Quantile normalization

Jeff Leek

Dependencies

This document depends on the following packages:

```
library(devtools)
library(Biobase)
library(preprocessCore)
```

To install these packages you can use the code (or if you are compiling the document, remove the `eval=FALSE` from the chunk.)

```
install.packages(c("devtools"))
source("http://www.bioconductor.org/biocLite.R")
biocLite(c("Biobase","preprocessCore"))
```

General principles

- Preprocessing and normalization take raw data and turn it into processed data
- These techniques are highly application specific
- I'll illustrate a very general technique here (quantile normalization (https://en.wikipedia.org/wiki/Quantile_normalization))
- Then I'll show one or two processing examples for specific data types

Load some data

We will use this expression set that combines two studies Transcriptome genetics using second generation sequencing in a Caucasian population. (http://www.ncbi.nlm.nih.gov/pubmed?term=20220756%5Buid%5D) and Understanding mechanisms underlying human gene expression variation with RNA sequencing. (http://www.ncbi.nlm.nih.gov/pubmed?term=20220758). These studies are different populations but we counted the same genes for both. Then we'll explore the differences.

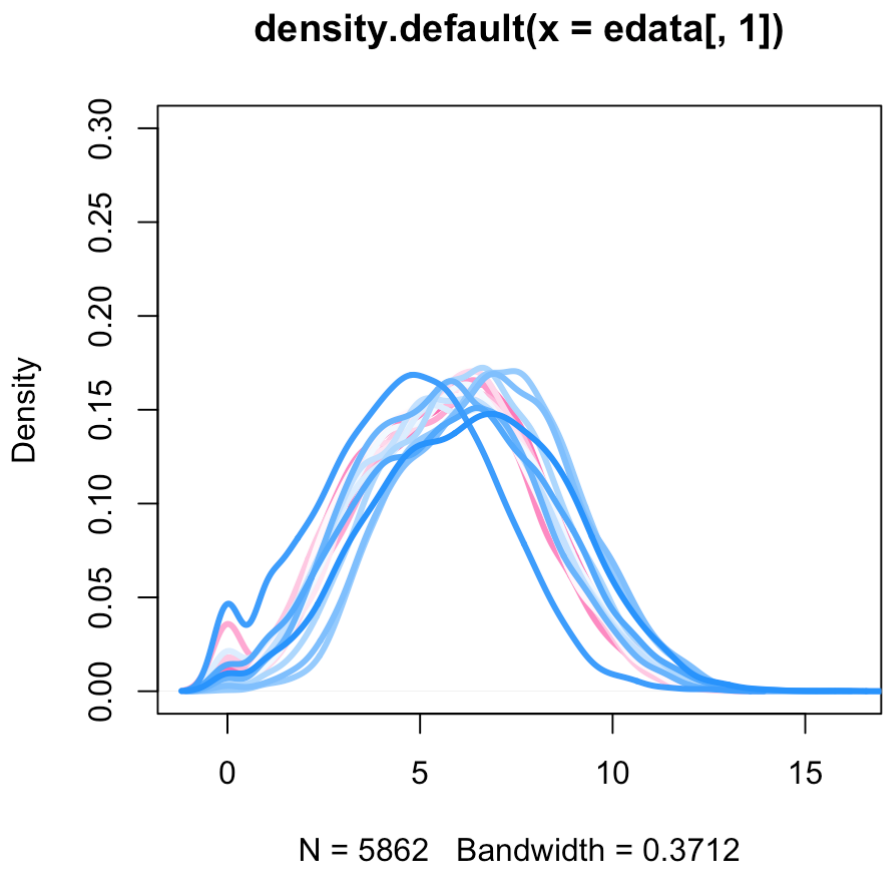
```
con =url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/montpick_eset.RData")
load(file=con)
close(con)
mp = montpick.eset
pdata=pData(mp)
edata=as.data.frame(exprs(mp))
fdata = fData(mp)
ls()
```

```
## [1] "con"           "edata"         "fdata"         "montpick.eset"
## [5] "mp"            "pdata"         "tropical"
```

Show distributions for log2 counts for several samples

Here we show density plots for the first 20 samples

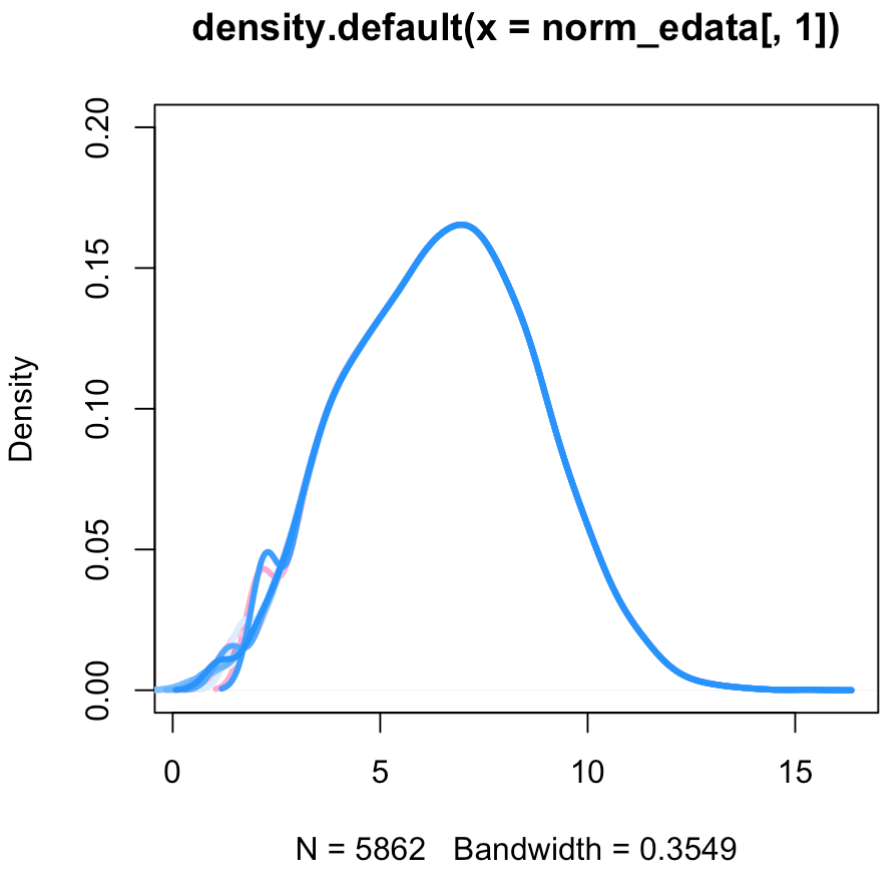
```
edata = log2(edata + 1)
edata = edata[rowMeans(edata) > 3, ]
colramp = colorRampPalette(c(3,"white",2))(20)
plot(density(edata[,1]),col=colramp[1],lwd=3,ylim=c(0,.30))
for(i in 2:20){lines(density(edata[,i]),lwd=3,col=colramp[i])}
```



Quantile normalization

Now we perform quantile normalization to make the distributions the same across samples. Note that near the tail the distributions aren't perfectly the same, but for the most part the distributions land right on top of each other.

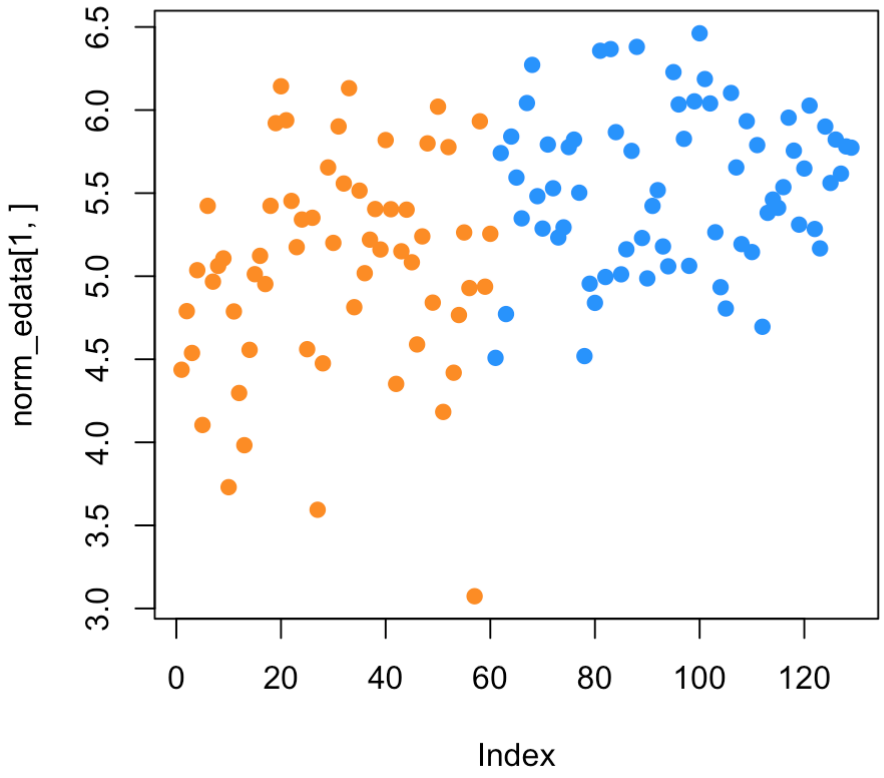
```
norm_edata = normalize.quantiles(as.matrix(edata))
plot(density(norm_edata[,1]),col=colramp[1],lwd=3,ylim=c(0,.20))
for(i in 2:20){lines(density(norm_edata[,i]),lwd=3,col=colramp[i])}
```



Matching distributions leaves variability

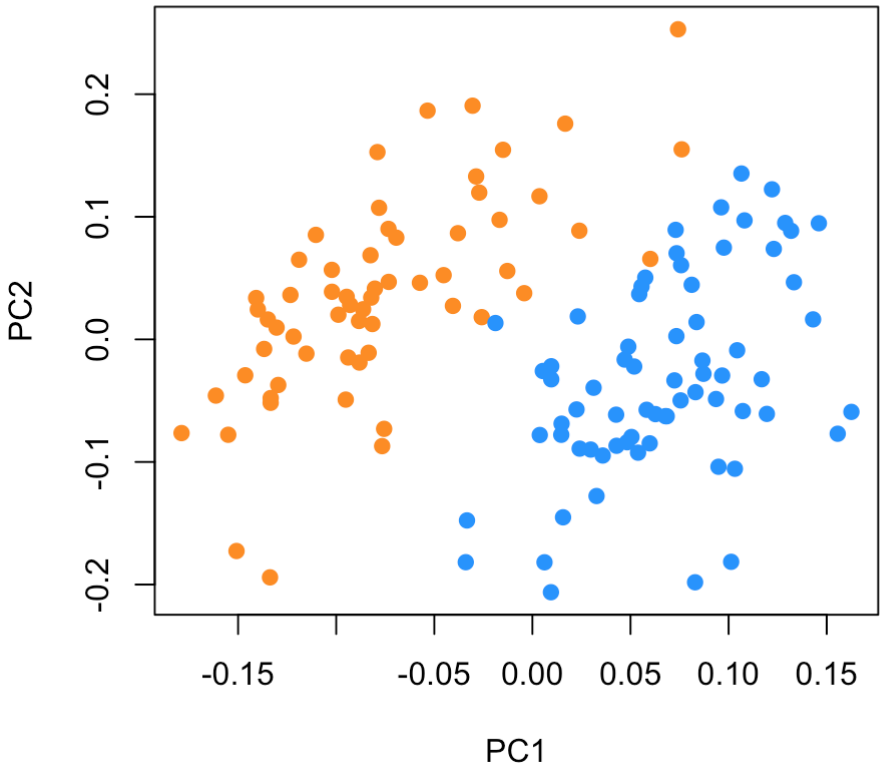
Normalization removes bulk differences due to technology. But there still may be differences you don't want after normalization. The only way to figure this out is to check. For example if we plot the quantile normalized data with the first

```
plot(norm_edata[1,],col=as.numeric(pdata$study))
```



You can see this in that the principal components still reflect variation between studies

```
svd1 = svd(norm_edata - rowMeans(norm_edata))
plot(svd1$v[,1],svd1$v[,2],xlab="PC1",ylab="PC2",
     col=as.numeric(pdata$study))
```



Further resources

Preprocessing and normalization are highly application specific. Here I'm going to point you to resources for several specific types of data sets that you can follow.

- Preprocessing gene expression microarrays
- Affymetrix - affy (http://www.bioconductor.org/packages/release/bioc/html/affy.html)
- Illumina - lumi (http://www.bioconductor.org/packages/release/bioc/html/lumi.html)
- Preprocessing methylation microarray data
- Illumina 450k - minfi (http://bioconductor.org/packages/release/bioc/html/minfi.html)
- Preprocessing RNA-seq data
- Gene count based models - Rsubread (http://bioconductor.org/packages/release/bioc/html/Rsubread.html), cqn (http://www.bioconductor.org/packages/release/bioc/html/cqn.html), edgeR User's guide (http://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf)
- Preprocessing Chip-seq data
- Peaks - DiffBind (http://bioconductor.org/packages/release/bioc/html/DiffBind.html)
- Preprocessing variant data
- Often not done in R, mostly for computational/historical reasons, but VariantTools (http://bioconductor.org/packages/release/bioc/html/VariantTools.html) is good.

Session information

Here is the session information

```
devtools::session_info()
```

```
##  setting  value
##  version  R version 3.2.1 (2015-06-18)
##  system   x86_64, darwin10.8.0
##  ui       RStudio (0.99.447)
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##
##  package           * version     date
##  acepack             1.3-3.3     2014-11-24
##  annotate            1.46.1      2015-07-11
##  AnnotationDbi     * 1.30.1      2015-04-26
##  assertthat          0.1         2013-12-06
##  BiasedUrn         * 1.06.1      2013-12-29
##  Biobase           * 2.28.0      2015-04-17
##  BiocGenerics      * 0.14.0      2015-04-17
##  BiocInstaller     * 1.18.4      2015-07-22
##  BiocParallel        1.2.20      2015-08-07
##  biomaRt             2.24.0      2015-04-17
##  Biostrings          2.36.3      2015-08-12
##  bitops              1.0-6       2013-08-17
##  bladderbatch      * 1.6.0       2015-08-26
##  broom             * 0.3.7       2015-05-06
##  caTools             1.17.1      2014-09-10
##  cluster             2.0.3       2015-07-21
##  colorspace          1.2-6       2015-03-11
##  corpcor             1.6.8       2015-07-08
##  curl                0.9.2       2015-08-08
##  DBI               * 0.3.1       2014-09-24
##  dendextend        * 1.1.0       2015-07-31
##  DESeq2            * 1.8.1       2015-05-02
##  devtools          * 1.8.0       2015-05-09
##  digest              0.6.8       2014-12-31
##  dplyr             * 0.4.3       2015-09-01
##  edge              * 2.1.0       2015-09-06
##  evaluate            0.7.2       2015-08-13
##  foreign             0.8-66      2015-08-19
##  formatR             1.2         2015-04-21
##  Formula           * 1.2-1       2015-04-07
##  futile.logger       1.4.1       2015-04-20
##  futile.options      1.0.0       2010-04-06
##  gdata               2.17.0      2015-07-04
##  genefilter        * 1.50.0      2015-04-17
##  geneLenDataBase   * 1.4.0       2015-09-06
##  geneplotter         1.46.0      2015-04-17
##  GenomeInfoDb      * 1.4.2       2015-08-15
##  GenomicAlignments   1.4.1       2015-04-24
##  GenomicFeatures     1.20.2      2015-08-14
##  GenomicRanges     * 1.20.5      2015-06-09
##  genstats          * 0.1.02      2015-09-05
##  ggplot2           * 1.0.1       2015-03-17
##  git2r               0.11.0      2015-08-12
##  GO.db               3.1.2       2015-09-06
```

```
##  goseq             * 1.20.0      2015-04-17
##  gplots            * 2.17.0      2015-05-02
##  gridExtra           2.0.0       2015-07-14
##  gtable              0.1.2       2012-12-05
##  gtools              3.5.0       2015-05-29
##  highr               0.5         2015-04-21
##  HistData          * 0.7-5       2014-04-26
##  Hmisc             * 3.16-0      2015-04-30
##  htmltools           0.2.6       2014-09-08
##  httr                1.0.0       2015-06-25
##  IRanges           * 2.2.7       2015-08-09
##  KernSmooth          2.23-15     2015-06-29
##  knitr             * 1.11        2015-08-14
##  lambda.r            1.1.7       2015-03-20
##  lattice           * 0.20-33     2015-07-14
##  latticeExtra        0.6-26      2013-08-15
##  lazyeval            0.1.10      2015-01-02
##  limma             * 3.24.15     2015-08-06
##  lme4                1.1-9       2015-08-20
##  locfit              1.5-9.1     2013-04-20
##  magrittr            1.5         2014-11-22
##  MASS              * 7.3-43      2015-07-16
##  Matrix            * 1.2-2       2015-07-08
##  MatrixEQTL        * 2.1.1       2015-02-03
##  memoise             0.2.1       2014-04-22
##  mgcv              * 1.8-7       2015-07-23
##  minqa               1.2.4       2014-10-09
##  mnormt              1.5-3       2015-05-25
##  munsell             0.4.2       2013-07-11
##  nlme              * 3.1-122     2015-08-19
##  nloptr              1.0.4       2014-08-04
##  nnet                7.3-10      2015-06-29
##  org.Hs.eg.db      * 3.1.2       2015-07-17
##  plyr                1.8.3       2015-06-12
##  preprocessCore    * 1.30.0      2015-04-17
##  proto               0.3-10      2012-12-22
##  psych               1.5.6       2015-07-08
##  qvalue            * 2.0.0       2015-04-17
##  R6                  2.1.1       2015-08-19
##  RColorBrewer        1.1-2       2014-12-07
##  Rcpp              * 0.12.0      2015-07-25
##  RcppArmadillo     * 0.5.400.2.0 2015-08-17
##  RCurl               1.95-4.7    2015-06-30
##  reshape2            1.4.1       2014-12-06
##  rmarkdown           0.7         2015-06-13
##  rpart               4.1-10      2015-06-29
##  Rsamtools           1.20.4      2015-06-01
##  RSkittleBrewer    * 1.1         2015-09-05
##  RSQLite           * 1.0.0       2014-10-25
##  rstudioapi          0.3.1       2015-04-07
##  rtracklayer         1.28.9      2015-08-19
##  rversions           1.0.2       2015-07-13
##  S4Vectors         * 0.6.5       2015-09-01
##  scales              0.3.0       2015-08-25
```

```
##  snm                 1.16.0      2015-04-17
##  snpStats          * 1.18.0      2015-04-17
##  stringi             0.5-5       2015-06-29
##  stringr             1.0.0       2015-04-30
##  survival          * 2.38-3      2015-07-02
##  sva               * 3.14.0      2015-04-17
##  tidyr               0.2.0       2014-12-05
##  UsingR            * 2.0-5       2015-08-06
##  whisker             0.3-2       2013-04-28
##  XML                 3.98-1.3    2015-06-30
##  xml2                0.1.2       2015-09-01
##  xtable              1.7-4       2014-09-12
##  XVector             0.8.0       2015-04-17
##  yaml                2.1.13      2014-06-12
##  zlibbioc            1.14.0      2015-04-17
##  source
##  CRAN (R 3.2.0)
##  Bioconductor
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  CRAN (R 3.2.0)
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.2)
##  Github (jdstorey/edge@a1947b5)
##  CRAN (R 3.2.2)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  Bioconductor
```

```
##  Bioconductor
##  Bioconductor
##  local
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.2)
##  Bioconductor
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.2)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  Bioconductor
##  CRAN (R 3.2.2)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.1)
##  Bioconductor
##  Github (alyssafrazee/RSkittleBrewer@0a96a20)
```

```
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  Bioconductor
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.2)
##  Bioconductor
##  Bioconductor
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.2)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.2)
##  CRAN (R 3.2.0)
##  Bioconductor
##  CRAN (R 3.2.0)
##  Bioconductor
```

It is also useful to compile the time the document was processed. This document was processed on: 2015-09-06.