The qsmooth user's guide

Kwame Okrah okrah.kwame@gene.com Stephanie C. Hicks shicks@jimmy.harvard.edu
Hector Corrado Bravo hcorrada@gmail.com Rafael A. Irizarry rafa@jimmy.harvard.edu

Modified: March 5, 2015. Compiled: December 7, 2015

Contents

1 Introduction

Normalization strategies that are based solely on the observed data without any external information typicallly make the assumption that: for each cell or tissue under study only a few genes change expression levels or that an equivalent number of genes increase and decrease across the different biological conditions [1].

These assumptions can be interpreted in different ways leading to different normalization procedures. For example, the mean expression level across genes within each sample should be the same across biological conditions [2]. Or that on average the distribution of gene expression within each sample should be the same across biological conditions [3]. Other normalization methods are based on *housekeeping genes* [4]. These are genes that are believed to be play a critical role in basic cellular pathways and as such should be expressed all the time at an equal rate independent of biological conditions. While these assumptions may be reasonable in certain experiments, they may not always hold [5, 6]. For example, mRNA content has been shown to fluctuate significantly during zebrafish early developmental stages [1]. It has also been shown that cells with high levels of c-Myc can amplify their global gene expression two to three times more than their low c-Myc counterparts [5].

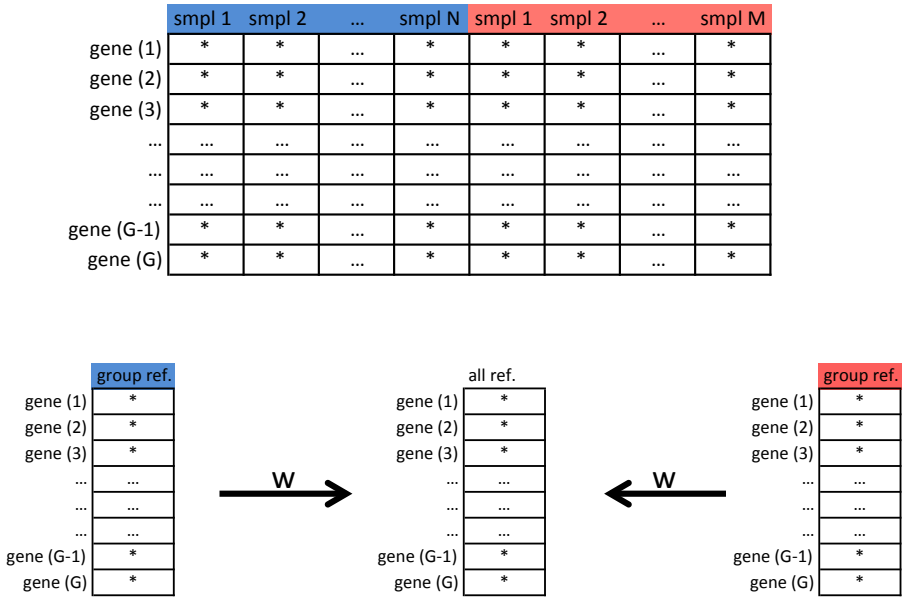
With the current improvements in technology and reduction in cost we are now able to relax some of these assumptions to allow for a more nuanced and information retaining normalization techniques. In this vignette we introduce, smooth quantile normalization (**qsmooth**), a generalization of quantile normalization [3] that makes the assumption that: all samples within the same biological group should have the same distribution.

Qsmooth first performs quantile normalization within each biological group and then shrinks the group quantiles towards the overall reference quantile depending on the variation between the group quantiles and the variation of quantiles within

the groups. The algorithm is described in Figure 1 below. Let gene(g) denote the g^{th} row after sorting each column in the data. For each row, gene(g), we compute the weight $w_{(g)} \in [0, 1]$. Where a 0 weight implies quantile normalization within groups and a weight of 1 indicates quantile normalization across the groups. The weight at each row depends on the between group sum of squares $\text{SSB}_{(g)}$ and total sum of squares $\text{SST}_{(g)}$, as follows:

$$w_{(g)} = \text{median}\{1 - \text{SSB}_{(i)}/\text{SST}_{(i)} |\ i = g - k, \ldots, g, \ldots, g + k\}, \quad (1)$$

where $k =$ floor(Total number of genes * 0.05). By using the rolling median, we borrow information from neighbouring genes.



$$w_{(g)} = \text{median}\{1 - \text{SSB}_{(i)} / \text{SST}_{(i)} \ |\ i \text{ in window centered at } g\}$$

Figure 1: **The qsmooth algorithm**

2 Rat bodymap

The **bodymapRat** package contains an ExpressionSet derived from the Raw FASTQ files obtained from Yu et al. (2013). It contains expression levels (RNAseq) on 11 organs, from male and female rats, at 4 developmental stages. We will use a subset of this data in this vignette.

For help with the bodymapRat R-package, there is a vignette available in the /vignettes folder.

2.1 Example 1

The first example is based a dataset which contains lung samples from 21 week old male and female rats. Four samples are from males and four samples are from females.

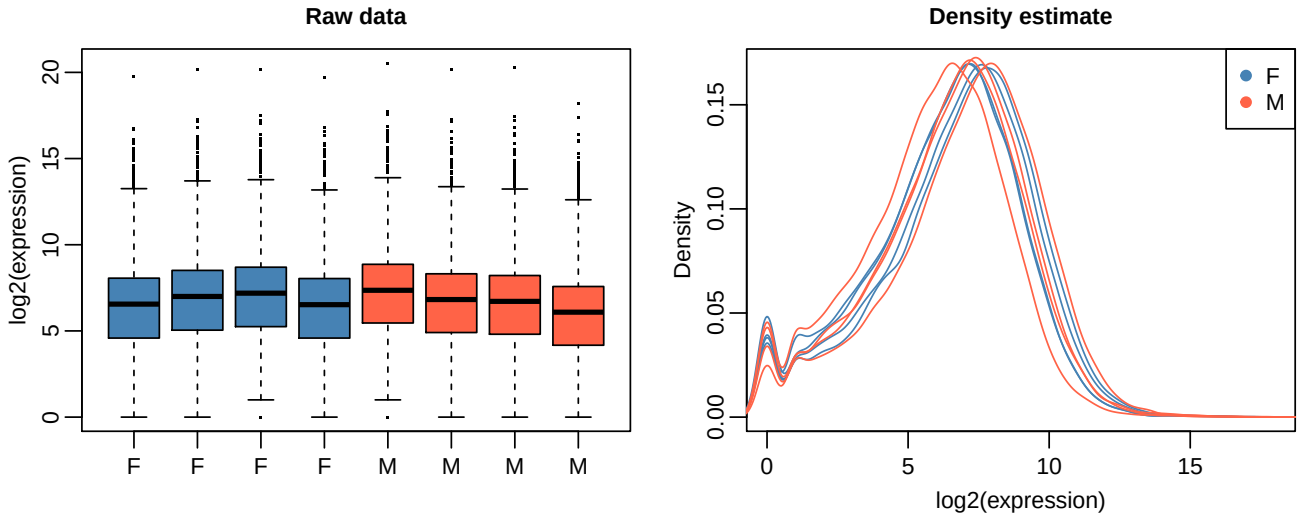
```
library(Biobase)
library(bodymapRat)

pd = pData(bodymapRat) # grab pheno data

# Subset samples from bodymapRat
sel = pd$organ %in% "Lung" # select lung samples
sel = sel & pd$stage == 21 # select stage 21 weeks
sel = sel & pd$techRep == 1 # select biological replicates

# Filter out low count genes
keep = rowMeans(exprs(bodymapRat)) > 10
data1 = bodymapRat[keep, sel]
```

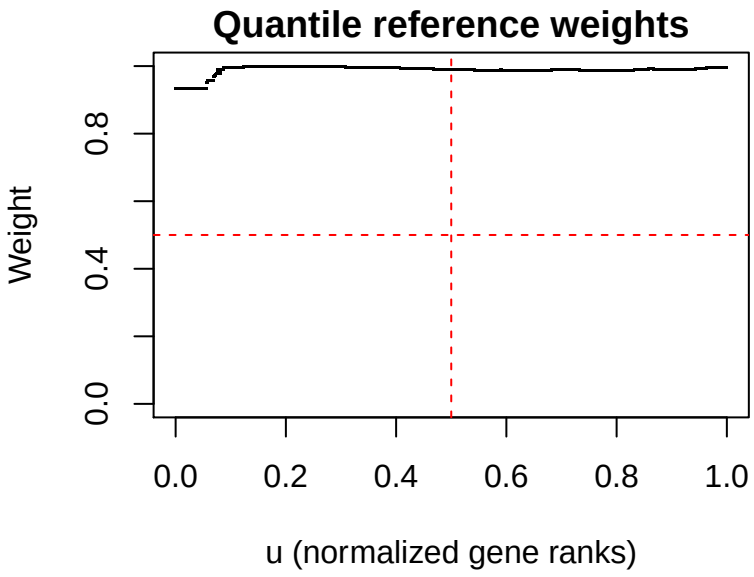
Below are the boxplots and the density estimate plots of the data after after adding 1 and log2 transforming the raw counts (ie. log2(counts+1)).



To run the **qsmooth** algorithm on the log transformed raw counts. We must specify sample groups. In this example we will use sex as the grouping factor.

We begin by loading **qsmooth** into R.

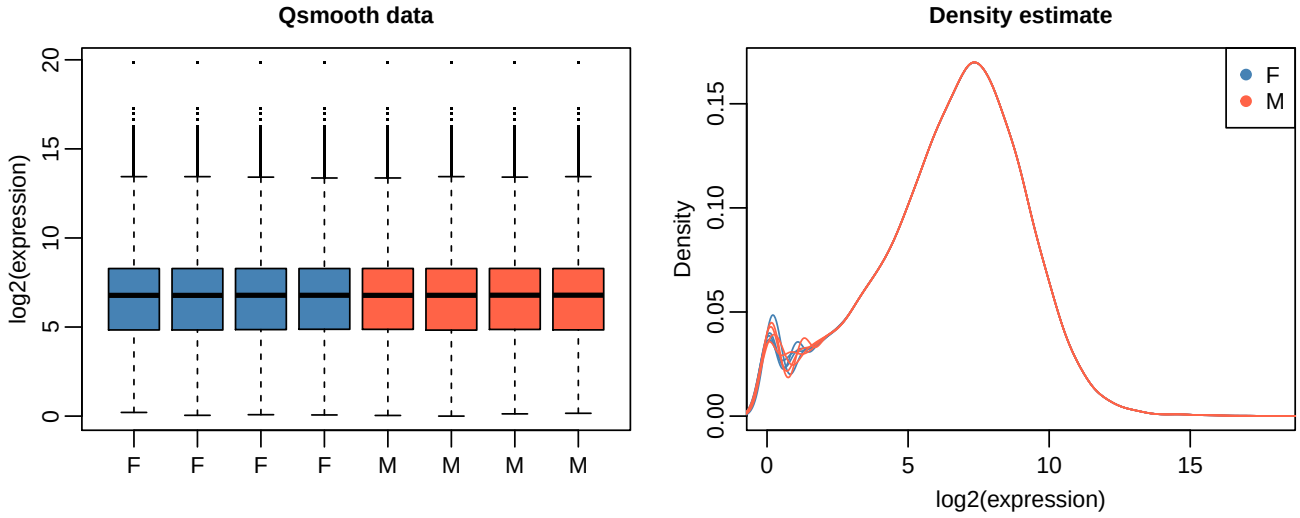
```
library(qsmooth)
norm.data1 = qsmooth(exprs=data1, groups=sex, plot=TRUE)
```



The parameter plot=TRUE indicates that we want to see the weigths of interpolation. Weights are computed for each quantile in the data set. A weight of 1 indicates full quantile normalization, where as a weight of 0 indicates quantile normalization within the groups. See Figure 1 for more details on the computation of the weights.

In this example the weights are mostly close to 1, indicating that the is no major difference between the quantiles from the female and male samples. Here the **qsmooth** algorithm outputs results that is identical (for practical purposes) to full quantile normalization.

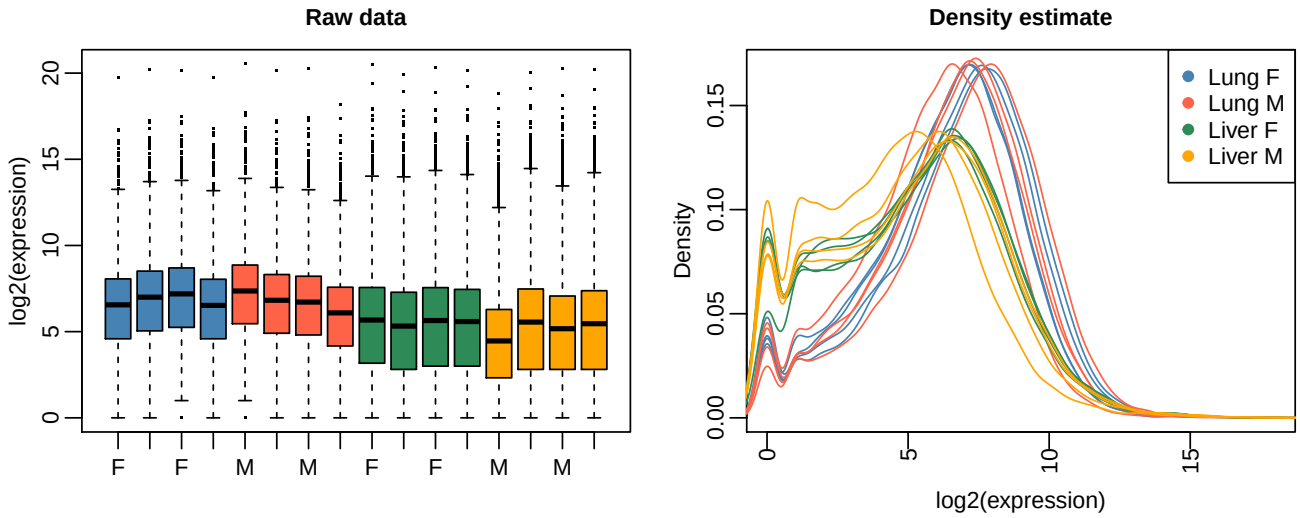
Below are the boxplots and density plots after normalization.



2.2 Example 2

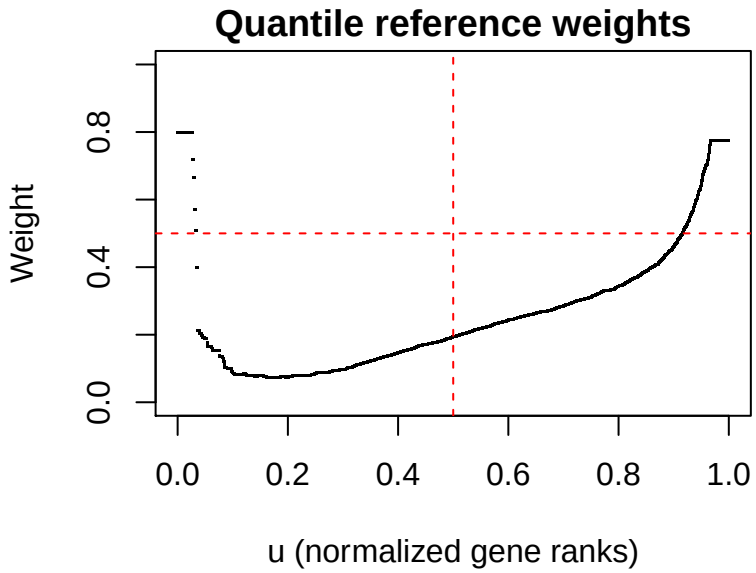
The second example is based a dataset which contains lung and liver samples from 21 week old male and female rats. Eight samples are from males and eight samples are from females.

Let's take a look at the raw data. Below is the boxplot and the density plot of the raw counts after after adding 1 followed by a log2 transformation.



We now run the qsmooth algorithm on the log transform raw counts. First we must specify sample groups. In this example we specify the groupings using sex and organ.

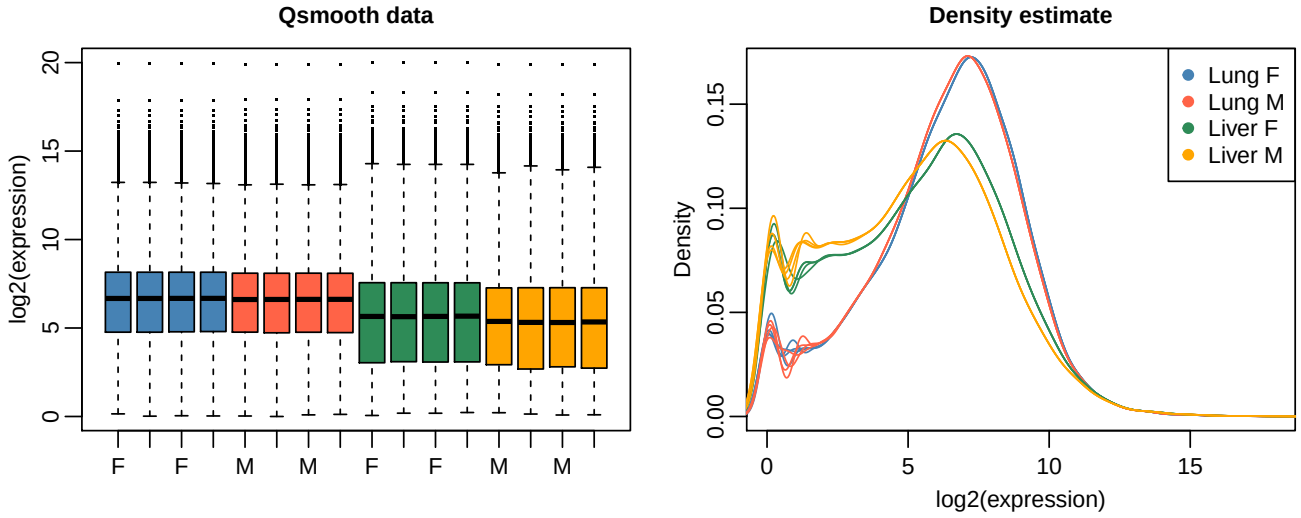
```
norm.data2 = qsmooth(exprs=data2, groups=paste0(sex, organ), plot=TRUE)
```



In this example the weights are mostly below 0.2 before the median (u = 0.5) and increase steadily to 0.8. Indicating that there is a difference in the empirical distributions of the samples across the groups. In this scenario full quantile

normalization is not appropriate.

Below are the boxplots and density plots after normalization. Note that within the liver samples males and females show a difference that is not in the lung samples.



3 The qsmooth function

The qsmooth function accepts five parameters.

1. exprs: for counts use log2(raw counts + 1)), for microarray use log2(raw intensities)
2. groups: groups to which samples belong (character vector)
3. norm.factors: scaling normalization factors (**optional**)
4. plot: plot weights? (default=FALSE) (**optional**)
5. plot: window window size for running median (defined as a fraction of the number of rows of exprs) (default=0.05)

The qsmooth function requires an expression matrix and a character vector or factor specifying which group a sample belongs. The plot parameter is optional. It specifies whether or not the weights should be plotted (See discussion on spike-in below). It is set to FALSE as defualt. The norm.factors allows the user to specify a vector of scaling factors that will be used to modify the expression data set prior to applying the qsmooth algorithm.

3.1 External RNA Control Consortium Spike-in Mixes

The External RNA Control Consortium (ERCC) is a collaborative group of academic, private, and public organizations hosted at the National Institutes of Standard and Technology (NIST) [7, 8]. The ERCC has developed a set of 92 mRNA controls (20-mer poly(A) tails) that can be used in gene expression platforms such as RNA-seq, DNA microarrays, and quantitative real-time reverse transcriptase PCR (qRT-PCR). The 92 mRNA transcripts are divided into 4 groups labelled A, B, C, and D. Each group contains 23 mRNA transcripts spanning a 10^6-fold concentration range. There are two ERCC control spike-in mixes: mix 1 and mix 2. The molar concentration ratios of mix 1 to mix 2 are 4, 1, 0.67, and 0.5 for group A, B, C, and D respectively. When the ERCC spike-in mix is used as a control in the experiment its measurements can be used as part of the data normalization process [5, 9].

In Figure 2 we show the distribution of the **true and known** concentration of each of the 92 "genes" in mix 1 and mix 2. Based on these plots we can make the assumption that the mix 1 and mix 2 "transcriptomes" have the same distribution (even though certain "genes" are differentially expressed).

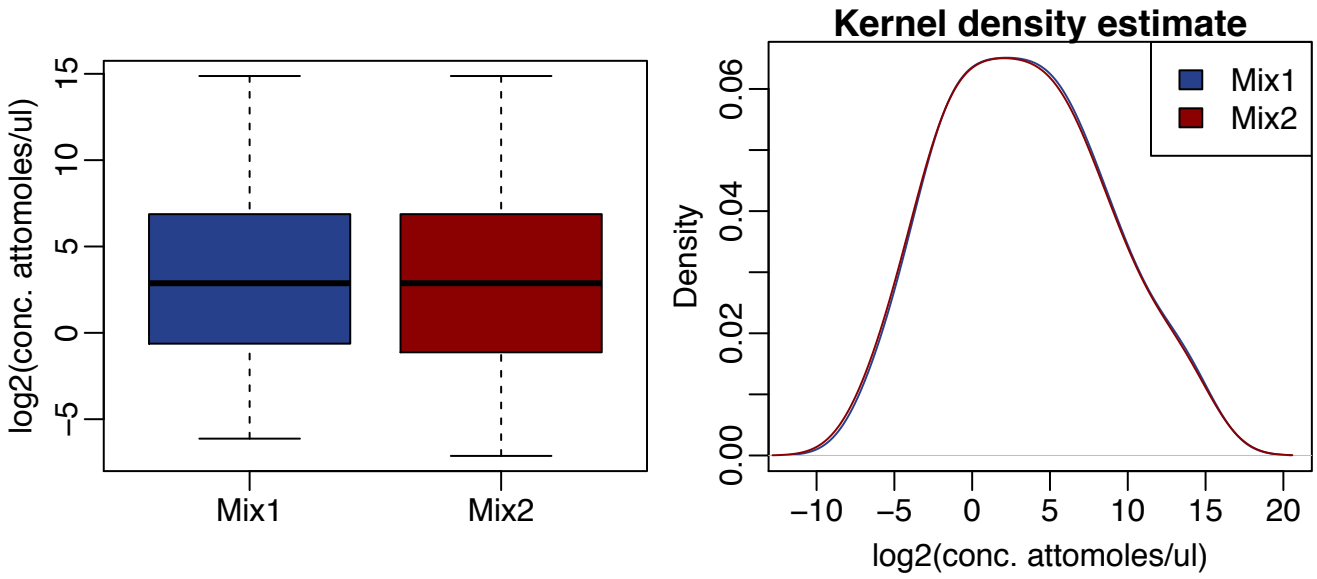


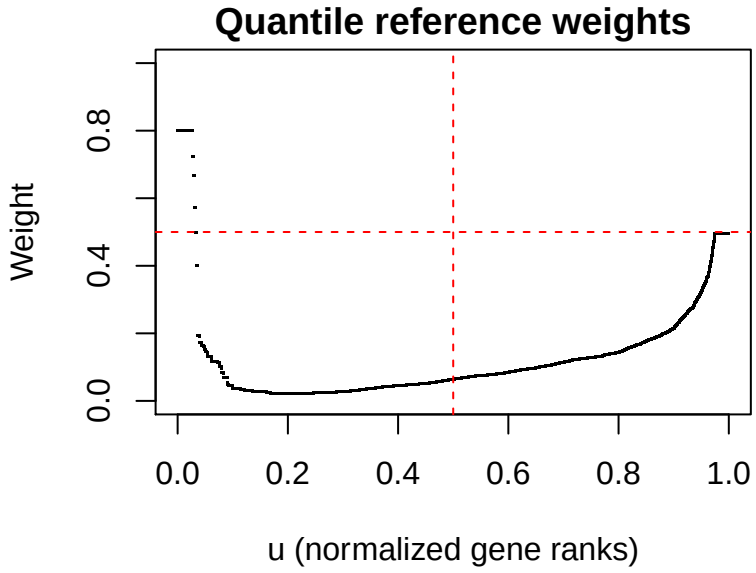
Figure 2: **ERCC spike-in mix 1 and mix 2**

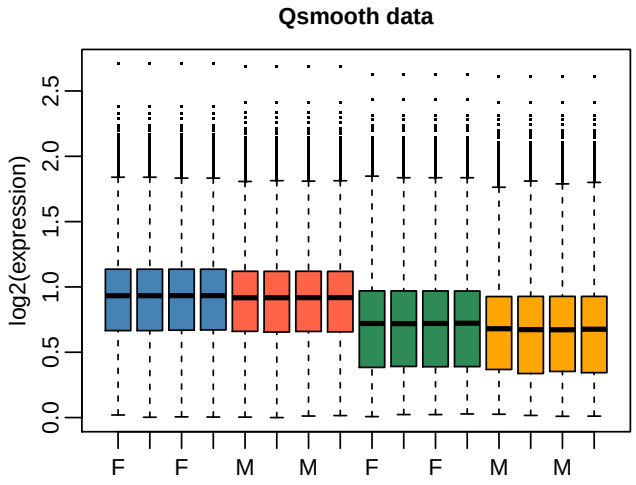
3.2 Pre-specified scaling factors

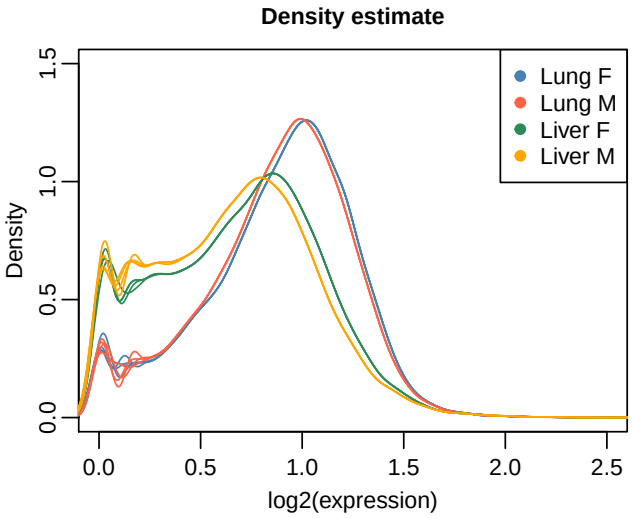
```
ercc = data2[grep("^ERCC", rownames(data2)), ]
dim(ercc)

## [1] 48 16

errcSF = apply(ercc, 2, median)
norm.data3 = qsmooth(exprs=t(t(data2)/errcSF), groups=paste0(sex, organ), plot=TRUE)
```







4 SessionInfo

```
sessionInfo()

## R version 3.2.1 (2015-06-18)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.10.5 (Yosemite)
##
## 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] parallel  stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] qsmooth_0.0.0.9000  bodymapRat_0.0.1    Biobase_2.28.0      BiocGenerics_0.14.0
## [5] knitr_1.11
##
## loaded via a namespace (and not attached):
## [1] BiocStyle_1.6.0 magrittr_1.5    formatR_1.2.1   tools_3.2.1     stringi_1.0-1
## [6] highr_0.5.1     stringr_1.0.0   evaluate_0.8
```

References

[1] Håvard Aanes, Cecilia Winata, Lars F Moen, Olga Østrup, Sinnakaruppan Mathavan, Philippe Collas, Torbjørn Rognes, and Peter Aleström. Normalization of rna-sequencing data from samples with varying mrna levels. *PloS one*, 9(2):e89158, 2014.

[2] Mark D Robinson, Alicia Oshlack, et al. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25, 2010.

[3] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[4] Eli Eisenberg and Erez Y Levanon. Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574, 2013.

[5] Jakob Lovén, David A Orlando, Alla A Sigova, Charles Y Lin, Peter B Rahl, Christopher B Burge, David L Levens, Tong Ihn Lee, and Richard A Young. Revisiting global gene expression analysis. *Cell*, 151(3):476–482, 2012.

[6] Stephanie C. Hicks and Rafael A. Irizarry. When to use quantile normalization? *bioRxiv*, 2014. doi:10.1101/012203.

[7] Shawn C Baker, Steven R Bauer, Richard P Beyer, James D Brenton, Bud Bromley, John Burrill, Helen Causton, Michael P Conley, Rosalie Elespuru, Michael Fero, et al. The external rna controls consortium: a progress report. *Nature methods*, 2(10):731–734, 2005.

[8] External RNA Controls Consortium et al. Proposed methods for testing and selecting the ercc external rna controls. *BMC genomics*, 6(1):150, 2005.

[9] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.