RNAseq_QC_v19

Sahar Mozaffari

3/26/2017

R Markdown for RNA-seq data

Genecount matrix:

- genes in rows, individuals/samples by lane and flowcell in columns

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##              10_100092_lane_1 10_100092_lane_2 10_106052_lane_3
## 1/2-SBSRNA4                7                8                1
## A1BG                      69               72               35
## A1BG-AS1                  23               21               23
## A1CF                       0                0                0
## A2LD1                    104              122               59
```

- There are a total of 2.3368×10^4 genes and 989 samples

Covariates

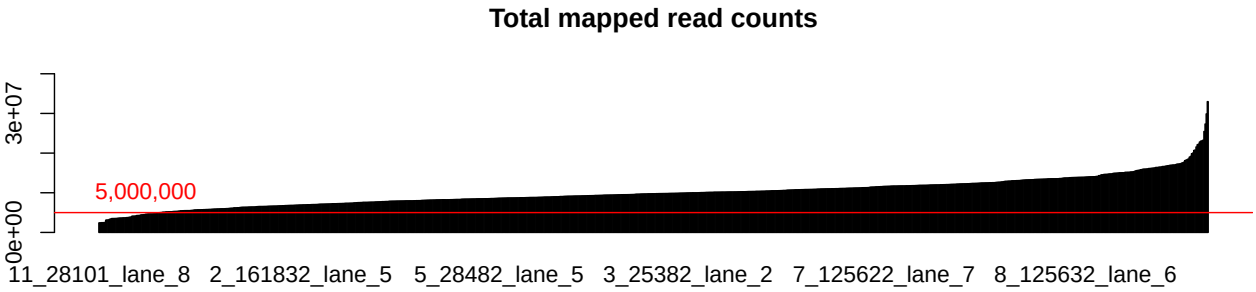
- covariate file has number of reads from total, maternal, and paternal; flowcell, findiv, lane, and adaptor index

```
##    FC_findiv_lane afterWASPwithXY Maternal Paternal Flowcell FINDIV Lane
## 1 1_106272_lane_3        12208829    76983    77587        1 106272    3
## 2 1_106272_lane_4        12046171    75846    75378        1 106272    4
## 3 1_106561_lane_7        13004762    88321    88301        1 106561    7
## 4 1_106561_lane_8        12951599    87880    87700        1 106561    8
## 5 1_106651_lane_5        11026633    71137    71303        1 106651    5
## 6 1_106651_lane_6        11196367    72035    72753        1 106651    6
##   Adaptor_index
## 1             8
## 2             8
## 3             9
## 4             9
## 5             2
```

```
## 6              2
```

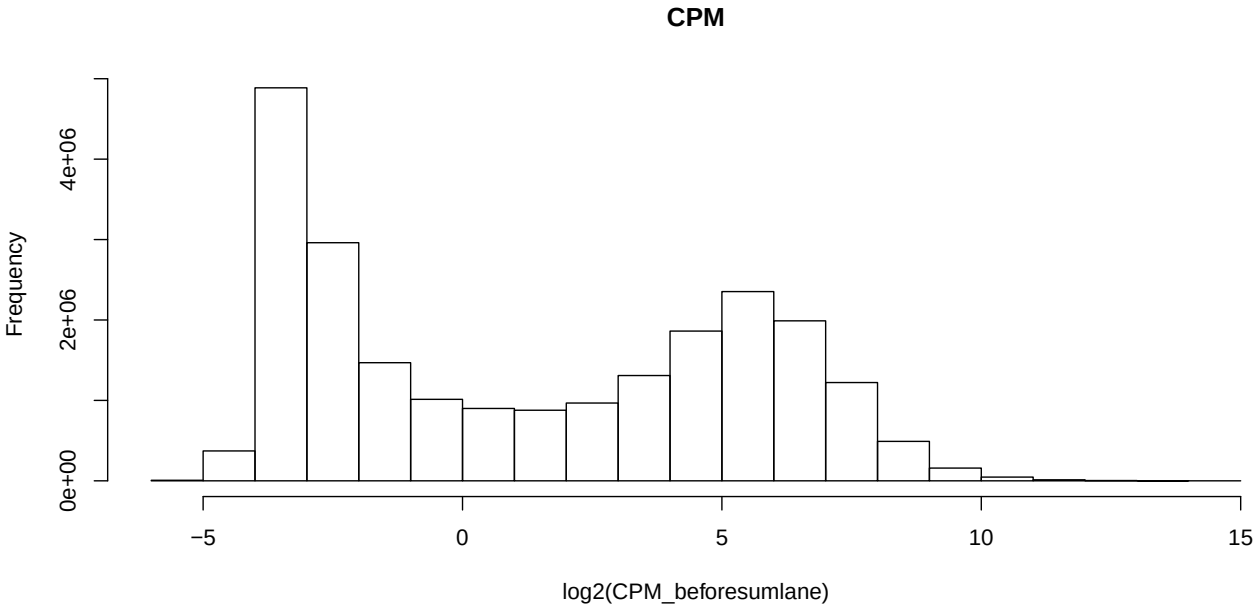
Number of lanes with enough reads, before combining replicates

```
## enough.reads
## FALSE  TRUE
##   278   703
```



* Before combining replicates, 703 out of total 981 have more than 10 million reads

- The distribution of Counts Per Million:



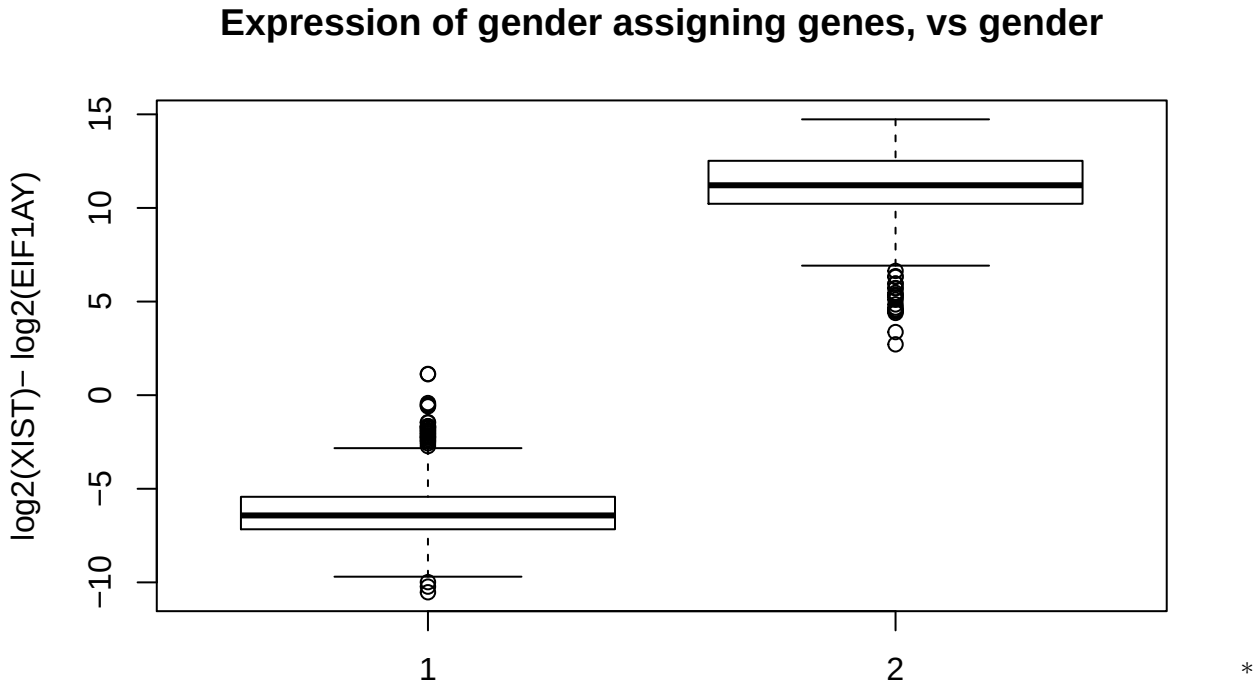
Sexcheck

Sex assigned by ratio of XIST to EIF1AY gene

```
## callSex
##   F   M
## 515 465
```

- According to expression of sex genes, there are 515 females and 465 males.

```
## gender
##   1   2
## 467 513
```



There are supposed to be: 0 females and 0 males. * The samples misassigned are: 1.1204378, 1.132239 *

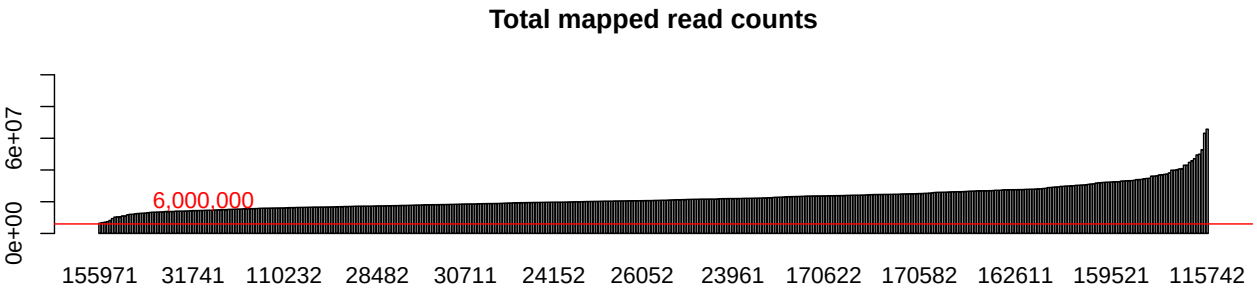
Combining technical replicates

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

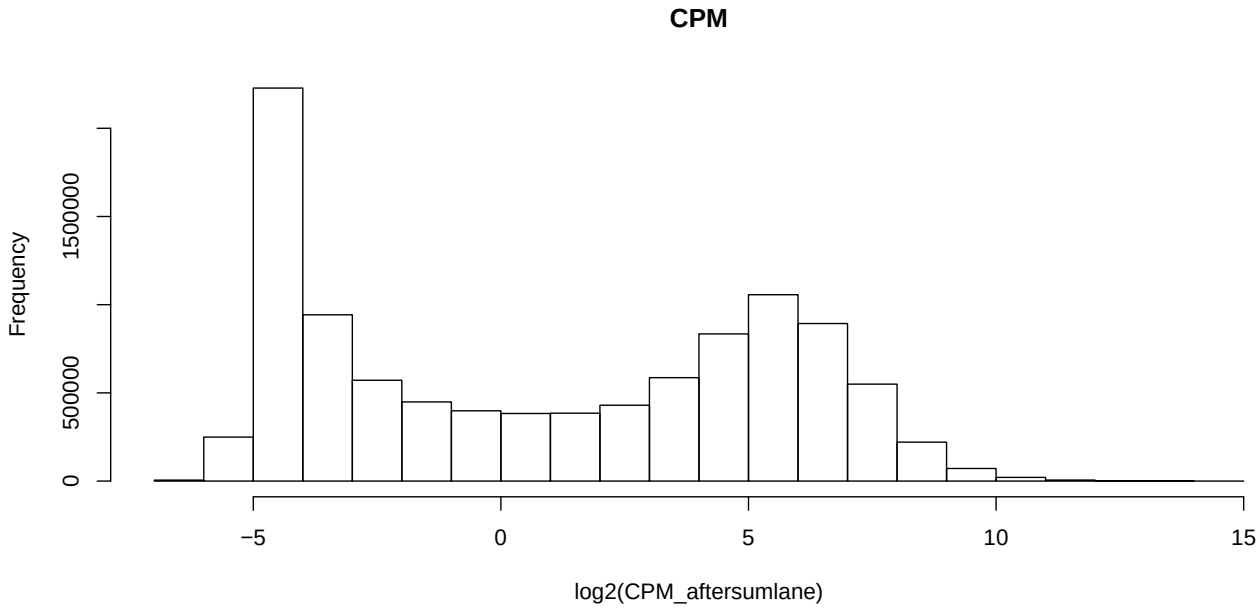
```
##             100092 100172 100182 100202 100372
## 1/2-SBSRNA4     15     16     21      7      9
## A1BG           141    160     87    141     87
## A1BG-AS1        44     50     98     49     39
## A1CF             0      0      2      1      1
## A2LD1          226    263    144    170    128
```

- combine total number of read covariate value

```
##        afterWASPwithXY  Unknown Maternal Paternal
## 100092        24964977 23492036   116281   109759
## 100172        32470844 30273899   132422   134851
## 100182        20511259 19224967    63746    63273
## 100202        23884898 22362200    69909    70220
## 100372        21869250 20596306    79428    79946
```



- The distribution of Counts Per Million after combining replicates:

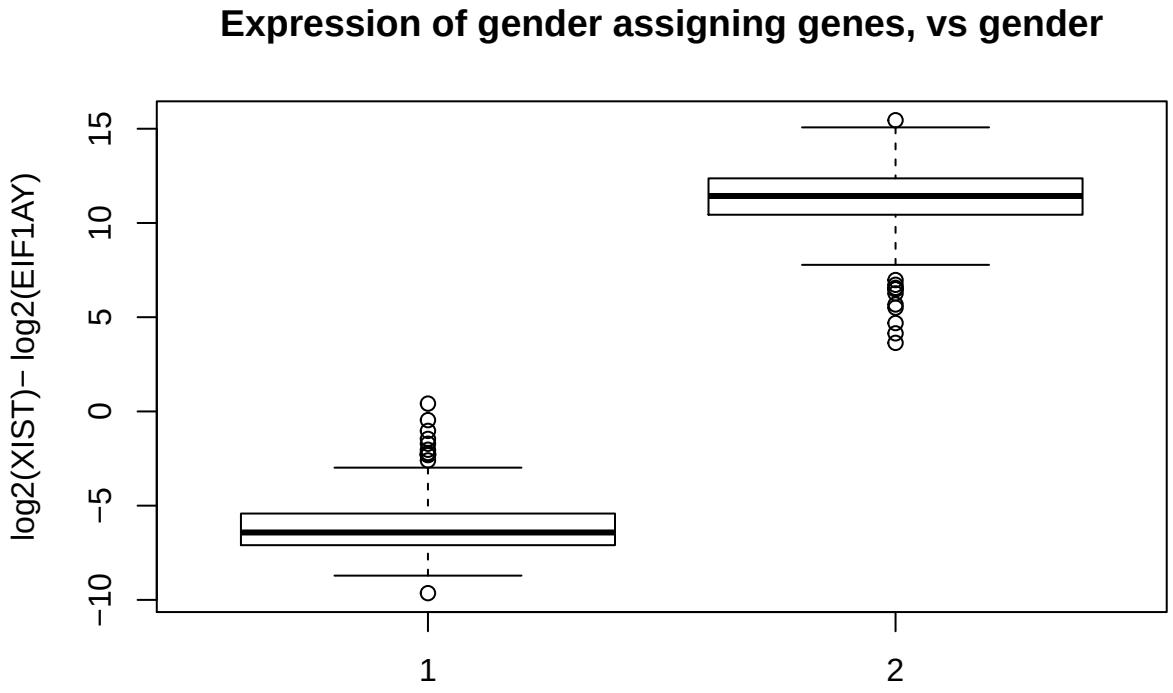


Checking sex after combining replicates

```
## 100092 100172 100182 100202 100372
##  22748  44884    128  18152  13284
```

```
## callSex
##   F   M
## 229 211
```

```
## gender
##   1   2
## 212 228
```



```
##    171351
## 0.4163166
```

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

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

- There are supposed to be: 229 females and 211 males.
- The samples misassigned are: 171351

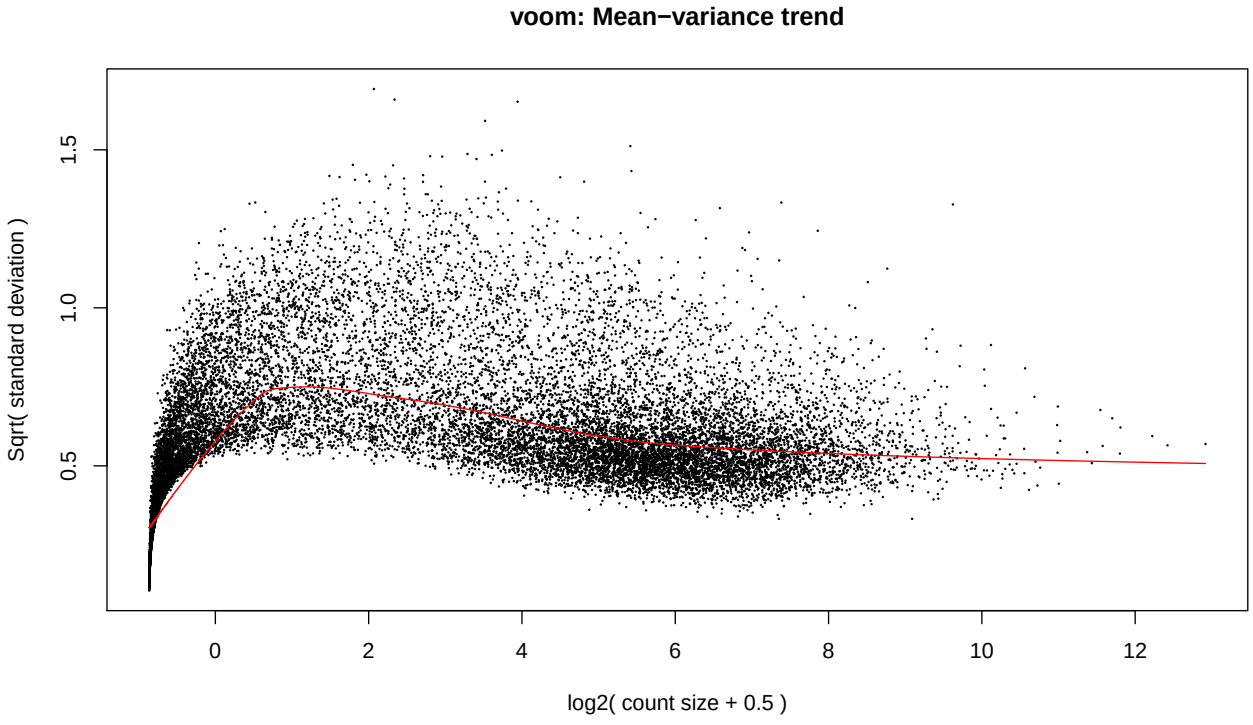
These 1 **individuals have wrong assigned sex- last** 0 **have quite a large error - remove from data**

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

- Total number of chromosome X genes: 2321, Y genes: 494, and mt genes: 37
- Number in data that are removed:
- X chromosome genes: 939
- Y genes: 92
- mt genes: 0

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
## 20336
```

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

```
## [1] 14014   440
```

```
dge <- DGEList(counts=aftersumlane.y.x.nolowexpressed)
dge <- calcNormFactors(dge)
logCPM <- cpm(dge, log=TRUE, prior.count=3)

x <- DGEList(counts=aftersumlane.y.x.nolowexpressed)
```

```
library(RColorBrewer)
nsamples <- ncol(x)
col <- brewer.pal(nsamples, "Paired")
par(mfrow=c(1,2))
plot(density(lcpm[,1]), col=col[1], lwd=2, ylim=c(0,0.21), las=2,
     main="", xlab="")
title(main="A. Raw data", xlab="Log-cpm")
abline(v=0, lty=3)
for (i in 2:nsamples){
 den <- density(lcpm[,i])
 lines(den$x, den$y, col=col[i], lwd=2)
}
legend("topright", samplenames, text.col=col, bty="n")
lcpm <- cpm(x, log=TRUE)
plot(density(lcpm[,1]), col=col[1], lwd=2, ylim=c(0,0.21), las=2,
     main="", xlab="")
title(main="B. Filtered data", xlab="Log-cpm")
abline(v=0, lty=3)
for (i in 2:nsamples){
   den <- density(lcpm[,i])
   lines(den$x, den$y, col=col[i], lwd=2)
}
legend("topright", samplenames, text.col=col, bty="n")

x <- calcNormFactors(x, method = "TMM")


x2 <- x
x2$samples$norm.factors <- 1
x2$counts[,1] <- ceiling(x2$counts[,1]*0.05)
x2$counts[,2] <- x2$counts[,2]*5

par(mfrow=c(1,2))
lcpm <- cpm(x2, log=TRUE)
boxplot(lcpm, las=2, col=col, main="")
```

```
title(main="A. Example: Unnormalised data",ylab="Log-cpm")
x2 <- calcNormFactors(x2)
x2$samples$norm.factors

## [1] 0.0547 6.1306 1.2293 1.1705 1.2149 1.0562 1.1459 1.2613 1.1170

lcpm <- cpm(x2, log=TRUE)
boxplot(lcpm, las=2, col=col, main="")
title(main="B. Example: Normalised data",ylab="Log-cpm")
```

Covariates:

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

```
##        sex readsafterWsex rnaconc  rin batch prep  conc length flowlane
## 100092   2       24964977   965.0  9.8     7    1  9.15    284        1
## 100172   2       32470844   192.0  9.2     6    2 14.49    295       12
## 100182   2       20511259   173.0  9.2     4    2 10.43    282       23
## 100202   2       23884898   835.1  9.6     3    2  4.78    282       34
## 100372   2       21869250   588.0 10.0     5    2 11.50    290       45
## 100582   2       22956471   191.0  9.2     4    2  2.55    270       56
##        index
## 100092     5
## 100172     3
## 100182    11
## 100202     1
## 100372     3
## 100582    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     33.24129 27.35754 25.96343 22.12024 19.85386
## Proportion of Variance  0.09302  0.06301  0.05675  0.04119  0.03318
## Cumulative Proportion   0.09302  0.15603  0.21278  0.25397  0.28716
##                             PC6      PC7      PC8      PC9     PC10
## Standard deviation     18.33217 16.76933 16.03703 15.30572 14.59360
## Proportion of Variance  0.02829  0.02367  0.02165  0.01972  0.01793
## Cumulative Proportion   0.31545  0.33912  0.36078  0.38050  0.39843
##                            PC11     PC12     PC13     PC14     PC15
## Standard deviation     13.84685 12.26243 12.05930 11.33701 10.63829
## Proportion of Variance  0.01614  0.01266  0.01224  0.01082  0.00953
## Cumulative Proportion   0.41457  0.42723  0.43947  0.45029  0.45982
##                            PC16     PC17     PC18     PC19     PC20
## Standard deviation     10.50933 9.726522 9.515969 9.452034 9.106759
## Proportion of Variance  0.00930 0.007960 0.007620 0.007520 0.006980
```

```
## Cumulative Proportion   0.46912 0.477080 0.484700 0.492230 0.499210
```

```
##              sex readsafterWsex      rnaconc          rin      batch
## PC1 7.461675e-01    0.610272032 8.366128e-01 1.162597e-08 0.72949675
## PC2 6.019484e-03    0.393162819 2.811595e-05 1.640758e-01 0.45859441
## PC3 6.387896e-02    0.786314507 3.164324e-03 2.196646e-10 0.02844282
## PC4 2.498006e-01    0.320979380 1.446660e-01 1.192671e-01 0.09023776
## PC5 4.812173e-01    0.006022885 4.406247e-02 7.976124e-08 0.27171573
## PC6 2.351190e-06    0.109274428 3.915483e-03 3.918529e-01 0.30012308
##          prep      conc    length  flowlane     index
## PC1 0.7058100 0.4120636 0.3251854 0.6346912 0.2531707
## PC2 0.4657934 0.9926793 0.4222635 0.3921315 0.8370803
## PC3 0.1534110 0.4177554 0.4325015 0.6736427 0.8050613
## PC4 0.1639986 0.6001914 0.6234873 0.3143513 0.5208386
## PC5 0.8862177 0.4321111 0.4931637 0.5968285 0.7716769
## PC6 0.5917626 0.1485907 0.2985333 0.3553152 0.2633359
```

Regress out RIN

- PCA for the second time:

```
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     32.27133 27.32248 24.80110 22.05164 19.05235
## Proportion of Variance  0.08916  0.06391  0.05266  0.04163  0.03108
## Cumulative Proportion   0.08916  0.15307  0.20573  0.24737  0.27844
##                             PC6      PC7      PC8      PC9     PC10
## Standard deviation     18.29421 16.74603 16.03586 15.25602 14.59032
## Proportion of Variance  0.02865  0.02401  0.02202  0.01993  0.01823
## Cumulative Proportion   0.30710  0.33111  0.35312  0.37305  0.39127
##                            PC11     PC12     PC13     PC14     PC15
## Standard deviation     13.79078 12.26210 12.04550 11.08834 10.60608
## Proportion of Variance  0.01628  0.01287  0.01242  0.01053  0.00963
## Cumulative Proportion   0.40755  0.42043  0.43285  0.44338  0.45301
##                            PC16     PC17     PC18    PC19     PC20
## Standard deviation     10.36545 9.658549 9.515786 9.44384 9.105702
## Proportion of Variance  0.00920 0.007990 0.007750 0.00764 0.007100
## Cumulative Proportion   0.46221 0.470190 0.477940 0.48558 0.492680
```

```
##              sex readsafterWsex      rnaconc rin     batch      prep
## PC1 8.170304e-01    0.424759896 0.3277364796   1 0.3801526 0.5015663
## PC2 9.829544e-03    0.452725880 0.0001975074   1 0.2966988 0.5880956
## PC3 1.319667e-02    0.754540349 0.0146677150   1 0.1541950 0.2190896
## PC4 1.482389e-01    0.469936170 0.1209255699   1 0.0984839 0.1385956
## PC5 1.099119e-02    0.009775702 0.1743022778   1 0.2696382 0.9812279
## PC6 2.704599e-05    0.037661108 0.0018203202   1 0.2116422 0.5684455
##           conc    length  flowlane     index
## PC1 0.43656298 0.3227036 0.5705046 0.2272362
## PC2 0.92323682 0.3660869 0.3911953 0.7884783
## PC3 0.39930822 0.5901317 0.5080746 0.9614151
## PC4 0.62462210 0.6272957 0.3740827 0.5087056
## PC5 0.38407627 0.4320041 0.5853797 0.4999666
## PC6 0.09688161 0.2195614 0.3845099 0.3206170
```

- RNA concentration correlated with PC1, regress that out:

Regress out RNA concentration

```
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     32.24057 26.94929 24.62274 21.99025 19.02102
## Proportion of Variance  0.08956  0.06258  0.05224  0.04167  0.03117
## Cumulative Proportion   0.08956  0.15214  0.20438  0.24605  0.27722
##                             PC6      PC7      PC8      PC9     PC10
## Standard deviation     18.09446 16.74305 15.95586 15.23245 14.51328
## Proportion of Variance  0.02821  0.02415  0.02194  0.01999  0.01815
## Cumulative Proportion   0.30543  0.32959  0.35152  0.37151  0.38966
##                            PC11     PC12     PC13     PC14     PC15
## Standard deviation     13.76876 12.26209 11.99411 11.01768 10.44901
## Proportion of Variance  0.01633  0.01296  0.01240  0.01046  0.00941
## Cumulative Proportion   0.40600  0.41895  0.43135  0.44181  0.45122
##                            PC16     PC17     PC18     PC19     PC20
## Standard deviation     10.36537 9.634447 9.509499 9.443461 9.076617
## Proportion of Variance  0.00926 0.008000 0.007790 0.007680 0.007100
## Cumulative Proportion   0.46047 0.468470 0.476260 0.483950 0.491050
```

```
##              sex readsafterWsex rnaconc       rin      batch      prep
## PC1 0.8222275167     0.46005837       1 0.9050733 0.37271112 0.5183737
## PC2 0.0077107269     0.54730571       1 0.6510773 0.25005579 0.6712510
## PC3 0.0020620375     0.69784701       1 0.7185407 0.17724060 0.1888091
## PC4 0.2857982073     0.38567880       1 0.8260379 0.09532483 0.1506635
## PC5 0.0021804683     0.01978712       1 0.8826540 0.33167407 0.9765077
## PC6 0.0004101819     0.02852654       1 0.6678846 0.18417384 0.5608234
##          conc    length  flowlane     index
## PC1 0.4987273 0.3339316 0.5651998 0.2346049
## PC2 0.5675197 0.2767820 0.3021863 0.7276342
## PC3 0.6288393 0.7451663 0.6442899 0.9508174
## PC4 0.7770922 0.5492231 0.4025706 0.5390870
## PC5 0.5546334 0.5413124 0.5227272 0.4408842
## PC6 0.1799236 0.2571663 0.4427586 0.3622930
```

- Flowcell/lane next correlated covariate with PC3

Using ComBat to regress out Flowcell/lane

- Batch next correlated covariate

Using combat to regress out batch

- When Flowcell/Lane is regressed out, it seems that the covariate batch becomes signficantly correlated with PC3 and moreso than Flowcell/Lane was before it was regressed out.

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

So use my_data moving forward for now.