LCL

Sahar Mozaffari

5/22/2017

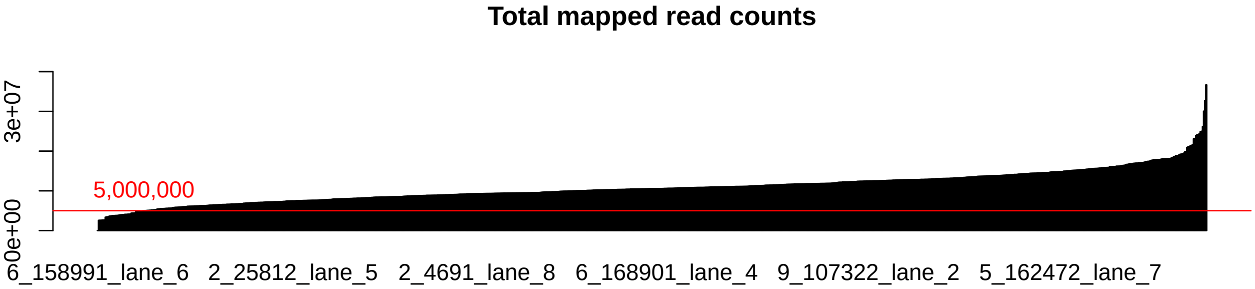
R Markdown for RNA-seq data

Genecount matrix:

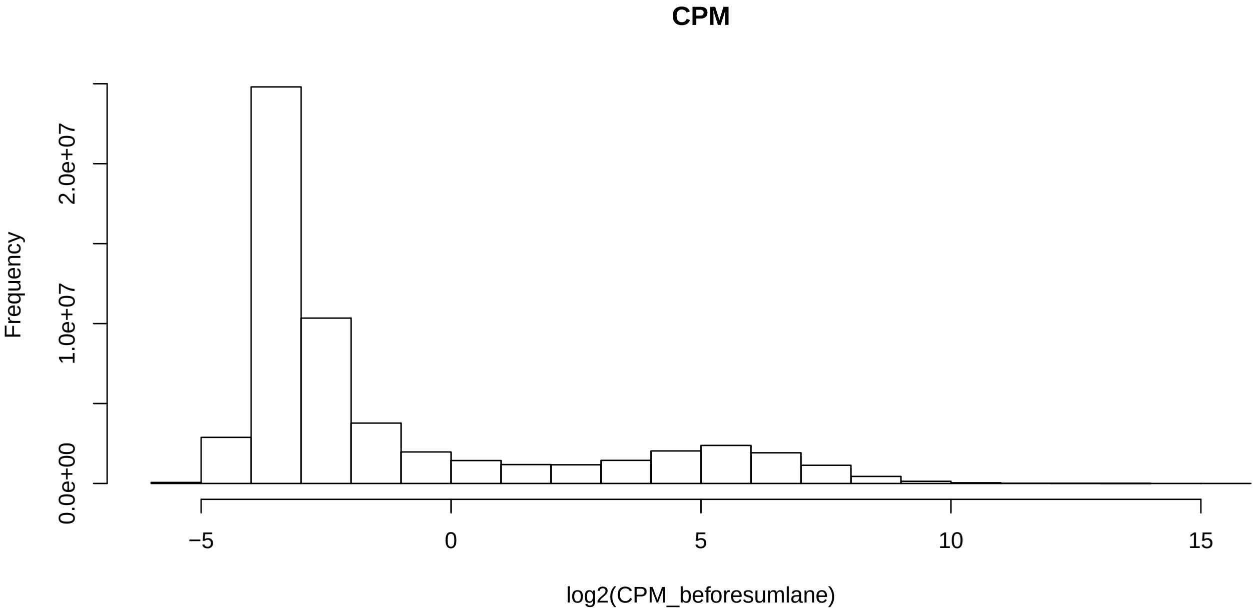
- genes in rows, individuals/samples by lane and flowcell in columns
- There are a total of 5.7819×10^4 genes and 989 samples

verifyBAMid found some sample swaps:

Number of lanes with enough reads, before combining replicates



- The distribution of Counts Per Million:



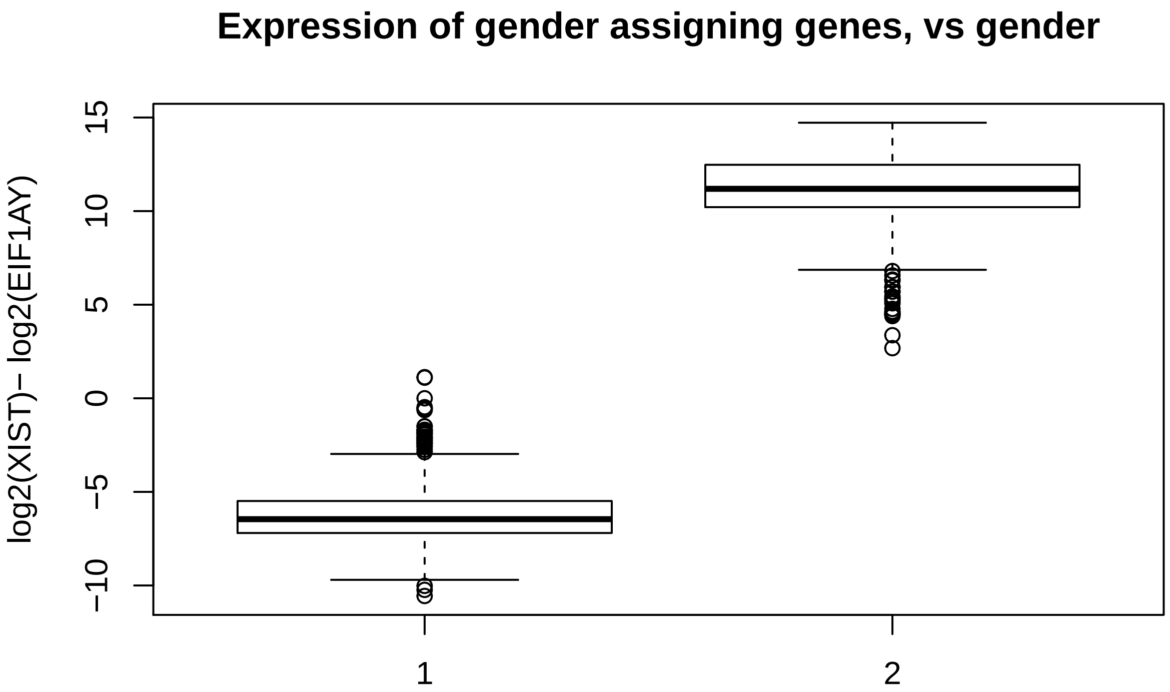
Sexcheck

Sex assigned by ratio of XIST to EIF1AY gene

```
## callSex
##   F   M 
## 521 468
```

- According to expression of sex genes, there are 521 females and 468 males.

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



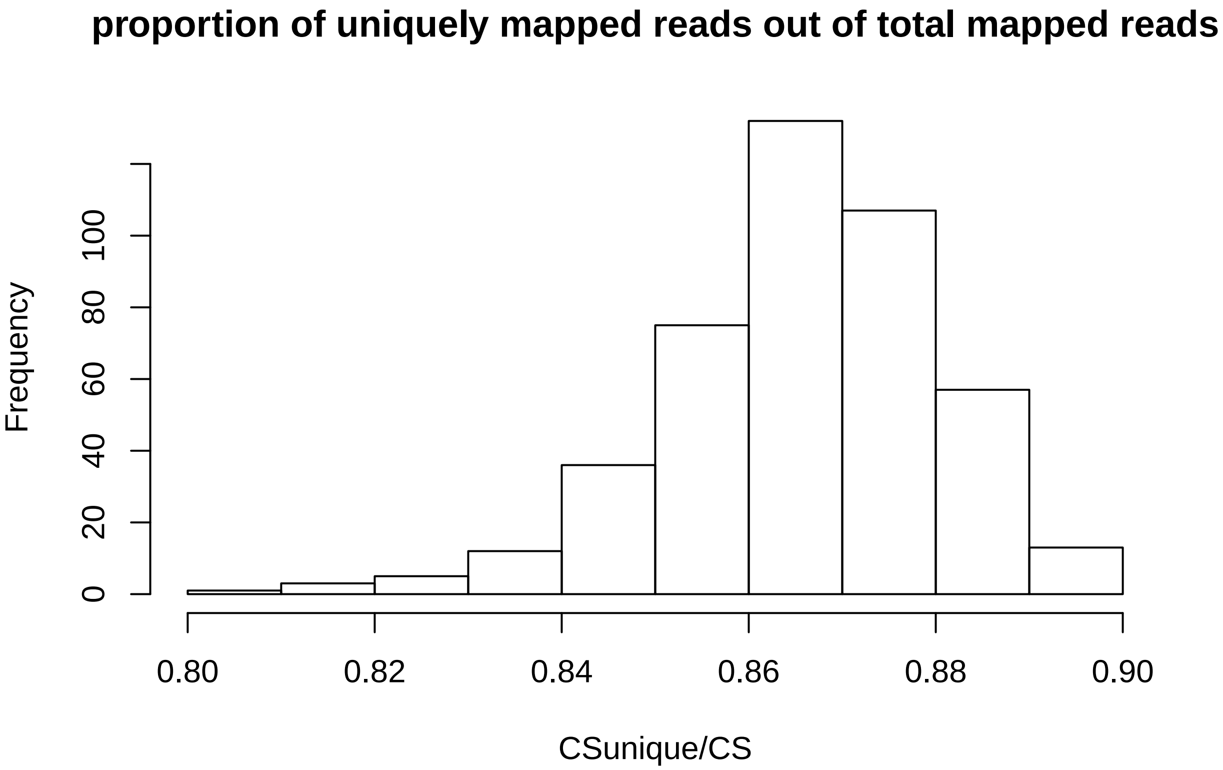
*

There are supposed to be: 0 females and 0 males. * The samples misassigned are: 1.1020687, 1.1278372

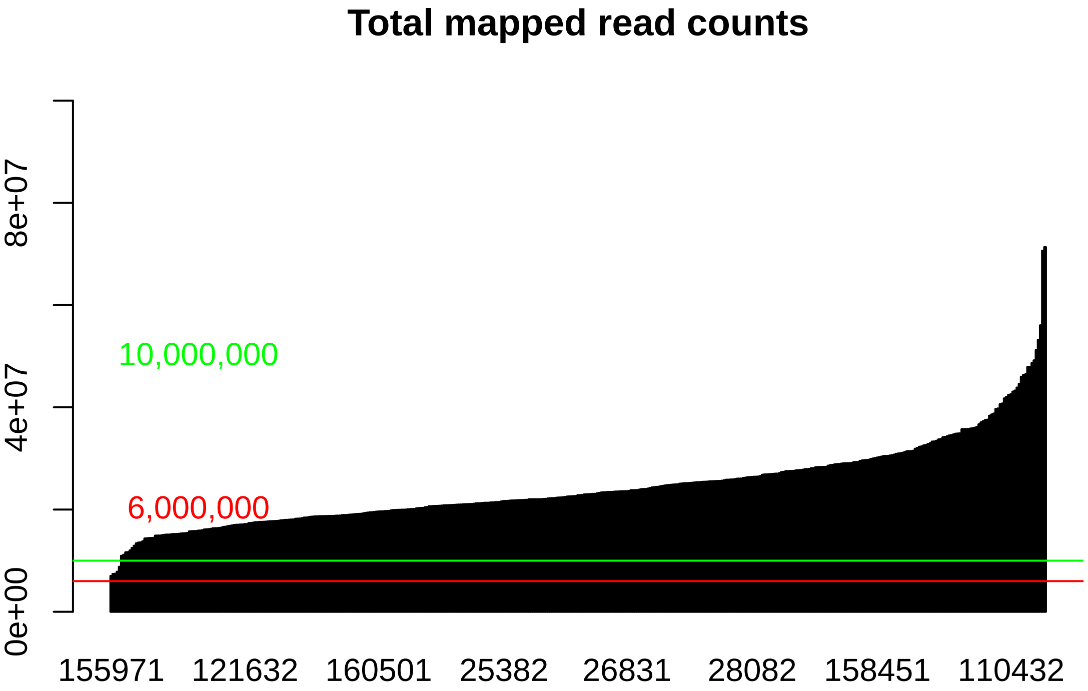
Combining technical replicates

- 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             48      51      63      29      47
## N_multimapping         27      40      23      21      18
## N_noFeature       1819891 3491256 1962976 1659087 1436061
## N_ambiguous       1407294 1655854 1011721 1290194 1210412
## ENSG00000223972.4       0       0       0       0       0
```

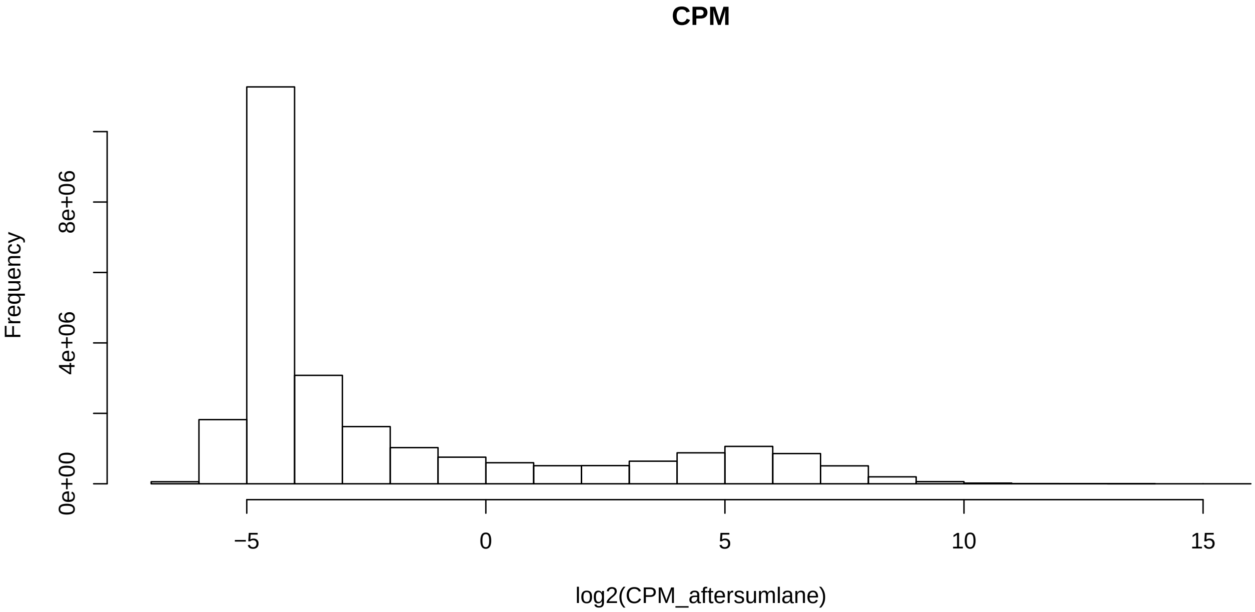


```
##                    100092 100172 100182 100202 100372
## ENSG00000223972.4       0      0      0      0      0
## ENSG00000227232.4       4     21     25     13      5
## ENSG00000243485.2       0      0      0      0      0
## ENSG00000237613.2       0      0      0      0      0
## ENSG00000268020.2       0      0      0      0      0
```



- combine total number of read covariate value

- After combining replicates, 108821, 155971, 158431, 159021, 163372 out of total 441 have more than 10 million reads
- The distribution of Counts Per Million after combining replicates:

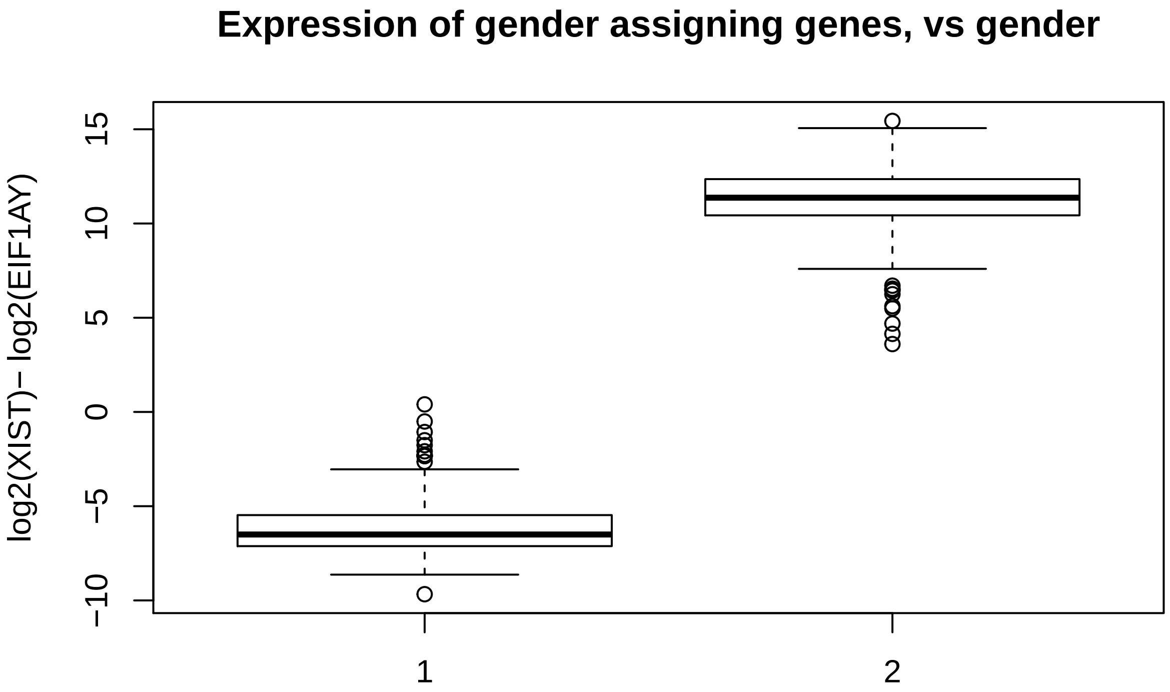


Checking sex after combining replicates

```
## 100092 100172 100182 100202 100372
##  22596  44548    128  18002  13194
```

```
## callSex
##   F   M
## 231 210
```

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



```
##    171351
## 0.4024327
```

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

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

- There are supposed to be: 231 females and 210 males.
- The samples misassigned are: 171351

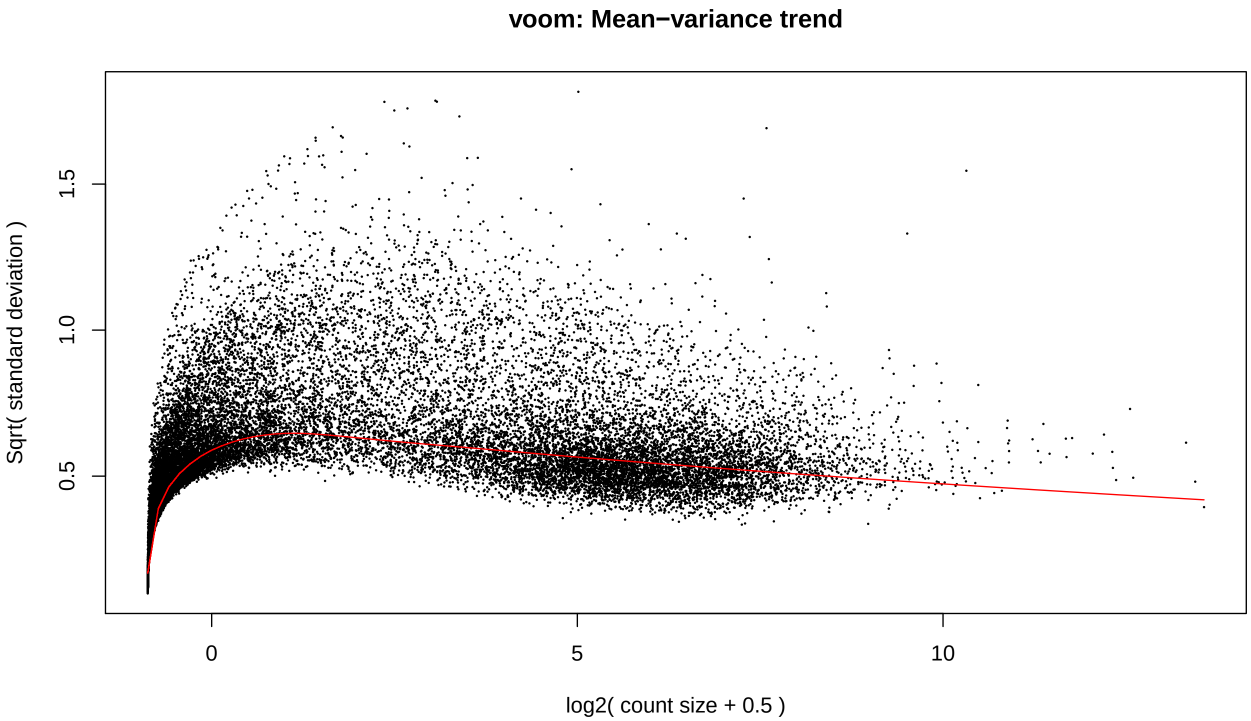
These 1 **individuals have wrong assigned sex- last** 0 **have quite a large error - remove from data**

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)==438)
```

```
##
## FALSE  TRUE
## 39965     2
```

```
keep.exprs <- rowSums(cpm>1)>=10
aftersumlane.y.x.nolowexpressed<- aftersumlane.y.x[keep.exprs, ]
dim(aftersumlane.y.x.nolowexpressed)
```

```
## [1] 16146   441
```

```
#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
## Standard deviation     7172.72957 4319.30465 3381.02648 2617.34050
## Proportion of Variance    0.42966    0.15581    0.09547    0.05721
## Cumulative Proportion     0.42966    0.58546    0.68093    0.73814
##                               PC5        PC6        PC7        PC8
## Standard deviation     1885.45678 1750.07985 1461.13841 1425.00207
## Proportion of Variance    0.02969    0.02558    0.01783    0.01696
## Cumulative Proportion     0.76783    0.79341    0.81124    0.82819
##                               PC9       PC10       PC11       PC12
## Standard deviation     1362.94769 1281.10118 1181.11850 1110.57991
## Proportion of Variance    0.01551    0.01371    0.01165    0.01030
## Cumulative Proportion     0.84371    0.85741    0.86906    0.87937
##                              PC13      PC14      PC15      PC16      PC17
## Standard deviation     1097.84390 991.18333 954.81236 876.55734 782.07363
## Proportion of Variance    0.01007   0.00820   0.00761   0.00642   0.00511
## Cumulative Proportion     0.88943   0.89764   0.90525   0.91167   0.91677
##                             PC18      PC19      PC20
## Standard deviation     768.78837 724.89311 653.06915
## Proportion of Variance   0.00494   0.00439   0.00356
## Cumulative Proportion    0.92171   0.92610   0.92966
```

```
##           sex      indiv      rnaconc          rin      batch      prep
## PC1 0.9590861 0.24339012 2.081491e-05 3.437025e-04 0.87093979 0.8296127
## PC2 0.9363389 0.53646916 1.316722e-01 2.857106e-08 0.46659602 0.9520625
## PC3 0.4872276 0.01758998 8.825878e-01 5.418935e-15 0.28007040 0.9159418
## PC4 0.8432731 0.35124788 4.414221e-07 2.091130e-01 0.05371794 0.1508468
## PC5 0.4719166 0.38721908 4.880774e-01 1.548595e-01 0.08948297 0.1422734
## PC6 0.5360782 0.41731569 4.929866e-01 8.962083e-03 0.10078382 0.7024798
##          conc     length  flowlane     index
## PC1 0.6884387 0.92593023 0.1803528 0.7429108
## PC2 0.6599178 0.61744619 0.9163519 0.2508214
## PC3 0.1935642 0.44675989 0.2354815 0.2429743
## PC4 0.1658708 0.82320809 0.8420896 0.2684072
## PC5 0.5125035 0.93449976 0.2180389 0.1899668
## PC6 0.3693728 0.05476976 0.8934433 0.2349370
```

Regress out RIN

- PCA for the second time:

```
##                               PC1       PC2       PC3       PC4
## Standard deviation     7076.77940 4196.77561 3114.71579 2609.26603
## Proportion of Variance    0.43449    0.15281    0.08417    0.05907
## Cumulative Proportion     0.43449    0.58730    0.67146    0.73053
##                               PC5       PC6       PC7       PC8
## Standard deviation     1880.35929 1732.13726 1461.12180 1424.83470
## Proportion of Variance    0.03068    0.02603    0.01852    0.01761
## Cumulative Proportion     0.76121    0.78724    0.80576    0.82337
##                               PC9      PC10      PC11      PC12
## Standard deviation     1362.10577 1279.85140 1146.74792 1109.74462
## Proportion of Variance    0.01610    0.01421    0.01141    0.01068
## Cumulative Proportion     0.83947    0.85368    0.86509    0.87577
##                              PC13      PC14     PC15      PC16      PC17
## Standard deviation     1095.29483 985.24386 954.1214 868.34345 781.89912
## Proportion of Variance    0.01041   0.00842   0.0079   0.00654   0.00530
## Cumulative Proportion     0.88618   0.89460   0.9025   0.90904   0.91435
##                             PC18      PC19      PC20
## Standard deviation     752.27905 711.66863 641.88345
## Proportion of Variance   0.00491   0.00439   0.00357
## Cumulative Proportion    0.91926   0.92365   0.92722
```

```
##           sex     indiv      rnaconc rin      batch      prep      conc
## PC1 0.9202675 0.3352821 1.650977e-04   1 0.95606344 0.7985363 0.7563713
## PC2 0.9745610 0.2356439 2.822910e-01   1 0.74592397 0.9784938 0.6091393
## PC3 0.2565326 0.0700735 2.781370e-01   1 0.48084912 0.9242861 0.1367589
## PC4 0.7233442 0.3183021 1.167720e-06   1 0.05138457 0.1561525 0.1181020
## PC5 0.5061609 0.3427249 5.282228e-01   1 0.09741891 0.1370830 0.4011098
## PC6 0.4539121 0.5204935 3.785008e-01   1 0.04827945 0.5321611 0.5609802
##         length  flowlane     index
## PC1 0.98847947 0.2695891 0.7802804
## PC2 0.61409895 0.5971489 0.1930546
## PC3 0.35103624 0.4975193 0.3416830
## PC4 0.73326174 0.8148754 0.2417144
## PC5 0.79190322 0.2407557 0.1971672
## PC6 0.07082453 0.8414347 0.1612958
```

- RNA concentration correlated with PC1, regress that out:

Regress out RNA concentration

```
##                               PC1       PC2       PC3       PC4
## Standard deviation     6964.70335 4191.22078 3110.90076 2538.15051
## Proportion of Variance    0.42885    0.15530    0.08556    0.05696
## Cumulative Proportion     0.42885    0.58415    0.66971    0.72666
##                               PC5       PC6       PC7       PC8
## Standard deviation     1879.41867 1730.47520 1460.39889 1423.32339
## Proportion of Variance    0.03123    0.02647    0.01886    0.01791
## Cumulative Proportion     0.75789    0.78437    0.80322    0.82113
##                               PC9      PC10      PC11      PC12
## Standard deviation     1361.64640 1273.99801 1146.63585 1108.80114
## Proportion of Variance    0.01639    0.01435    0.01162    0.01087
## Cumulative Proportion     0.83753    0.85187    0.86350    0.87437
##                              PC13      PC14      PC15      PC16      PC17
## Standard deviation     1089.88980 972.37650 949.70225 867.50960 776.24317
## Proportion of Variance    0.01050   0.00836   0.00797   0.00665   0.00533
## Cumulative Proportion     0.88487   0.89323   0.90120   0.90786   0.91318
##                             PC18      PC19      PC20
## Standard deviation     752.25413 705.88306 641.86029
## Proportion of Variance   0.00500   0.00441   0.00364
## Cumulative Proportion    0.91819   0.92259   0.92623
```

```
##           sex      indiv rnaconc       rin      batch      prep      conc
## PC1 0.7095491 0.25085258       1 0.5847933 0.95891400 0.7686799 0.4776841
## PC2 0.9105483 0.26004504       1 0.8747322 0.75110012 0.9616655 0.5234426
## PC3 0.2837501 0.07371833       1 0.8850482 0.45440176 0.9594069 0.1001467
## PC4 0.4903978 0.21913463       1 0.4616774 0.04389163 0.1529939 0.3419889
## PC5 0.5408285 0.37050234       1 0.9191692 0.10174002 0.1407114 0.4410337
## PC6 0.4867467 0.55808576       1 0.8890841 0.04468745 0.5259814 0.6220092
##         length  flowlane     index
## PC1 0.87981941 0.4810320 0.8769885
## PC2 0.58556564 0.5268655 0.2021623
## PC3 0.32550465 0.4358327 0.3386687
## PC4 0.91820997 0.4110586 0.1908825
## PC5 0.81943828 0.2770148 0.1805520
## PC6 0.07448965 0.9265217 0.1802423
```

Looks good!

05-18-16 decided to only do TMM normalization but not regress out effects:

So use my_data moving forward for now.

```
#5-18-16 decided to only do TMM normalization but not regress out effects:
# SO use my_data moving forward for now.

#my_data<- r.residual.int2
maternal <- maternalrun1[-c(1:4),]
paternal <- paternalrun1[-c(1:4),]
#maternal<- read.table("~/star_overhang_v19_genecount_maternalaltcountReadsPerGene.out.tab", check.name
#paternal<- read.table("~/star_overhang_v19_genecount_paternalaltcountReadsPerGene.out.tab", check.name

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)
```

```r
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


#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)]

aftersumlaneprop <- aftersumlane[genes,colnames(newer)]

propmat <- newmat2/aftersumlaneprop
proppat <- newpat2/aftersumlaneprop

newermat2 <- (propmat*newer)
newerpat2 <- (proppat*newer)

newermat2[is.na(newermat2)] <- 0
newerpat2[is.na(newerpat2)] <- 0

newermat2[newermat2==-Inf] <- 0
newerpat2[newerpat2==-Inf] <- 0

justnormalizedmat <- newermat2
justnormalizedpat <- newerpat2
#write.table(newerpat2, "~/Paternal_gene_normalized_v19_08.02.17.txt", quote =F, row.names = T, col.nam
#write.table(newermat2, "~/Maternal_gene_normalized_v19_08.02.17.txt", quote =F, row.names = T, col.nam
```

```r
#5-18-16 decided to only do TMM normalization but not regress out effects:
# SO use my_data moving forward for now.
```

```
my_data<- r.residual.int2
maternal <- maternalrun1[-c(1:4),]
paternal <- paternalrun1[-c(1:4),]
#maternal<- read.table("~/star_overhang_v19_genecount_maternalaltcountReadsPerGene.out.tab", check.name
#paternal<- read.table("~/star_overhang_v19_genecount_paternalaltcountReadsPerGene.out.tab", check.name

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


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

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

```
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)
```

```
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)]

aftersumlaneprop <- aftersumlane[genes,colnames(newer)]

propmat <- newmat2/aftersumlaneprop
proppat <- newpat2/aftersumlaneprop

newermat2 <- (propmat*newer)
```

```r
newerpat2 <- (proppat*newer)

newermat2[is.na(newermat2)] <- 0
newerpat2[is.na(newerpat2)] <- 0

newermat2[newermat2==-Inf] <- 0
newerpat2[newerpat2==-Inf] <- 0


#write.table(newerpat2, "~/Paternal_gene_normalized_v19_08.02.17.txt", quote =F, row.names = T, col.nam
#write.table(newermat2, "~/Maternal_gene_normalized_v19_08.02.17.txt", quote =F, row.names = T, col.nam
```

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

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

```r
sessionInfo()
```

```
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X El Capitan 10.11.6
##
## 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] sva_3.22.0        genefilter_1.56.0 mgcv_1.8-17       nlme_3.1-131
## [5] edgeR_3.16.5      limma_3.30.13     reshape2_1.4.2    dplyr_0.7.2
## [9] plyr_1.8.4
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.12         bindr_0.1            bitops_1.0-6
##  [4] tools_3.3.2          bit_1.1-12           digest_0.6.12
##  [7] memoise_1.1.0        RSQLite_2.0          annotate_1.52.1
## [10] evaluate_0.10.1      tibble_1.3.3         lattice_0.20-35
## [13] pkgconfig_2.0.1      rlang_0.1.1          Matrix_1.2-10
## [16] DBI_0.7              yaml_2.1.14          parallel_3.3.2
## [19] bindrcpp_0.2         stringr_1.2.0        knitr_1.16
## [22] IRanges_2.8.2        S4Vectors_0.12.2     bit64_0.9-7
## [25] locfit_1.5-9.1       stats4_3.3.2         rprojroot_1.2
## [28] grid_3.3.2           glue_1.1.1           Biobase_2.34.0
## [31] R6_2.2.2             AnnotationDbi_1.36.2 survival_2.41-3
## [34] XML_3.98-1.9         rmarkdown_1.6        blob_1.1.0
## [37] magrittr_1.5         splines_3.3.2        backports_1.1.0
## [40] htmltools_0.3.6      BiocGenerics_0.20.0  assertthat_0.2.0
## [43] xtable_1.8-2         stringi_1.1.5        RCurl_1.95-4.8
```