RNAseq analysis

Maarten van Iterson

6/13/2017

A DESeqData obtject is created from the airway data.

```
library(airway)
```

```
## Loading required package: SummarizedExperiment
## Loading required package: GenomicRanges
## Loading required package: stats4
## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
## The following objects are masked from 'package:stats':
##
##     IQR, mad, xtabs
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colnames,
##     do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, lengths, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff,
##     sort, table, tapply, union, unique, unsplit, which, which.max,
##     which.min
## Loading required package: S4Vectors
##
## Attaching package: 'S4Vectors'
## The following objects are masked from 'package:base':
##
##     colMeans, colSums, expand.grid, rowMeans, rowSums
## Loading required package: IRanges
## Loading required package: GenomeInfoDb
## Loading required package: Biobase
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
```

```
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
library(DESeq2)
data("airway")
airway$dex <- relevel(airway$dex, "untrt")
dds <- DESeqDataSet(airway, design = ~ cell + dex) #add formula
dds
```

```
## class: DESeqDataSet
## dim: 64102 8
## metadata(2): '' version
## assays(1): counts
## rownames(64102): ENSG00000000003 ENSG00000000005 ... LRG_98 LRG_99
## rowData names(0):
## colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521
## colData names(9): SampleName cell ... Sample BioSample
```

Filtering un- or lowly expressed genes using counts per million is advocated by the developers of edgeR[@] another package for the differential expression analysis (see section 2.6 Filtering).

```
cpm <- 1e6*counts(dds)/colSums(counts(dds))
keep <- rowSums(cpm>1) >= 4
dds <- dds[keep, ]
dds
```

```
## class: DESeqDataSet
## dim: 14360 8
## metadata(2): '' version
## assays(1): counts
## rownames(14360): ENSG00000000003 ENSG00000000419 ...
##   ENSG00000273356 ENSG00000273373
## rowData names(0):
## colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521
## colData names(9): SampleName cell ... Sample BioSample
```

Differential expression analysis is performed using the DESeq2-packages. An FDR of 5% is used to determine differential expressed genes.

```
library(org.Hs.eg.db)
```

```
## Loading required package: AnnotationDbi
```

```
##
```

```
dds <- DESeq(dds)
```

```
## estimating size factors
```

```
## estimating dispersions
```

```
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

```
## fitting model and testing
```

```
res <- results(dds, alpha = 0.05)
res$Symbol <- mapIds(org.Hs.eg.db, rownames(res), "SYMBOL", "ENSEMBL")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
res[order(res$padj),]
```

```
## log2 fold change (MAP): dex trt vs untrt
## Wald test p-value: dex trt vs untrt
## DataFrame with 14360 rows and 7 columns
##                    baseMean log2FoldChange      lfcSE          stat
##                   <numeric>      <numeric>  <numeric>     <numeric>
## ENSG00000152583    997.5202       4.293616  0.1721530      24.94071
## ENSG00000165995    495.4311       3.174093  0.1274643      24.90182
## ENSG00000101347  12708.7527       3.604035  0.1489683      24.19329
## ENSG00000120129   3411.4330       2.858802  0.1185387      24.11704
## ENSG00000189221   2342.8234       3.216087  0.1366041      23.54312
## ...                     ...            ...        ...           ...
## ENSG00000009307 15828.50403  -4.327685e-05 0.07700615 -5.619921e-04
## ENSG00000123728   546.27594  -4.533937e-05 0.10468659 -4.330962e-04
## ENSG00000135722    75.27579   9.042885e-05 0.19991225  4.523427e-04
## ENSG00000173531   228.88746  -7.440328e-05 0.15275451 -4.870775e-04
## ENSG00000180673    23.73033   2.341427e-06 0.30301019  7.727221e-06
##                        pvalue          padj      Symbol
##                     <numeric>     <numeric> <character>
## ENSG00000152583 2.693499e-137 3.867864e-133     SPARCL1
## ENSG00000165995 7.110821e-137 5.105569e-133      CACNB2
## ENSG00000101347 2.617388e-129 1.252856e-125      SAMHD1
## ENSG00000120129 1.656548e-128 5.947008e-125       DUSP1
## ENSG00000189221 1.476513e-122 4.240545e-119        MAOA
## ...                       ...           ...         ...
## ENSG00000009307     0.9995516     0.9997241       CSDE1
## ENSG00000123728     0.9996544     0.9997241       RAP2C
## ENSG00000135722     0.9996391     0.9997241       FBXL8
## ENSG00000173531     0.9996114     0.9997241        MST1
## ENSG00000180673     0.9999938     0.9999938          NA
```

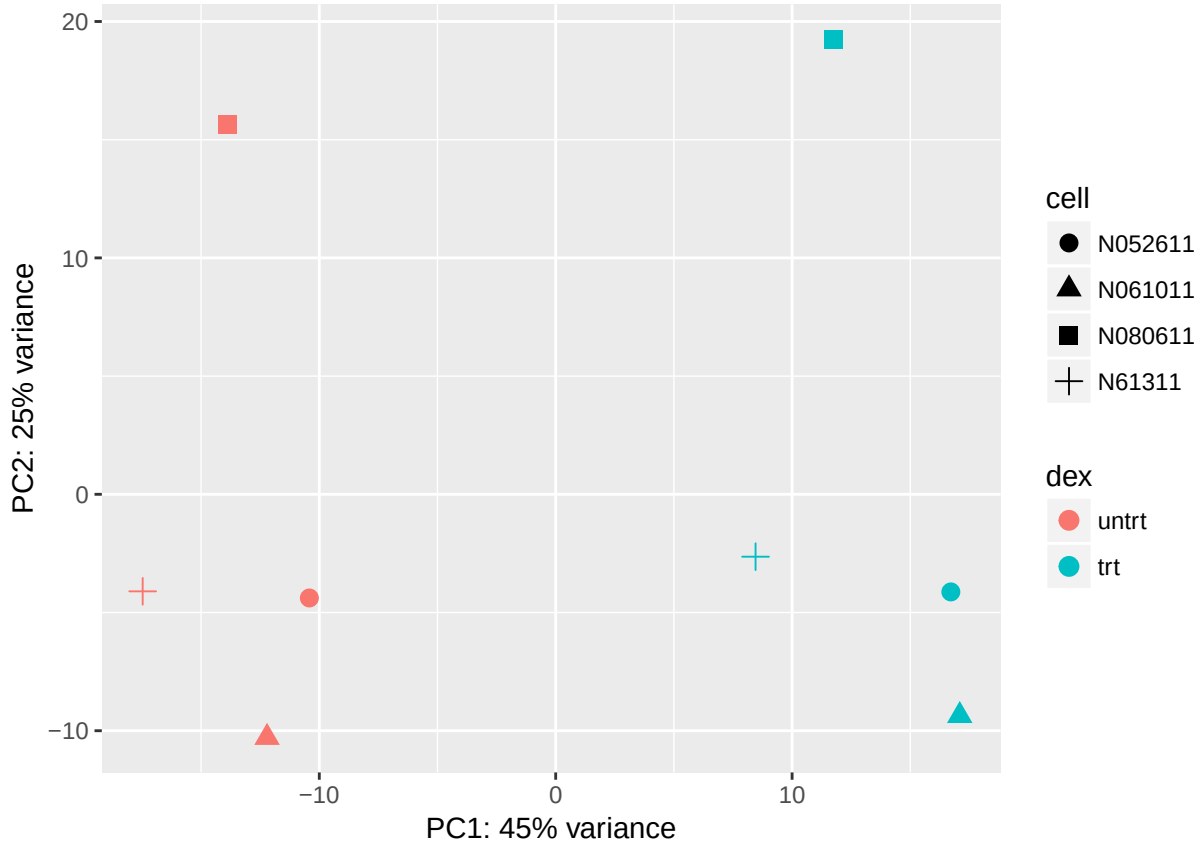
A principal component plot for exploratory analysis.

```
library(vsn)
library(ggplot2)
rld <- rlog(dds, blind = FALSE)
pcaData <- plotPCA(rld, intgroup = c( "dex", "cell"), returnData = TRUE)
pcaData
```

```
##                    PC1        PC2           group   dex    cell       name
## SRR1039508 -17.470328  -4.104236  untrt : N61311 untrt  N61311 SRR1039508
## SRR1039509   8.454183  -2.638643    trt : N61311   trt  N61311 SRR1039509
## SRR1039512 -10.419877  -4.384220 untrt : N052611 untrt N052611 SRR1039512
## SRR1039513  16.715160  -4.128324   trt : N052611   trt N052611 SRR1039513
## SRR1039516 -13.893510  15.656023 untrt : N080611 untrt N080611 SRR1039516
## SRR1039517  11.732798  19.249447   trt : N080611   trt N080611 SRR1039517
## SRR1039520 -12.209613 -10.287692 untrt : N061011 untrt N061011 SRR1039520
## SRR1039521  17.091186  -9.362354   trt : N061011   trt N061011 SRR1039521
```

```
percentVar <- round(100 * attr(pcaData, "percentVar"))
ggplot(pcaData, aes(x = PC1, y = PC2, color = dex, shape = cell)) +
  geom_point(size =3) +
  xlab(paste0("PC1: ", percentVar[1], "% variance")) +
```

```
  ylab(paste0("PC2: ", percentVar[2], "% variance")) +
  coord_fixed()
```



Overview of R and package version used in this analysis.

```
sessionInfo()
```

```
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.2 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [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  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] ggplot2_2.2.1              vsn_3.42.3
##  [3] org.Hs.eg.db_3.4.0         AnnotationDbi_1.36.2
##  [5] DESeq2_1.14.1              airway_0.108.0
```

```
##  [7] SummarizedExperiment_1.4.0 Biobase_2.34.0
##  [9] GenomicRanges_1.26.4       GenomeInfoDb_1.10.3
## [11] IRanges_2.8.2              S4Vectors_0.12.2
## [13] BiocGenerics_0.20.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.11          locfit_1.5-9.1        lattice_0.20-35
##  [4] RevoUtilsMath_10.0.0  rprojroot_1.2         digest_0.6.12
##  [7] plyr_1.8.4            backports_1.1.0       acepack_1.4.1
## [10] RSQLite_1.1-2         evaluate_0.10         BiocInstaller_1.24.0
## [13] zlibbioc_1.20.0       rlang_0.1.1           lazyeval_0.2.0
## [16] data.table_1.10.4     annotate_1.52.1       rpart_4.1-11
## [19] Matrix_1.2-10         checkmate_1.8.2       preprocessCore_1.36.0
## [22] rmarkdown_1.5         labeling_0.3          splines_3.3.2
## [25] BiocParallel_1.8.2    geneplotter_1.52.0    stringr_1.2.0
## [28] foreign_0.8-68        htmlwidgets_0.8       RCurl_1.95-4.8
## [31] munsell_0.4.3         base64enc_0.1-3       htmltools_0.3.6
## [34] nnet_7.3-12           tibble_1.3.3          gridExtra_2.2.1
## [37] htmlTable_1.9         Hmisc_4.0-3           XML_3.98-1.7
## [40] bitops_1.0-6          grid_3.3.2            xtable_1.8-2
## [43] gtable_0.2.0          affy_1.52.0           DBI_0.6-1
## [46] magrittr_1.5          scales_0.4.1          stringi_1.1.5
## [49] XVector_0.14.1        genefilter_1.56.0     affyio_1.44.0
## [52] limma_3.30.13         latticeExtra_0.6-28   Formula_1.2-1
## [55] RColorBrewer_1.1-2    tools_3.3.2           survival_2.41-3
## [58] yaml_2.1.14           colorspace_1.3-2      cluster_2.0.6
## [61] memoise_1.1.0         knitr_1.16
```