YARN: Robust Multi-Tissue RNA-Seq Preprocessing and Normalization

Joseph N. Paulson & John Quackenbush

2017-10-30

YARN - Yet Another RNa-seq package

The goal of yarn is to expedite large RNA-seq analyses using a combination of previously developed tools. Yarn is meant to make it easier for the user to perform accurate comparison of conditions by leveraging many Bioconductor tools and various statistical and normalization techniques while accounting for the large heterogeneity and sparsity found in very large RNA-seq experiments.

Installation

You can install yarn from github with:

```
source("http://bioconductor.org/biocLite.R")
biocLite("yarn")
```

Quick Introduction

If you're here to grab the GTEx version 6.0 data then look no further than this gist that uses yarn to download all the data and preprocess it for you.

Preprocessing

Below are a few of the functions we can use to preprocess a large RNA-seq experiment. We follow a particular procedure where we:

1. Filter poor quality samples
2. Merge samples of similar conditions for increased power
3. Filter genes while preserving tissue or group specificity
4. Normalize while accounting for global differences in tissue distribution

We will make use of the `skin` dataset for examples. The `skin` dataset is a small sample of the full GTEx data that can be downloaded using the `downloadGTEx` function. The `skin` dataset looks like this:

```
skin
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 40824 features, 20 samples
##   element names: exprs
## protocolData: none
## phenoData
##   rowNames: GTEX-OHPL-0008-SM-4E3I9 GTEX-145MN-1526-SM-5SI9T ...
##     GTEX-144FL-0626-SM-5LU43 (20 total)
##   varLabels: SAMPID SMATSSCR ... DTHHRDY (65 total)
##   varMetadata: labelDescription
## featureData
```

```
##   featureNames: 48350 48365 ... 7565 (40824 total)
##   fvarLabels: ensembl_gene_id hgnc_symbol ... gene_biotype (6
##     total)
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
```

This is a basic workflow. Details will be fleshed out:

0. First always remember to have the library loaded.

```
library(yarn)
```

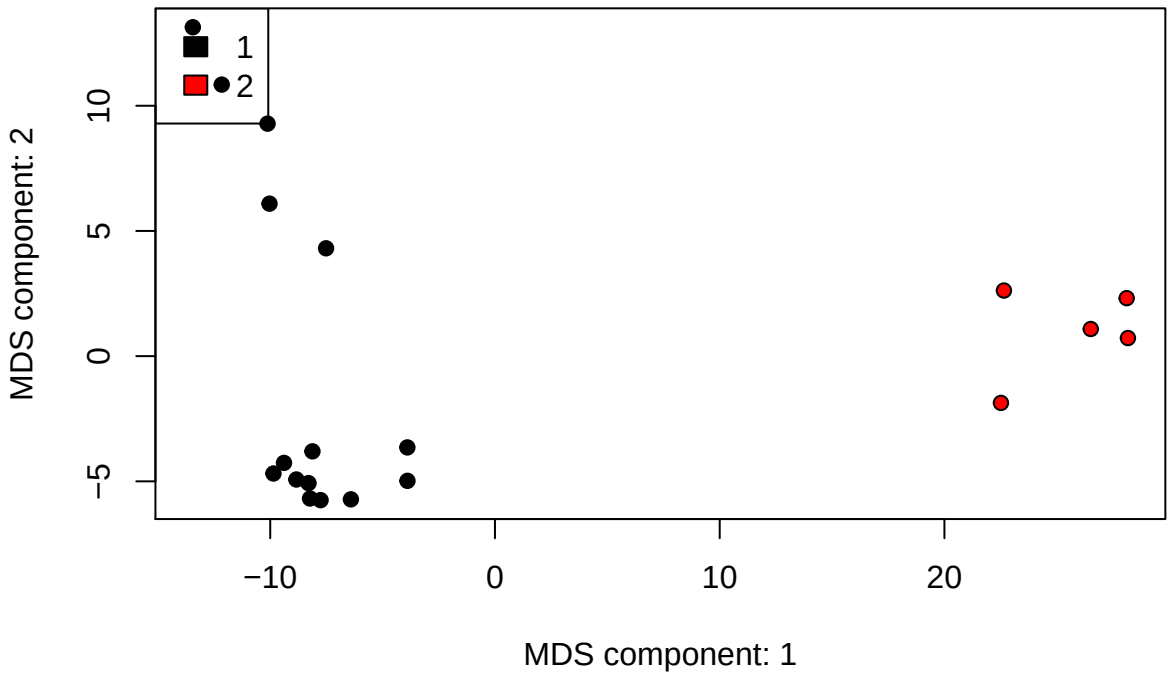
1. Download the GTEx gene count data as an ExpressionSet object or load the sample skin dataset.

For computational reasons we load the sample skin data instead of having the user download the

```
library(yarn)
data(skin)
```

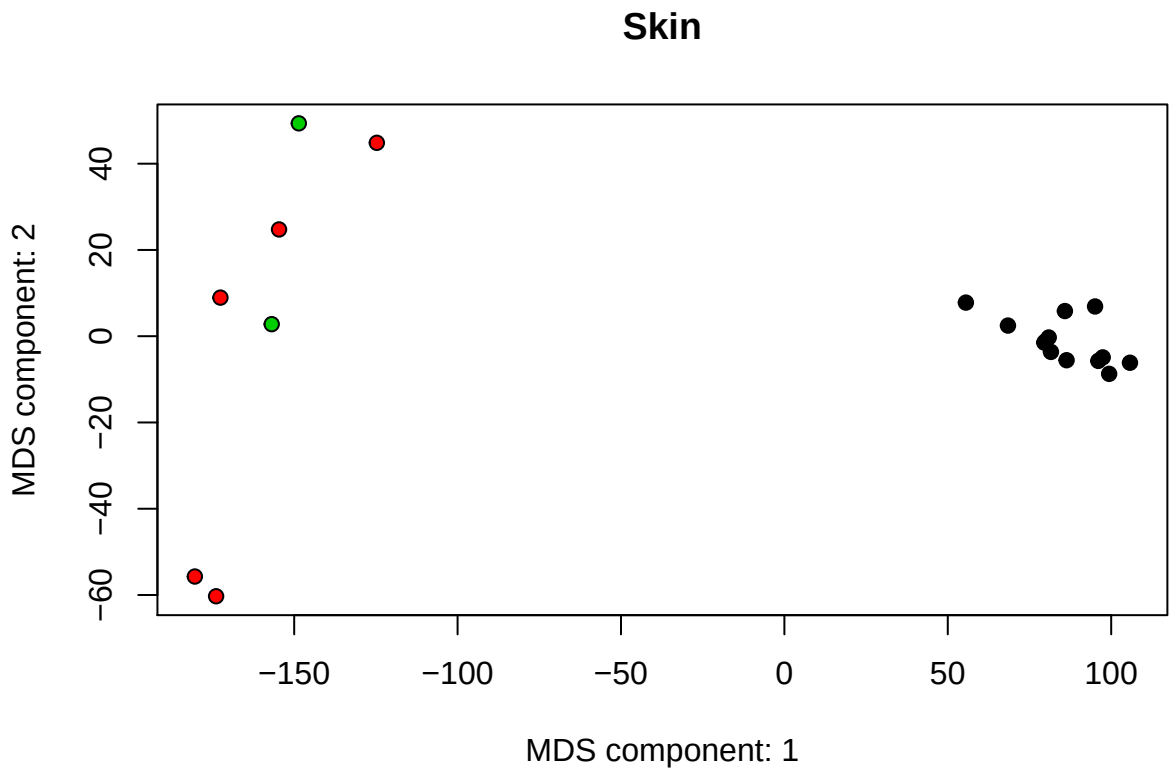
2. Check mis-annotation of gender or other phenotypes using group-specific genes

```
checkMisAnnotation(skin,"GENDER",controlGenes="Y",legendPosition="topleft")
```



3. Decide what sub-groups should be merged

```
checkTissuesToMerge(skin,"SMTS","SMTSD")
```



4. Filter lowly expressed genes

```
skin_filtered = filterLowGenes(skin,"SMTSD")
dim(skin)
```

```
## Features  Samples
##    40824       20
```

```
dim(skin_filtered)
```

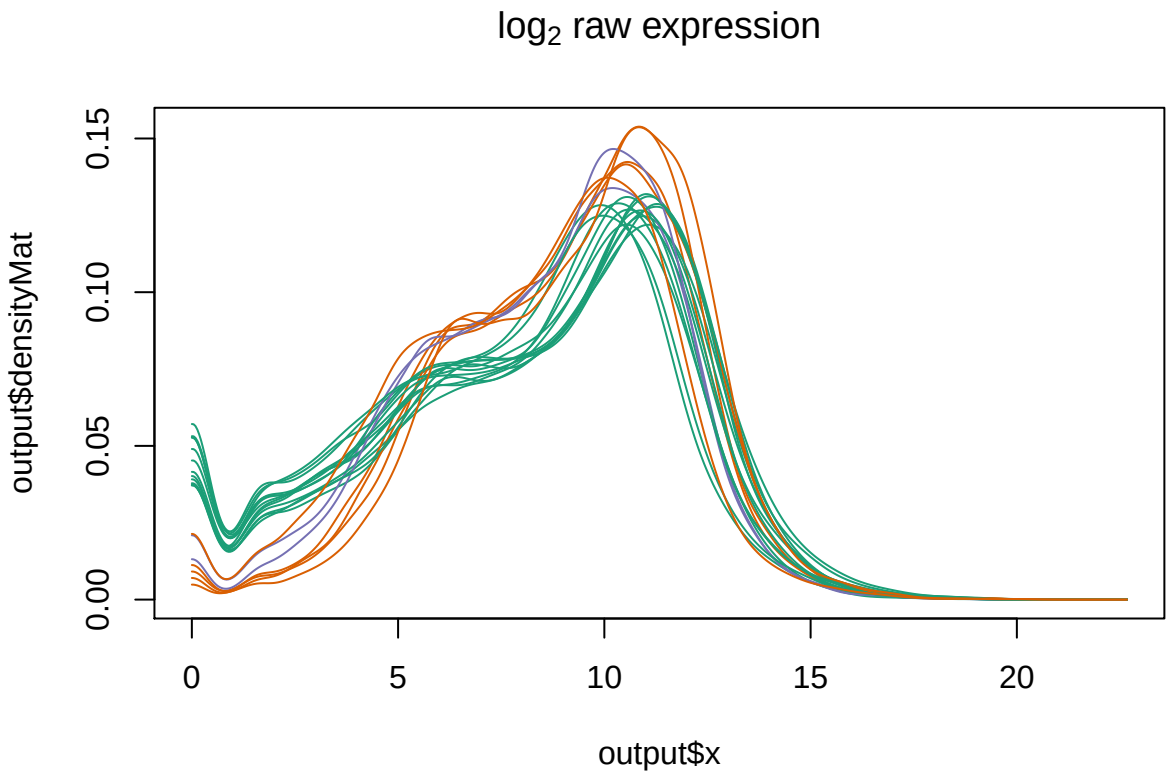
```
## Features  Samples
##    19933       20
```

Or group specific genes

```
tmp = filterGenes(skin,labels=c("X","Y","MT"),featureName = "chromosome_name")
# Keep only the sex names
tmp = filterGenes(skin,labels=c("X","Y","MT"),featureName = "chromosome_name",keepOnly=TRUE)
```

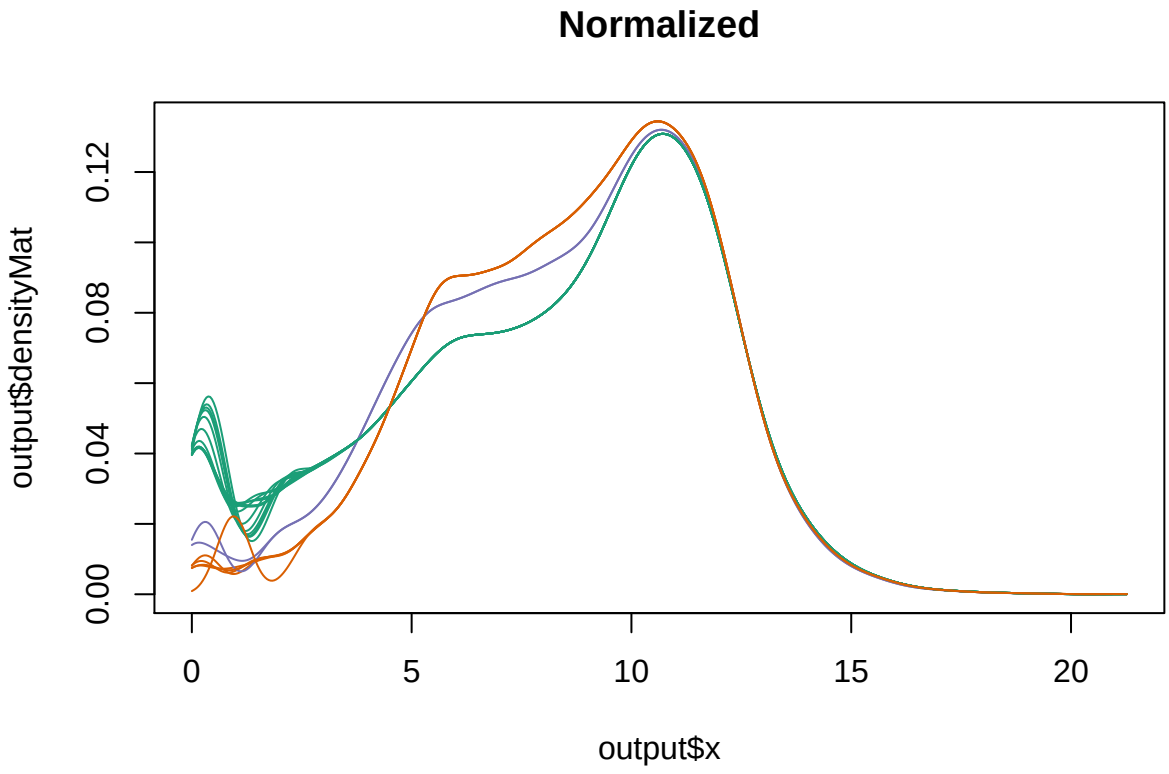
5. Normalize in a tissue or group-aware manner

```
plotDensity(skin_filtered,"SMTSD",main=expression('log'[2]*' raw expression'))
```




```
skin_filtered = normalizeTissueAware(skin_filtered,"SMTSD")
plotDensity(skin_filtered,"SMTSD",normalized=TRUE,main="Normalized")
```

```
## normalizedMatrix is assumed to already be log-transformed
```

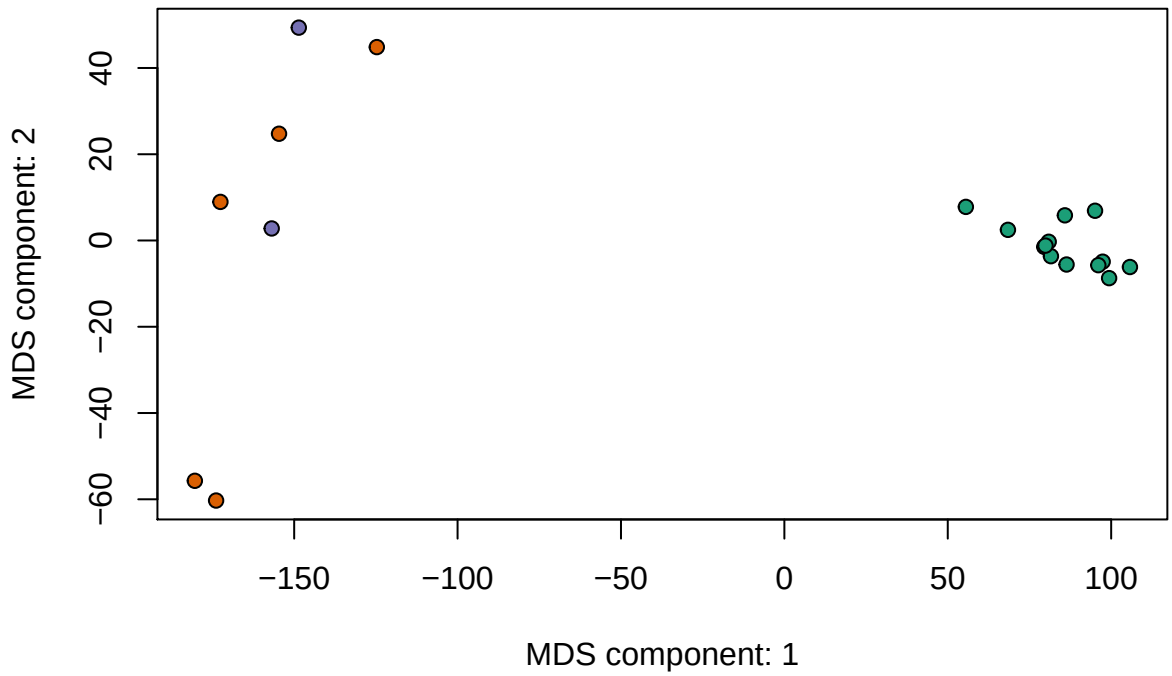


Helper functions

Other than `checkMisAnnotation` and `checkTissuesToMerge` we provide a few plotting function. We include, `plotCMDS`, `plotDensity`, `plotHeatmap`.

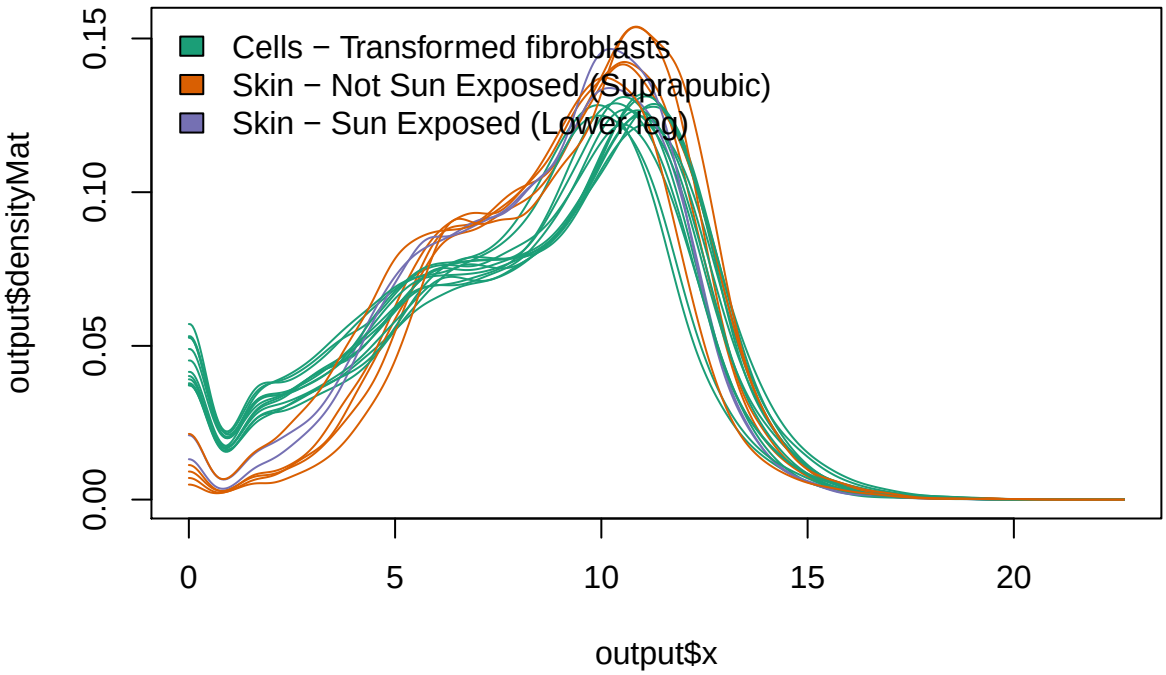
`plotCMDS` - PCoA / Classical Multi-Dimensional Scaling of the most variable genes.

```
data(skin)
res = plotCMDS(skin,pch=21,bg=factor(pData(skin)$SMTSD))
```



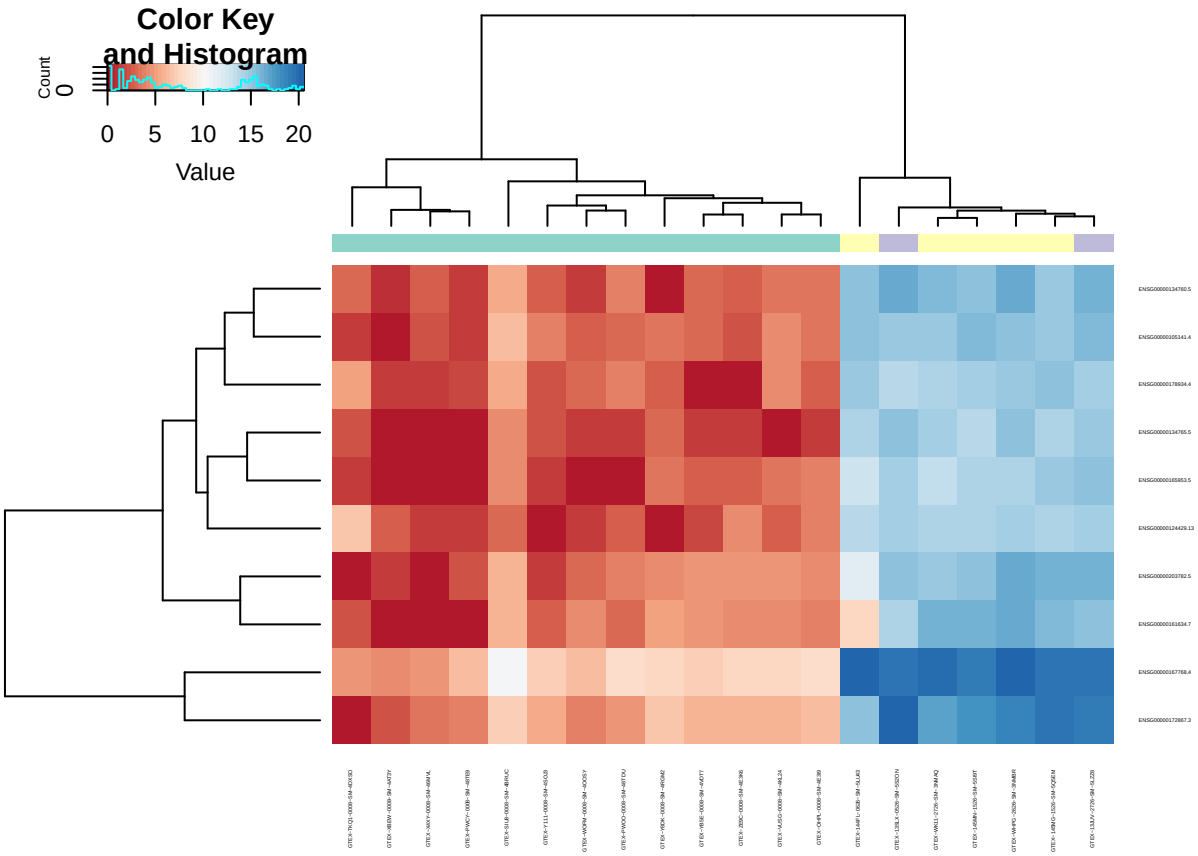
`plotDensity` - Density plots colored by phenotype of choosing. Allows for inspection of global trend differences.

```
filtData = filterLowGenes(skin,"SMTSD")
plotDensity(filtData,groups="SMTSD",legendPos="topleft")
```



`plotHeatmap` - Heatmap of the most variable genes.

```
library(RColorBrewer)
tissues = pData(skin)$SMTSD
heatmapColColors=brewer.pal(12,"Set3")[as.integer(factor(tissues))]
heatmapCols = colorRampPalette(brewer.pal(9, "RdBu"))(50)
plotHeatmap(skin,normalized=FALSE,log=TRUE,trace="none",n=10,
 col = heatmapCols,ColSideColors = heatmapColColors,cexRow = 0.25,cexCol = 0.25)
```



Information

```
sessionInfo()
```

```
## R version 3.4.2 (2017-09-28)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.3 LTS
##
## Matrix products: default
## BLAS: /home/biocbuild/bbs-3.6-bioc/R/lib/libRblas.so
## LAPACK: /home/biocbuild/bbs-3.6-bioc/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
```

```
## [1] RColorBrewer_1.1-2  yarn_1.4.0          Biobase_2.38.0
## [4] BiocGenerics_0.24.0
##
## loaded via a namespace (and not attached):
##   [1] nlme_3.1-131               bitops_1.0-6
##   [3] matrixStats_0.52.2         bit64_0.9-7
##   [5] httr_1.3.1                 doParallel_1.0.11
##   [7] progress_1.1.2             rprojroot_1.2
##   [9] GenomeInfoDb_1.14.0        tools_3.4.2
##  [11] backports_1.1.1            doRNG_1.6.6
##  [13] nor1mix_1.2-3              R6_2.2.2
##  [15] KernSmooth_2.23-15         DBI_0.7
##  [17] lazyeval_0.2.1             colorspace_1.3-2
##  [19] prettyunits_1.0.2          RMySQL_0.10.13
##  [21] base64_2.0                 bit_1.1-12
##  [23] compiler_3.4.2             preprocessCore_1.40.0
##  [25] xml2_1.1.1                 DelayedArray_0.4.0
##  [27] pkgmaker_0.22              rtracklayer_1.38.0
##  [29] caTools_1.17.1             scales_0.5.0
##  [31] readr_1.1.1                quadprog_1.5-5
##  [33] genefilter_1.60.0          Rsamtools_1.30.0
##  [35] stringr_1.2.0              digest_0.6.12
##  [37] illuminaio_0.20.0          rmarkdown_1.6
##  [39] siggenes_1.52.0            GEOquery_2.46.0
##  [41] XVector_0.18.0             pkgconfig_2.0.1
##  [43] htmltools_0.3.6            limma_3.34.0
##  [45] rlang_0.1.2                RSQLite_2.0
##  [47] bindr_0.1                  quantro_1.12.0
##  [49] BiocParallel_1.12.0        mclust_5.3
##  [51] gtools_3.5.0               dplyr_0.7.4
##  [53] RCurl_1.95-4.8             magrittr_1.5
##  [55] GenomeInfoDbData_0.99.1    Matrix_1.2-11
##  [57] Rcpp_0.12.13               munsell_0.4.3
##  [59] S4Vectors_0.16.0           stringi_1.1.5
##  [61] yaml_2.1.14                edgeR_3.20.0
##  [63] MASS_7.3-47                SummarizedExperiment_1.8.0
##  [65] zlibbioc_1.24.0            gplots_3.0.1
##  [67] plyr_1.8.4                 bumphunter_1.20.0
##  [69] grid_3.4.2                 minfi_1.24.0
##  [71] blob_1.1.0                 gdata_2.18.0
##  [73] lattice_0.20-35            Biostrings_2.46.0
##  [75] splines_3.4.2              multtest_2.34.0
##  [77] GenomicFeatures_1.30.0     annotate_1.56.0
##  [79] hms_0.3                    locfit_1.5-9.1
##  [81] knitr_1.17                 beanplot_1.2
##  [83] GenomicRanges_1.30.0       rngtools_1.2.4
##  [85] codetools_0.2-15           biomaRt_2.34.0
##  [87] stats4_3.4.2               glue_1.2.0
##  [89] XML_3.98-1.9               evaluate_0.10.1
##  [91] downloader_0.4             data.table_1.10.4-3
##  [93] foreach_1.4.3              purrr_0.2.4
##  [95] tidyr_0.7.2                gtable_0.2.0
##  [97] openssl_0.9.7              reshape_0.8.7
##  [99] assertthat_0.2.0           ggplot2_2.2.1
```

```
## [101] xtable_1.8-2               survival_2.41-3
## [103] tibble_1.3.4               iterators_1.0.8
## [105] GenomicAlignments_1.14.0   AnnotationDbi_1.40.0
## [107] registry_0.3               memoise_1.1.0
## [109] IRanges_2.12.0             bindrcpp_0.2
```