The qlasso user's guide

Kwame Okrah kwame.okrah@gmail.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: April 20, 2015

Contents

1 Introduction

Add introduction here.

2 Getting Started

Load the qlasso package in R.

```r
library(quantro)
library(HTShape)
library(qlasso)
```

3 Data

3.1 Pickrell Data Example

Load an example data set. Here we use the Pickrell data set. (but we can change this to whatever).

```
data(examplesData)
names(examplesData)

## [1] "bottomly"             "pickrell"             "tcruziExtracellular"
## [4] "seqc"                 "seqc.ercc"

counts = examplesData$pickrell$exprs
groups = examplesData$pickrell$cond
```

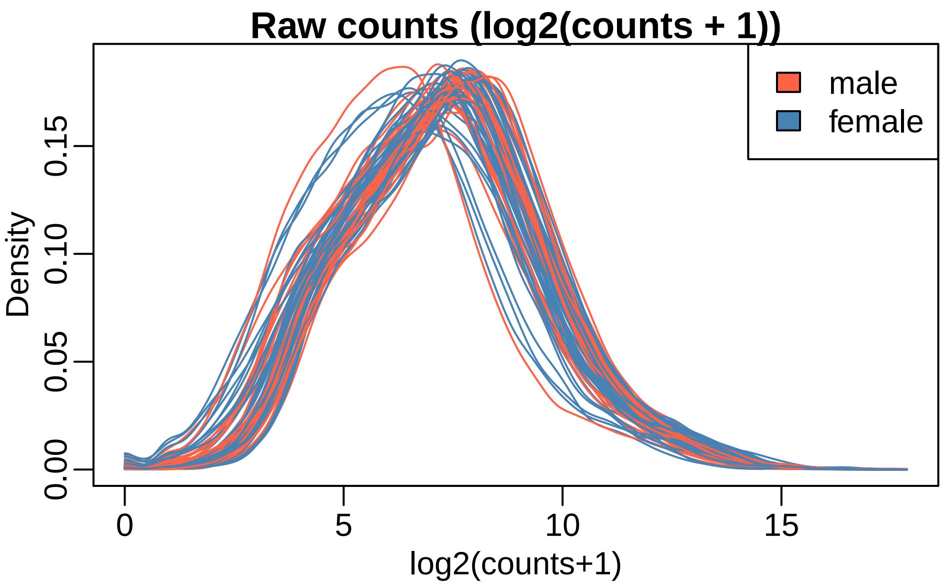
4 Exploratory Data Analysis

In this section we will look at summary plots of the raw data.

First, we will filter out genes with low counts with the `filterCounts` function. This function will only retain genes whose counts per million (cpm) exceeds 1 (can be changed, see the `thresh` parameter) in a given number of samples (see the `minSamples` parameter).

```
(minSamples <- min(table(groups)))

## [1] 29

dim(counts)

## [1] 38415    69

counts = filterCounts(counts, thresh=1, minSamples=minSamples)
dim(counts)

## [1] 17471    69
```

Density plots

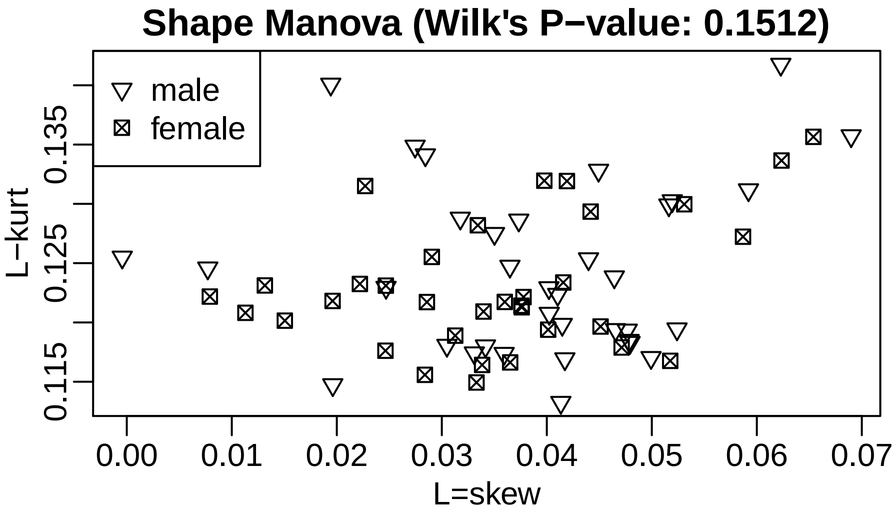


4.1 Samples shape assessment

In this section we will formally test whether the transcriptome shapes (densities) differ due to a factor of interest. In this case sex. We will use both quantro and HTShape for this test and compare results.

4.1.1 L-ratios manova stat.

First we will use the `shapeManova` function in HTShape (see HTShape for more details). This method first summarizes each sample in the data set with scale-free skewness and kurtosis coefficients (L-skew and L-kurt). These shape esitmates are based on the theory of L-moments (cite:Hosking1990, Okrah2015). We perform a multivariate analysis of variance based on the shape (L-skew, L-kurt) esitmates (see xxx for more details).



```
## $WL
## [1] 0.9451745
##
## $Fstat
## [1] 1.887206
##
## $df1
## [1] 2
##
## $df2
## [1] 132
##
## $pval
## [1] 0.1512349
```

The pvalue of 0.151 indicates that sex and transcriptome shape are not related.

4.1.2 Quantro stat.

Use qauntro for the same test.

```
(qtest <- quantro(log2(counts+1), groups, verbose=FALSE, B=500))

## quantro: Test for global differences in distributions
##    nGroups:  2
##    nTotSamples:  69
##    nSamplesinGroups:  40 29
##    anovaPval:  0.50626
##    quantroStat:  2.01579
##    quantroPvalPerm:  0.134
```

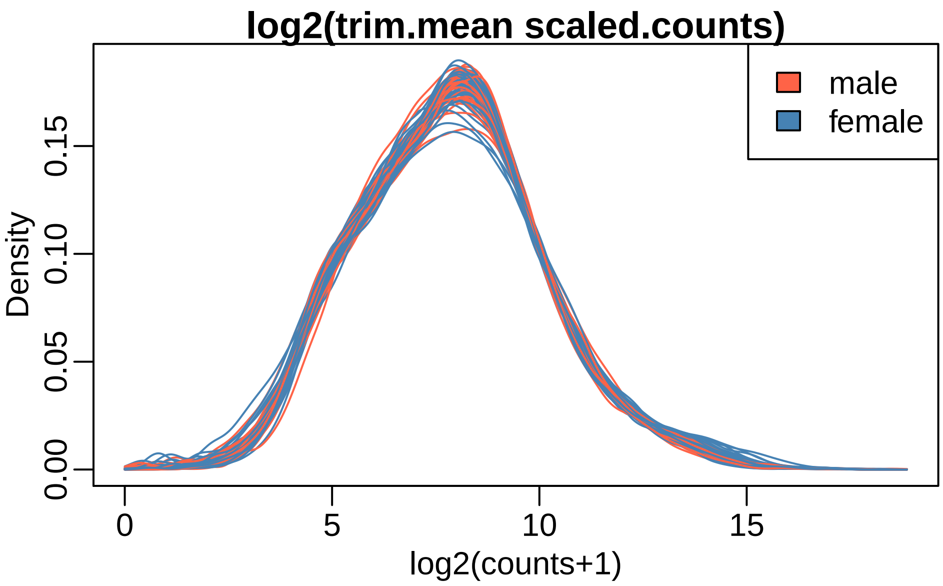
Conclusions are the same as shapeManova.

5 Using qlasso for normalization

5.1 Scale samples

Scale samples using trimmed mean. Trim off top and bottom 0.25 quantiles. Other methods can be used (eg. AH, median, mean).

Density plots

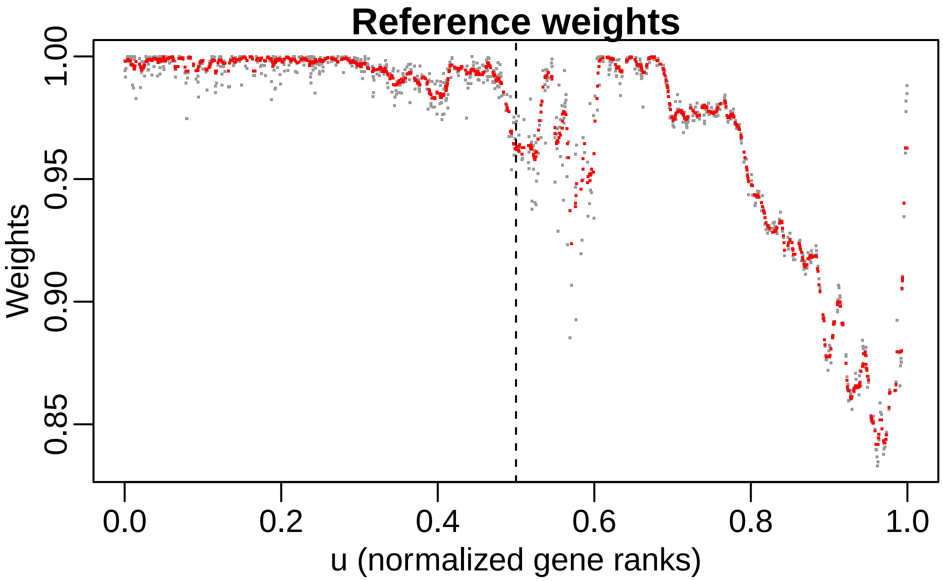


5.2 Computing quantiles

The sample quantiles of the raw data, reference quantile, and shrinkage weights can be computed using the `qstats()` function. The reference quantile can be computed as a average across sample quantiles (as in full quantile normalization) or can be obtained by taking the median across reference quantiles. The `refType` parameter specifies which type of reference quantile to use.

```
qs = qstats(exprs=log2(scaled.counts), groups=groups,
            refType="mean", groupLoc="mean", window=99)
```

plots weights

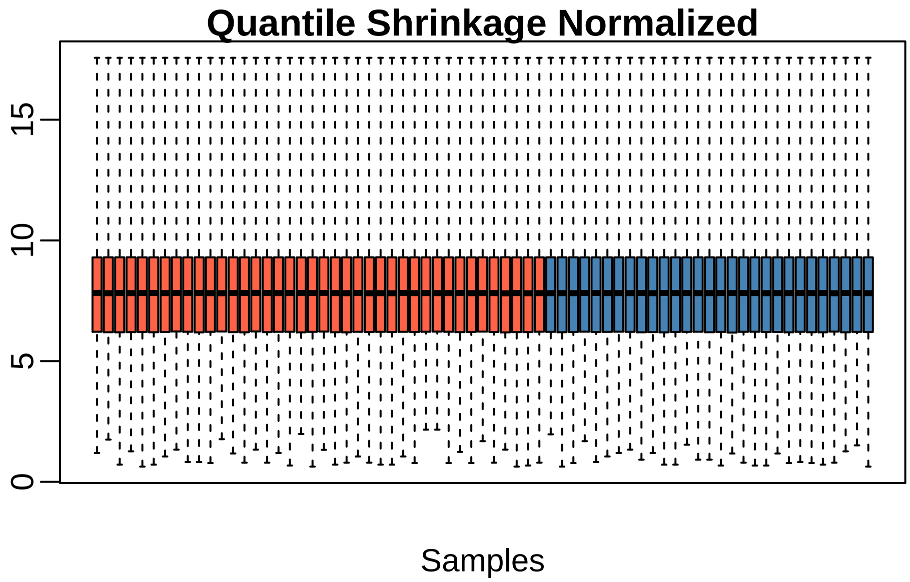


5.3 qshrink normalized values

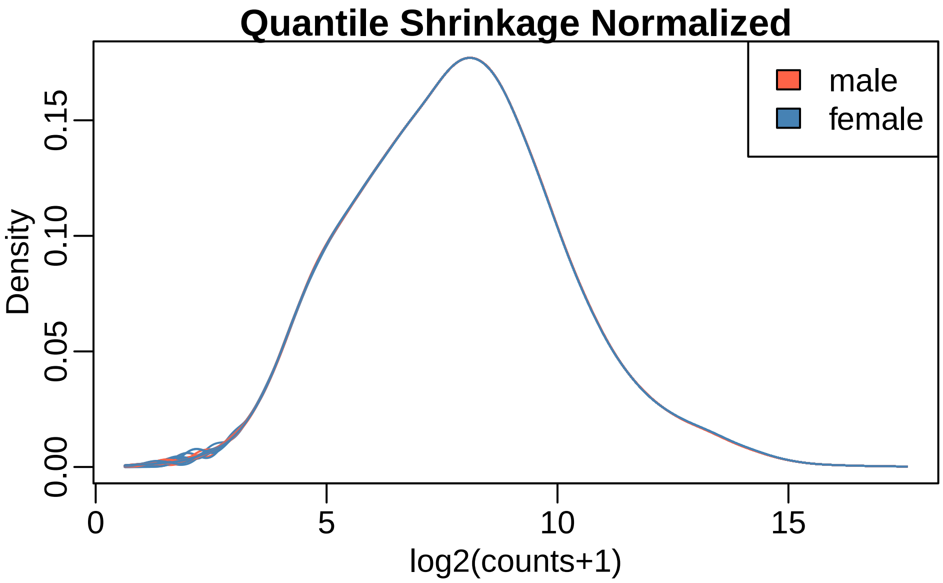
The normalized values are computed using the qshrink function. This function is based on the resultys of qstats. We do not need to call qstats. It was shown above for demonstration.

```
normExprs = qshrink(exprs=log2(scaled.counts), groups=groups,
                    refType="mean", groupLoc="mean", window=99)
```

Boxplots



Density plots



6 SessionInfo

```
sessionInfo()

## R version 3.1.2 (2014-10-31)
## Platform: x86_64-apple-darwin10.8.0 (64-bit)
##
## 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] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] qlasso_0.0.0.9000 HTShape_1.0.0     quantro_1.0.0     knitr_1.9
##
## loaded via a namespace (and not attached):
##  [1] annotate_1.44.0       AnnotationDbi_1.28.2  base64_1.1
##  [4] beanplot_1.2          Biobase_2.26.0        BiocGenerics_0.12.1
##  [7] BiocStyle_1.4.1       Biostrings_2.34.1     bumphunter_1.6.0
## [10] codetools_0.2-11      colorspace_1.2-6      compiler_3.1.2
## [13] DBI_0.3.1             digest_0.6.8          doParallel_1.0.8
## [16] doRNG_1.6             evaluate_0.6          foreach_1.4.2
## [19] formatR_1.1           genefilter_1.48.1     GenomeInfoDb_1.2.5
## [22] GenomicRanges_1.18.4  ggplot2_1.0.1         grid_3.1.2
## [25] gtable_0.1.2          highr_0.4.1           illuminaio_0.8.0
## [28] IRanges_2.0.1         iterators_1.0.7       lattice_0.20-31
## [31] limma_3.22.7          locfit_1.5-9.1        MASS_7.3-40
## [34] matrixStats_0.14.0    mclust_5.0.0          minfi_1.12.0
## [37] multtest_2.22.0       munsell_0.4.2         nlme_3.1-120
## [40] nor1mix_1.2-0         parallel_3.1.2        pkgmaker_0.22
## [43] plyr_1.8.1            preprocessCore_1.28.0 proto_0.3-10
## [46] quadprog_1.5-5        RColorBrewer_1.1-2    Rcpp_0.11.5
## [49] registry_0.2          reshape_0.8.5         reshape2_1.4.1
## [52] rngtools_1.2.4        RSQLite_1.0.0         S4Vectors_0.4.0
## [55] scales_0.2.4          siggenes_1.40.0       splines_3.1.2
## [58] stats4_3.1.2          stringr_0.6.2         survival_2.38-1
## [61] tools_3.1.2           XML_3.98-1.1          xtable_1.7-4
## [64] XVector_0.6.0         zlibbioc_1.12.0
```