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

```
##
## 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
```

- There are a total of

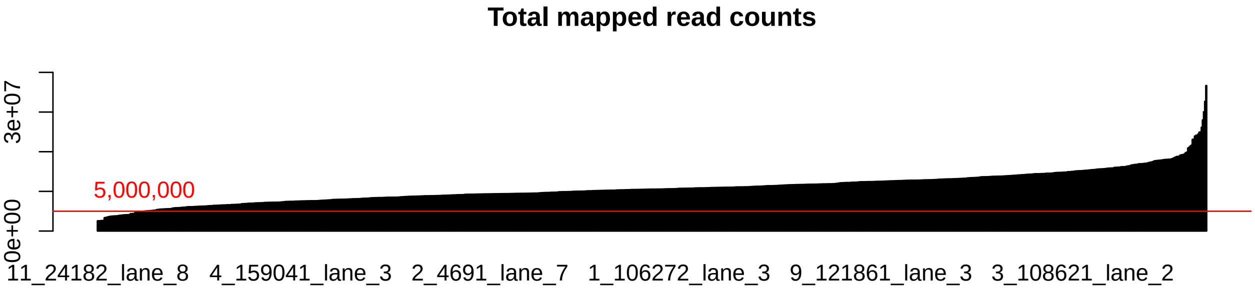
  $$genes and$$

  samples

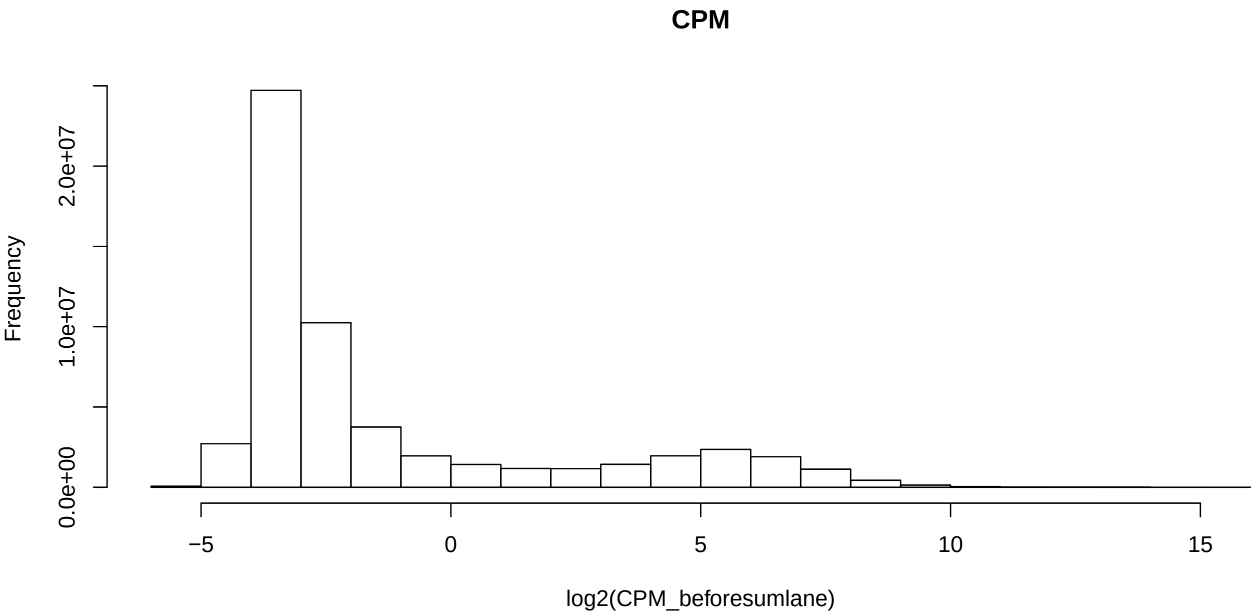
Covariates

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

Number of lanes with enough reads, before combining replicates



- The distribution of Counts Per Million:



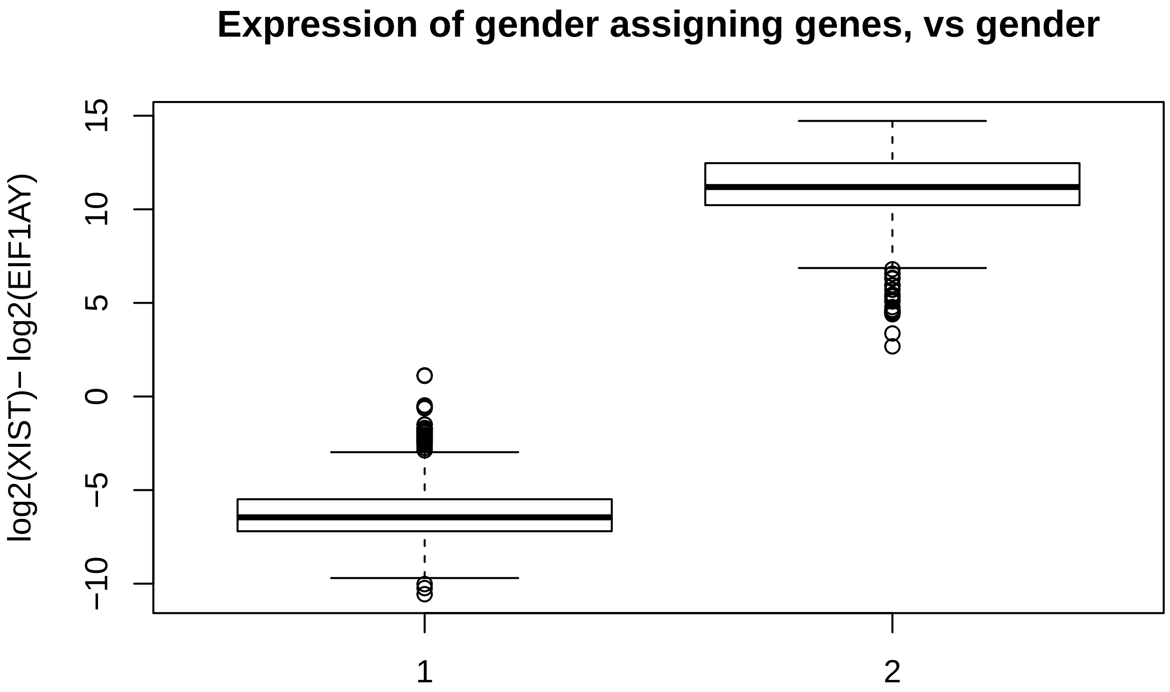
Sexcheck

Sex assigned by ratio of XIST to EIF1AY gene

```
## callSex
##   F   M
## 515 464
```

- According to expression of sex genes, there are 515 females and 464 males.

```
## gender
##   1   2
## 466 513
```



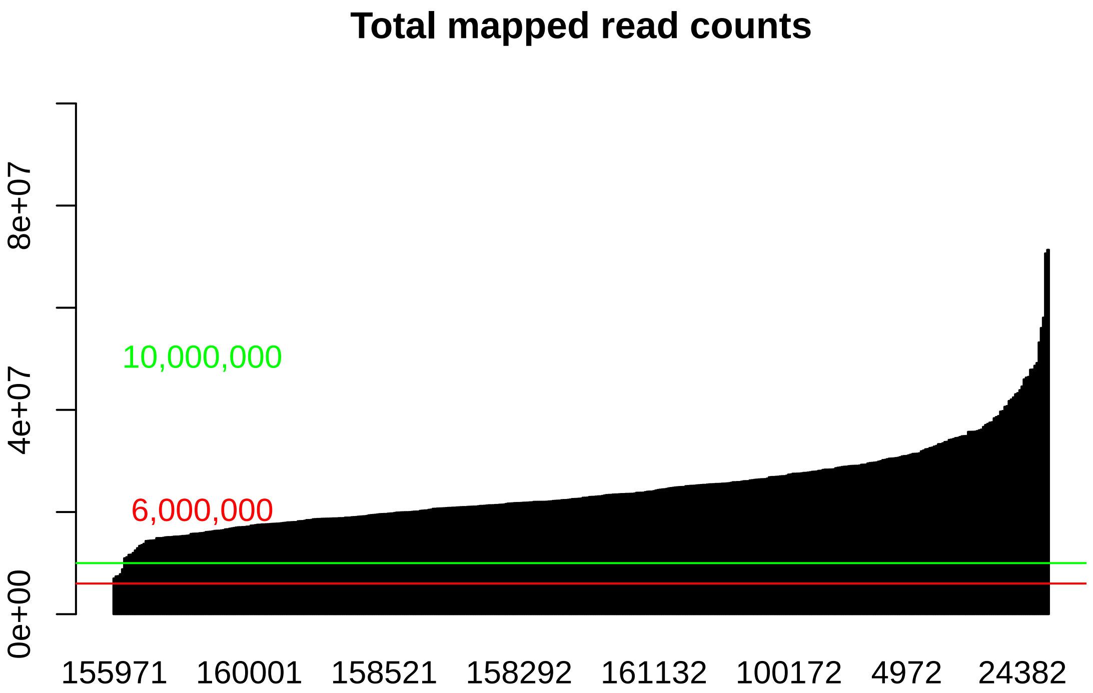
*

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

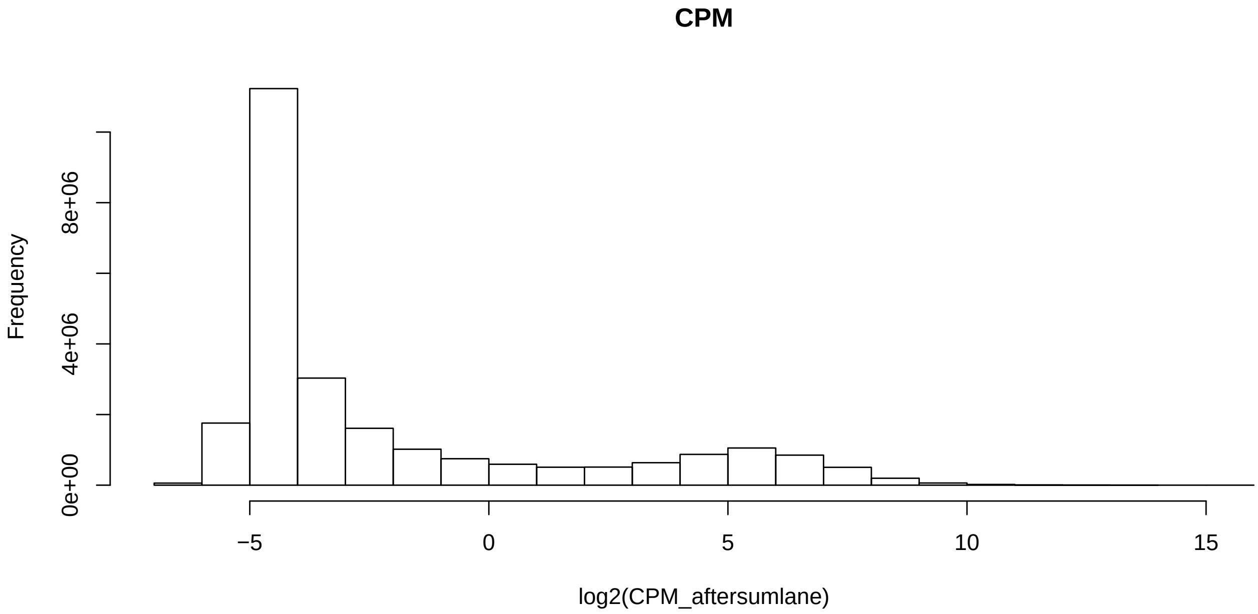
Combining technical replicates

- 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
## 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 437 have more than 10 million reads
- The distribution of Counts Per Million after combining replicates:

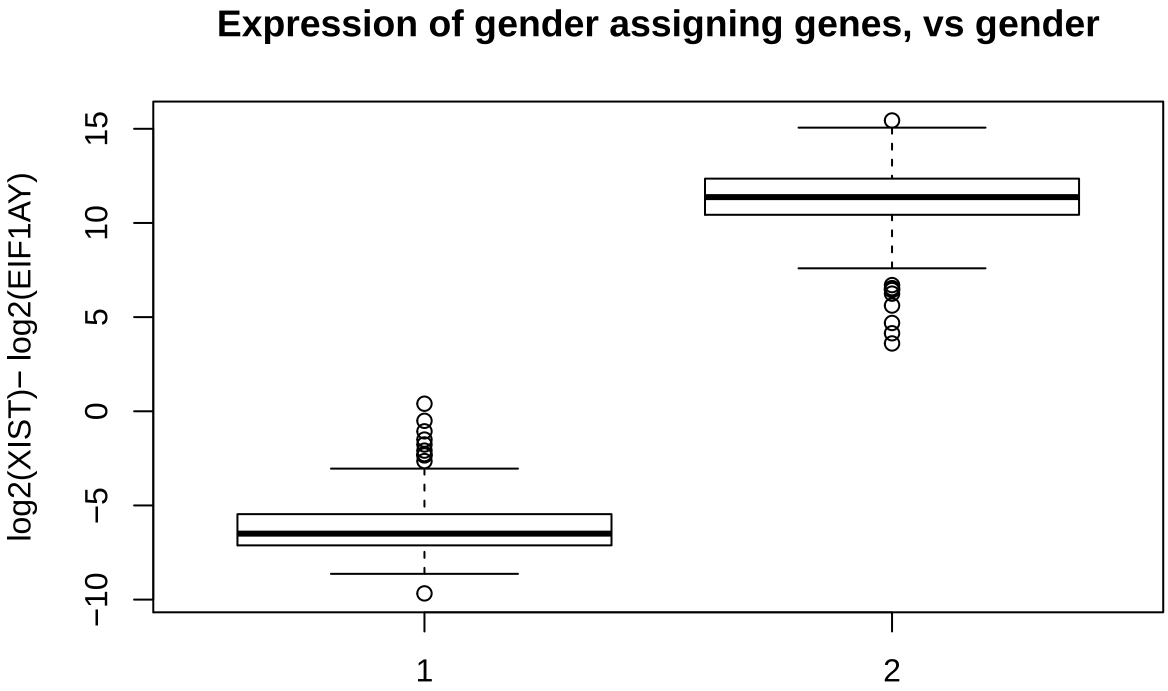


Checking sex after combining replicates

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

```
## callSex
##   F   M
## 228 209
```

```
## gender
##   1   2
## 210 227
```



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

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

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

- There are supposed to be: 228 females and 209 males.
- The samples misassigned are: 171351

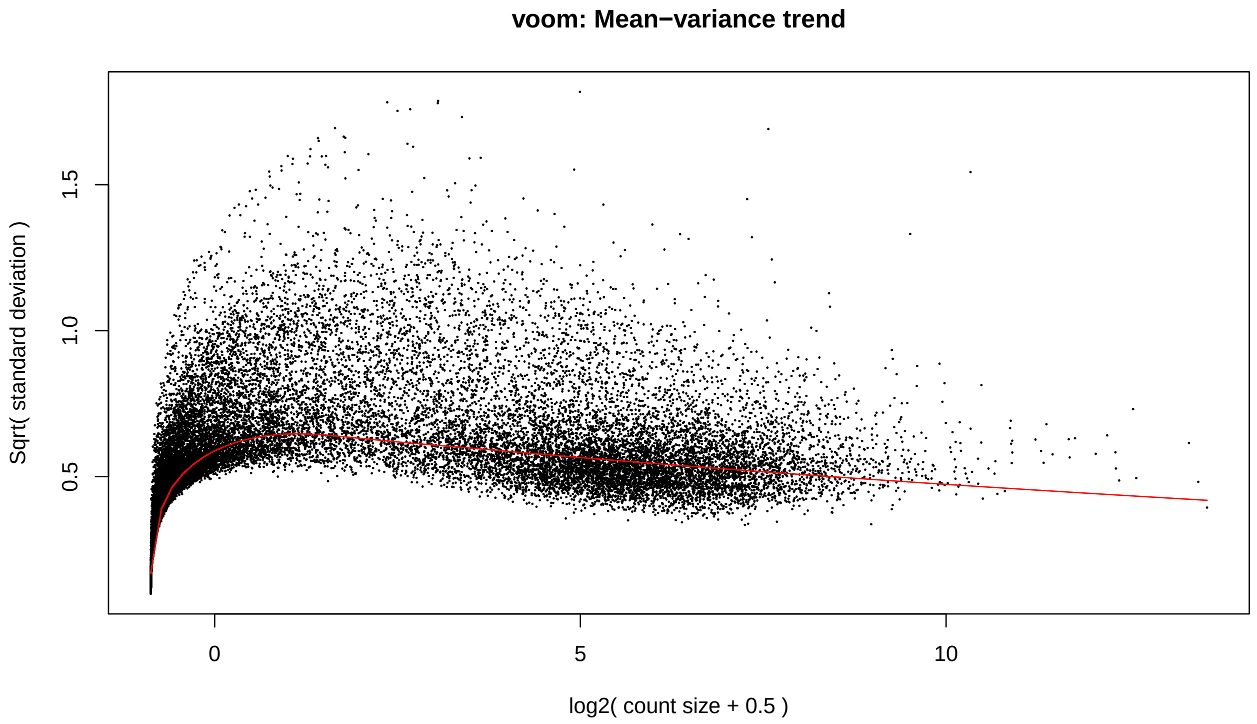
These 1 individuals have wrong assigned sex- last 0 have quite a large error - remove from data

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
## 39922
```

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

```
## [1] 16136   437
```

```
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)
```

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     35.41812 31.76757 27.94463 24.18587 22.24206
## Proportion of Variance  0.07939  0.06387  0.04942  0.03702  0.03131
## Cumulative Proportion   0.07939  0.14325  0.19267  0.22969  0.26100
##                             PC6      PC7      PC8      PC9     PC10
## Standard deviation     20.67450 18.72388 17.12134 16.90055 16.01524
## Proportion of Variance  0.02705  0.02219  0.01855  0.01808  0.01623
## Cumulative Proportion   0.28805  0.31023  0.32878  0.34686  0.36309
##                            PC11     PC12     PC13     PC14     PC15
## Standard deviation     15.24835 13.59474 13.32444 12.79806 11.85018
## Proportion of Variance  0.01471  0.01170  0.01124  0.01037  0.00889
## Cumulative Proportion   0.37781  0.38950  0.40074  0.41110  0.41999
##                            PC16     PC17     PC18     PC19     PC20
## Standard deviation     11.75609 11.22994 10.62009 10.40081 10.37510
## Proportion of Variance  0.00875  0.00798  0.00714  0.00685  0.00681
## Cumulative Proportion   0.42874  0.43672  0.44385  0.45070  0.45751
```

```
##              sex       indiv      rnaconc          rin      batch
## PC1 6.513265e-01 0.698869356 6.899276e-01 1.049767e-09 0.85742729
## PC2 1.790702e-01 0.046211615 8.402046e-05 1.973394e-01 0.48813150
## PC3 2.348008e-02 0.001571595 7.054750e-04 4.111750e-09 0.05346471
## PC4 6.147037e-01 0.001661544 5.291299e-01 7.791240e-04 0.03656515
## PC5 1.350091e-02 0.018622586 3.892536e-03 8.523410e-05 0.80270876
## PC6 1.081193e-05 0.486208479 6.052579e-01 3.849170e-03 0.89211387
##          prep      conc    length     flowlane     index
## PC1 0.8183122 0.4027546 0.3542952 0.3957010195 0.2957258
## PC2 0.3753408 0.9807508 0.6284207 0.8439696474 0.6322683
## PC3 0.1292098 0.4484404 0.5208400 0.4856957994 0.7374124
## PC4 0.1824927 0.8173726 0.2027166 0.0002287784 0.6409979
## PC5 0.7171645 0.1391021 0.3296896 0.0569535608 0.6367982
## PC6 0.6892299 0.4886875 0.6124205 0.6195475674 0.3199101
```

Regress out RIN

- PCA for the second time:

```
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     34.25595 31.70550 26.92064 23.85135 21.76247
## Proportion of Variance  0.07542  0.06461  0.04658  0.03657  0.03044
## Cumulative Proportion   0.07542  0.14004  0.18662  0.22318  0.25362
##                             PC6      PC7      PC8      PC9     PC10
## Standard deviation     20.31469 18.70891 17.10327 16.78235 16.01419
## Proportion of Variance  0.02653  0.02250  0.01880  0.01810  0.01648
## Cumulative Proportion   0.28015  0.30265  0.32145  0.33955  0.35603
##                            PC11     PC12     PC13     PC14     PC15
## Standard deviation     15.21384 13.54795 13.31083 12.64484 11.84076
## Proportion of Variance  0.01488  0.01180  0.01139  0.01028  0.00901
## Cumulative Proportion   0.37091  0.38271  0.39410  0.40437  0.41339
##                            PC16     PC17     PC18     PC19     PC20
## Standard deviation     11.51338 11.22992 10.51761 10.38763 10.34311
## Proportion of Variance  0.00852  0.00811  0.00711  0.00694  0.00688
## Cumulative Proportion   0.42191  0.43001  0.43712  0.44406  0.45093
```

```
##              sex        indiv      rnaconc rin      batch      prep
## PC1 7.332690e-01 0.9228922553 0.2798877440   1 0.48147412 0.5410789
## PC2 1.850910e-01 0.0268682574 0.0004069091   1 0.35323015 0.4626249
## PC3 4.010800e-03 0.0029163098 0.0060821410   1 0.25614628 0.2166499
## PC4 6.751760e-01 0.0008232266 0.3457257486   1 0.04112761 0.1151413
## PC5 4.187634e-03 0.1246094410 0.0047963656   1 0.60139140 0.9041386
## PC6 1.244163e-05 0.3898688371 0.8709066302   1 0.75911876 0.8014772
##            conc    length    flowlane     index
## PC1 0.41933408 0.3465555 0.307877222 0.2789148
## PC2 0.94951898 0.5472460 0.746255446 0.5513417
## PC3 0.50607671 0.7842689 0.077308963 0.8707139
## PC4 0.86788300 0.2435481 0.003388772 0.5819470
## PC5 0.06179794 0.1621325 0.021522970 0.5820186
## PC6 0.99111064 0.9650718 0.878589324 0.3296931
```

- RNA concentration correlated with PC1, regress that out:

Regress out RNA concentration

```
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     34.21924 31.29410 26.69079 23.82782 21.56696
## Proportion of Variance  0.07572  0.06333  0.04607  0.03671  0.03008
## Cumulative Proportion   0.07572  0.13905  0.18512  0.22183  0.25191
##                             PC6      PC7      PC8      PC9     PC10
## Standard deviation     20.31402 18.69467 16.98864 16.78042 15.98693
## Proportion of Variance  0.02668  0.02260  0.01866  0.01821  0.01653
## Cumulative Proportion   0.27860  0.30120  0.31986  0.33807  0.35459
##                            PC11     PC12     PC13     PC14     PC15
## Standard deviation     15.10747 13.50450 13.30088 12.64380 11.61790
## Proportion of Variance  0.01476  0.01179  0.01144  0.01034  0.00873
## Cumulative Proportion   0.36935  0.38115  0.39259  0.40293  0.41165
##                            PC16     PC17     PC18     PC19     PC20
## Standard deviation     11.45186 11.22151 10.50531 10.38557 10.33114
## Proportion of Variance  0.00848  0.00814  0.00714  0.00697  0.00690
## Cumulative Proportion   0.42013  0.42828  0.43541  0.44239  0.44929
```

```
##              sex        indiv rnaconc       rin      batch      prep
## PC1 7.295649e-01 0.8393292524       1 0.9019603 0.45853704 0.5754405
## PC2 1.445587e-01 0.0234410402       1 0.6454014 0.34943939 0.5026558
## PC3 9.330629e-04 0.0040824265       1 0.6977848 0.26017566 0.1914628
## PC4 8.673397e-01 0.0006536795       1 0.8966409 0.04149224 0.1231403
## PC5 1.297070e-02 0.0518054459       1 0.6908020 0.64120484 0.8910105
## PC6 1.233753e-05 0.3685228926       1 0.9799428 0.75916806 0.8030553
##          conc    length    flowlane     index
## PC1 0.4774958 0.3530477 0.340353191 0.2797756
## PC2 0.6676675 0.4744978 0.588291064 0.5120150
## PC3 0.7469614 0.9053290 0.035187577 0.9662151
## PC4 0.7604857 0.2139380 0.001576899 0.6065381
## PC5 0.1166778 0.2120782 0.081274212 0.5970024
## PC6 0.9949085 0.9629063 0.890712238 0.3313577
```

- 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.