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 9, 2015

Contents

1 Introduction

In the analysis of high-throughput gene expression data, normalization strategies based solely on observed data without any external information often make the following assumption: for each sample in the study only a minority of genes are expected to be differentially epxressed or that an equivalent number of genes increase and decrease across the different biological conditions [1].

This assumption can be interpreted in different ways leading to different normalization procedures. For example, in one normalization procedure, the method assumes the mean expression level across genes should be the same across samples [2]. In contrast, quantile normalization assumes the statistical distribution of gene expression is the same across all samples [3]. Other normalization methods based on *housekeeping genes* assume genes play a critical role in basic cellular pathways and as such should be expressed all the time at an equal rate independent of samples [4]. While these assumptions may be reasonable in certain experiments, they may not always be appropriate [5, 6]. For example, mRNA content has been shown to fluctuate significantly during zebrafish early developmental stages [1]. Similarily, cells expressing high levels of c-Myc undergo transcriptional amplification causing a 2 to 3 fold change in global gene expression compared to cells expressiong low c-Myc [5]. In these cases with global changes of gene expression between biological conditions such as in cancer, transcriptional amplification or early developmental stages of zebrafish, quantile normalization is not an appropriate normalization method. In these cases, we can consider a more relaxed assumption about the data, namely that the statistical distribution of each sample should be the same within biological conditions or groups (compared to the more stringent assumption of quantile normalization, which states the statistical distribution is the same across all samples).

In this vignette we introduce a generalization of quantile normalization, referred to as **smooth quantile normalization** (**qsmooth**), which is a weighted average of the two types of assumptions about the data. The qsmooth R-package contains the qsmooth() function, which computes a weight at every quantile that compares the variability between groups relative to within groups. In one extreme quantile normalization is applied and in the other extreme quantile normalization within each biological condition is applied. The weight shrinks the group-level quantile normalized data towards the overall reference quantiles if variability between groups is sufficiently smaller than the variability within 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 weight of 0 implies quantile normalization within groups is applied and a weight of 1 indicates quantile normalization is applied. 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). The number 0.05 is a flexible parameter that can be altered to change the window of the number of genes considered. In this way, we can use a rolling median to borrow information from neighboring genes in the weight.

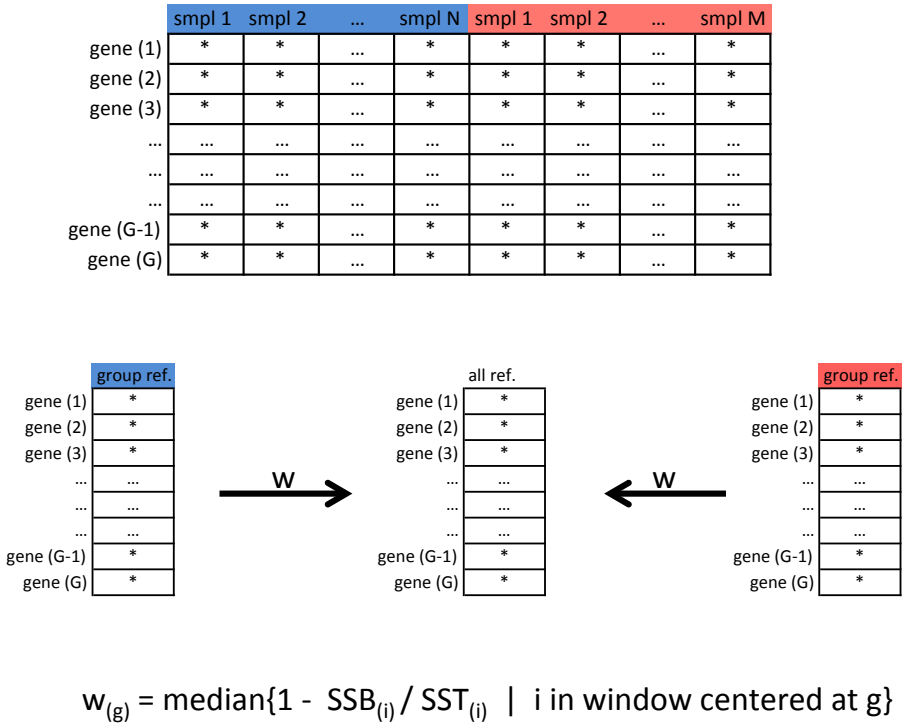


Figure 1: **The qsmooth algorithm**

2 Getting Started

Load the package in R

```
library(qsmooth)
```

3 Data

The **bodymapRat** package contains an `ExpressionSet` of 652 RNA-Seq samples from a comprehensive rat transcriptomic BodyMap study. This data was derived from the raw FASTQ files obtained from Yu et al. (2013) [7]. It contains expression levels from 11 organs in male and female rats at 4 developmental stages. We will use a subset of this data in this vignette.

The R-package bodymapRat can be installed from GitHub (`https://github.com/stephaniehicks/bodymapRat`) using the R package **devtools**.

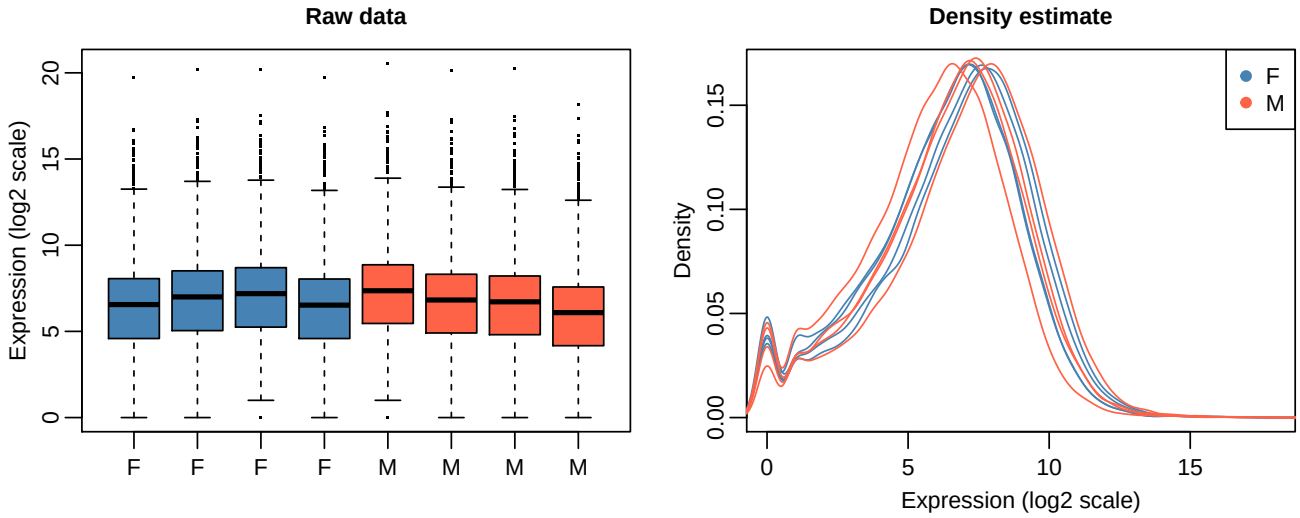
```
library(devtools)
install_github("stephaniehicks/bodymapRat")
```

3.1 bodymapRat 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.

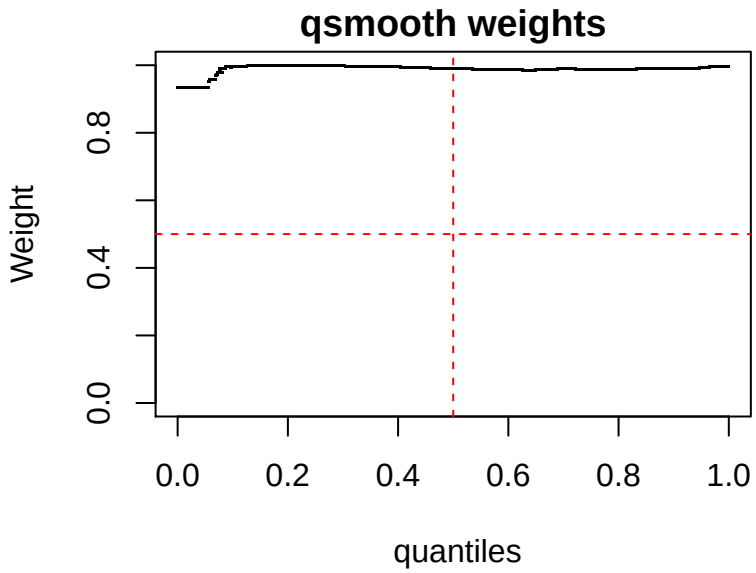
```
library(Biobase)
library(bodymapRat)

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 plots of the data after after adding 1 and transforming the raw counts on the `log2()` scale (i.e. `log2(counts+1)`).



To run the qsmooth() algorithm on the log transformed raw counts, we must specify the group-level information for each sample. In this example we will use gender as the group-level factor.

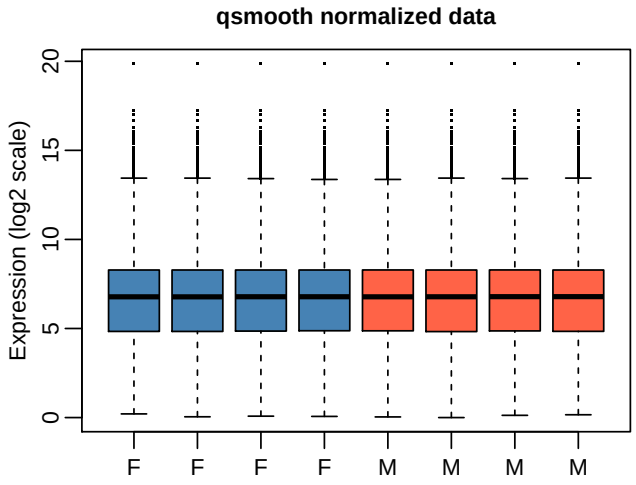
```
norm.data1 = qsmooth(exprs=data1, groups=sex, plot=TRUE)
```

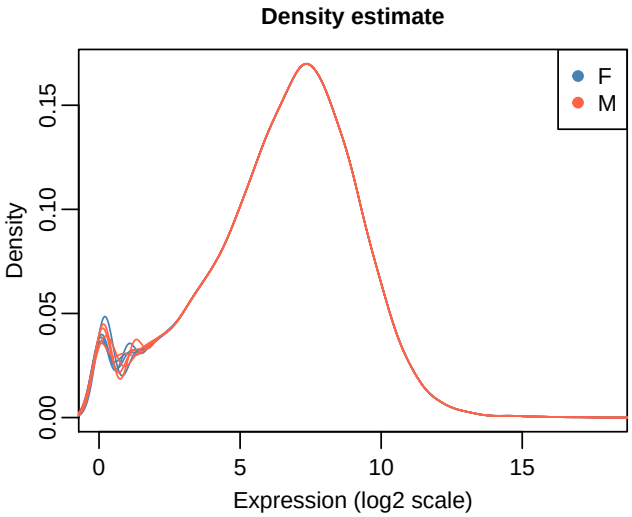


The parameter plot=TRUE indicates that we want to see the weight of interpolation. Weights are computed for each quantile in the data set. A weight of 1 indicates quantile normalization is applied, where as a weight of 0 indicates quantile normalization within the groups is applied. See Figure 1 for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

In this example, the weights are close to 1 across all the quantiles indicating that there is no major difference between the group-level quantiles in the female and male rats. Here, the qsmooth() algorithm returns a normalized data set that is nearly identical (for practical purposes) to the conventional quantile normalization.

Below are the boxplots and density plots after applying the qsmooth() normalization.

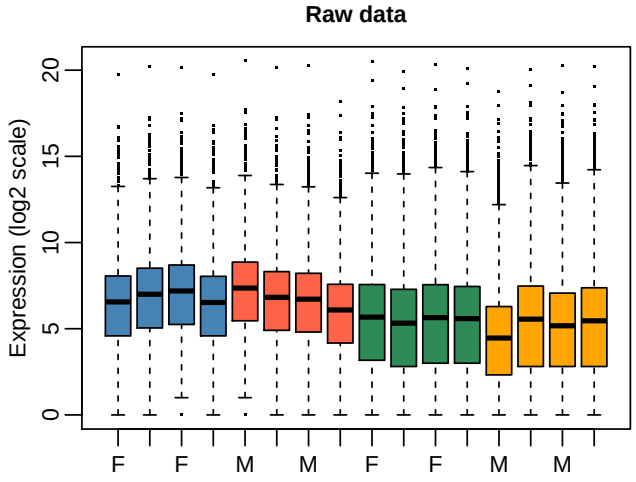


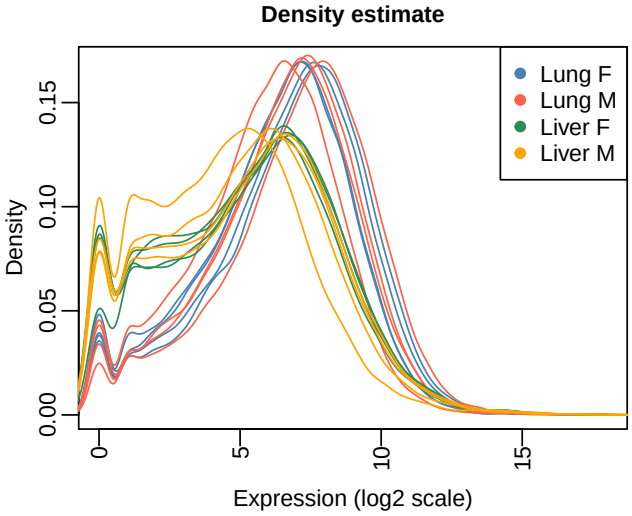


3.2 bodymapRat example 2

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

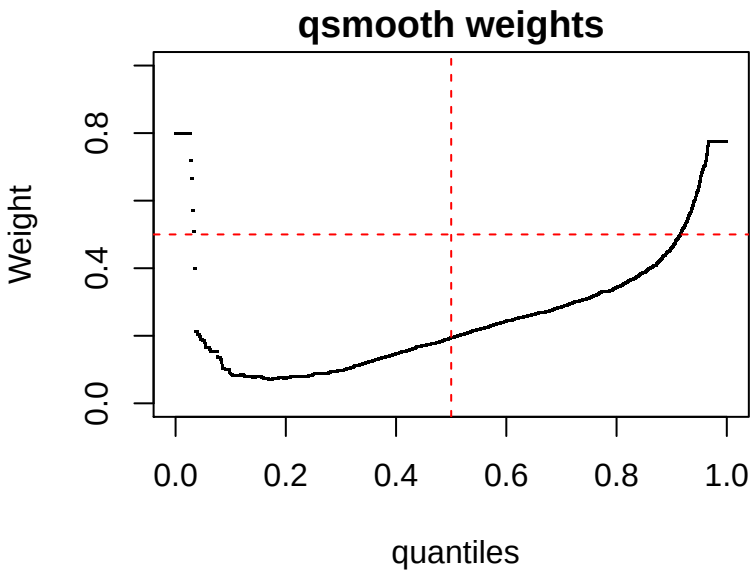
Below are the boxplots and the density plots of the raw data after after adding 1 and transforming the raw counts on the `log2()` scale (i.e. `log2(counts+1)`).





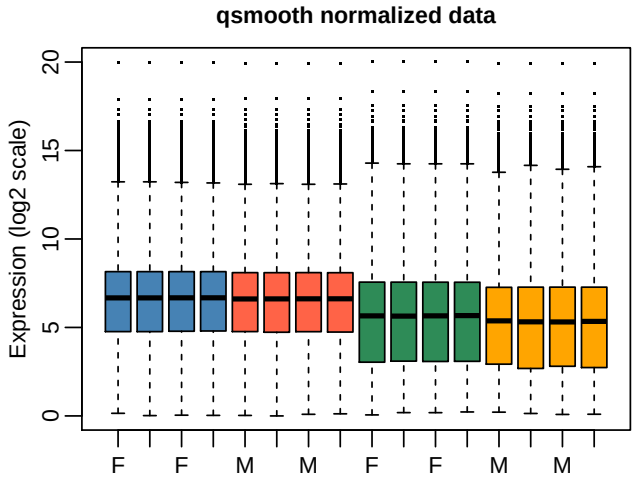
To run the `qsmooth()` algorithm on the log transformed raw counts, we must specify the group-level information for each sample. In this example we will use gender and organ as the group-level factor.

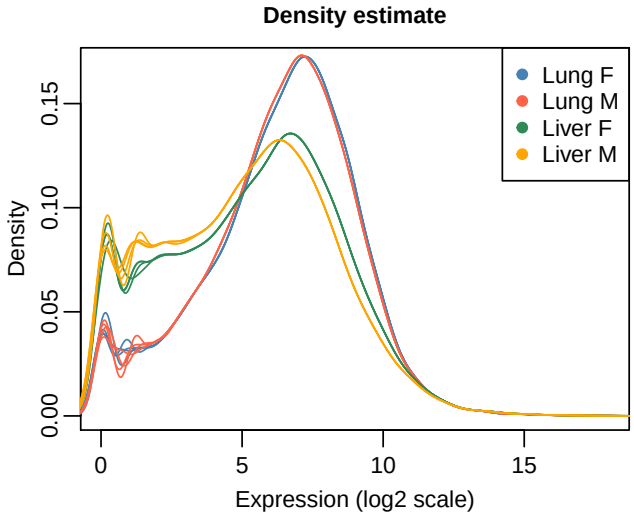
```
norm.data2 = qsmooth(exprs=data2, groups=paste0(sex, organ), plot=TRUE)
```



In this example, the weights are mostly below 0.2 before the median quantile (0.5) and increase steadily to 0.8. This indicates there is a difference in the statistical distributions of the samples between the groups. In this case, the conventional quantile normalization is not appropriate.

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





4 Additional information for 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 the group-level information for that sample. The plot parameter is optional. It specifies whether or not the weights should be plotted. It is set to FALSE as default. 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 (see discussion on spike-in below).

4.1 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) [8, 9]. 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, 10].

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).

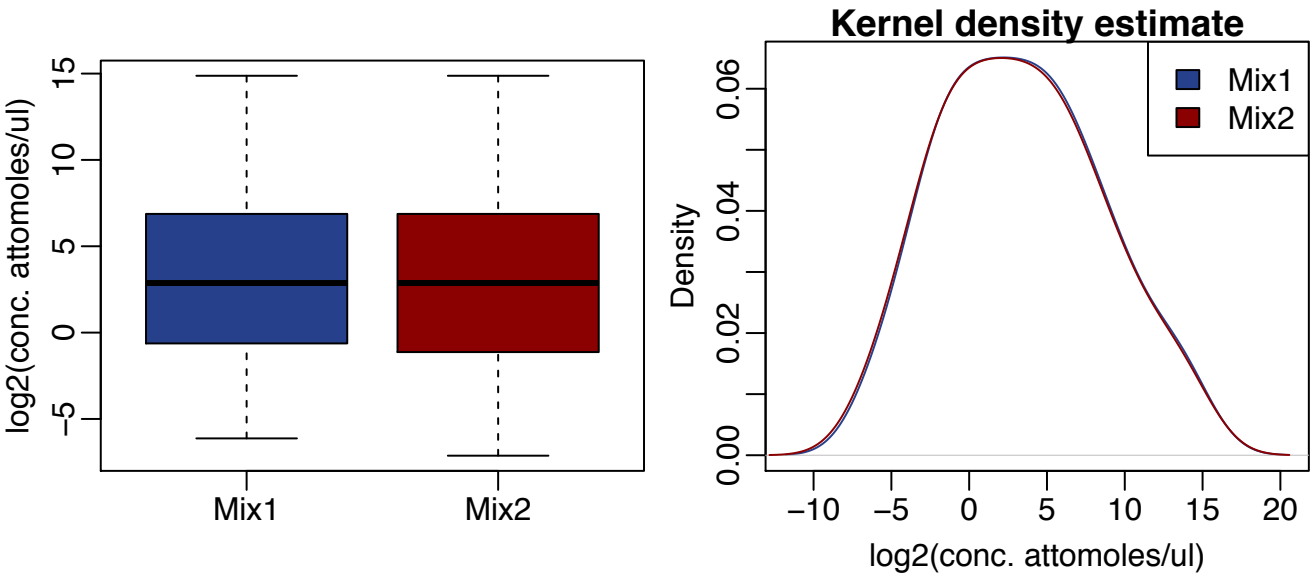
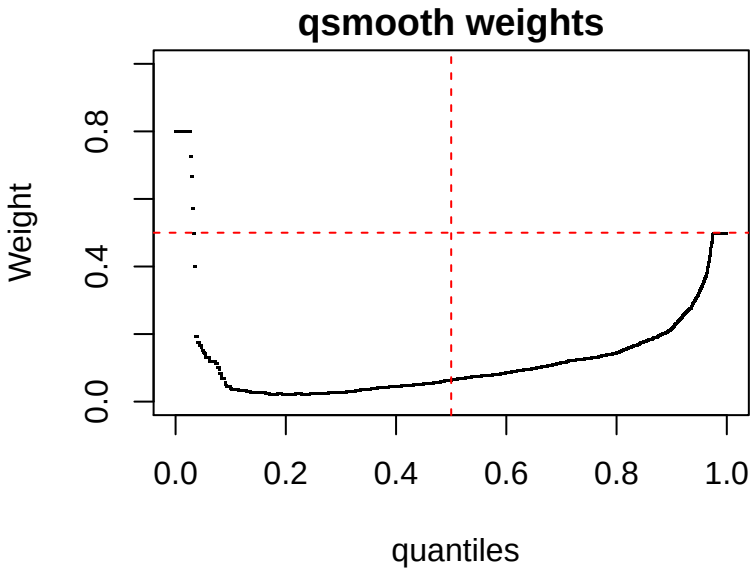


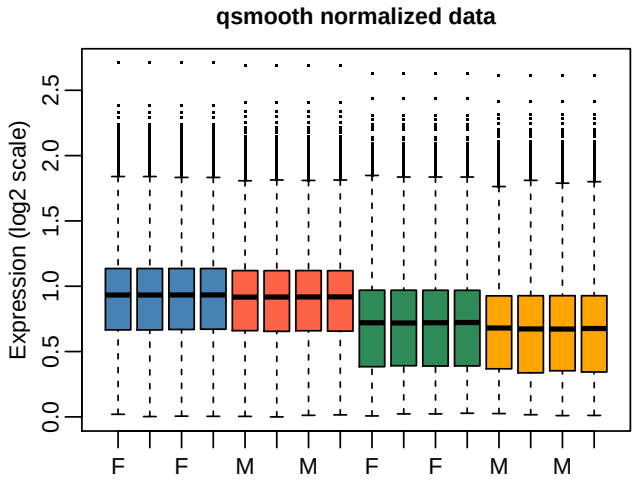
Figure 2: **ERCC spike-in mix 1 and mix 2**

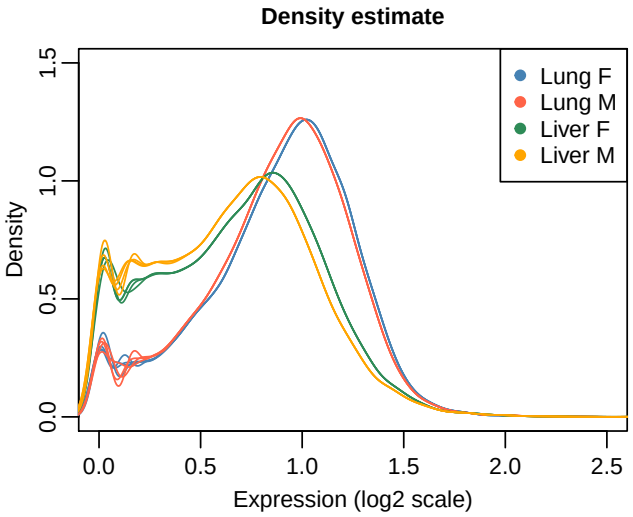
4.2 Pre-specified scaling factors

This is an example of incorporating the ERCC spike-ins into qsmooth() as 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)
```







5 SessionInfo

```
sessionInfo()

## R version 3.2.2 (2015-08-14)
```

```
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.11.1 (El Capitan)
##
## 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 datasets  utils     methods   base
##
## other attached packages:
## [1] bodymapRat_0.0.1    Biobase_2.30.0      BiocGenerics_0.16.1 qsmooth_0.0.0.9000
## [5] knitr_1.11
##
## loaded via a namespace (and not attached):
## [1] BiocStyle_1.8.0 magrittr_1.5    formatR_1.2.1   tools_3.2.2     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] Ying Yu, James C Fuscoe, Chen Zhao, Chao Guo, Meiwen Jia, Tao Qing, Desmond I Bannon, Lee Lancashire, Wenjun Bao, Tingting Du, Heng Luo, Zhenqiang Su, Wendell D Jones, Carrie L Moland, William S Branham, Feng Qian, Baitang Ning, Yan Li, Huixiao Hong, Lei Guo, Nan Mei, Tieliu Shi, Kevin Y Wang, Russell D Wolfinger, Yuri Nikolsky, Stephen J Walker, Penelope Duerksen-Hughes, Christopher E Mason, Weida Tong, Jean Thierry-Mieg, Danielle Thierry-Mieg, Leming Shi, and Charles Wang. A rat rna-seq transcriptomic bodymap across 11 organs and 4 developmental stages. *Nat Commun*, 5:3230, 2014. doi:10.1038/ncomms4230.

[8] 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.

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

[10] 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.