HSCR QC and Filtering

Liz Vincent

September 6, 2017

Import Data

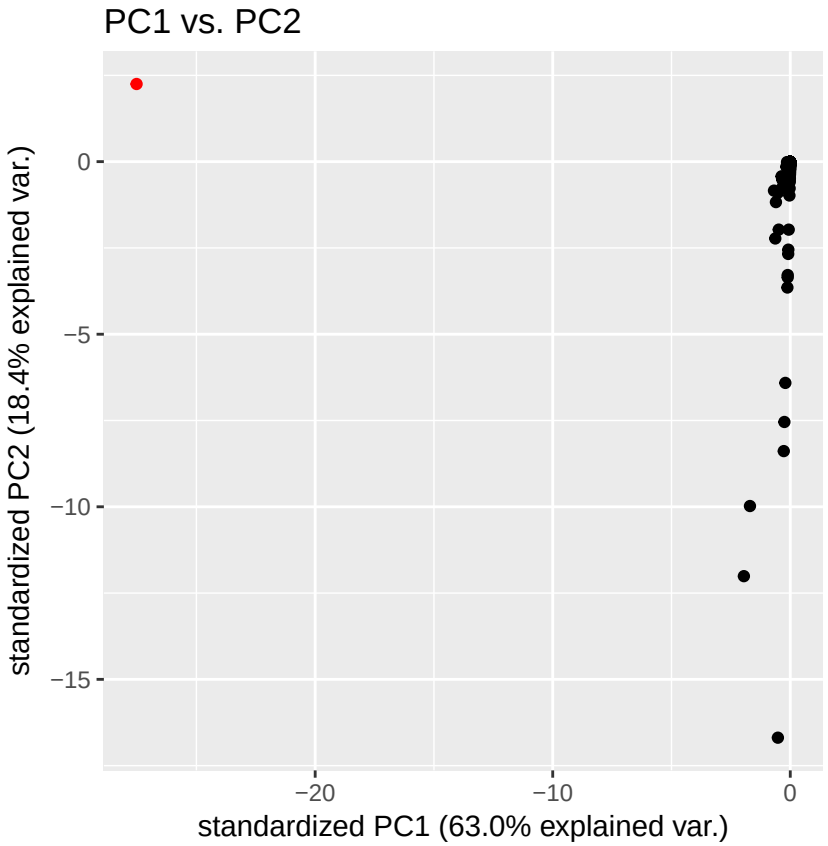
Data from Cuffnorm run on all 768 cells

```
# Gene FPKMs
fpkms<-read.delim("genes.fpkm_table",row.names=1,stringsAsFactors = F)
colnames(fpkms)<-unlist(strsplit(as.character(colnames(fpkms)),"_0$"))

# Isoform FPKMs
isoform_fpkms<-read.delim("isoforms.fpkm_table",row.names=1,stringsAsFactors = F)
colnames(isoform_fpkms)<-unlist(strsplit(as.character(colnames(isoform_fpkms)),"_0$"))

# Sample Annotation
sample_ann<-read.delim("samples.table",row.names=1,stringsAsFactors = F)
rownames(sample_ann)<-unlist(strsplit(as.character(rownames(sample_ann)),"_0$"))
master_cell_sheet<-read.delim("sample_info.txt",stringsAsFactors=F,row.names=1)
sample_info<-merge(sample_ann,master_cell_sheet,by='row.names')
rownames(sample_info)<-sample_info[,1]
sample_info<-sample_info[,-1]

# Gene Annotation
gene_ann<-read.delim("genes.attr_table",row.names=1,stringsAsFactors = F)

fd<-new("AnnotatedDataFrame",data=gene_ann)
pd<-new("AnnotatedDataFrame",data=sample_info)

# Create cell data set object
dat.relative <- newCellDataSet(cellData=as.matrix(fpkms),
                       phenoData=pd,
                       featureData=fd)
```

Remove Outliers

Iteratively run PCA and manually remove outliers

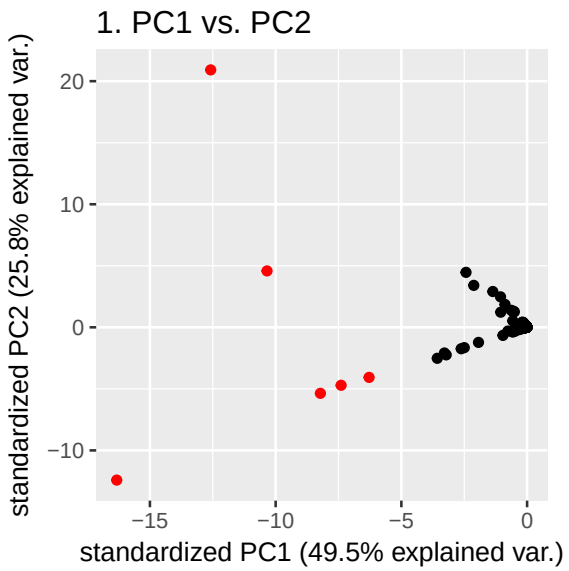
```
# PCA on FPKM values
dat.relative.pca<-prcomp(t(exprs(dat.relative)),scale=F,center=F)
```

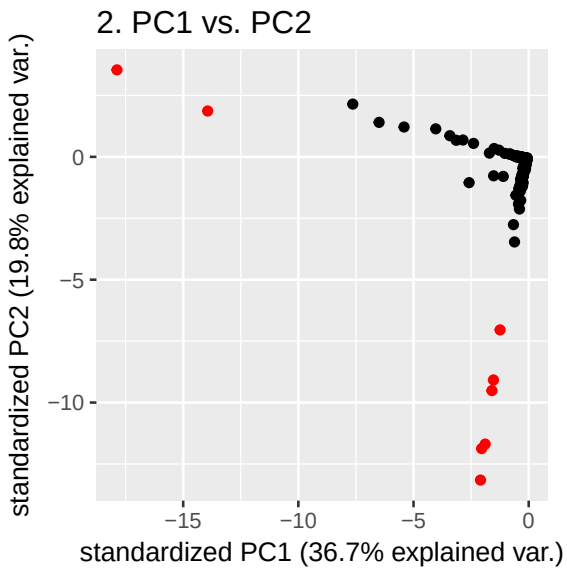


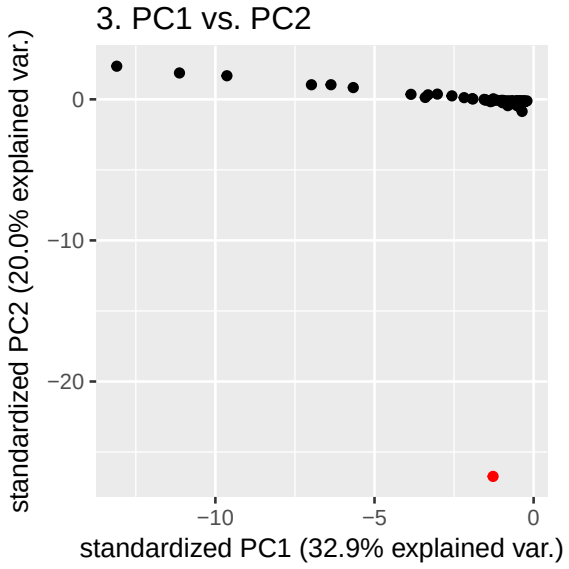
Remove PC 1 and 2 outliers and rerun PCA

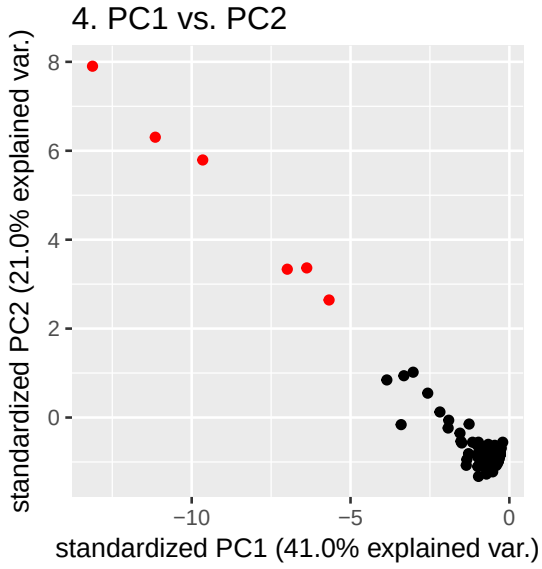
```
remove<-names(which(dat.relative.pca$x[,1] < -2e+06))
dat.relative.filtered<-dat.relative[,!(row.names(pData(dat.relative)) %in% remove)]

dat.relative.filtered.pca<-prcomp(t(exprs(dat.relative.filtered)),scale=F,center=F)
```

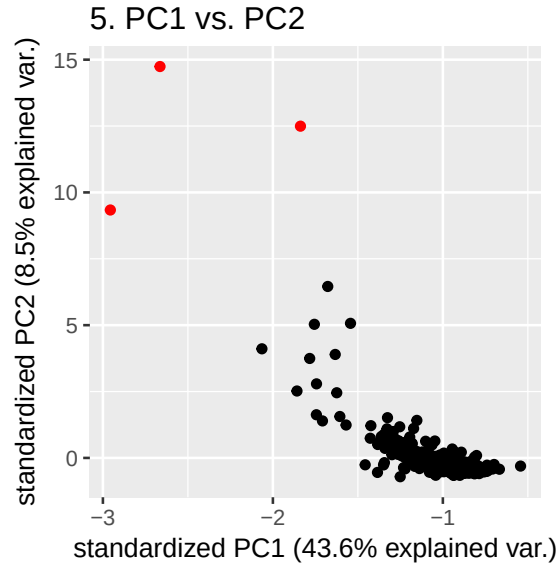
Repeat iteratively until there are no obvious outliers

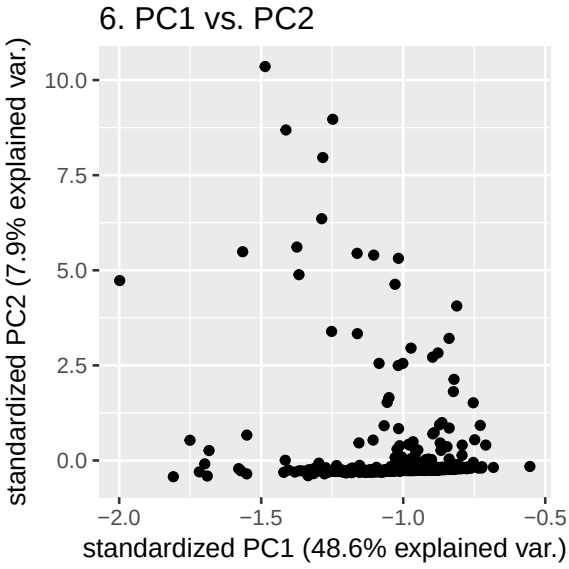

1. PC1 vs. PC2
standardized PC2 (25.8% explained var.)
standardized PC1 (49.5% explained var.)


2. PC1 vs. PC2
standardized PC2 (19.8% explained var.)
standardized PC1 (36.7% explained var.)


3. PC1 vs. PC2
standardized PC2 (20.0% explained var.)
standardized PC1 (32.9% explained var.)


4. PC1 vs. PC2
standardized PC2 (21.0% explained var.)
standardized PC1 (41.0% explained var.)


5. PC1 vs. PC2
standardized PC2 (8.5% explained var.)
standardized PC1 (43.6% explained var.)


6. PC1 vs. PC2
standardized PC2 (7.9% explained var.)
standardized PC1 (48.6% explained var.)

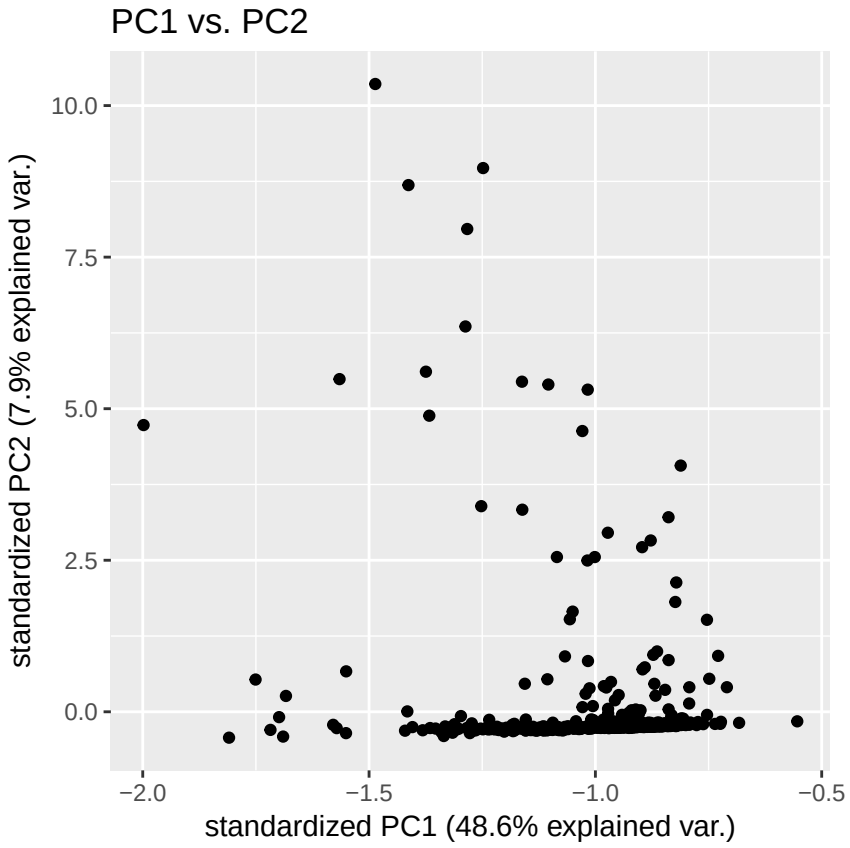
Rerun Cuffnorm on remaining 743 cells and import data

```
# Gene FPKMs
fpkms<-read.delim("outliers_removed_genes.fpkm_table",row.names=1,stringsAsFactors = F)
colnames(fpkms)<-unlist(strsplit(as.character(colnames(fpkms)),"_0$"))

# Isoform FPKMs
isoform_fpkms<-read.delim("outliers_removed_isoforms.fpkm_table",row.names=1,stringsAsFactors = F)
colnames(isoform_fpkms)<-unlist(strsplit(as.character(colnames(isoform_fpkms)),"_0$"))

# Sample Annotation
sample_ann<-read.delim("outliers_removed_samples.table",row.names=1,stringsAsFactors = F)
rownames(sample_ann)<-unlist(strsplit(as.character(rownames(sample_ann)),"_0$"))
sample_info<-merge(sample_ann,master_cell_sheet,by='row.names')
rownames(sample_info)<-sample_info[,1]
sample_info<-sample_info[,-1]

# Gene Annotation
gene_ann<-read.delim("outliers_removed_genes.attr_table",row.names=1,stringsAsFactors = F)

fd<-new("AnnotatedDataFrame",data=gene_ann)
pd<-new("AnnotatedDataFrame",data=sample_info)

# Create cell data set object
dat.relative.743 <- newCellDataSet(cellData=as.matrix(fpkms),
                       phenoData=pd,
                       featureData=fd)
```

Remove Outliers

Run PCA on second round of cuffnorm data

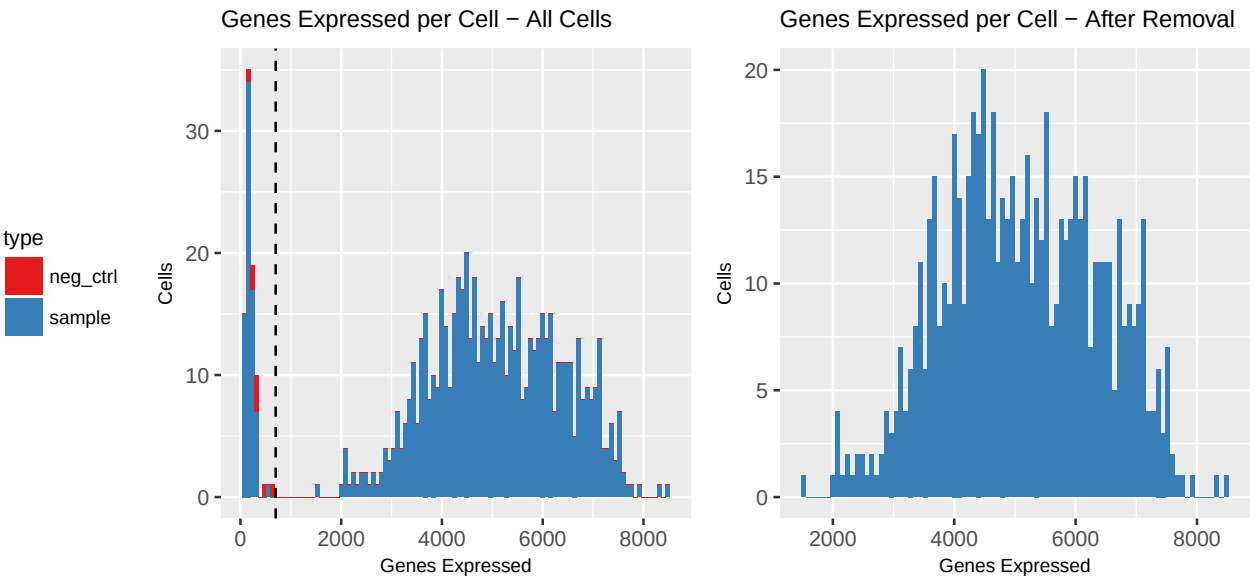
```
# PCA on FPKM values
dat.relative.pca<-prcomp(t(exprs(dat.relative.743)),scale=F,center=F)
```



Remove cells expressing very few genes

Remove cells with fewer detected genes than the negative control wells.

```
dat.relative.743<-detectGenes(dat.relative.743,min_expr=0.000001)
# Detect any non-zero genes

remove<-rownames(pData(dat.relative.743)[pData(dat.relative.743)$num_genes_expressed < 700,])
dat.relative.filtered<-dat.relative.743[,!(row.names(pData(dat.relative.743)) %in% remove)]
```

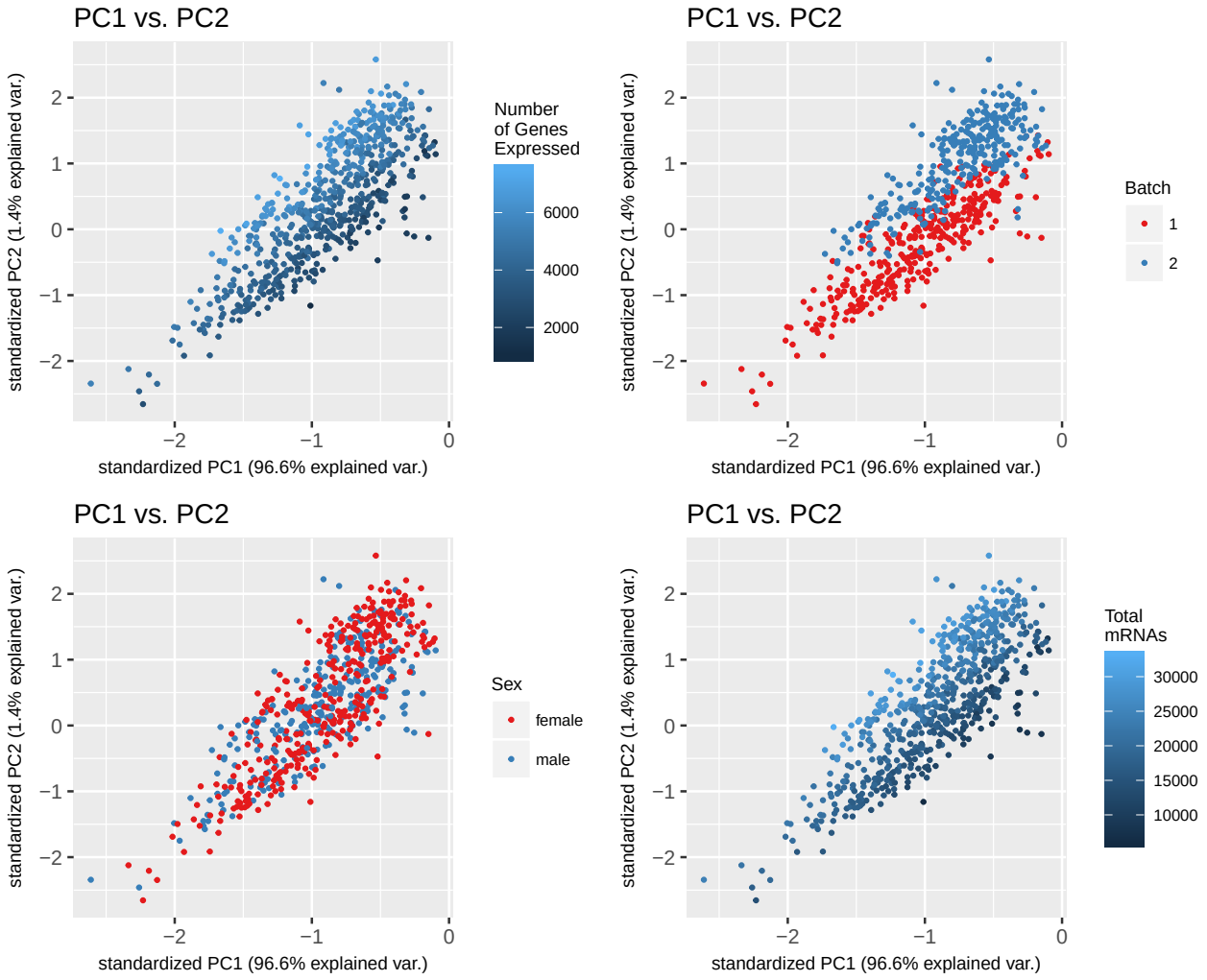


Convert FPKM to CPC

```
isoform_t_estimate<-estimate_t(isoform_fpkms)

fpkm_matrix_adj<-relative2abs(dat.relative.filtered,cores=detectCores()-1,t_estimate = isoform_t_estima

# Create new cell data set with CPC values
dat.filtered <- newCellDataSet(as.matrix(fpkm_matrix_adj),
                       phenoData = pd[rownames(pd) %in% colnames(fpkm_matrix_adj)],
                       featureData=fd,
                       expressionFamily=negbinomial.size(),
                       lowerDetectionLimit=1)

#Add and format metadata
pData(dat.filtered)$Total_mRNAs <- colSums(round(exprs(dat.filtered)))
pData(dat.filtered)$mean_expr<-esApply(dat.filtered,2,function(x){mean(x)})
pData(dat.filtered)$sd_expr<-esApply(dat.filtered,2,function(x){sd(x)})
pData(dat.filtered)$genotype<-factor(pData(dat.filtered)$genotype)
pData(dat.filtered)$sex<-factor(pData(dat.filtered)$sex)
pData(dat.filtered)$batch<-factor(pData(dat.filtered)$batch)

dat.filtered<-detectGenes(dat.filtered,min_expr=0.1)
dat.filtered@dim_reduce_type<-"DDRTree"
dat.filtered@auxOrderingData<-new.env()

fData(dat.filtered)$gene_id<-rownames(fData(dat.filtered))
#Otherwise gene_id is a factor, now it's a character
fData(dat.filtered)$mean_expr<-esApply(dat.filtered,1,function(x){mean(x)})
fData(dat.filtered)$sd_expr<-esApply(dat.filtered,1,function(x){sd(x)})
fData(dat.filtered)$bcv<-(fData(dat.filtered)$sd_expr/fData(dat.filtered)$mean_expr)**2
fData(dat.filtered)$percent_detection<-
  (fData(dat.filtered)$num_cells_expressed/dim(dat.filtered)[2])*100
```

PCA on Cleaned Data

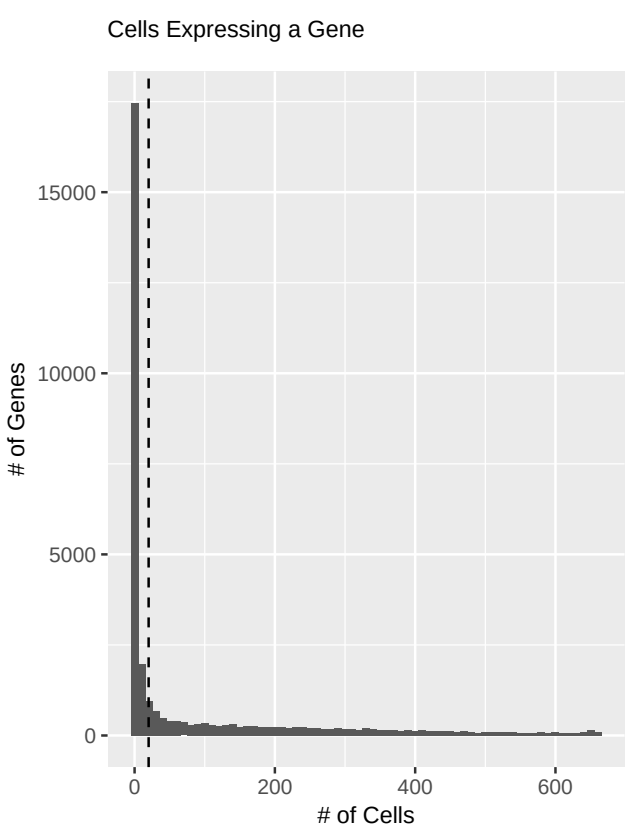
```
dat.filtered.pca<-prcomp(t(exprs(dat.filtered)),scale=F,center=F)
```

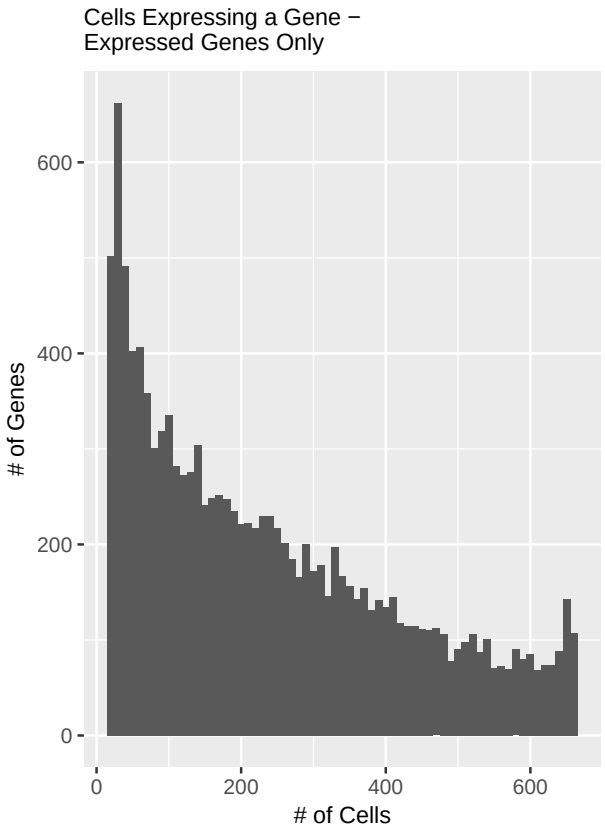


Determine Expressed Genes

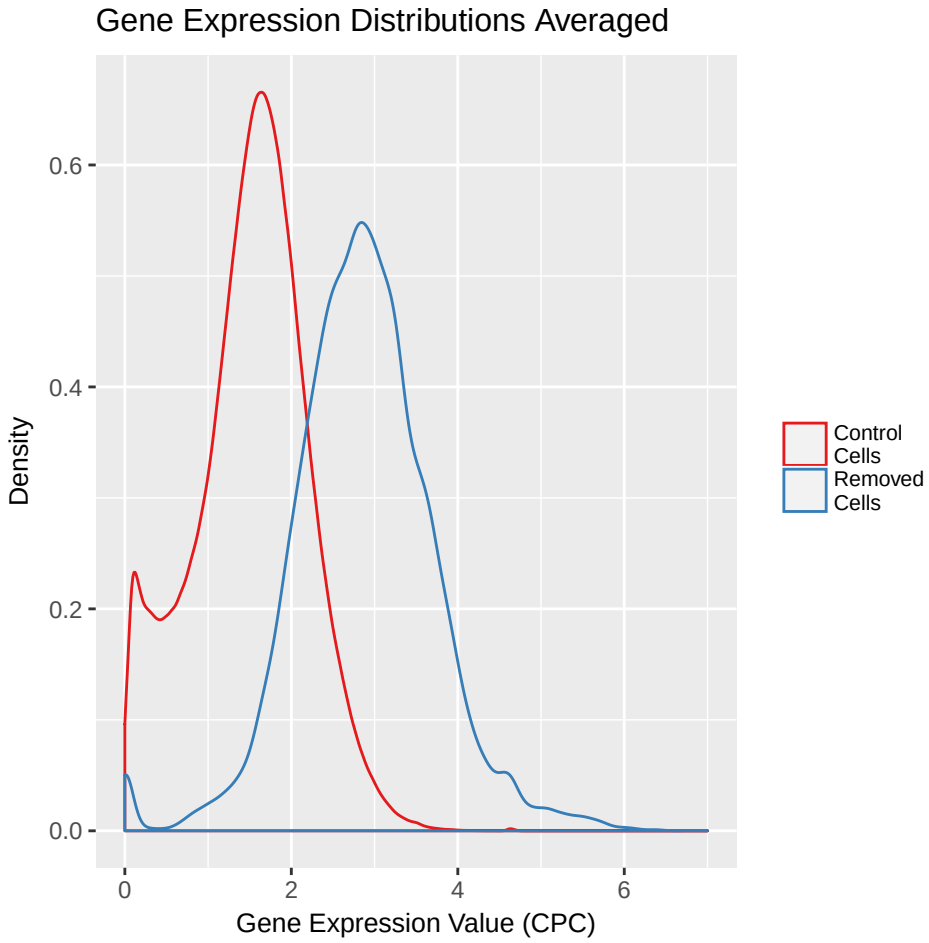
Determine which genes are expressed in a large enough number of cells to be meaningful. A gene is considered expressed if it is detected at a minimum value of 0.000001 FPKM in at least 20 cells (approximately 10% of cells in one condition), with a mean expression level of 0.01 CPC. 12,470 genes are expressed according to this criteria.

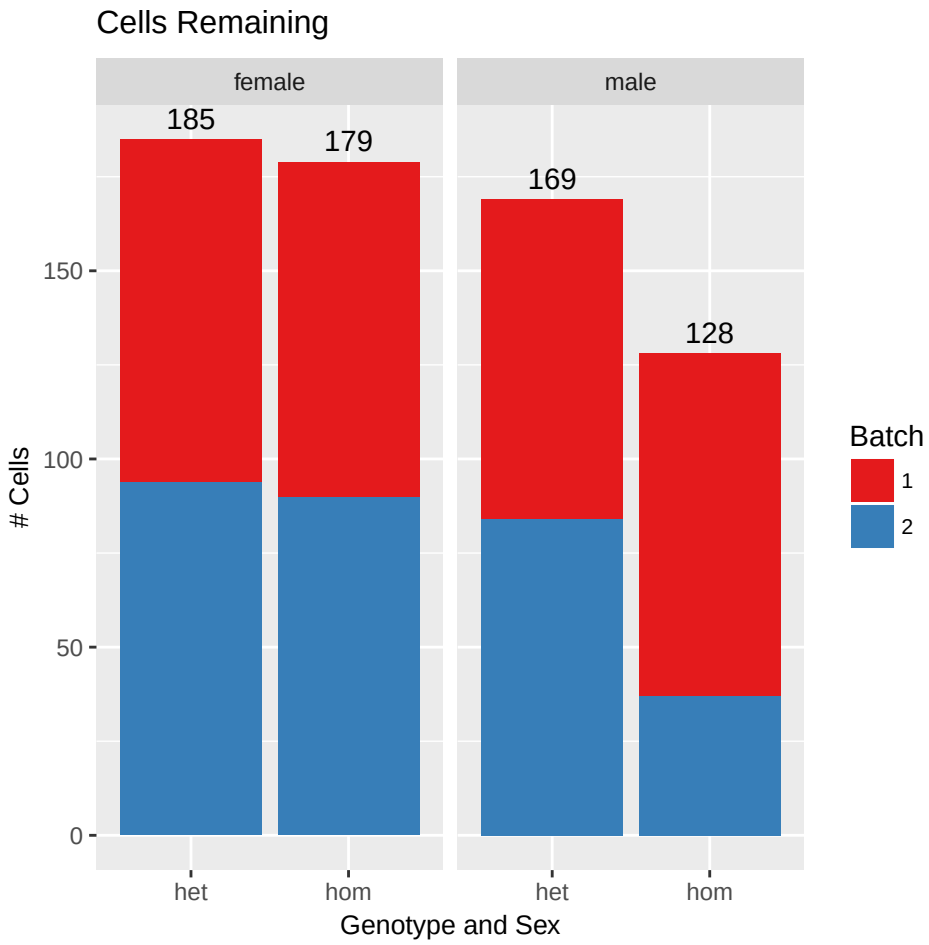
```
expressed_genes<-rownames(fData(dat.filtered)[fData(dat.filtered)$num_cells_expressed >= 20 & fData(dat
#12,470 genes
```


Cells Expressing a Gene
15000
10000
5000
0
# of Genes
0
200
400
600
# of Cells


Cells Expressing a Gene −
Expressed Genes Only
600
400
200
0
# of Genes
0
200
400
600
# of Cells

Plots of Cleaned Cell Data Set





Session Info

```
## R version 3.4.1 (2017-06-30)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
##  [1] grid      splines   stats4    parallel  stats     graphics  grDevices
##  [8] utils     datasets  methods   base
##
## other attached packages:
##  [1] Hmisc_4.0-3         Formula_1.2-2       survival_2.41-3
##  [4] lattice_0.20-35     pheatmap_1.0.8      mclust_5.3
##  [7] corrplot_0.77       slackr_1.4.2        ggbiplot_0.55
## [10] scales_0.5.0        plyr_1.8.4          gridExtra_2.2.1
```

```
## [13] reshape2_1.4.2      stringi_1.1.5       stringr_1.2.0
## [16] tsne_0.1-3          monocle_2.4.0       DDRTree_0.1.5
## [19] irlba_2.2.1         VGAM_1.0-4          ggplot2_2.2.1
## [22] Biobase_2.36.2      BiocGenerics_0.22.0 Matrix_1.2-11
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.12           assertthat_0.2.0       rprojroot_1.2
##  [4] digest_0.6.12          slam_0.1-40            R6_2.2.2
##  [7] backports_1.1.0        acepack_1.4.1          qlcMatrix_0.9.5
## [10] evaluate_0.10.1        httr_1.3.1             rlang_0.1.2
## [13] lazyeval_0.2.0         data.table_1.10.4      rpart_4.1-11
## [16] combinat_0.0-8         checkmate_1.8.3        rmarkdown_1.6
## [19] labeling_0.3           Rtsne_0.13             foreign_0.8-69
## [22] htmlwidgets_0.9        igraph_1.1.2           munsell_0.4.3
## [25] compiler_3.4.1         pkgconfig_2.0.1        base64enc_0.1-3
## [28] htmltools_0.3.6        nnet_7.3-12            htmlTable_1.9
## [31] tibble_1.3.4           matrixStats_0.52.2     dplyr_0.7.2
## [34] densityClust_0.2.1     jsonlite_1.5           gtable_0.2.0
## [37] magrittr_1.5           bindrcpp_0.2           limma_3.32.5
## [40] latticeExtra_0.6-28    fastICA_1.2-1          RColorBrewer_1.1-2
## [43] tools_3.4.1            glue_1.1.1             HSMMSingleCell_0.110.0
## [46] yaml_2.1.14            colorspace_1.3-2       cluster_2.0.6
## [49] knitr_1.17             bindr_0.1
```