RNA-seq

Sahar Mozaffari

August 8, 2016

R Markdown for RNA-seq data

Genecount matrix:

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

```
##             10_100092_lane_1 10_100092_lane_2 10_106052_lane_3
## 1/2-SBSRNA4                7                8                2
## A1BG                      69               72               33
## A1BG-AS1                  23               24               22
## A1CF                       0                0                0
## A2LD1                    102              121               59
```

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

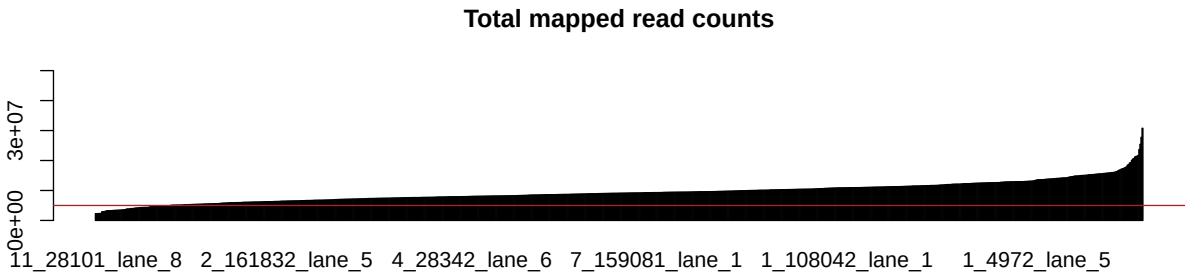
Covariates

- 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        11346119    69194    70271        1 106272    3
## 2 1_106272_lane_4        11193584    67396    69173        1 106272    4
## 3 1_106561_lane_7        12132310    78373    80493        1 106561    7
## 4 1_106561_lane_8        12066147    77896    80476        1 106561    8
## 5 1_106651_lane_5        10223689    62548    64463        1 106651    5
## 6 1_106651_lane_6        10374263    63660    65029        1 106651    6
##   Adaptor_index
## 1             8
## 2             8
## 3             9
## 4             9
## 5             2
## 6             2
```

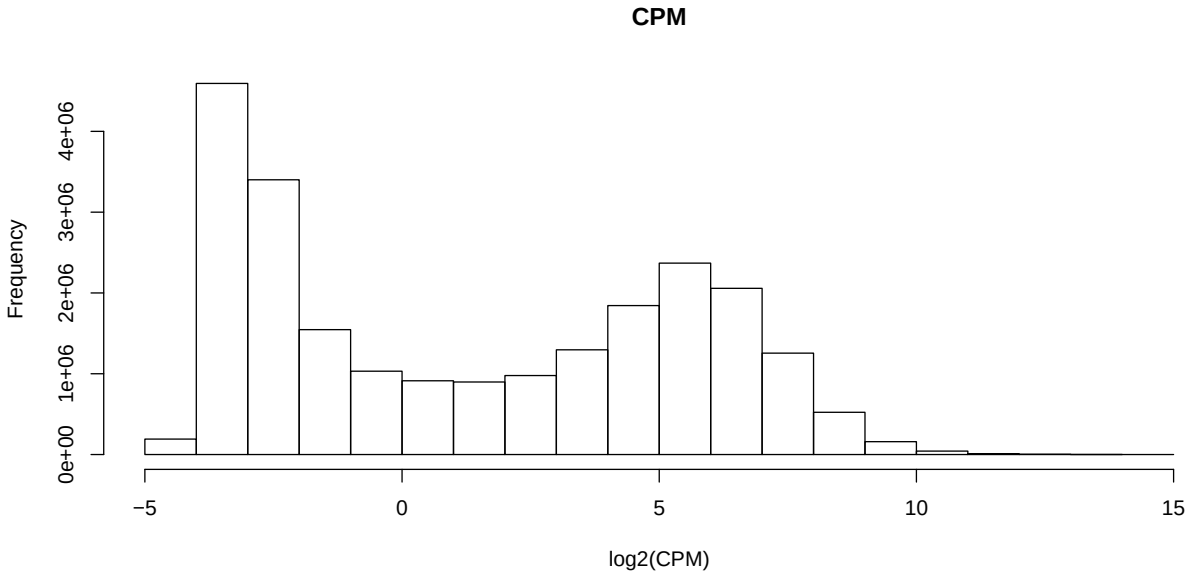
Number of lanes with enough reads, before combining replicates

```
## enough.reads
## FALSE  TRUE
##   355   634
```



* Before combining replicates, 634 out of total 989 have more than 10 million reads

- The distribution of Counts Per Million:



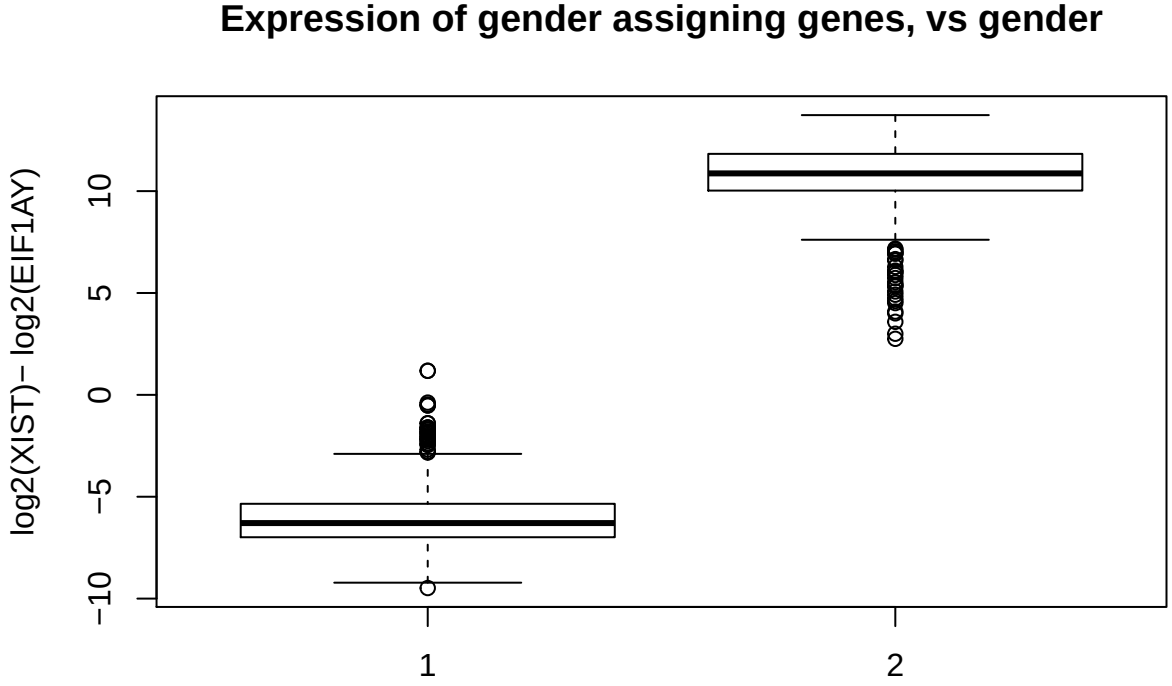
Sexcheck

Sex assigned by ratio of XIST to EIF1AY gene

```
## callSex
##   F   M
## 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: 1.1826263, 1.1776429

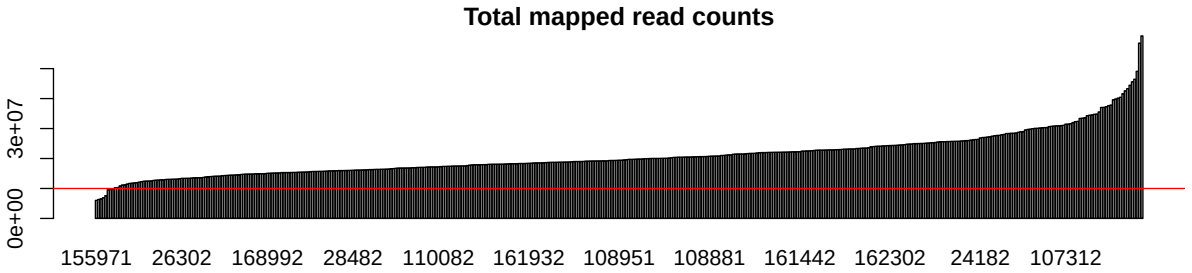
Combining technical replicates

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

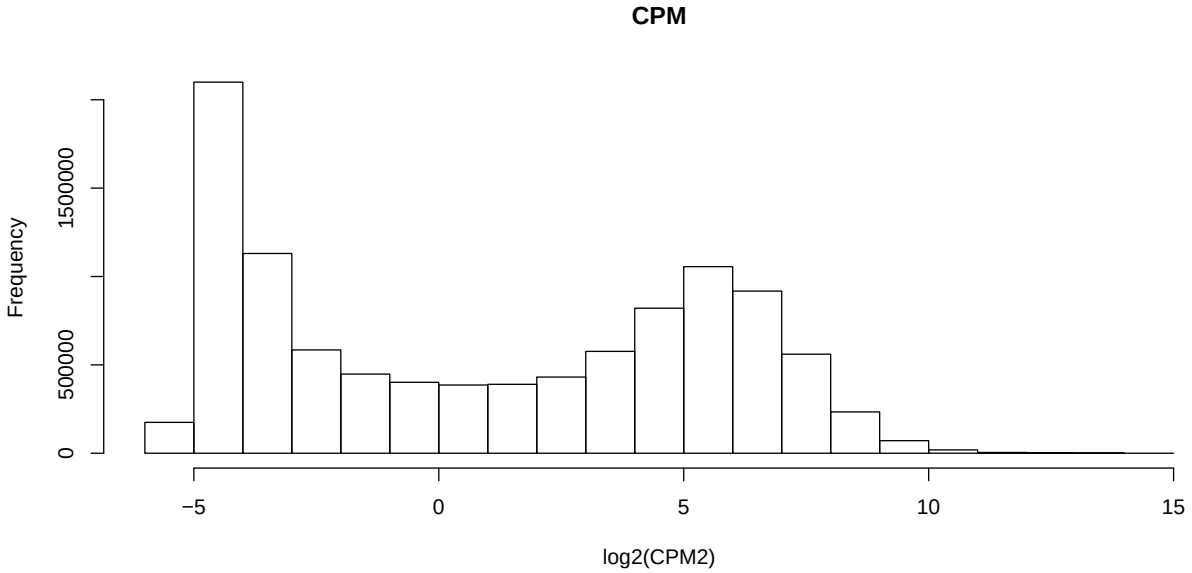
```
##              100092 100172 100182 100202 100372
## 1/2-SBSRNA4      15     16     21      7      9
## A1BG            141    160     85    139     87
## A1BG-AS1         47     49     91     47     33
## A1CF              0      0      1      1      1
## A2LD1           223    262    145    172    129
```

- combine total number of read covariate value

```
##        afterWASPwithXY  Unknown Maternal Paternal
## 100092        23229439 21672695    99014   104902
## 100172        29994465 27743032   118249   115154
## 100182        19029613 17717850    55757    56596
## 100202        22093953 20546423    60973    59554
## 100372        20354387 19046981    72203    71041
```



- The distribution of Counts Per Million after combining replicates:

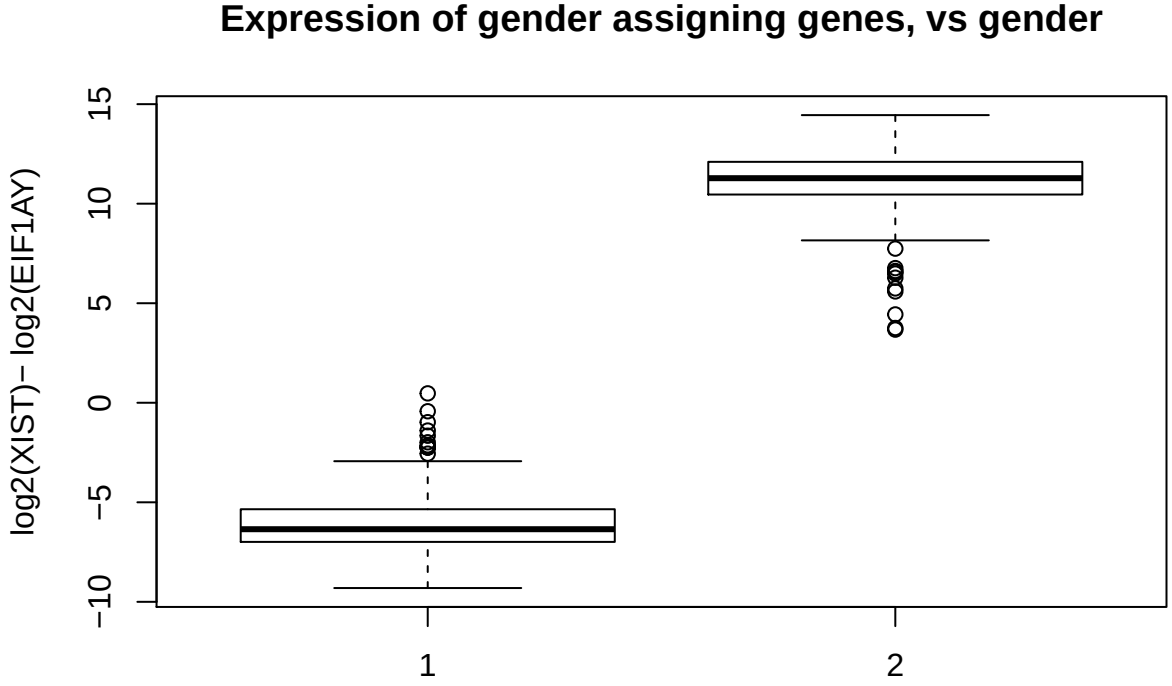


Checking sex after combining replicates

```
## 100092 100172 100182 100202 100372
##  11344  22376     64   9062   6623
```

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

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




```
##    171351
## 0.4716795
```

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

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

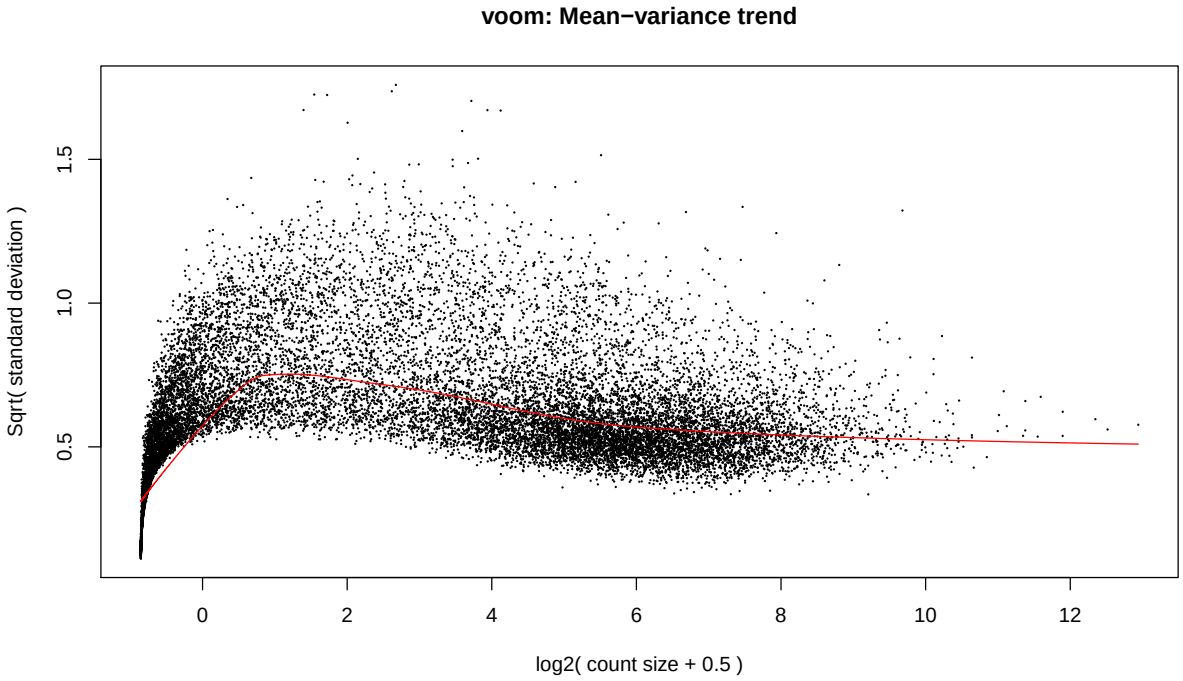
These 1 **individuals have wrong assigned sex- last** 0 **have quite a large error - remove from data**

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

- Total number of chromosome X genes: 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.



Covariates:

```
## Warning in cbind(uflowcells, c(1:96)): number of rows of result is not a
## multiple of vector length (arg 2)
```

```
##        sex readsafterWsex rnaconc  rin   batch    prep  conc length
## 100092   2       23229439   965.0  9.8 Batch_7     Amy  9.15    284
## 100172   2       29994465   192.0  9.2 Batch_6 Katelyn 14.49    295
## 100182   2       19029613   173.0  9.2 Batch_4 Katelyn 10.43    282
## 100202   2       22093953   835.1  9.6 Batch_3 Katelyn  4.78    282
## 100372   2       20354387   588.0 10.0 Batch_5 Katelyn 11.50    290
## 100582   2       21048363   191.0  9.2 Batch_4 Katelyn  2.55    270
##        flowlane
## 100092        1
## 100172        2
## 100182        3
## 100202        4
## 100372        5
## 100582        6
```

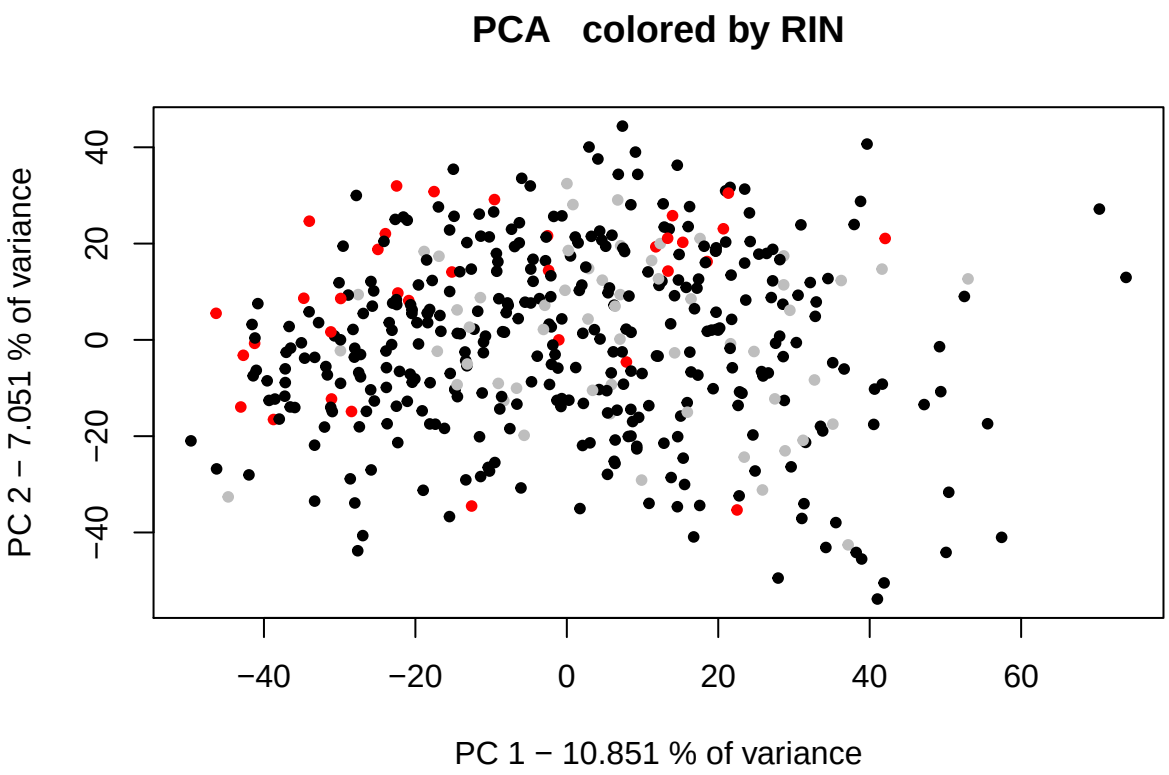
PCA:

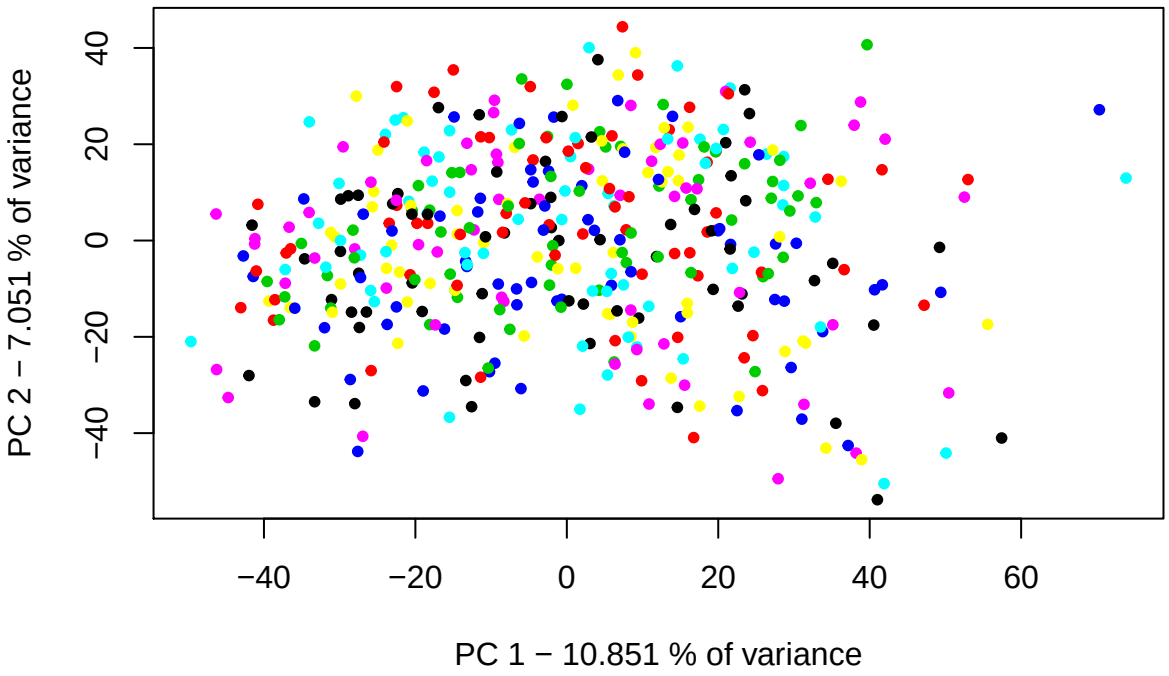
- First PCA showing variation of Proportion of Variance in PCs and correlation with covariates:

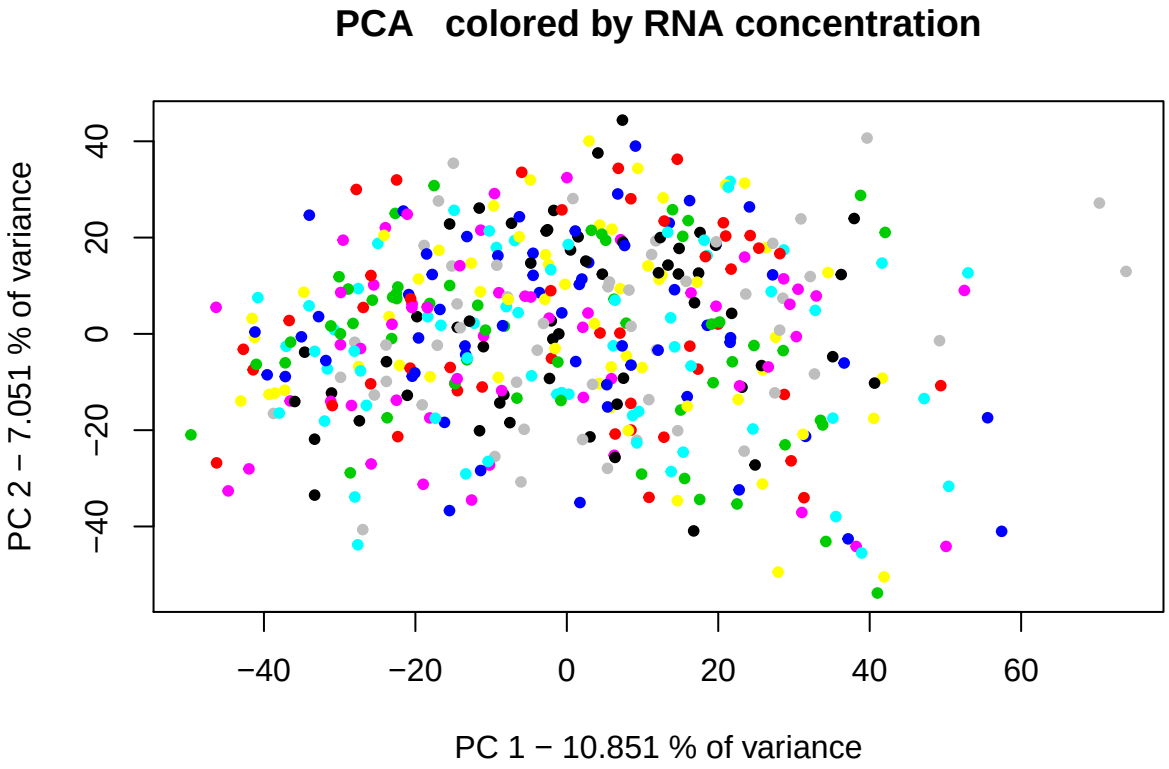
```
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     23.22190 18.71940 17.56678 14.29876 13.49912
## Proportion of Variance  0.10851  0.07051  0.06209  0.04114  0.03667
## Cumulative Proportion   0.10851  0.17902  0.24111  0.28225  0.31891
##                             PC6      PC7      PC8      PC9     PC10
## Standard deviation     13.09929 11.61324 11.18910 10.34799 9.977759
## Proportion of Variance  0.03453  0.02714  0.02519  0.02155 0.020030
## Cumulative Proportion   0.35344  0.38058  0.40577  0.42732 0.447350
##                            PC11     PC12     PC13    PC14     PC15
## Standard deviation     9.253744 8.531616 7.979752 7.79005 7.332295
## Proportion of Variance 0.017230 0.014650 0.012810 0.01221 0.010820
## Cumulative Proportion  0.464580 0.479220 0.492040 0.50425 0.515070
##                            PC16    PC17     PC18     PC19     PC20
## Standard deviation     7.222637 6.87586 6.434937 6.269148 6.216599
## Proportion of Variance 0.010500 0.00951 0.008330 0.007910 0.007780
## Cumulative Proportion  0.525560 0.53507 0.543410 0.551310 0.559090
```

```
##             sex readsafterWsex      rnaconc          rin        batch
## PC1 0.286074029     0.39604387 4.903692e-02 1.247020e-13 9.986023e-01
## PC2 0.007809376     0.69089613 2.616591e-07 3.338799e-02 6.136944e-02
## PC3 0.583834256     0.54250243 2.057221e-01 2.155387e-08 8.432566e-02
## PC4 0.024650421     0.58729348 6.699572e-02 4.827607e-01 7.517541e-01
## PC5 0.301539559     0.04201178 9.643179e-01 3.663313e-01 1.897829e-08
## PC6 0.418434663     0.47251266 6.885592e-02 3.561120e-04 7.598583e-05
##          prep      conc    length     flowlane
## PC1 0.9624044 0.7781436 0.4572799 8.774088e-01
## PC2 0.6144594 0.5789479 0.7354802 5.861031e-01
## PC3 0.1174324 0.2644392 0.1479965 3.872180e-01
## PC4 0.2013856 0.2955532 0.9618431 8.971290e-01
## PC5 0.8408432 0.8529469 0.4682677 4.080742e-08
## PC6 0.8553633 0.4582875 0.8694907 7.318871e-03
```

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


PCA colored by RIN
PC 2 − 7.051 % of variance
PC 1 − 10.851 % of variance
−40
−20
0
20
40
60


PCA colored by Batch
PC 2 − 7.051 % of variance
PC 1 − 10.851 % of variance
−40
−20
0
20
40
60



TMM Normalization

Regress out RIN

- PCA for the second time:

```
##                             PC1        PC2        PC3        PC4
## Standard deviation     3681.30810 2192.52808 1808.34515 1456.45960
## Proportion of Variance    0.33335    0.11825    0.08044    0.05218
## Cumulative Proportion     0.33335    0.45160    0.53204    0.58422
##                             PC5        PC6        PC7        PC8
## Standard deviation     1411.67120 1335.56741 1154.44538 1024.49488
## Proportion of Variance    0.04902    0.04388    0.03278    0.02582
## Cumulative Proportion     0.63324    0.67711    0.70989    0.73571
##                             PC9       PC10       PC11       PC12       PC13
## Standard deviation     895.77551 869.44138 835.48849 787.29861 708.28681
## Proportion of Variance   0.01974   0.01859   0.01717   0.01525   0.01234
## Cumulative Proportion    0.75545   0.77404   0.79121   0.80646   0.81880
##                            PC14       PC15       PC16       PC17       PC18
## Standard deviation     645.87925 607.14333 552.19778 529.90022 498.83002
## Proportion of Variance   0.01026   0.00907   0.00750   0.00691   0.00612
## Cumulative Proportion    0.82906   0.83813   0.84563   0.85254   0.85866
##                            PC19       PC20
## Standard deviation     484.13095 450.90439
## Proportion of Variance   0.00577   0.00500
## Cumulative Proportion    0.86442   0.86942
```

```
##             sex readsafterWsex      rnaconc rin        batch       prep
## PC1 0.710466221      0.3450285 0.0003671717   1 1.746323e-01 0.44018256
## PC2 0.846229527      0.4088719 0.0421882736   1 6.186388e-01 0.94162516
## PC3 0.514307153      0.0413807 0.1380579300   1 6.790165e-08 0.73027238
## PC4 0.981588132      0.2191356 0.0278504470   1 8.626355e-15 0.00261069
## PC5 0.005242481      0.2015974 0.3991259013   1 7.592884e-01 0.60520490
## PC6 0.141692103      0.9475836 0.1323932826   1 1.168121e-01 0.23076313
##           conc    length     flowlane
## PC1 0.09693398 0.8637653 5.576591e-01
## PC2 0.83773523 0.4974114 4.176807e-02
## PC3 0.46418591 0.1236003 9.054259e-11
## PC4 0.02854740 0.2113875 4.888440e-05
## PC5 0.62423325 0.7407385 5.107836e-01
## PC6 0.05094636 0.5702389 2.567381e-03
```

- RNA concentration correlated with PC1, regress that out:

Regress out RNA concentration

```
##                             PC1        PC2        PC3        PC4
## Standard deviation     3629.34542 2182.26290 1803.81088 1448.91329
## Proportion of Variance    0.32869    0.11883    0.08119    0.05239
## Cumulative Proportion     0.32869    0.44752    0.52871    0.58110
##                             PC5        PC6        PC7        PC8
## Standard deviation     1410.36949 1331.93569 1153.01423 1017.94550
## Proportion of Variance    0.04964    0.04427    0.03317    0.02586
## Cumulative Proportion     0.63073    0.67500    0.70817    0.73403
##                             PC9      PC10      PC11      PC12      PC13
## Standard deviation     895.72684 860.77801 830.31297 776.04406 704.81270
## Proportion of Variance   0.02002   0.01849   0.01720   0.01503   0.01240
## Cumulative Proportion    0.75405   0.77254   0.78974   0.80477   0.81717
##                            PC14      PC15      PC16      PC17      PC18
## Standard deviation     643.30771 606.31116 550.15652 529.63711 498.81612
## Proportion of Variance   0.01033   0.00917   0.00755   0.00700   0.00621
## Cumulative Proportion    0.82749   0.83667   0.84422   0.85122   0.85743
##                            PC19      PC20
## Standard deviation     483.49440 448.63556
## Proportion of Variance   0.00583   0.00502
## Cumulative Proportion    0.86326   0.86828
```

```
##             sex readsafterWsex rnaconc       rin        batch        prep
## PC1 0.530485825     0.41668478       1 0.6070078 1.270879e-01 0.366464875
## PC2 0.725276482     0.35160573       1 0.7670745 7.308082e-01 0.952769878
## PC3 0.583138477     0.05036608       1 0.8283371 1.415403e-08 0.720061287
## PC4 0.957926896     0.27547102       1 0.7628347 7.428174e-14 0.002996456
## PC5 0.003983229     0.17807646       1 0.8889621 8.229442e-01 0.703819844
## PC6 0.189210460     0.98656596       1 0.8169436 9.256497e-02 0.179699174
##           conc    length     flowlane
## PC1 0.19388306 0.9640716 5.942457e-01
## PC2 0.67206727 0.4403754 2.368434e-02
## PC3 0.58606316 0.1373546 9.431097e-11
```

```
## PC4 0.05300606 0.1880555 1.204088e-04
## PC5 0.63336248 0.6349945 5.159995e-01
## PC6 0.06145174 0.6588725 1.876998e-03
```

- Flowcell/lane next correlated covariate with PC3

Using ComBat to regress out Flowcell/lane

```
## Found 96 batches
## Note: one batch has only one sample, setting mean.only=TRUE
## Adjusting for 0 covariate(s) or covariate level(s)
## Standardizing Data across genes
## Fitting L/S model and finding priors
## Finding parametric adjustments
## Adjusting the Data
```

```
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     121.45277 16.86813 14.54441 13.00434 12.64189
## Proportion of Variance   0.73071  0.01409  0.01048  0.00838  0.00792
## Cumulative Proportion    0.73071  0.74480  0.75528  0.76366  0.77157
##                             PC6      PC7      PC8      PC9     PC10
## Standard deviation     10.86140 10.47591 9.874994 9.490985 8.929503
## Proportion of Variance  0.00584  0.00544 0.004830 0.004460 0.003950
## Cumulative Proportion   0.77742  0.78285 0.787690 0.792150 0.796100
##                            PC11     PC12     PC13     PC14     PC15
## Standard deviation     8.575337 8.304656 8.010527 7.636989 7.353862
## Proportion of Variance 0.003640 0.003420 0.003180 0.002890 0.002680
## Cumulative Proportion  0.799740 0.803160 0.806340 0.809220 0.811900
##                            PC16     PC17    PC18     PC19     PC20
## Standard deviation     6.782836 6.557206 6.50384 6.315012 6.118709
## Proportion of Variance 0.002280 0.002130 0.00210 0.001980 0.001850
## Cumulative Proportion  0.814180 0.816310 0.81841 0.820380 0.822240
```

```
##            sex readsafterWsex   rnaconc       rin        batch      prep
## PC1 0.03134467     0.94425520 0.7824299 0.7463571 9.408932e-01 0.6438475
## PC2 0.18717626     0.08552335 0.9566704 0.8391850 9.058382e-01 0.7363298
## PC3 0.69958322     0.40974307 0.8796377 0.8172939 1.182476e-12 0.4509422
## PC4 0.10712253     0.60970331 0.9046164 0.9908486 1.999514e-01 0.3418424
## PC5 0.05193530     0.89498456 0.8893693 0.2915825 3.349313e-01 0.4750463
## PC6 0.14399333     0.69478731 0.9992082 0.9112996 2.067818e-01 0.1537456
##          conc    length    flowlane
## PC1 0.4343443 0.8171893 1.000000000
## PC2 0.9474869 0.4968486 1.000000000
## PC3 0.6894773 0.1221366 0.006195037
## PC4 0.1258602 0.4929055 0.999988370
## PC5 0.7658362 0.3754967 0.999896779
## PC6 0.1133086 0.6581632 0.999999998
```

- Batch next correlated covariate

Using combat to regress out batch

```
## Found 7 batches
## Adjusting for 0 covariate(s) or covariate level(s)
## Standardizing Data across genes
## Fitting L/S model and finding priors
## Finding parametric adjustments
## Adjusting the Data
```

```
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     122.11916 17.00457 13.88136 13.11797 12.68923
## Proportion of Variance   0.73875  0.01432  0.00955  0.00852  0.00798
## Cumulative Proportion    0.73875  0.75307  0.76262  0.77114  0.77912
##                             PC6      PC7     PC8      PC9     PC10
## Standard deviation     10.88160 10.48145 9.93665 9.228583 8.889343
## Proportion of Variance  0.00587  0.00544 0.00489 0.004220 0.003910
## Cumulative Proportion   0.78498  0.79042 0.79532 0.799530 0.803450
##                            PC11     PC12     PC13     PC14    PC15
## Standard deviation     8.422005 7.997646 7.514267 7.455482 6.90154
## Proportion of Variance 0.003510 0.003170 0.002800 0.002750 0.00236
## Cumulative Proportion  0.806960 0.810130 0.812930 0.815680 0.81804
##                            PC16     PC17     PC18     PC19     PC20
## Standard deviation     6.680149 6.338149 6.152824 6.066591 5.834355
## Proportion of Variance 0.002210 0.001990 0.001880 0.001820 0.001690
## Cumulative Proportion  0.820250 0.822240 0.824120 0.825940 0.827630
```

```
##             sex readsafterWsex   rnaconc       rin     batch      prep
## PC1 0.02865202      0.4369847 0.4884797 0.1492605 0.9993174 0.8892796
## PC2 0.18163941      0.1041112 0.9662802 0.7836372 0.9998743 0.7269213
## PC3 0.87484461      0.1190649 0.6514550 0.7582721 0.6801085 0.5162837
## PC4 0.03815123      0.9067845 0.9216612 0.6349610 0.8797597 0.3676187
## PC5 0.11391904      0.6487265 0.9023712 0.3344888 0.9703151 0.4360296
## PC6 0.12212617      0.9514518 0.8746134 0.9306397 0.9988935 0.3343205
##          conc     length  flowlane
## PC1 0.4576944 0.99810711 1.0000000
## PC2 0.8439857 0.55809632 1.0000000
## PC3 0.2480309 0.05479392 0.2955241
## PC4 0.2270133 0.90603763 0.9999144
## PC5 0.3068196 0.45288684 0.9970049
## PC6 0.2971924 0.81007726 1.0000000
```

- 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.

```
##  [1] "106451" "106581" "106742" "108211" "159521" "160462" "160591"
```

```
##  [8] "161011" "162112" "173492"
```

```
##  num [1:20187, 1:431] 0 0 0 0 0 0 0 0 2 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:20187] "1/2-SBSRNA4" "A1BG" "A1BG-AS1" "A1CF" ...
##   ..$ : chr [1:431] "100092" "106052" "106092" "106202" ...
```

```
## [1] 20187   431
```

```
##  num [1:20187, 1:431] 15 141 47 0 223 6 0 0 301 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:20187] "1/2-SBSRNA4" "A1BG" "A1BG-AS1" "A1CF" ...
##   ..$ : chr [1:431] "100092" "100172" "100182" "100202" ...
```

```
## [1] 20187   431
```