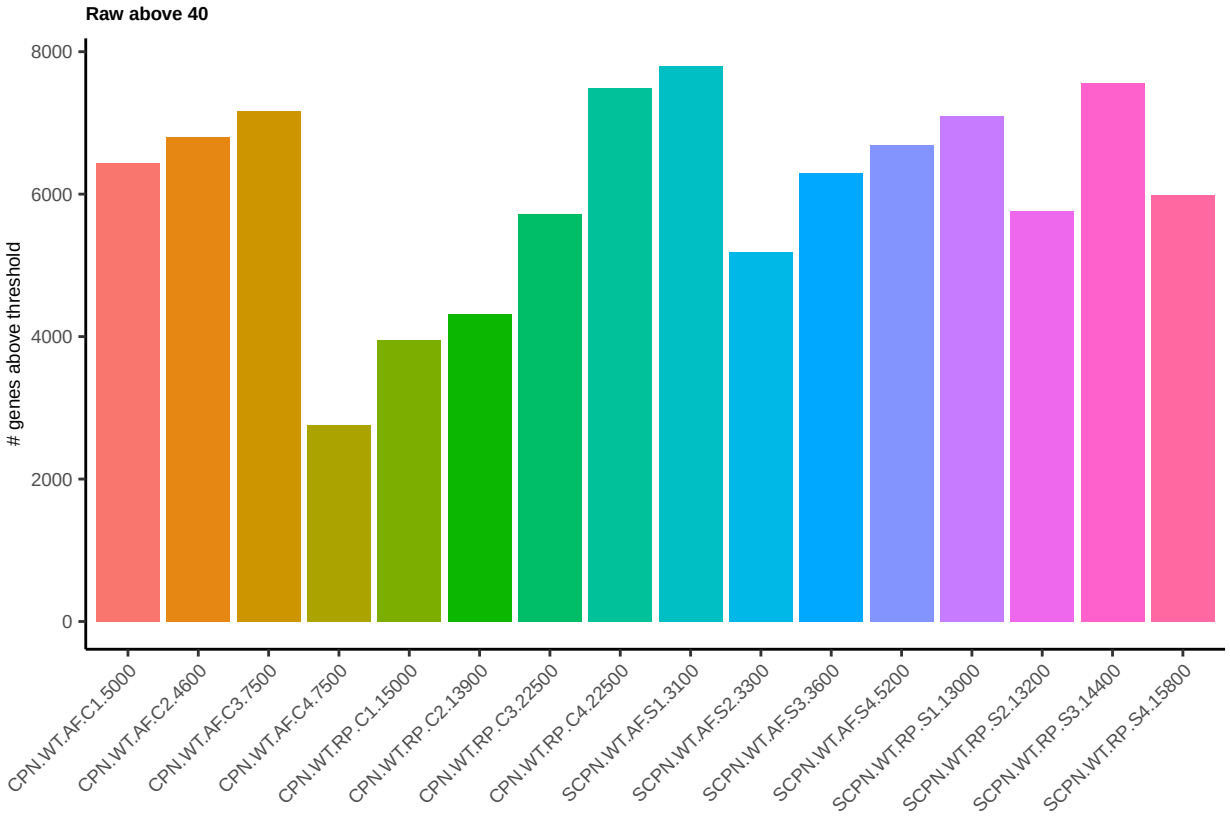


###Plot the distribtion of rawCounts in each sample:



###Plot the number of genes in each sample above rawCounts=10, 20, 40

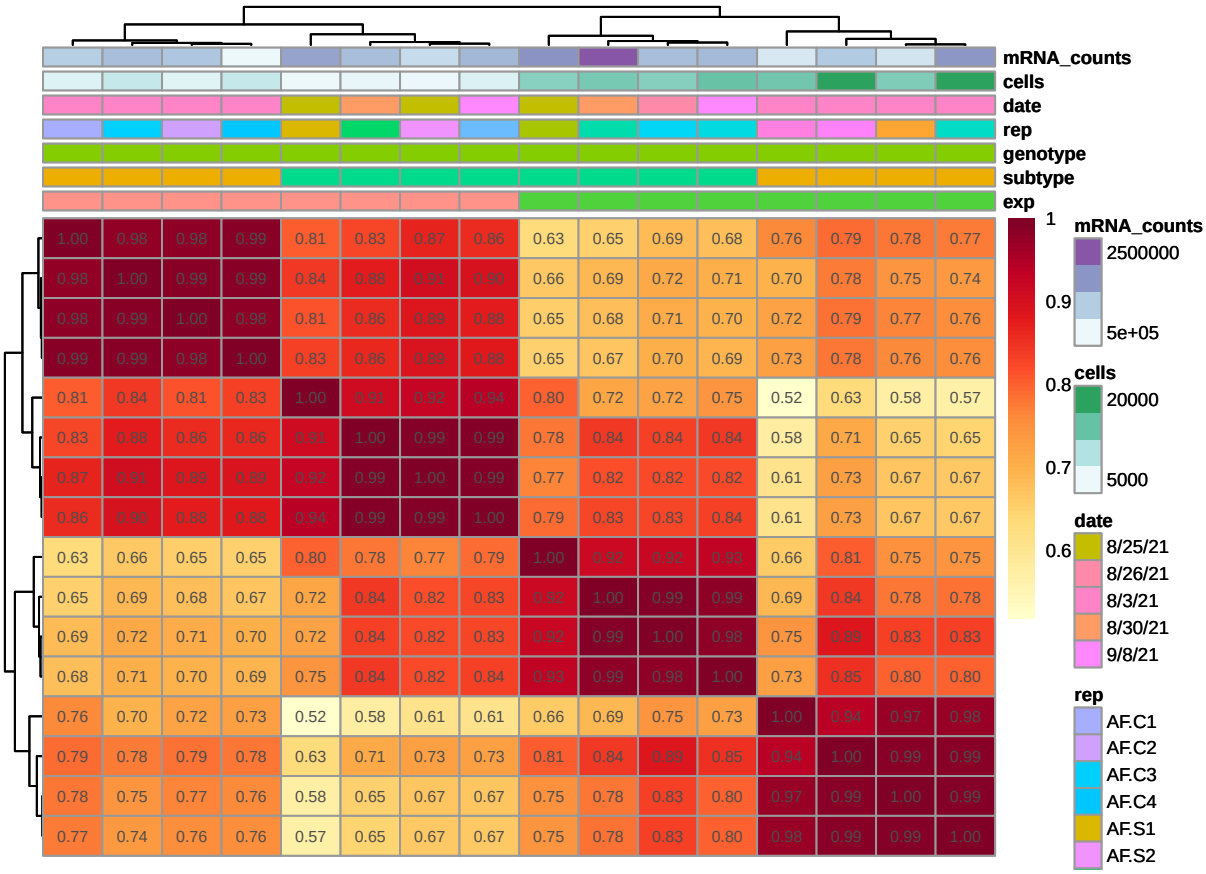


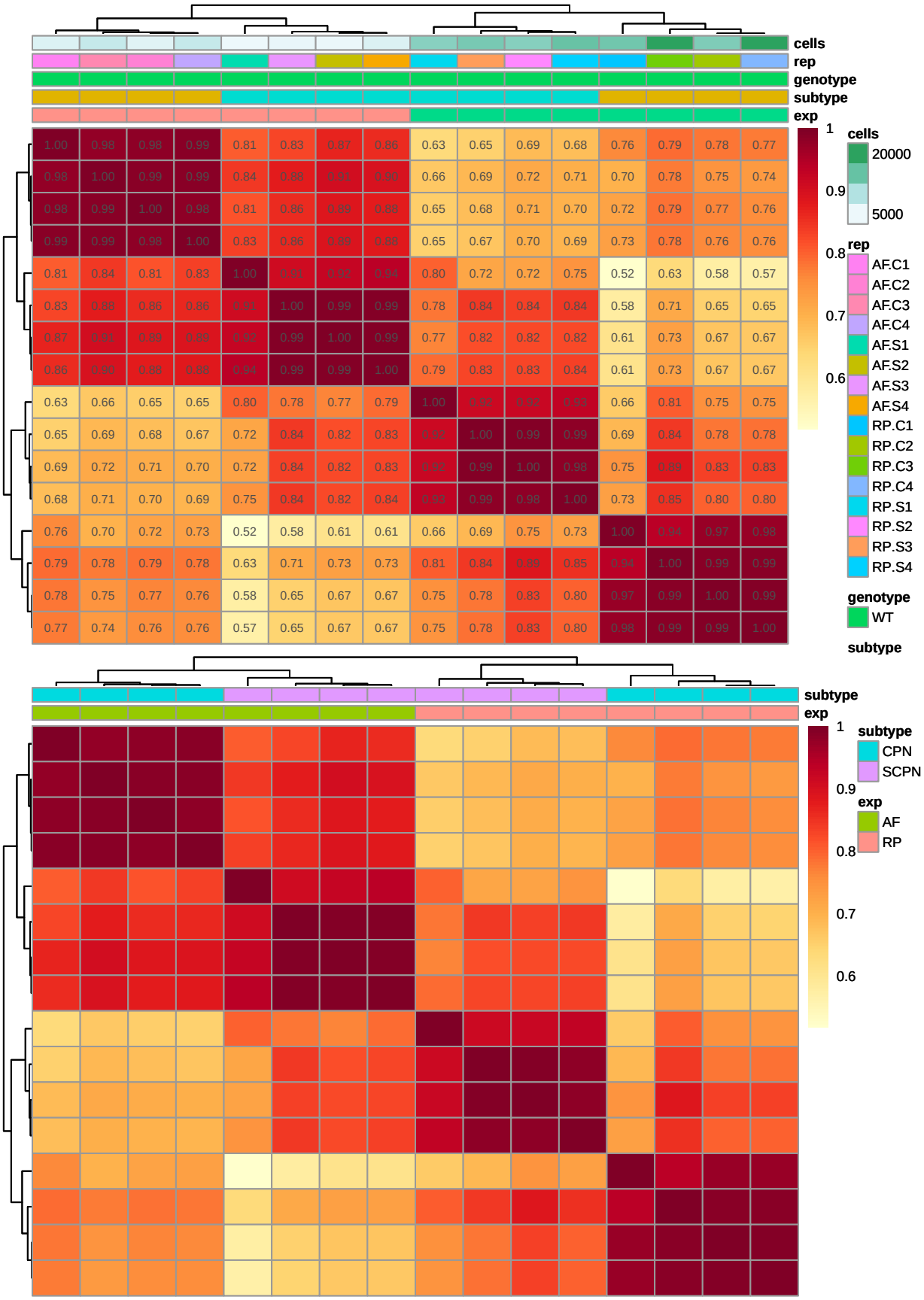


###Code to filter a TPM-table either for genes where 1) all samples above a TPM-based filter (filterTPMs) ###Or 2) genes above a rawCount threshold (defult=10) in a given fraction of samples (default=1/4).

###Calculate and plot Pearson's R correlation coefficient for all genes above a certain raw threshold (here, 10). ###In a certain fraction of samples (here, 1/4 of samples) ###Leave mRNA counts and date in the report, remove for the figure version for clarity.

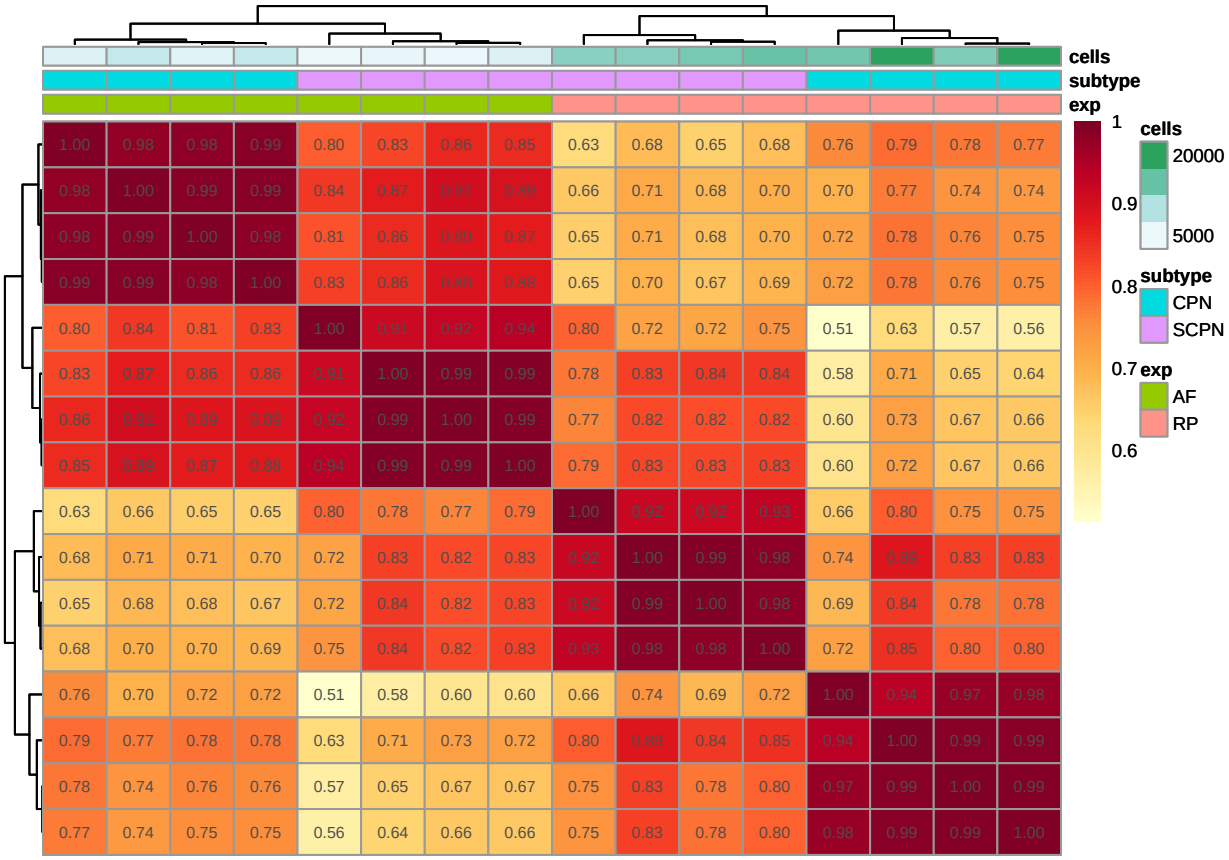
```
## [1] 11572
```

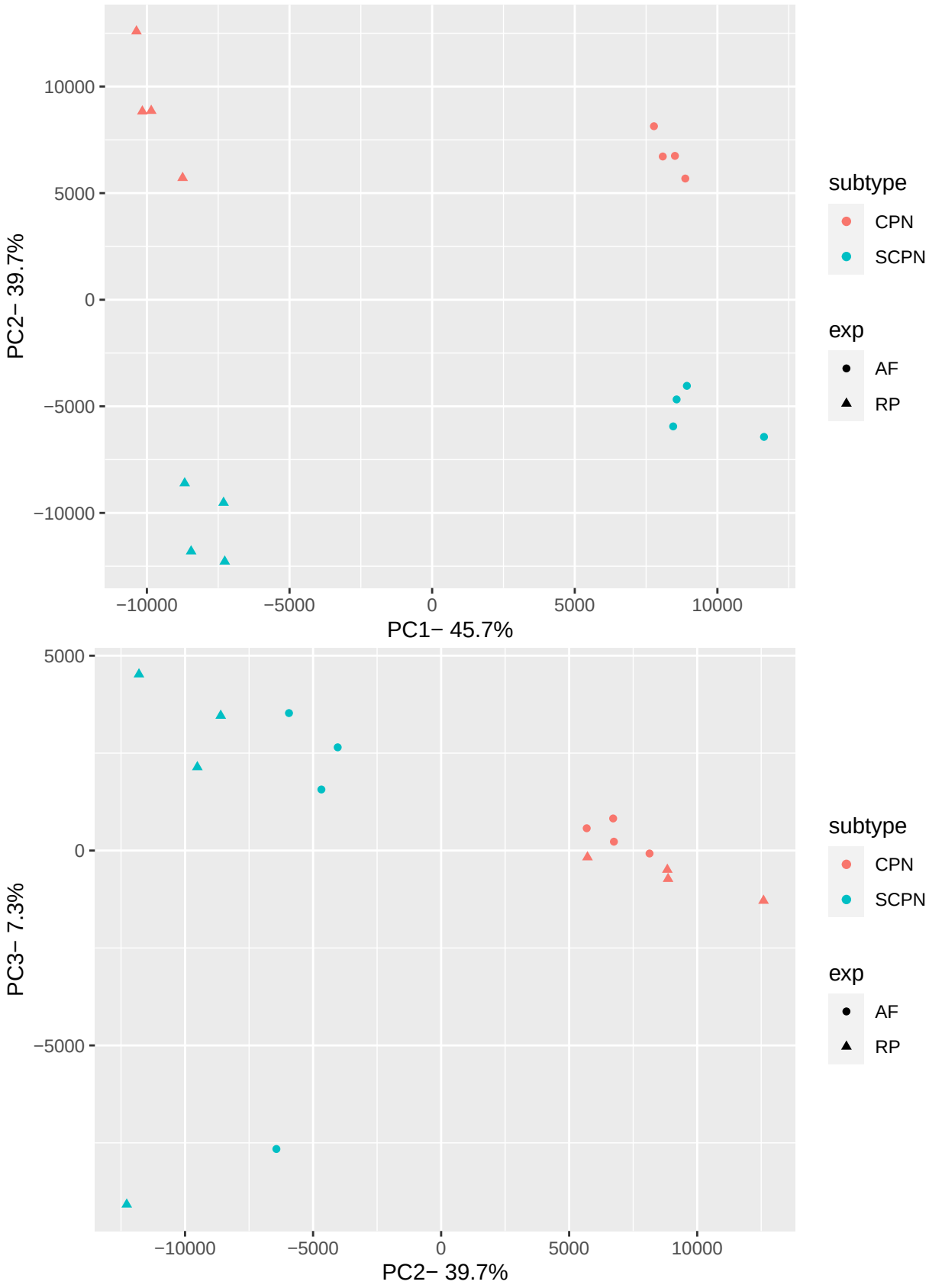

mRNA_counts
cells
date
rep
genotype
subtype
exp
1.00 0.98 0.98 0.99 0.81 0.83 0.87 0.86 0.63 0.65 0.69 0.68 0.76 0.79 0.78 0.77
0.98 1.00 0.99 0.99 0.84 0.88 0.91 0.90 0.66 0.69 0.72 0.71 0.70 0.78 0.75 0.74
0.98 0.99 1.00 0.98 0.81 0.86 0.89 0.88 0.65 0.68 0.71 0.70 0.72 0.79 0.77 0.76
0.99 0.99 0.98 1.00 0.83 0.86 0.89 0.88 0.65 0.67 0.70 0.69 0.73 0.78 0.76 0.76
0.81 0.84 0.81 0.83 1.00 0.91 0.92 0.94 0.80 0.72 0.72 0.75 0.52 0.63 0.58 0.57
0.83 0.88 0.86 0.86 0.91 1.00 0.99 0.99 0.78 0.84 0.84 0.84 0.58 0.71 0.65 0.65
0.87 0.91 0.89 0.89 0.92 0.99 1.00 0.99 0.77 0.82 0.82 0.82 0.61 0.73 0.67 0.67
0.86 0.90 0.88 0.88 0.94 0.99 0.99 1.00 0.79 0.83 0.83 0.84 0.61 0.73 0.67 0.67
0.63 0.66 0.65 0.65 0.80 0.78 0.77 0.79 1.00 0.92 0.92 0.93 0.66 0.81 0.75 0.75
0.65 0.69 0.68 0.67 0.72 0.84 0.82 0.83 0.92 1.00 0.99 0.99 0.69 0.84 0.78 0.78
0.69 0.72 0.71 0.70 0.72 0.84 0.82 0.83 0.92 0.99 1.00 0.98 0.75 0.89 0.83 0.83
0.68 0.71 0.70 0.69 0.75 0.84 0.82 0.84 0.93 0.99 0.98 1.00 0.73 0.85 0.80 0.80
0.76 0.70 0.72 0.73 0.52 0.58 0.61 0.61 0.66 0.69 0.75 0.73 1.00 0.94 0.97 0.98
0.79 0.78 0.79 0.78 0.63 0.71 0.73 0.73 0.81 0.84 0.89 0.85 0.94 1.00 0.99 0.99
0.78 0.75 0.77 0.76 0.58 0.65 0.67 0.67 0.75 0.78 0.83 0.80 0.97 0.99 1.00 0.99
0.77 0.74 0.76 0.76 0.57 0.65 0.67 0.67 0.75 0.78 0.83 0.80 0.98 0.99 0.99 1.00
1
0.9
0.8
0.7
0.6
mRNA_counts
2500000
5e+05
cells
20000
5000
date
8/25/21
8/26/21
8/3/21
8/30/21
9/8/21
rep
AF.C1
AF.C2
AF.C3
AF.C4
AF.S1
AF.S2

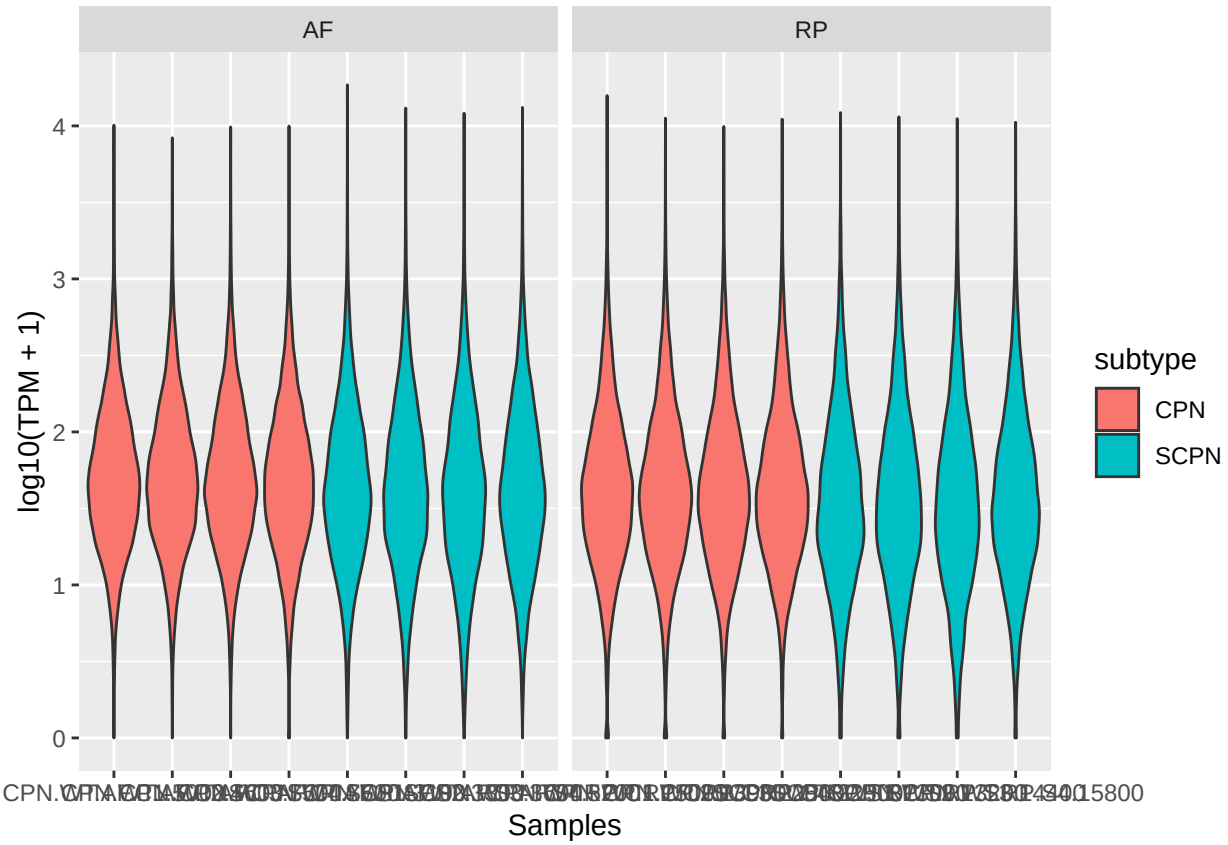

cells
rep
genotype
subtype
exp
1.00 0.98 0.98 0.99 0.81 0.83 0.87 0.86 0.63 0.65 0.69 0.68 0.76 0.79 0.78 0.77
0.98 1.00 0.99 0.99 0.84 0.88 0.91 0.90 0.66 0.69 0.72 0.71 0.70 0.78 0.75 0.74
0.98 0.99 1.00 0.98 0.81 0.86 0.89 0.88 0.65 0.68 0.71 0.70 0.72 0.79 0.77 0.76
0.99 0.99 0.98 1.00 0.83 0.86 0.89 0.88 0.65 0.67 0.70 0.69 0.73 0.78 0.76 0.76
0.81 0.84 0.81 0.83 1.00 0.91 0.92 0.94 0.80 0.72 0.72 0.75 0.52 0.63 0.58 0.57
0.83 0.88 0.86 0.86 0.91 1.00 0.99 0.99 0.78 0.84 0.84 0.84 0.58 0.71 0.65 0.65
0.87 0.91 0.89 0.89 0.92 0.99 1.00 0.99 0.77 0.82 0.82 0.82 0.61 0.73 0.67 0.67
0.86 0.90 0.88 0.88 0.94 0.99 0.99 1.00 0.79 0.83 0.83 0.84 0.61 0.73 0.67 0.67
0.63 0.66 0.65 0.65 0.80 0.78 0.77 0.79 1.00 0.92 0.92 0.93 0.66 0.81 0.75 0.75
0.65 0.69 0.68 0.67 0.72 0.84 0.82 0.83 0.92 1.00 0.99 0.99 0.69 0.84 0.78 0.78
0.69 0.72 0.71 0.70 0.72 0.84 0.82 0.83 0.92 0.99 1.00 0.98 0.75 0.89 0.83 0.83
0.68 0.71 0.70 0.69 0.75 0.84 0.82 0.84 0.93 0.99 0.98 1.00 0.73 0.85 0.80 0.80
0.76 0.70 0.72 0.73 0.52 0.58 0.61 0.61 0.66 0.69 0.75 0.73 1.00 0.94 0.97 0.98
0.79 0.78 0.79 0.78 0.63 0.71 0.73 0.73 0.81 0.84 0.89 0.85 0.94 1.00 0.99 0.99
0.78 0.75 0.77 0.76 0.58 0.65 0.67 0.67 0.75 0.78 0.83 0.80 0.97 0.99 1.00 0.99
0.77 0.74 0.76 0.76 0.57 0.65 0.67 0.67 0.75 0.78 0.83 0.80 0.98 0.99 0.99 1.00
1
0.9
0.8
0.7
0.6
cells
20000
5000
rep
AF.C1
AF.C2
AF.C3
AF.C4
AF.S1
AF.S2
AF.S3
AF.S4
RP.C1
RP.C2
RP.C3
RP.C4
RP.S1
RP.S2
RP.S3
RP.S4
genotype
WT
subtype
subtype
exp
1
0.9
0.8
0.7
0.6
subtype
CPN
SCPN
exp
AF
RP

#Filter out samples with fewer than 150000 mRNA-aligned reads (rounded to nearest 1000)

```
## [1] 9967
```




10000
5000
0
−5000
−10000
PC2− 39.7%
−10000
−5000
0
5000
10000
PC1− 45.7%
subtype
CPN
SCPN
exp
AF
RP
5000
0
−5000
PC3− 7.3%
−10000
−5000
0
5000
10000
PC2− 39.7%
subtype
CPN
SCPN
exp
AF
RP

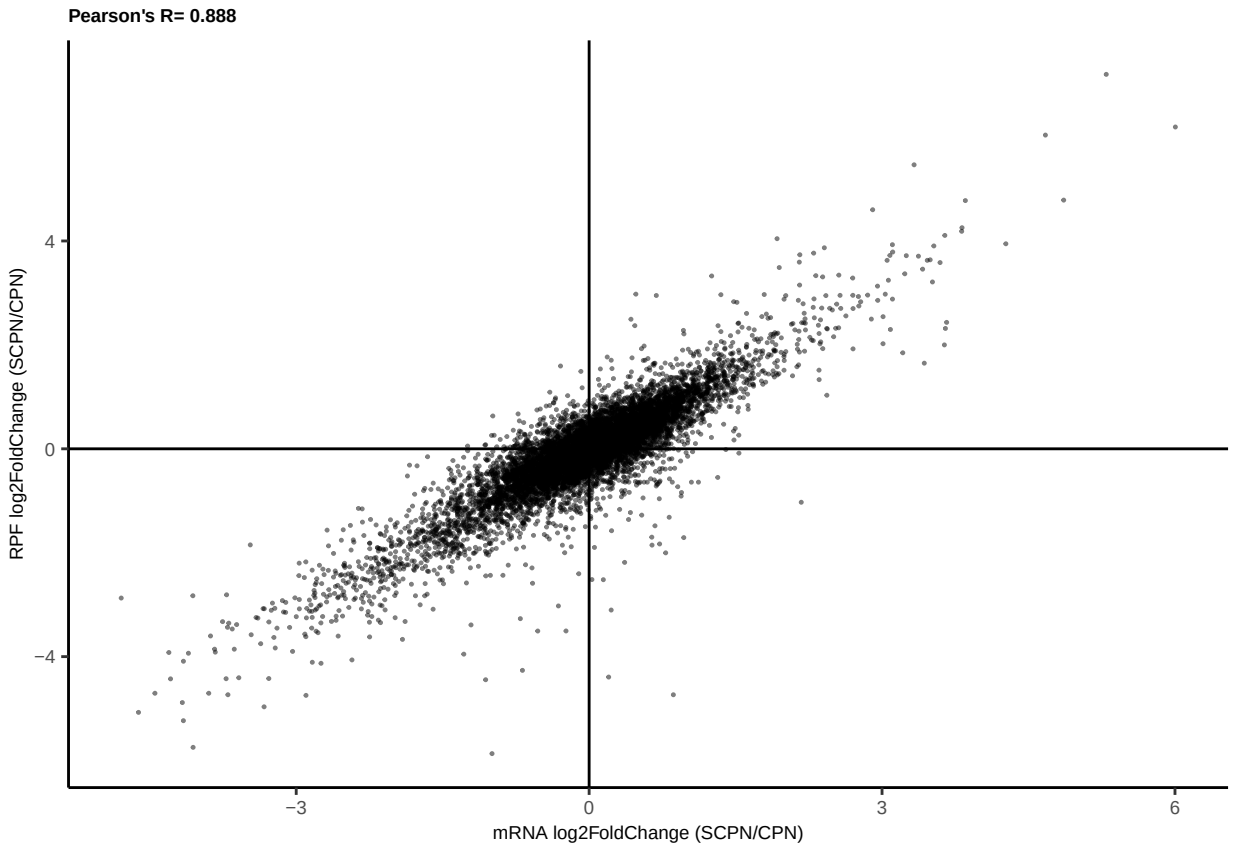


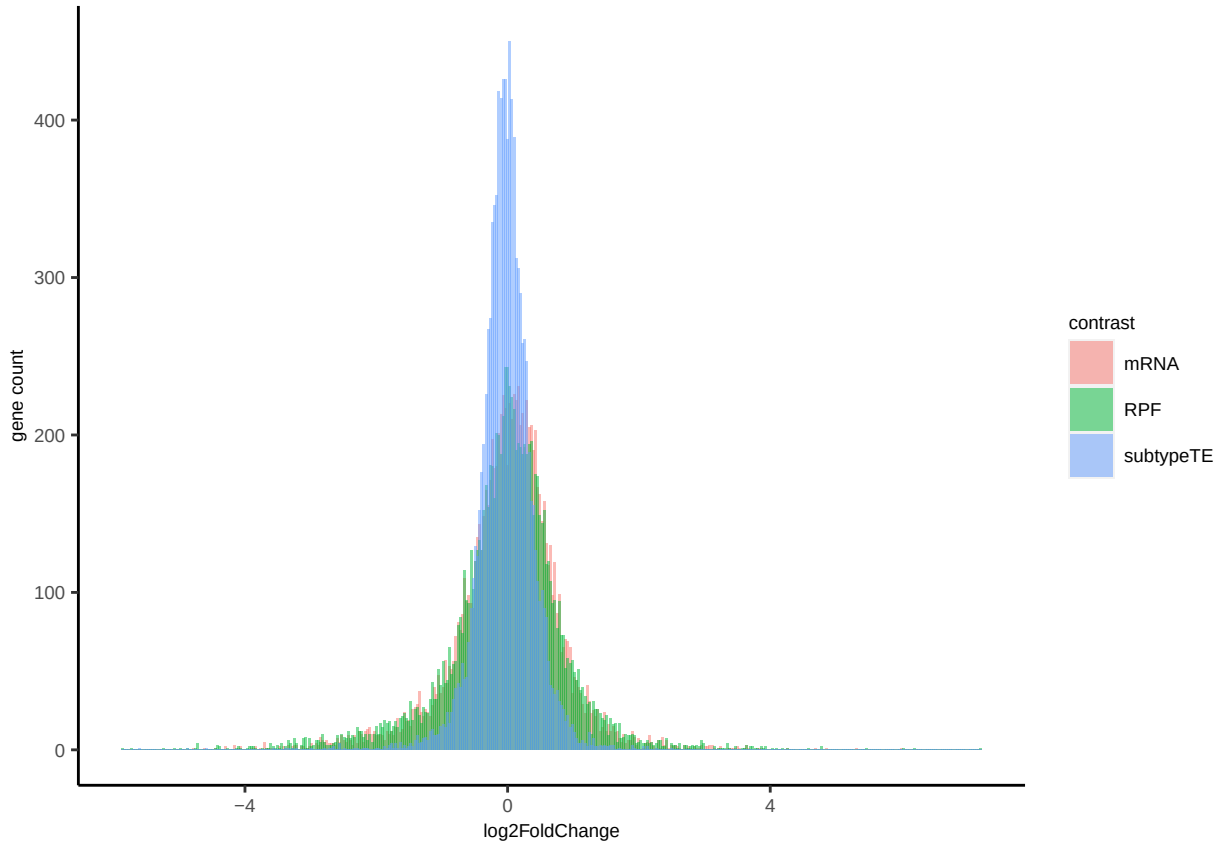
###Make the counts matrix for DESeq2.

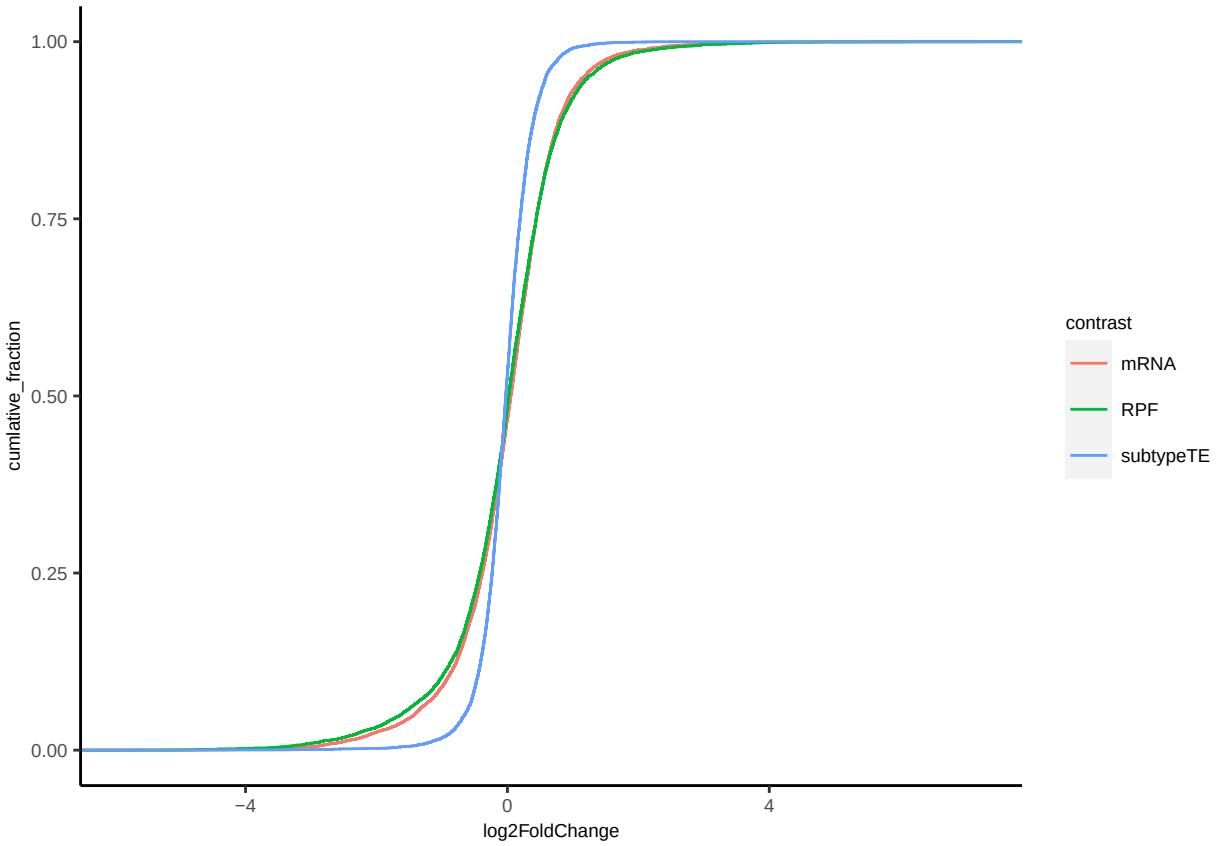
###Run DESeq2, including an exp:subtype interaction term. The genes with this ###interaction term significant are the ones differentially translated. ###However, this includes exclusively translated genes, and buffered genes. ###Need to run DESeq2 separately on the AF (RNA) and RP (Ribo-Seq) data sets ###To be able to parse out translational regulation from buffering. ###Update 4/21/21: running one full model: ~exp+subtype+exp:subtype ###The main effect of exp is translation efficiency (TE) in CPN ###The main effect of exp+exp:subtype interaction is SCPN TE ###The main effect of subtype is the RP/AF ratio in CPN ###The main effect of subtype+exp:subtype interaction is RP/AF in SCPN ###Awkwardly renaming the results objects to match previous names ###Using genes that are significant in the subtype main effect as "RNA" ###and genes significant in the exp comparison as "Ribo". Significant in both ###is "RNA+Ribo" ###Write out the Geneids and log2FC for pre-ranked GSEA analysis

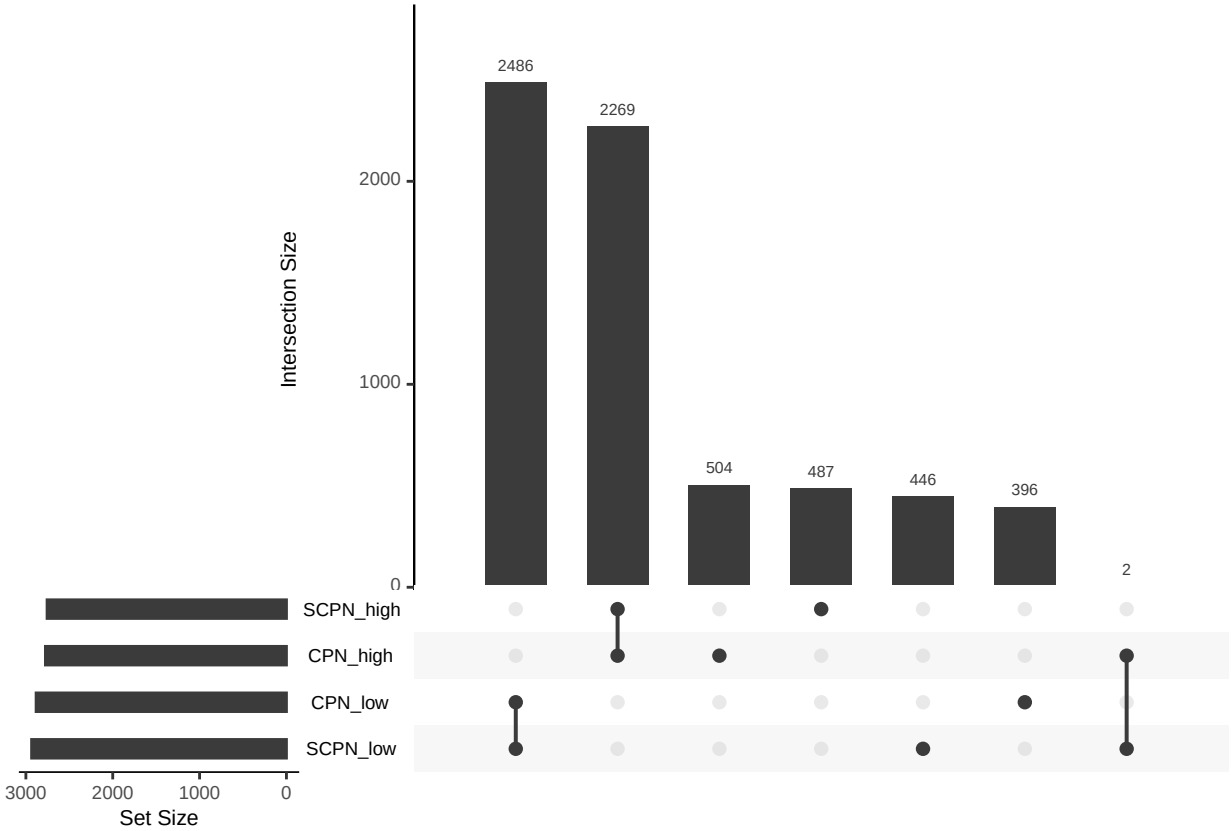
```
## Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in
## design formula are characters, converting to factors
```

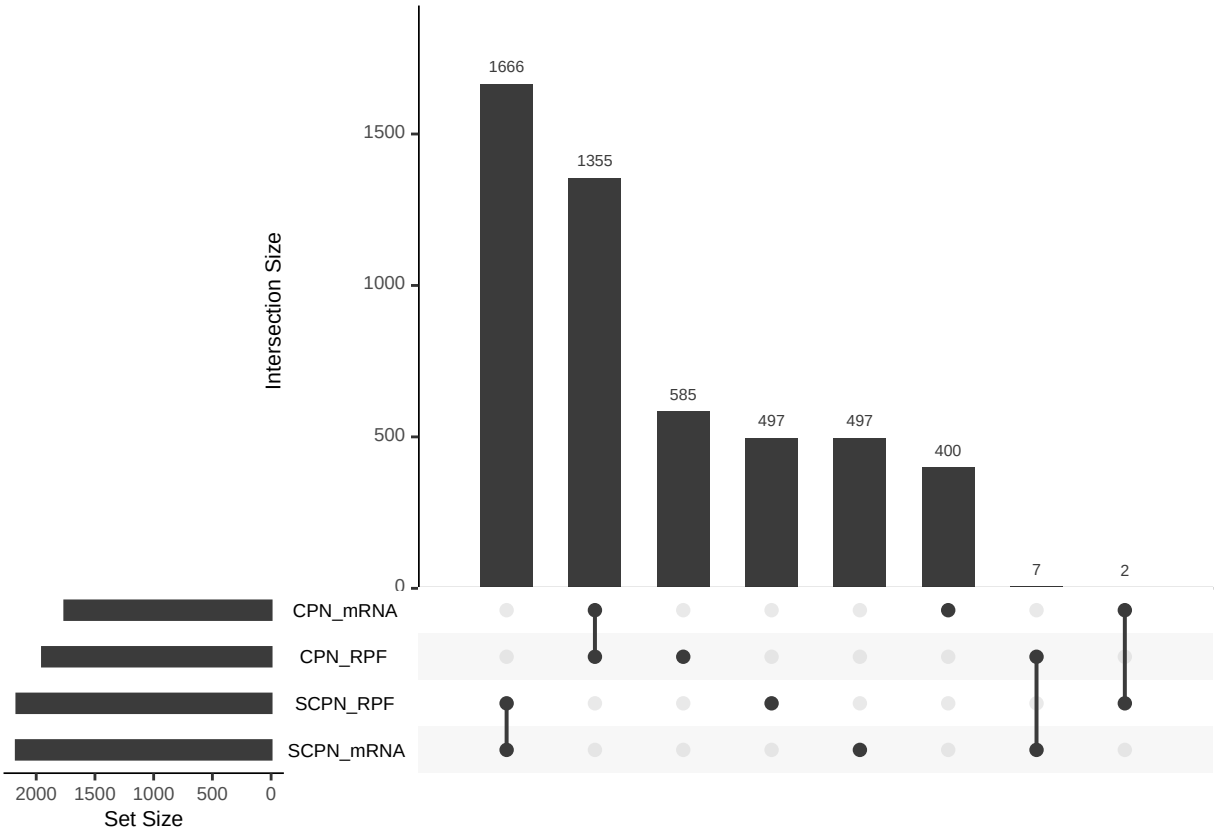
```
## Warning: Ignoring unknown parameters: aes
## Ignoring unknown parameters: aes
## Ignoring unknown parameters: aes
```

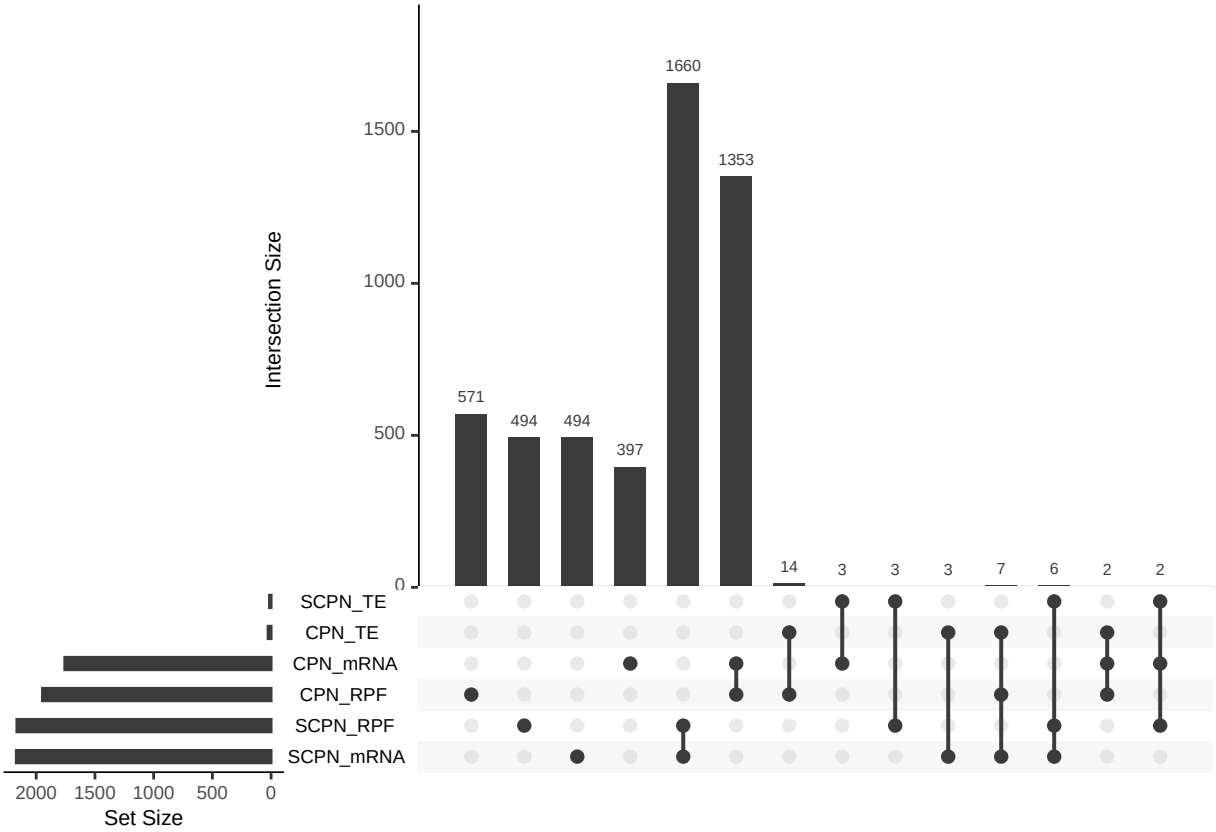

Pearson's R= 0.888
RPF log2FoldChange (SCPN/CPN)
4
0
−4
−3
0
3
6
mRNA log2FoldChange (SCPN/CPN)


gene count
400
300
200
100
0
−4
0
4
log2FoldChange
contrast
mRNA
RPF
subtypeTE


1.00
0.75
0.50
0.25
0.00
cumlative_fraction
−4
0
4
log2FoldChange
contrast
mRNA
RPF
subtypeTE


2486
2269
504
487
446
396
2
2000
1000
0
Intersection Size
SCPN_high
CPN_high
CPN_low
SCPN_low
3000
2000
1000
0
Set Size






```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

```
## pdf
##   2
```