RNA-seq LCL_nodup

Sahar Mozaffari

2/7/2018

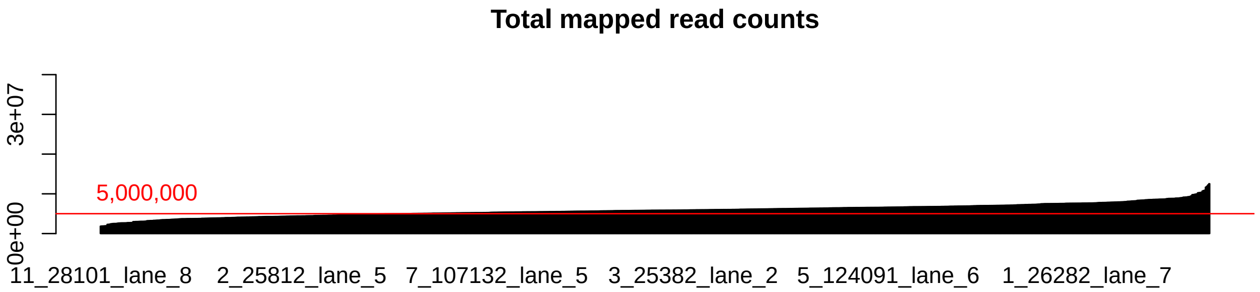
R Markdown for RNA-seq data

Genecount matrix:

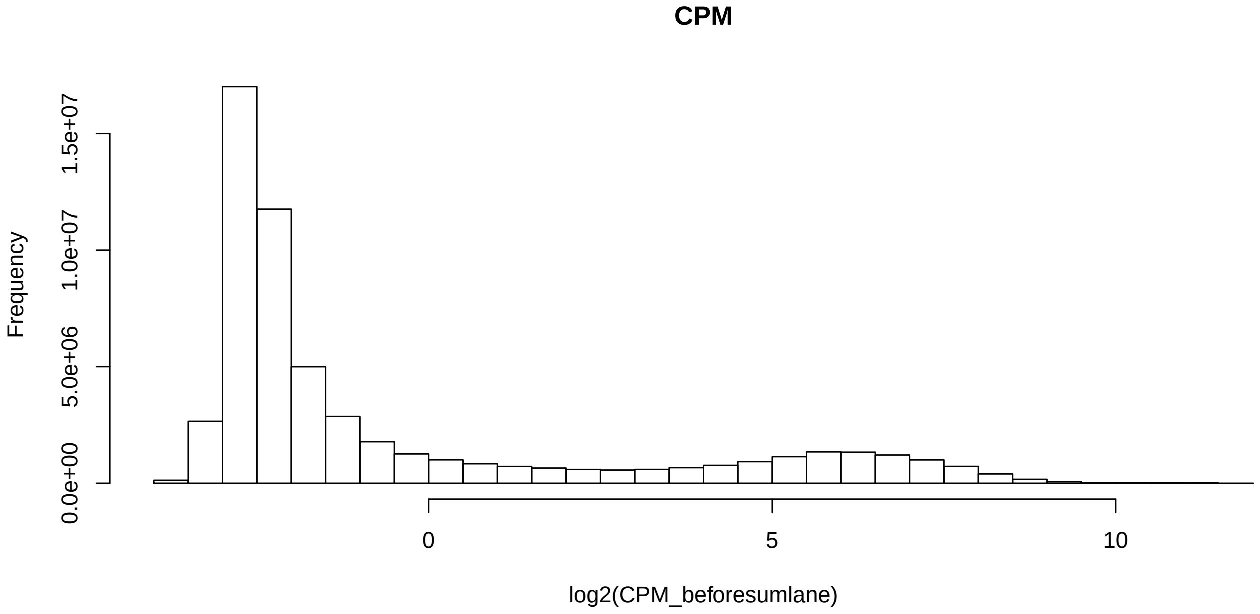
- genes in rows, individuals/samples by lane and flowcell in columns
- There are a total of 5.7819×10^4 genes and 989 samples

verifyBAMid found some sample swaps:

Number of lanes with enough reads, before combining replicates



- The distribution of Counts Per Million:



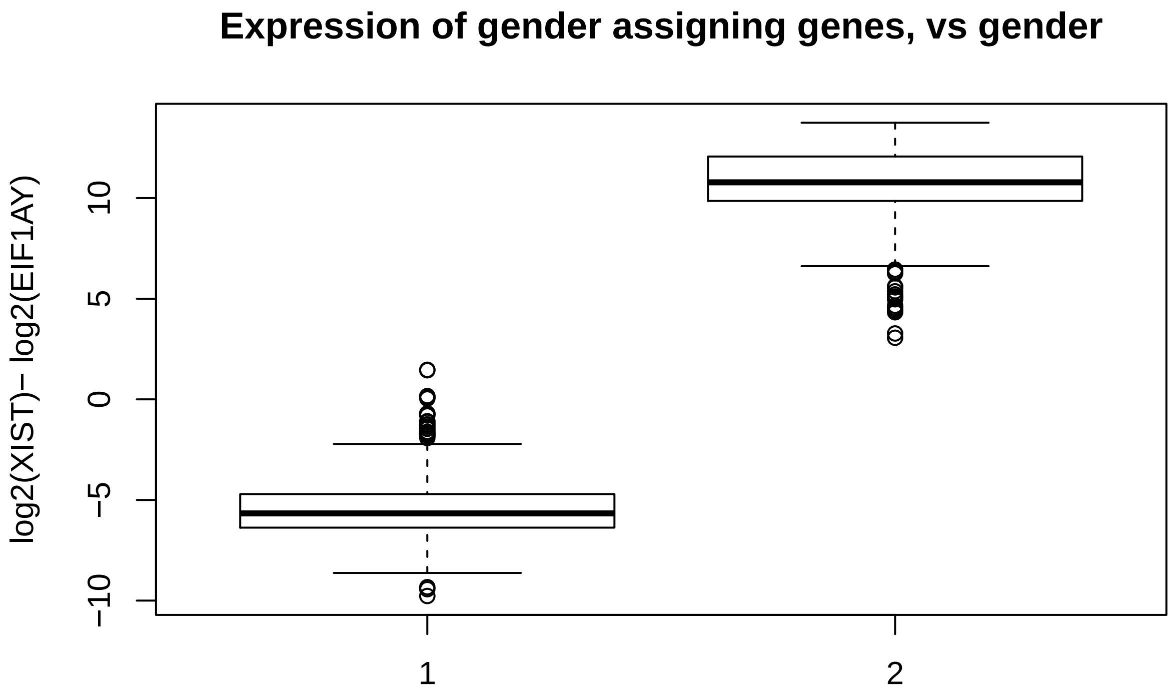
Sexcheck

Sex assigned by ratio of XIST to EIF1AY gene

```
## callSex
##   F   M
## 525 464
```

- According to expression of sex genes, there are 525 females and 464 males.

```
## gender
##   1   2
## 470 519
```



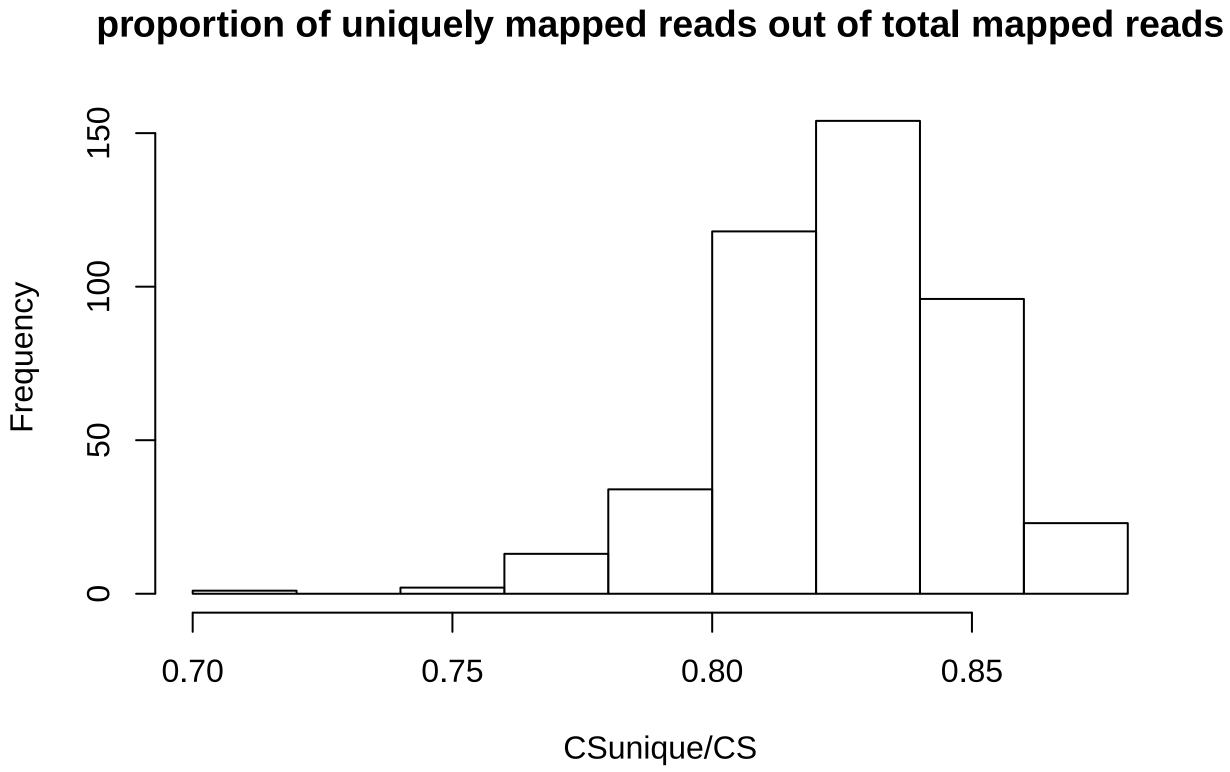
*

There are supposed to be: 0 females and 0 males. * The samples misassigned are: 1.4444432, 1.467126, 0.1632929, 0.1272232, 0.042

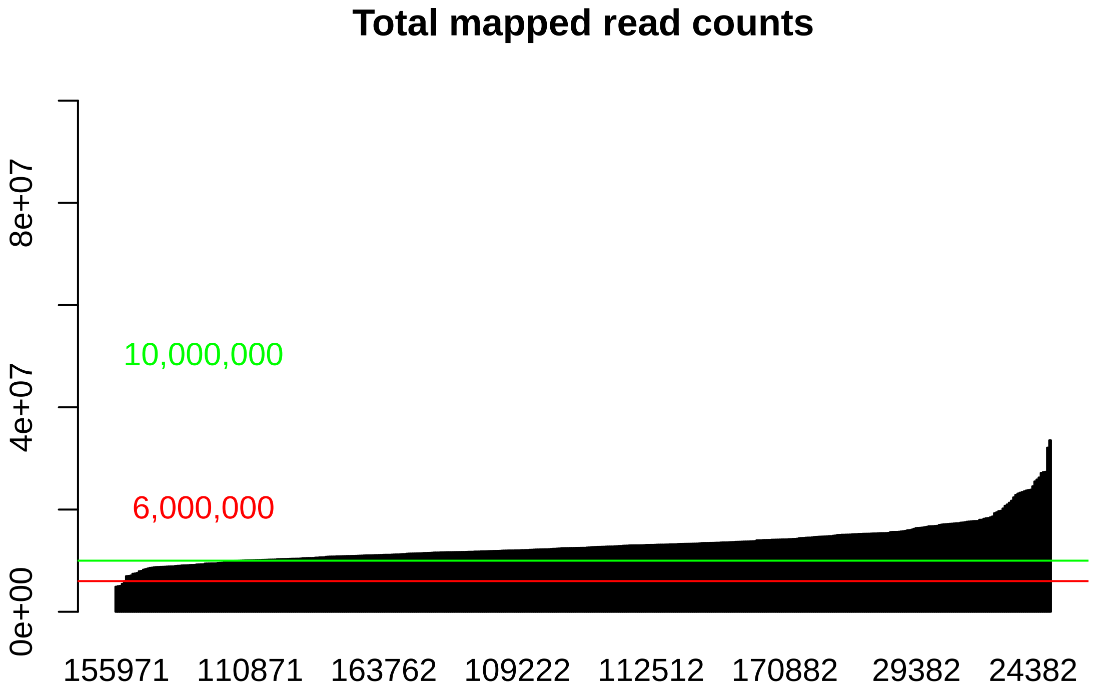
Combining technical replicates

- gene count matrix combined across lanes/flowcells so that each individual has one sum value of gene expression for each gene

```
##                      100092  100172  100182  100202  100372
## N_unmapped               31      40      17      23      27
## N_multimapping           21      35      20      18      17
## N_noFeature         1629626 3175667 1815864 1515082 1247025
## N_ambiguous          711964  873262  610006  683836  598228
## ENSG00000223972.4         0       0       0       0       0
```

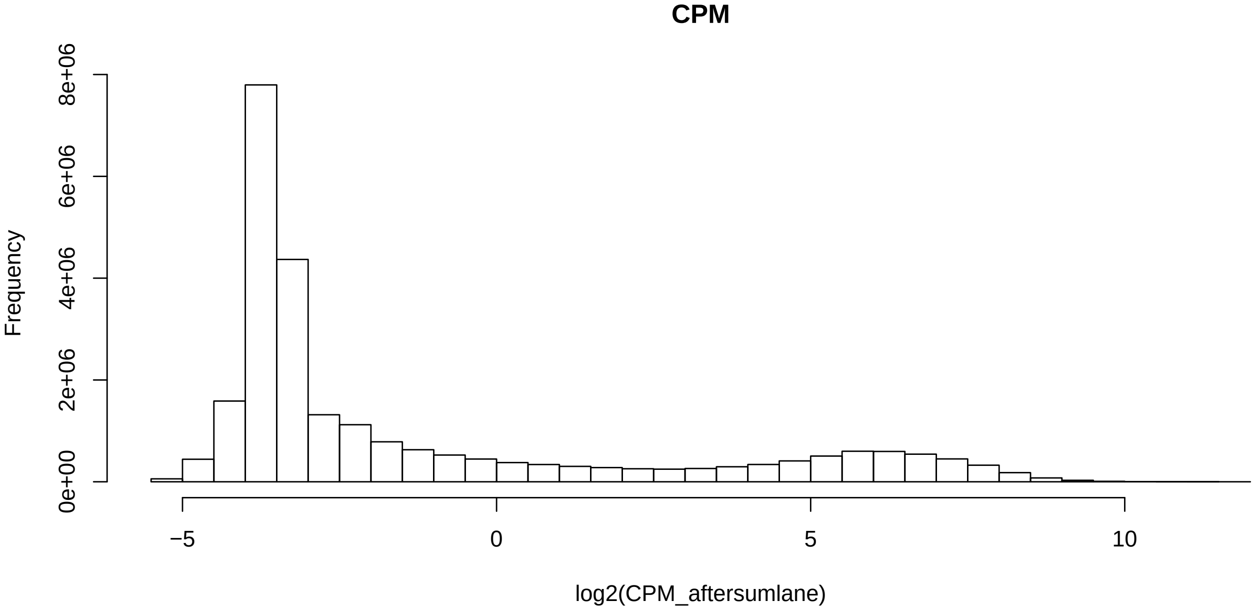


```
##                    100092 100172 100182 100202 100372
## ENSG00000223972.4       0      0      0      0      0
## ENSG00000227232.4       3     20     23     11      5
## ENSG00000243485.2       0      0      0      0      0
## ENSG00000237613.2       0      0      0      0      0
## ENSG00000268020.2       0      0      0      0      0
```



- combine total number of read covariate value

- After combining replicates, 106052, 106411, 106621, 107032, 107122, 107522, 108052, 108821, 110052, 110342, 110472, 110961 out of total 441 have more than 10 million reads
- The distribution of Counts Per Million after combining replicates:

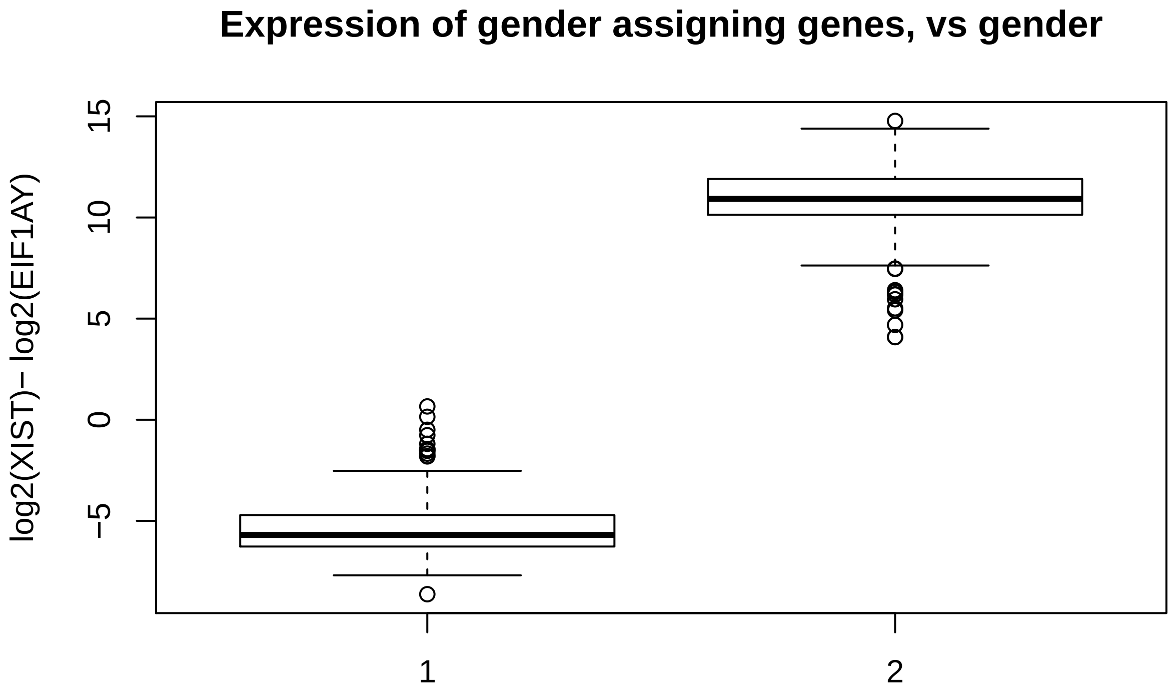


Checking sex after combining replicates

```
## 100092 100172 100182 100202 100372
##  15346  24888    128  13110   9678
```

```
## callSex
##   F   M
## 232 209
```

```
## gender
##   1   2
## 211 230
```



```
##    160581    171351
## 0.1451077 0.6587376
```

```
## character(0)
```

```
## named integer(0)
```

- There are supposed to be: 232 females and 209 males.
- The samples misassigned are: 160581, 171351

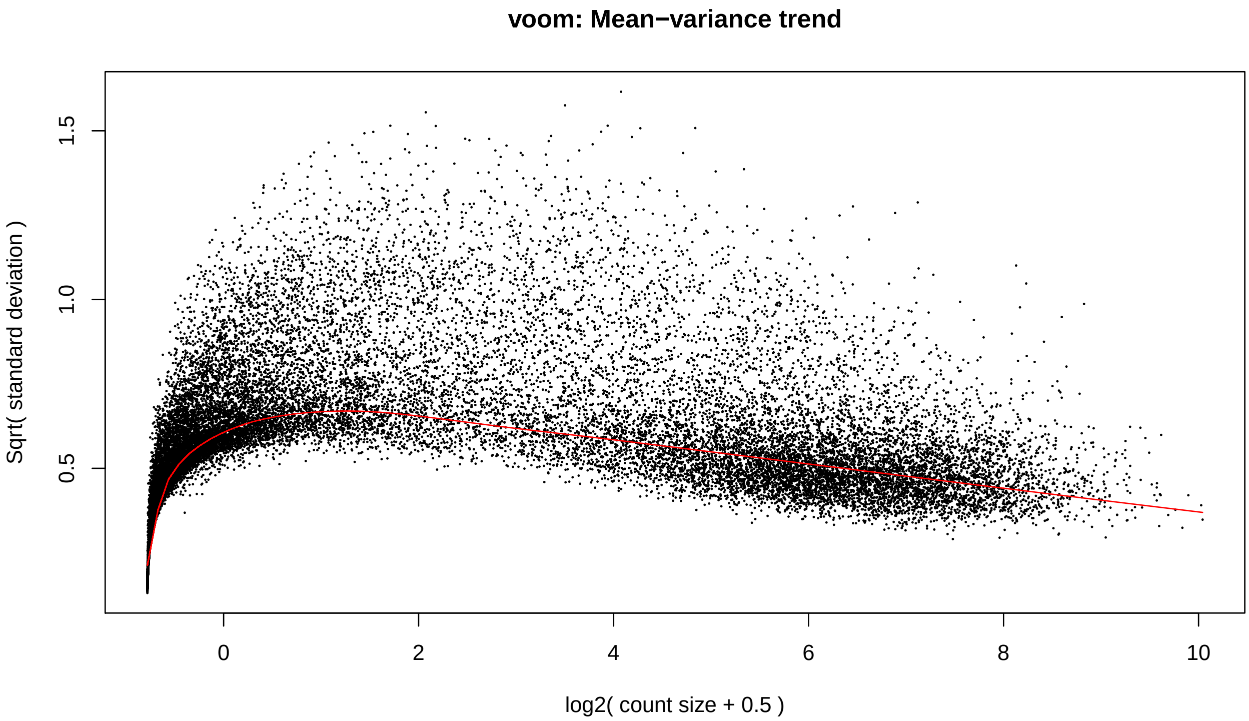
These 2 individuals have wrong assigned sex- last 0 have quite a large error - remove from data

Removing X and Y chromosome (and mitochondrial) genes –(and genes not expressed in anyone)–

- Total number of chromosome X genes: 2392, Y genes: 495, and mt genes: 37
- Number in data that are removed:
- X chromosome genes: 2392
- Y genes: 495
- mt genes: 37

Analysis after combining replicates

Normalization mean-variance trend looks strange because I didn't remove lowly expressed genes. This is of the expression before combining lanes - but removing those with few reads and who didn't pass sex check.



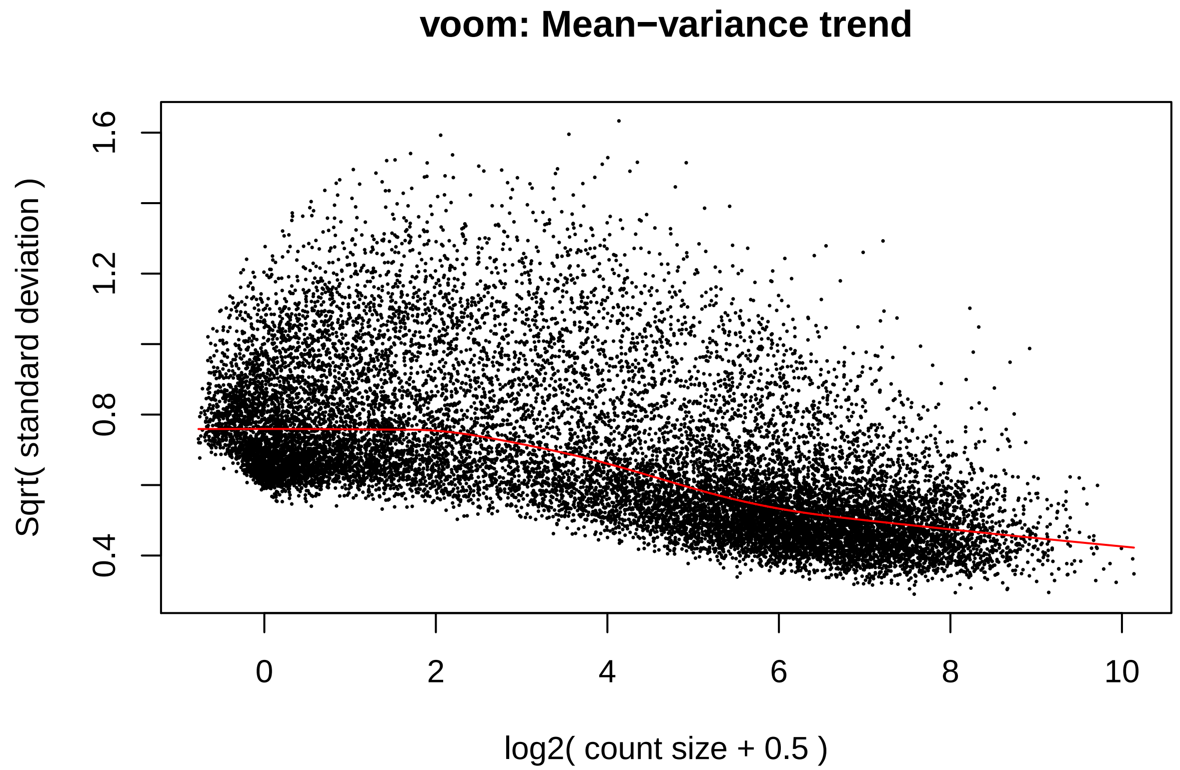
```
cpm <- cpm(aftersumlane.y.x)
lcpm <- cpm(aftersumlane.y.x, log=TRUE)
table(rowSums(aftersumlane.y.x==0)==441)
```

```
##
## FALSE
## 39879
```

```
keep.exprs <- rowSums(cpm>1)>=20
aftersumlane.y.x.nolowexpressed<- aftersumlane.y.x[keep.exprs, ]
dim(aftersumlane.y.x.nolowexpressed)
```

```
## [1] 17167   441
```

```
voom.after.CPM_aftersumlane.y.x.nolowexpressed <- voom(cpm(aftersumlane.y.x.nolowexpressed),plot=TRUE, 
```


voom: Mean−variance trend

```
#dge <- DGEList(counts=aftersumlane.y.x.nolowexpressed)
#dge <- calcNormFactors(dge)
#logCPM <- cpm(dge, log=TRUE, prior.count=1)
#x <- DGEList(counts=aftersumlane.y.x.nolowexpressed)
```

Covariates:

```
## Warning in cbind(findivs, flowlane): number of rows of result is not a
## multiple of vector length (arg 1)
```

```
## Warning in cbind(uflowcells, c(1:98)): number of rows of result is not a
## multiple of vector length (arg 2)
```

```
##        sex indiv rnaconc  rin batch prep  conc length flowlane index
## 100092   2     1   965.0  9.8     7    1  9.15    284        1     5
## 100172   2     2   192.0  9.2     6    2 14.49    295       12     3
## 100182   2     3   173.0  9.2     4    2 10.43    282       23    11
## 100202   2     4   835.1  9.6     3    2  4.78    282       34     1
## 100372   2     5   588.0 10.0     5    2 11.50    290       45     3
## 100582   2     6   191.0  9.2     4    2  2.55    270       56    10
```

- RIN, Batch and RNA concentration were significant, so plot by first two PC's:

TMM Normalization

PCA:

- First PCA showing variation of Proportion of Variance in PCs and correlation with covariates:

```
##                               PC1        PC2       PC3       PC4       PC5
## Standard deviation     1185.09940 1071.81116 819.71250 805.68342 718.30733
## Proportion of Variance    0.16032    0.13113   0.07670   0.07410   0.05890
## Cumulative Proportion     0.16032    0.29145   0.36815   0.44225   0.50115
##                              PC6       PC7       PC8       PC9      PC10
## Standard deviation     576.04420 549.25046 499.68841 465.56665 434.68243
## Proportion of Variance   0.03788   0.03444   0.02850   0.02474   0.02157
## Cumulative Proportion    0.53903   0.57346   0.60197   0.62671   0.64828
##                             PC11      PC12      PC13      PC14      PC15
## Standard deviation     400.83696 391.31989 367.40241 320.63575 314.11655
## Proportion of Variance   0.01834   0.01748   0.01541   0.01174   0.01126
## Cumulative Proportion    0.66662   0.68410   0.69951   0.71124   0.72250
##                             PC16      PC17      PC18      PC19      PC20
## Standard deviation     304.13400 290.08741 265.56550 258.96503 258.02266
## Proportion of Variance   0.01056   0.00961   0.00805   0.00766   0.00760
## Cumulative Proportion    0.73306   0.74267   0.75072   0.75837   0.76597
```

```
##            sex        indiv     rnaconc          rin        batch
## PC1 0.86109325 2.382435e-02 0.030924497 2.199444e-18 0.0353469461
## PC2 0.07215488 7.052233e-01 0.647587853 5.346449e-06 0.3535922840
## PC3 0.60497425 6.913702e-05 0.087499925 9.385409e-03 0.0010519343
## PC4 0.26812084 9.127805e-01 0.000706660 9.565436e-02 0.0007579598
## PC5 0.14100100 3.553721e-02 0.005960488 5.762529e-03 0.0044752043
## PC6 0.22461646 8.904616e-01 0.007416836 3.939635e-04 0.0001933334
##           prep       conc      length   flowlane     index
## PC1 0.14115877 0.91859786 0.260253735 0.09745677 0.2018023
## PC2 0.08681012 0.60204783 0.004406093 0.67179214 0.3301767
## PC3 0.18244331 0.55532553 0.646298265 0.21931876 0.9850921
## PC4 0.12886478 0.18503214 0.147902308 0.75206863 0.9411107
## PC5 0.07735660 0.35425621 0.397196749 0.58212010 0.4147948
## PC6 0.05453146 0.00503713 0.266463803 0.91662854 0.4139972
```

Regress out RIN

- PCA for the second time:

```
##                               PC1        PC2       PC3       PC4       PC5
## Standard deviation     1121.99923 1018.66327 813.62752 799.69182 709.24825
## Proportion of Variance    0.14932    0.12308   0.07852   0.07585   0.05967
## Cumulative Proportion     0.14932    0.27240   0.35092   0.42677   0.48644
##                              PC6       PC7       PC8       PC9      PC10
## Standard deviation     565.88493 546.84783 499.42809 462.33572 430.44400
## Proportion of Variance   0.03798   0.03547   0.02959   0.02535   0.02198
## Cumulative Proportion    0.52442   0.55989   0.58948   0.61483   0.63681
##                             PC11      PC12      PC13      PC14      PC15
## Standard deviation     400.81346 389.42512 365.81415 317.84517 312.80428
## Proportion of Variance   0.01906   0.01799   0.01587   0.01198   0.01161
## Cumulative Proportion    0.65586   0.67385   0.68972   0.70171   0.71331
##                             PC16      PC17      PC18      PC19      PC20
## Standard deviation     304.10773 289.52238 263.89807 258.38708 250.31588
## Proportion of Variance   0.01097   0.00994   0.00826   0.00792   0.00743
## Cumulative Proportion    0.72428   0.73422   0.74248   0.75040   0.75784
```

```
##            sex        indiv      rnaconc rin        batch       prep
## PC1 0.58336485 0.0864835347 2.961741e-01   1 7.259531e-02 0.04613540
## PC2 0.03827899 0.9823782656 4.772691e-01   1 4.646050e-01 0.28605593
## PC3 0.89704728 0.0004472253 7.406611e-01   1 2.035947e-01 0.67964746
## PC4 0.28460681 0.1676887394 2.217156e-05   1 4.868339e-05 0.08762717
## PC5 0.11282027 0.0188994071 3.267726e-02   1 2.007956e-03 0.06446631
## PC6 0.15848324 0.6616556810 3.406680e-04   1 7.608521e-05 0.03235760
##             conc      length  flowlane     index
## PC1 0.5474112956 0.009878271 0.1960142 0.0861196
## PC2 0.9243452734 0.106290269 0.9231959 0.9441549
## PC3 0.8624445815 0.313493925 0.3546657 0.8965722
## PC4 0.2213172544 0.240954171 0.4563575 0.9095229
## PC5 0.3962807896 0.444377537 0.5465349 0.4729891
## PC6 0.0007395261 0.117367982 0.5639258 0.3006163
```

- RNA concentration correlated with PC1, regress that out:

Regress out RNA concentration

```
##                               PC1        PC2       PC3       PC4       PC5
## Standard deviation     1120.71062 1018.12850 813.58210 785.05431 705.24836
## Proportion of Variance    0.15015    0.12392   0.07913   0.07368   0.05946
## Cumulative Proportion     0.15015    0.27407   0.35320   0.42688   0.48634
##                              PC6       PC7       PC8       PC9      PC10
## Standard deviation     557.97101 546.16105 499.24072 460.93141 427.06406
## Proportion of Variance   0.03722   0.03566   0.02980   0.02540   0.02180
## Cumulative Proportion    0.52356   0.55922   0.58901   0.61441   0.63622
##                             PC11      PC12      PC13      PC14      PC15
## Standard deviation     399.31072 386.25395 365.74080 317.72367 309.97922
## Proportion of Variance   0.01906   0.01784   0.01599   0.01207   0.01149
## Cumulative Proportion    0.65528   0.67311   0.68910   0.70117   0.71266
##                             PC16      PC17      PC18      PC19      PC20
## Standard deviation     303.71154 287.13620 262.96069 255.98950 250.27400
## Proportion of Variance   0.01103   0.00986   0.00827   0.00783   0.00749
## Cumulative Proportion    0.72369   0.73354   0.74181   0.74964   0.75713
```

```
##            sex        indiv rnaconc       rin        batch       prep
## PC1 0.61194105 0.0773136534       1 0.8880552 6.582215e-02 0.04432801
## PC2 0.03558051 0.9404499268       1 0.9246948 4.976321e-01 0.29837204
## PC3 0.87012431 0.0005629413       1 0.9828028 2.664856e-01 0.72871536
## PC4 0.20306175 0.1206037669       1 0.5762441 1.501273e-05 0.06235506
## PC5 0.05749523 0.0224414659       1 0.7355372 5.656690e-03 0.08589515
## PC6 0.23361560 0.9122959168       1 0.6184533 9.894627e-06 0.02447593
##            conc     length  flowlane      index
## PC1 0.631574866 0.01033055 0.2339968 0.07908991
## PC2 0.990288582 0.10736749 0.8783607 0.91050131
## PC3 0.809843663 0.29425804 0.3621235 0.90739704
## PC4 0.448482960 0.19913159 0.2235263 0.89554162
## PC5 0.265364672 0.30336973 0.3207269 0.54643230
## PC6 0.007808397 0.29709384 0.7713029 0.37299978
```

Looks good!

- RNA concentration correlated with PC1, regress that out:

Regress out RNA concentration

```
##                              PC1        PC2       PC3       PC4       PC5
## Standard deviation     1116.7700 1017.63507 812.90620 770.52359 698.52491
## Proportion of Variance    0.1506    0.12505   0.07979   0.07169   0.05892
## Cumulative Proportion     0.1506    0.27565   0.35544   0.42713   0.48605
##                              PC6       PC7       PC8       PC9      PC10
## Standard deviation     547.16973 544.61837 497.11880 460.07274 423.68826
## Proportion of Variance   0.03615   0.03582   0.02984   0.02556   0.02168
## Cumulative Proportion    0.52220   0.55802   0.58786   0.61342   0.63510
##                             PC11      PC12      PC13      PC14      PC15
## Standard deviation     397.39754 385.95192 362.56841 317.47143 309.81587
## Proportion of Variance   0.01907   0.01799   0.01587   0.01217   0.01159
## Cumulative Proportion    0.65417   0.67215   0.68803   0.70020   0.71179
##                             PC16      PC17      PC18      PC19      PC20
## Standard deviation     302.03520 285.65986 262.57277 255.61015 249.13343
## Proportion of Variance   0.01102   0.00985   0.00833   0.00789   0.00749
## Cumulative Proportion    0.72280   0.73266   0.74098   0.74887   0.75636
```

```
##            sex       indiv   rnaconc       rin batch      prep        conc
## PC1 0.58695933 0.085408652 0.9942220 0.8094497     1 0.2431194 0.404592314
## PC2 0.03296384 0.936756509 0.9978459 0.9568373     1 0.4540860 0.895480403
## PC3 0.83415033 0.001059323 0.9977542 0.9513224     1 0.8815069 0.642552111
## PC4 0.08091049 0.106075642 0.9870601 0.6995260     1 0.8949053 0.805189813
## PC5 0.13560049 0.009662546 0.9896419 0.6607217     1 0.6540475 0.633393006
## PC6 0.67678441 0.005956784 0.9910880 0.7113192     1 0.4690498 0.002920905
##         length  flowlane      index
## PC1 0.01771284 0.2670354 0.07321143
## PC2 0.11389536 0.8914513 0.94033815
## PC3 0.23392411 0.4279710 0.87906178
## PC4 0.29211468 0.1932696 0.96692485
## PC5 0.27912609 0.3637523 0.57312230
## PC6 0.03428248 0.1817418 0.26485522
```

```r
my_data<- r.residual.int3
maternal <- maternalrun1[-c(1:4),]
paternal <- paternalrun1[-c(1:4),]

mat <- as.matrix(maternal)
findivs <- sapply(strsplit(colnames(mat), "_"), "[", 2)
colnames(mat) <- findivs
nms <- colnames(mat)
aftersumlanemat <- as.data.frame(mat %*% sapply(unique(nms), "==", nms))
aftersumgenes <- gsub("\\..*","",rownames(aftersumlanemat))
rownames(aftersumlanemat) <- aftersumgenes

pat <- as.matrix(paternal)
findivs <- sapply(strsplit(colnames(pat), "_"), "[", 2)
colnames(pat) <- findivs
nms <- colnames(pat)
aftersumlanepat <- as.data.frame(pat %*% sapply(unique(nms), "==", nms))
aftersumgenes <- gsub("\\..*","",rownames(aftersumlanemat))
findivs<- colnames(aftersumlanepat)
rownames(aftersumlanepat) <- aftersumgenes
```

```r
#No parent of origin information for:
colnames(my_data)[which(!colnames(my_data)%in%findivs)]
```

```
## character(0)
```

```r
missing <- which(!colnames(my_data)%in%findivs)
#beforenames <- colsplit(string=colnames(my_data)[missing], pattern="_", names=c("FC", "findiv", "lanet
#findivsmissing <- beforenames$findiv

#Removed from qc:
findivs[which(!findivs%in%colnames(my_data))]
```

```
## character(0)
```

```r
removed <- which(!findivs%in%colnames(my_data))
#beforenames <- colsplit(string=findivs[removed], pattern="_", names=c("FC", "findiv", "lanetext","lane
#findivstoremove <- beforenames$findiv

newer <- my_data
genes<- rownames(newer)
newmat <- aftersumlanemat[genes,]
newpat <- aftersumlanepat[genes,]
newmat2 <- newmat[,colnames(newer)]
newpat2 <- newpat[,colnames(newer)]
write.table(newpat2, paste("~/Documents/POexpressionpaper/Analyses/Data/Paternal_gene_notnormalized_nodu
write.table(newmat2, paste("~/Documents/POexpressionpaper/Analyses/Data/Maternal_gene_notnormalized_nodu
```

```r
mean((rowSums(newmat2)+rowSums(newpat2))/rowSums(aftersumlane.y.x.nolowexpressed))
```

```
## [1] 0.01802522
```

```r
mean((colSums(newmat2)+colSums(newpat2))/colSums(aftersumlane.y.x.nolowexpressed))
```

```
## [1] 0.01385943
```

```
## character(0)
```

```
## character(0)
```

```r
sessionInfo()
```

```
## R version 3.4.2 (2017-09-28)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13.3
##
## 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] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] edgeR_3.20.5   limma_3.34.5   reshape2_1.4.3 dplyr_0.7.4
## [5] plyr_1.8.4
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.14     knitr_1.18       bindr_0.1        magrittr_1.5
##  [5] lattice_0.20-35  R6_2.2.2         rlang_0.1.6      stringr_1.2.0
##  [9] tools_3.4.2      grid_3.4.2       htmltools_0.3.6  yaml_2.1.16
## [13] rprojroot_1.3-1  digest_0.6.13    assertthat_0.2.0 tibble_1.4.1
## [17] bindrcpp_0.2     glue_1.2.0       evaluate_0.10.1  rmarkdown_1.8
## [21] stringi_1.1.6    compiler_3.4.2   pillar_1.0.1     backports_1.1.2
## [25] locfit_1.5-9.1   pkgconfig_2.0.1
```