# The qsmooth user's guide

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

Modified: March 5, 2015. Compiled: November 16, 2015

## Contents

## 1 Introduction

In the advent of high-throughput technologies (such as microarray and RNA-seq) crude assumptions were made in order to pre-process and normalize the data [1]. These crude assumptions were needed because of large technical variablilty and very few samples sizes. For some data sets the normalization techniques based on these crude assumptions lead to a significant loss in biological information [1, 2]. With the current improvements in technology and reduction in cost we are now able to relax some of the previous 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 shape.

**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 alogorithm 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 $(SSB_{(g)})$ and total sum of squares $(SST_{(g)})$, as follow median$\{1 - SSB_{(g)}/SST_{(g)} | g = g - k, g, g + k\}$, where $k$ =floor(total number of genes* 0.05).

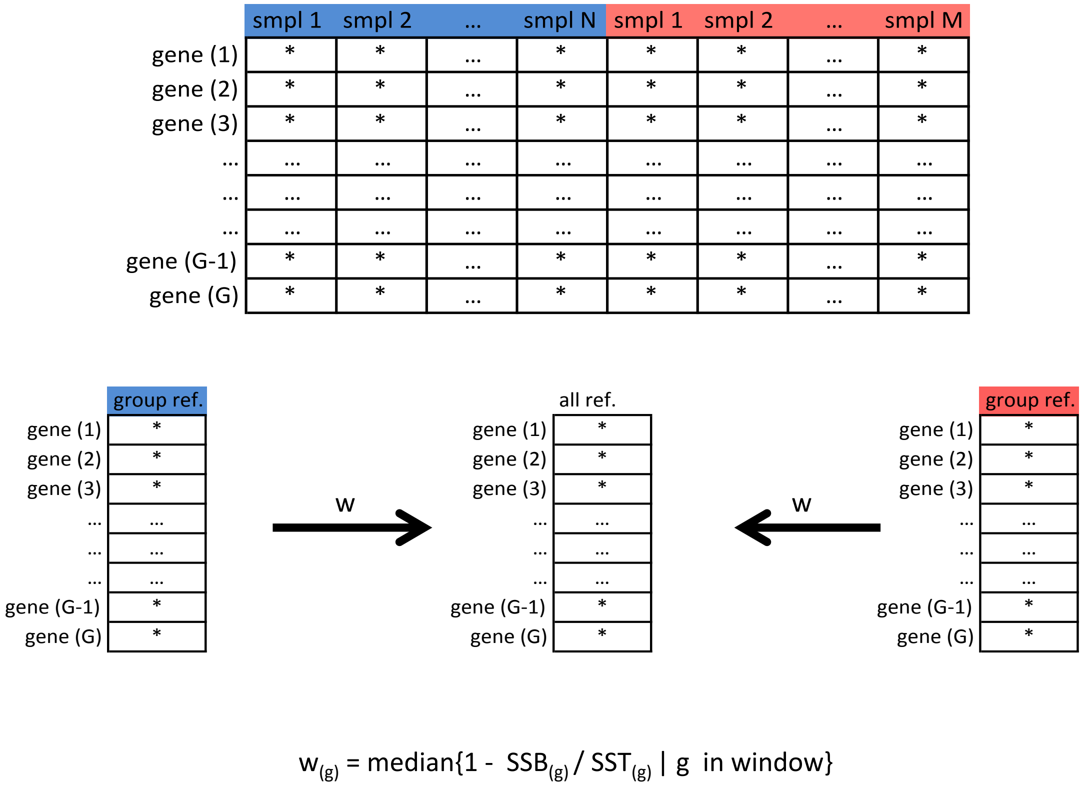


Figure 1: **The qsmooth algorithm.** At each quantile compute $R^2$.

## 2 Rat bodymap

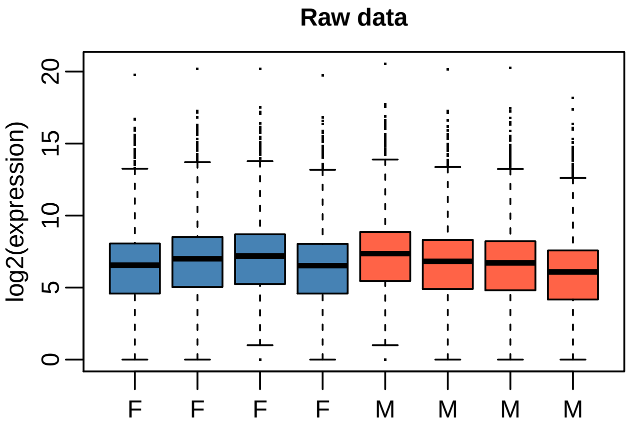
### 2.1 Data 1

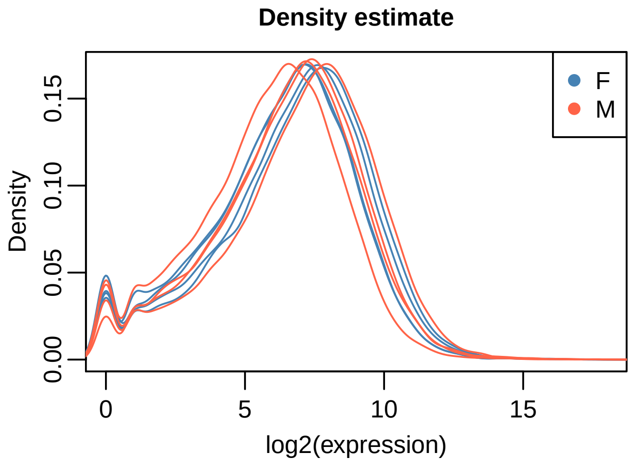
We begin by loading the **qsmooth** into R.

```
library(qsmooth)
```

The first example is based a data set (data1) which contains lung samples from 21 week old male and female rats. Four samples are from males and four samples are from females.

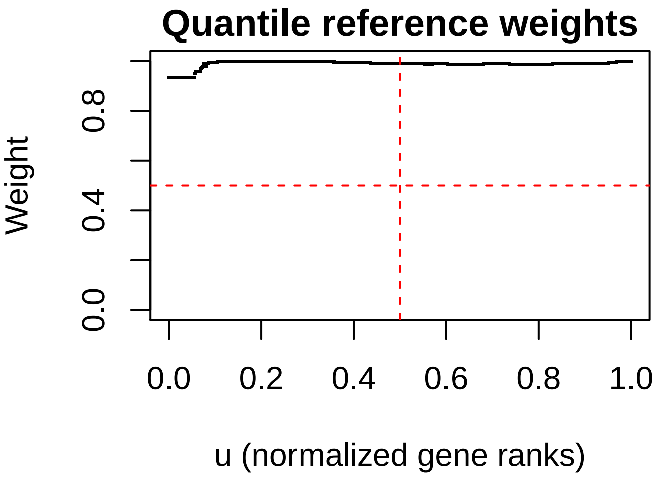
Below are the boxplots and the density estimate plots of the raw counts after after adding 1 and followed by a log2 transformation (ie. log2(counts+1)).





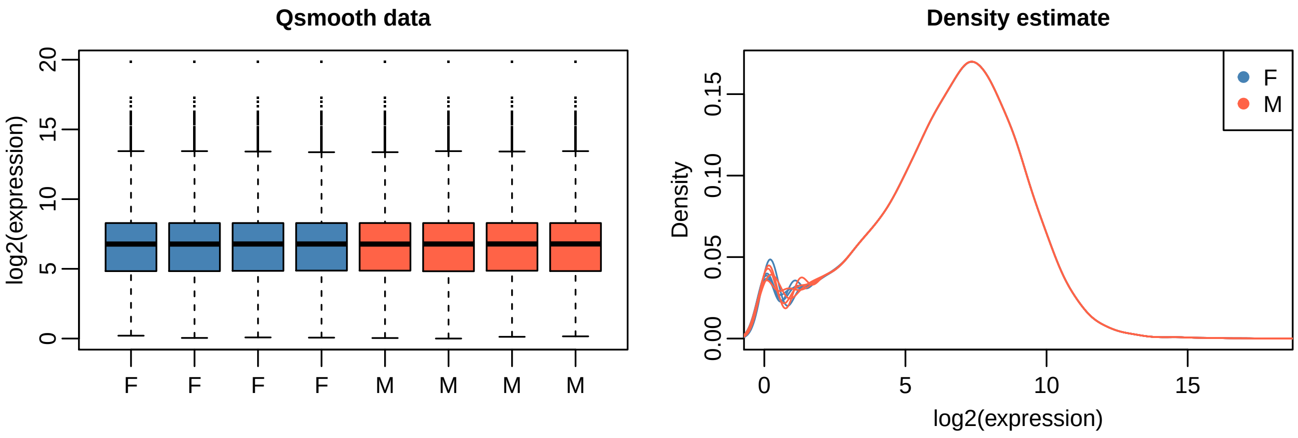
To run the **qsmooth** algorithm on the log transformed raw counts. We must specify sample groups. In this example we will use sex as the grouping factor.

```
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.

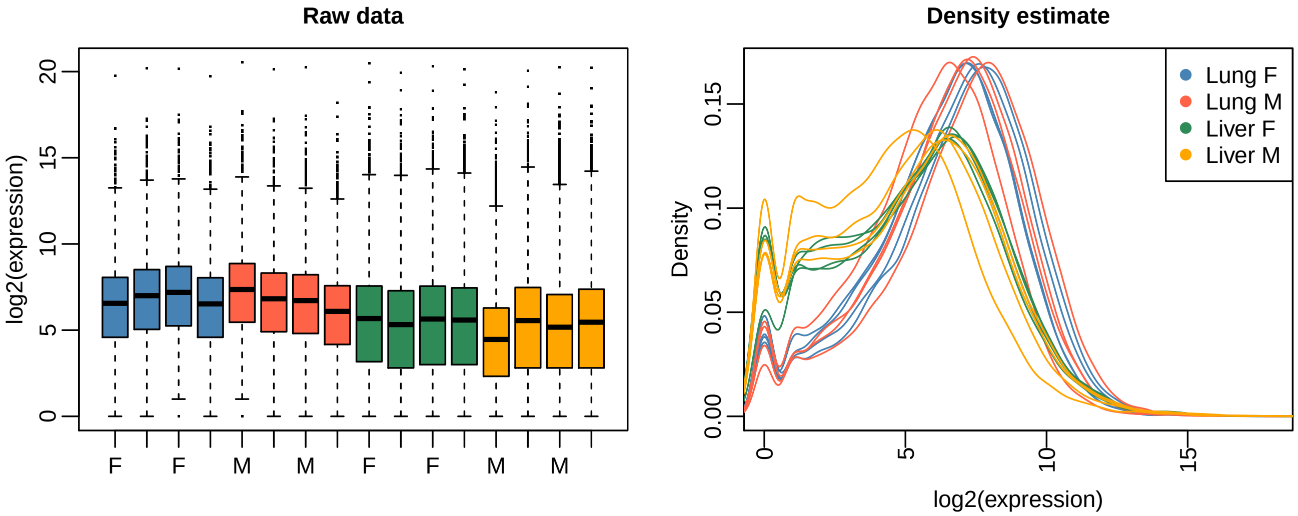
In this example the weights are mostly close to 1, indicating that the is no major difference between the quantiles from the female and male samples. Here the **qsmooth** algorithm outputs results that is identical (for practical purposes) to full quantile normalization.



## 2.2 Example 2

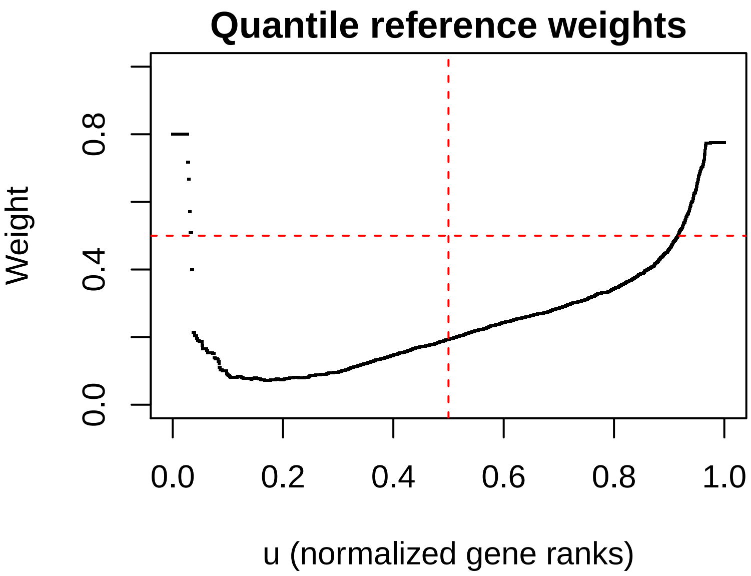
The first examples consists of a data set which contains lung samples from 21 week old male and female rats. Four samples from males and a four samples from females.

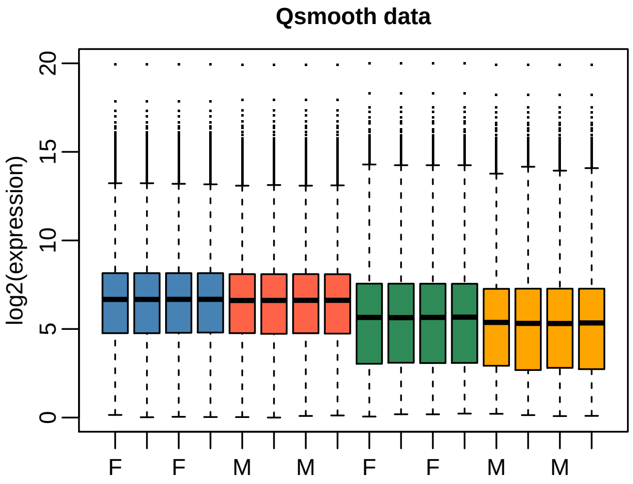
Let's take a look at the raw data. Below is the boxplot and the density plot of the raw counts after after adding 1 followed by a log2 transformation.

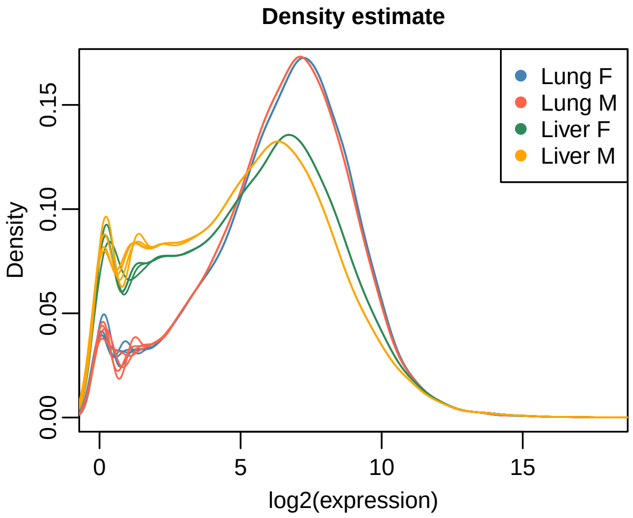


We now run the qsmooth algorithm on the log transform raw counts. First we must specify sample groups. In this example we specify the groupings using sex

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







## 3 Qsmooth

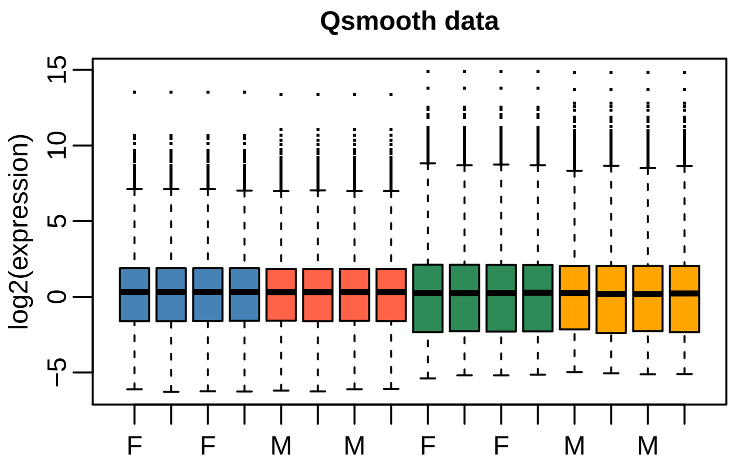
The qsmooth function accepts four parameters. Two are required and the other two are optional. The qsmooth function requires an expression matrix and a character vector or factor specifying which group a sample belongs, for the exprs and groups parameters respectively. The plot parameter is optional. It specifies whether or not the weights should be plotted. It is set to FALSE be 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 Pre-specified scaling factors

```
norm.data3 = qsmooth(exprs=data2, groups=paste0(sex, organ),
                     norm.factors = apply(data2, 2, mean))
```

```
par(mar=c(4, 3, 2, 0.5), mgp=c(1.5, 0.5, 0), cex.axis=0.8, cex.lab=0.8, cex.main=0.8)

# boxplot
boxplot(norm.data3, col=col, main="Qsmooth data",
        names=sex, ylab="log2(expression)", pch=".",
        cex.axis=0.8)
```



## 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] bodymapRat_0.0.1    Biobase_2.28.0      BiocGenerics_0.14.0 qsmooth_0.0.0.9000
## [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] 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.

[2] Stephanie Hicks and Rafael Irizarry. quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome Biology*, 16(1):117, 2015.

[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.