RNA-Seq workflow template: Some Descriptive Title

Project ID: RNAseq_Pi_Name_Organism_Jun2014
      Project PI: First Last (first.last@inst.edu)
      Author of Report: First Last (first.last@inst.edu)

April 24, 2017

Contents

1 Introduction .......................... 1

2 Sample definitions and environment settings ....... 2
   2.1 Environment settings and input data .......... 2
   2.2 Required packages and resources ........... 2
   2.3 Experiment definition provided by targets file .... 2

3 Read preprocessing ........................ 3
   3.1 FASTQ quality report ................... 3

4 Alignments .................................. 3
   4.1 Read mapping with Bowtie2/Tophat2 .......... 3
   4.2 Read and alignment stats ................. 4
   4.3 Create symbolic links for viewing BAM files in IGV .... 4

5 Read quantification per annotation range .......... 4
   5.1 Read counting with summarizeOverlaps in parallel mode using multiple cores .... 4
   5.2 Sample-wise correlation analysis .......... 5

6 Analysis of differentially expressed genes with edgeR .... 6
   6.1 GO term enrichment analysis of DEGs .......... 8
       6.1.1 Obtain gene-to-GO mappings ........... 8
       6.1.2 Batch GO term enrichment analysis ... 9
       6.1.3 Plot batch GO term results .......... 9

7 Clustering and heat maps ..................... 10

8 Version Information ........................ 11

9 Funding .................................... 12

10 References ................................ 12

1 Introduction

This report describes the analysis of an RNA-Seq project from Dr. First Last’s lab which studies the gene expression changes of ... in organism .... The experimental design is as follows...
2 Sample definitions and environment settings

2.1 Environment settings and input data

Typically, the user wants to record here the sources and versions of the reference genome sequence along with the corresponding annotations. In the provided sample data set all data inputs are stored in a data subdirectory and all results will be written to a separate results directory, while the systemPipeRNASeq.Rnw script and the targets file are expected to be located in the parent directory. The R session is expected to run from this parent directory.

To run this sample report, mini sample FASTQ and reference genome files can be downloaded from here. The chosen data set SRP010938 contains 18 paired-end (PE) read sets from *Arabidopsis thaliana* (Howard et al., 2013). To minimize processing time during testing, each FASTQ file has been subsetted to 90,000-100,000 randomly sampled PE reads that map to the first 100,000 nucleotides of each chromosome of the *A. thalina* genome. The corresponding reference genome sequence (FASTA) and its GFF annotation files (provided in the same download) have been truncated accordingly. This way the entire test sample data set is less than 200MB in storage space. A PE read set has been chosen for this test data set for flexibility, because it can be used for testing both types of analysis routines requiring either SE (single end) reads or PE reads.

2.2 Required packages and resources

The systemPipeR package needs to be loaded to perform the analysis steps shown in this report (Girke, 2014).

```
library(systemPipeR)
```

If applicable load custom functions not provided by systemPipeR

```
source("systemPipeRNAseq_Fct.R")
```

2.3 Experiment definition provided by targets file

The targets file defines all FASTQ files and sample comparisons of the analysis workflow.

```
targetspath <- system.file("extdata", "targets.txt", package="systemPipeR")
targets <- read.delim(targetspath, comment.char = "]")[,1:4]
targets
```

<table>
<thead>
<tr>
<th>FileName</th>
<th>SampleName</th>
<th>Factor</th>
<th>SampleLong</th>
</tr>
</thead>
<tbody>
<tr>
<td>./data/SRR446027_1.fastq</td>
<td>M1A</td>
<td>M1</td>
<td>Mock.1h.A</td>
</tr>
<tr>
<td>./data/SRR446028_1.fastq</td>
<td>M1B</td>
<td>M1</td>
<td>Mock.1h.B</td>
</tr>
<tr>
<td>./data/SRR446029_1.fastq</td>
<td>A1A</td>
<td>A1</td>
<td>Avr.1h.A</td>
</tr>
<tr>
<td>./data/SRR446030_1.fastq</td>
<td>A1B</td>
<td>A1</td>
<td>Avr.1h.B</td>
</tr>
<tr>
<td>./data/SRR446031_1.fastq</td>
<td>V1A</td>
<td>V1</td>
<td>Vir.1h.A</td>
</tr>
<tr>
<td>./data/SRR446032_1.fastq</td>
<td>V1B</td>
<td>V1</td>
<td>Vir.1h.B</td>
</tr>
<tr>
<td>./data/SRR446033_1.fastq</td>
<td>M6A</td>
<td>M6</td>
<td>Mock.6h.A</td>
</tr>
<tr>
<td>./data/SRR446034_1.fastq</td>
<td>M6B</td>
<td>M6</td>
<td>Mock.6h.B</td>
</tr>
<tr>
<td>./data/SRR446035_1.fastq</td>
<td>A6A</td>
<td>A6</td>
<td>Avr.6h.A</td>
</tr>
<tr>
<td>./data/SRR446036_1.fastq</td>
<td>A6B</td>
<td>A6</td>
<td>Avr.6h.B</td>
</tr>
<tr>
<td>./data/SRR446037_1.fastq</td>
<td>V6A</td>
<td>V6</td>
<td>Vir.6h.A</td>
</tr>
<tr>
<td>./data/SRR446038_1.fastq</td>
<td>V6B</td>
<td>V6</td>
<td>Vir.6h.B</td>
</tr>
<tr>
<td>./data/SRR446039_1.fastq</td>
<td>M12A</td>
<td>M12</td>
<td>Mock.12h.A</td>
</tr>
<tr>
<td>./data/SRR446040_1.fastq</td>
<td>M12B</td>
<td>M12</td>
<td>Mock.12h.B</td>
</tr>
<tr>
<td>./data/SRR446041_1.fastq</td>
<td>A12A</td>
<td>A12</td>
<td>Avr.12h.A</td>
</tr>
<tr>
<td>./data/SRR446042_1.fastq</td>
<td>A12B</td>
<td>A12</td>
<td>Avr.12h.B</td>
</tr>
<tr>
<td>./data/SRR446043_1.fastq</td>
<td>V12A</td>
<td>V12</td>
<td>Vir.12h.A</td>
</tr>
</tbody>
</table>
3 Read preprocessing

3.1 FASTQ quality report

The following seeFastq and seeFastqPlot functions generate and plot a series of useful quality statistics for a set of FASTQ files including per cycle quality box plots, base proportions, base-level quality trends, relative k-mer diversity, length and occurrence distribution of reads, number of reads above quality cutoffs and mean quality distribution. The results are written to a PDF file named fastqReport.pdf.

```r
args <- systemArgs(sysma="tophat.param", mytargets="targets.txt")
fqlist <- seeFastq(fastq=infile1(args), batchsize=100000, klength=8)
pdf("./results/fastqReport.pdf", height=18, width=4*length(fqlist))
seeFastqPlot(fqlist)
dev.off()
```

Figure 1: QC report for 18 FASTQ files.

4 Alignments

4.1 Read mapping with Bowtie2/Tophat2

The NGS reads of this project will be aligned against the reference genome sequence using Bowtie2/TopHat2 (Kim et al., 2013; Langmead and Salzberg, 2012). The parameter settings of the aligner are defined in the tophat.param file.

```r
args <- systemArgs(sysma="tophat.param", mytargets="targets.txt")
sysargs(args)[1] # Command-line parameters for first FASTQ file
```

Submission of alignment jobs to compute cluster, here using 72 CPU cores (18 qsub processes each with 4 CPU cores).

```r
moduleLoad(modules(args))
system("bowtie2-build ./data/tair10.fasta ./data/tair10.fasta")
resources <- list(walltime="20:00:00", nodes=paste0("1:ppn", cores(args)), memory="10gb")
reg <- clusterRun(args, conffile=".BatchJobs.R", template="torque.tmpl", Njobs=18, runid="01", resourceList=resources)
```
4.2 Read and alignment stats

The following provides an overview of the number of reads in each sample and how many of them aligned to the reference.

```r
countDFeByg <- summarizeOverlaps(eByg, x, mode="Union",
  ignore.strand=TRUE,
  inter.feature=FALSE, singleEnd=TRUE)
countDFeByg <- sapply(seq(along=counteByg), function(x) assays(counteByg[[x]])$counts)
```

4.3 Create symbolic links for viewing BAM files in IGV

The `symLink2bam` function creates symbolic links to view the BAM alignment files in a genome browser such as IGV. The corresponding URLs are written to a file with a path specified under `urlfile`, here `IGVurl.txt`.

```r
countDFeByg <- summarizeOverlaps(eByg, x, mode="Union",
  ignore.strand=TRUE,
  inter.feature=FALSE, singleEnd=TRUE)
countDFeByg <- sapply(seq(along=counteByg), function(x) assays(counteByg[[x]])$counts)
```

5 Read quantification per annotation range

5.1 Read counting with `summarizeOverlaps` in parallel mode using multiple cores

Reads overlapping with annotation ranges of interest are counted for each sample using the `summarizeOverlaps` function (Lawrence et al., 2013). The read counting is preformed for exonic gene regions in a non-strand-specific manner while ignoring overlaps among different genes. Subsequently, the expression count values are normalized by reads per kp per million mapped reads (RPKM). The raw read count table (`countDFeByg.xls`) and the corresponding RPKM table (`rpkmDFeByg.xls`) are written to separate files in the `results` directory of this project. Parallelization is achieved with the `BiocParallel` package, here using 8 CPU cores.

```r
txdb <- loadDb("./data/tair10.sqlite")
eByg <- exonBy(txdb, by="gene")
bfl <- BamFileList(outpaths(args), yieldSize=50000, index=character())
multicoreParam <- MulticoreParam(workers=8); register(multicoreParam); registered()
counteByg <- bplapply(bfl, function(x) summarizeOverlaps(eByg, x, mode="Union",
  ignore.strand=TRUE,
  inter.feature=FALSE, singleEnd=TRUE))
countDFeByg <- sapply(seq(along=counteByg), function(x) assays(counteByg[[x]])$counts)
```
SystemPipeR RNA-Seq Workflow

5 Read quantification per annotation range

```r
rownames(countDFeByg) <- names(rowRanges(counteByg[[1]])); colnames(countDFeByg) <- names(bfl)
rpkmDFeByg <- apply(countDFeByg, 2, function(x) returnRPKM(counts=x, ranges=eByg))
write.table(countDFeByg, "results/countDFeByg.xls", col.names=NA, quote=FALSE, sep="\t")
write.table(rpkmDFeByg, "results/rpkmDFeByg.xls", col.names=NA, quote=FALSE, sep="\t")
```

Sample of data slice of count table

```r
read.delim("results/countDFeByg.xls", row.names=1, check.names=FALSE)[1:4,1:5]
```

Sample of data slice of RPKM table

```r
read.delim("results/rpkmDFeByg.xls", row.names=1, check.names=FALSE)[1:4,1:4]
```

Note, for most statistical differential expression or abundance analysis methods, such as edgeR or DESeq2, the raw count values should be used as input. The usage of RPKM values should be restricted to specialty applications required by some users, e.g. manually comparing the expression levels among different genes or features.

5.2 Sample-wise correlation analysis

The following computes the sample-wise Spearman correlation coefficients from the rlog transformed expression values generated with the DESeq2 package. After transformation to a distance matrix, hierarchical clustering is performed with the hclust function and the result is plotted as a dendrogram (sample_tree.pdf).

```r
library(DESeq2, quietly=TRUE); library(ape, warn.conflicts=FALSE)
countDF <- as.matrix(read.table("./results/countDFeByg.xls"))
colData <- data.frame(row.names=targetsin(args)$SampleName, condition=targetsin(args)$Factor)
dds <- DESeqDataSetFromMatrix(countData = countDF, colData = colData, design = ~ condition)
d <- cor(assay(rlog(dds)), method="spearman")
hc <- hclust(dist(1-d))
pdf("results/sample_tree.pdf")
plot.phylo(as.phylo(hc), type="p", edge.col="blue", edge.width=2, show.node.label=TRUE, no.margin=TRUE)
dev.off()
```
6 Analysis of differentially expressed genes with edgeR

The analysis of differentially expressed genes (DEGs) is performed with the glm method from the edgeR package (Robinson et al., 2010). The sample comparisons used by this analysis are defined in the header lines of the targets file starting with <CMP>.

```r
library(edgeR)
countDF <- read.delim("countDFeByg.xls", row.names=1, check.names=FALSE)
targets <- read.delim("targets.txt", comment="#")
cmp <- readComp(file="targets.txt", format="matrix", delim="-")
edgeDF <- run_edgeR(countDF=countDF, targets=targets, cmp=cmp[[1]], independent=FALSE, mdsplot="")

Add custom functional descriptions. Skip this step if desc.xls is not available.

desc <- read.delim("data/desc.xls")
desc <- desc[!duplicated(desc[,1]),]
descv <- as.character(desc[,2]); names(descv) <- as.character(desc[,1])
edgeDF <- data.frame(edgeDF, Desc=descv[rownames(edgeDF)], check.names=FALSE)
write.table(edgeDF, "/results/edgeRglm_allcomp.xls", quote=FALSE, sep="\t", col.names = NA)
```

Filter and plot DEG results for up and down regulated genes. The definition of 'up' and 'down' is given in the corresponding help file. To open it, type ?filterDEGs in the R console.

```r
degDF <- read.delim("results/edgeRglm_allcomp.xls", row.names=1, check.names=FALSE)
pdf("results/DEGcounts.pdf")
DEG_list <- filterDEGs(degDF=degDF, filter=c(Fold=2, FDR=1))
dev.off()
```
write.table(DEG_list$Summary, "./results/DEGcounts.xls", quote=FALSE, sep="\t", row.names=FALSE)

The function `overLapper` can compute Venn intersects for large numbers of sample sets (up to 20 or more) and `vennPlot` can plot 2-5 way Venn diagrams. A useful feature is the possibility to combine the counts from several Venn comparisons with the same number of sample sets in a single Venn diagram (here for 4 up and down DEG sets).

```r
vennsetup <- overLapper(DEG_list$Up[6:9], type="vennsets")
vennsetdown <- overLapper(DEG_list$Down[6:9], type="vennsets")

pdf("results/vennplot.pdf")

vennPlot(list(vennsetup, vennsetdown), mymain="", mysub="", colmode=2, ccol=c("blue", "red"))

dev.off()
```

Figure 3: Up and down regulated DEGs with FDR of 1%.
6.1 GO term enrichment analysis of DEGs

6.1.1 Obtain gene-to-GO mappings

The following shows how to obtain gene-to-GO mappings from biomaRt (here for A. thaliana) and how to organize them for the downstream GO term enrichment analysis. Alternatively, the gene-to-GO mappings can be obtained for many organisms from Bioconductor’s *.db genome annotation packages or GO annotation files provided by various genome databases. For each annotation this relatively slow preprocessing step needs to be performed only once. Subsequently, the preprocessed data can be loaded with the load function as shown in the next subsection.

```r
library("biomaRt")
listMarts() # To choose BioMart database
m <- useMart("ENSEMBL_MART_PLANT"); listDatasets(m)
m <- useMart("ENSEMBL_MART_PLANT", dataset="athaliana_eg_gene")
listAttributes(m) # Choose data types you want to download
go <- getBM(attributes=c("go_accession", "tair_locus", "go_namespace_1003"), mart=m)
go <- go[go[,3]!="",]; go[,3] <- as.character(go[,3])
```

Figure 4: Venn Diagram for 4 Up and Down DEG Sets.
6 Analysis of differentially expressed genes with edgeR

```r
go[1:4,]
dir.create("./data/GO")
write.table(go, "data/GO/GOannotationsBiomart_mod.txt", quote=FALSE, row.names=FALSE, col.names=FALSE, sep="
")
catdb <- makeCATdb(myfile="data/GO/GOannotationsBiomart_mod.txt", lib=NULL, org="", colno=c(1,2,3), idconv=NULL)
save(catdb, file="data/GO/catdb.RData")
```

### 6.1.2 Batch GO term enrichment analysis

Apply the enrichment analysis to the DEG sets obtained the above differential expression analysis. Note, in the following example the FDR filter is set here to an unreasonably high value, simply because of the small size of the toy data set used in this vignette. Batch enrichment analysis of many gene sets is performed with the \texttt{GOCluster_Report} function. When \texttt{method="all"}, it returns all GO terms passing the p-value cutoff specified under the \texttt{cutoff} arguments. When \texttt{method="slim"}, it returns only the GO terms specified under the \texttt{myslimv} argument. The given example shows how a GO slim vector for a specific organism can be obtained from BioMart.

```r
load("data/GO/catdb.RData")
DEG_list <- filterDEGs(degDF=edgeDF, filter=c(Fold=2, FDR=50), plot=FALSE)
up_down <- DEG_list$UporDown; names(up_down) <- paste(names(up_down), "_up_down", sep="")
up <- DEG_list$Up; names(up) <- paste(names(up), "_up", sep="")
down <- DEG_list$Down; names(down) <- paste(names(down), "_down", sep="")
DEGlist <- c(up_down, up, down)
DEGlist <- DEGlist[apply(DEGlist, length) > 0]
BatchResult <- GOCluster_Report(catdb=catdb, setlist=DEGlist, method="all", id_type="gene", CLSZ=2, cutoff=0.9, gocats=c("MF", "BP", "CC"), recordSpecGO=NULL)
library("biomaRt"); m <- useMart("ENSEMBL_MART_PLANT", dataset="athaliana_eg_gene")
goslimvec <- as.character(getBM(attributes=c("goslim_goa_accession"), mart=m)[,1])
BatchResultslim <- GOCluster_Report(catdb=catdb, setlist=DEGlist, method="slim", id_type="gene", myslimv=goslimvec)
```

### 6.1.3 Plot batch GO term results

The \texttt{data.frame} generated by \texttt{GOCluster_Report} can be plotted with the \texttt{goBarplot} function. Because of the variable size of the sample sets, it may not always be desirable to show the results from different DEG sets in the same bar plot. Plotting single sample sets is achieved by subsetting the input data frame as shown in the first line of the following example.

```r
gos <- BatchResultslim[grep("M6-V6_up_down", BatchResultslim$CLID), ]
gos <- BatchResultslim
pdf("GOSlimbarplotMF.pdf", height=8, width=10); goBarplot(gos, gocat="MF"); dev.off()
goBarplot(gos, gocat="BP")
goBarplot(gos, gocat="CC")
```
7 Clustering and heat maps

The following example performs hierarchical clustering on the rlog transformed expression matrix subsetted by the DEGs identified in the above differential expression analysis. It uses a Pearson correlation-based distance measure and complete linkage for cluster joining.

```r
library(pheatmap)
geneids <- unique(as.character(unlist(DEG_list[[1]])))
y <- assay(rlog(dds))[geneids, ]
pdf("heatmap1.pdf")
pheatmap(y, scale="row", clustering_distance_rows="correlation", clustering_distance_cols="correlation")
```

Figure 5: GO Slim Barplot for MF Ontology.
Figure 6: Heat map with hierarchical clustering dendrograms of DEGs.

8 Version Information

toLatex(sessionInfo())

- R version 3.4.0 (2017-04-21), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, 
  LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_ADDRESS=C, LC_TELEPHONE=C, 
  LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Running under: Ubuntu 16.04.2 LTS
- Matrix products: default
- BLAS: /home/biocbuild/bbs-3.5-bioc/R/lib/libRblas.so
- LAPACK: /home/biocbuild/bbs-3.5-bioc/R/lib/libRlapack.so
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: Biobase 2.36.0, BiocGenerics 0.22.0, BiocParallel 1.10.0, BiocStyle 2.4.0, Biostrings 2.44.0, 
  DESeq2 1.16.0, DelayedArray 0.2.0, GenomInfoDb 1.12.0, GenomicAlignments 1.12.0, GenomicRanges 1.28.0, 
  IRanges 2.10.0, Rsamtools 1.28.0, S4Vectors 0.14.0, ShortRead 1.34.0, SummarizedExperiment 1.6.0, 
  XVector 0.16.0, ape 4.1, ggplot2 2.2.1, knitr 1.15.1, matrixStats 0.52.2, systemPipeR 1.10.0
- Loaded via a namespace (and not attached): AnnotationDbi 1.38.0, AnnotationForge 1.18.0, BBmisc 1.11, 
  BatchJobs 1.6, Category 2.42.0, DBI 0.6-1, Formula 1.2-1, GO.db 3.4.1, GOstats 2.42.0, GSEABase 1.38.0,
GenomeInfoDbData 0.99.0, GenomicFeatures 1.28.0, Hmisc 4.0-2, Matrix 1.2-9, RBGL 1.52.0, RColorBrewer 1.1-2, RCurl 1.95-4.8, RSQLite 1.1-2, Rcpp 0.12.10, XML 3.98-1.6, acepack 1.4.1, annotate 1.54.0, backports 1.0.5, base64enc 0.1-3, biomaRt 2.32.0, bitops 1.0-6, brew 1.0-6, checkmate 1.8.2, cluster 2.0.6, codeTools 0.2-15, colorspace 1.3-2, compiler 3.4.0, data.table 1.10.4, digest 0.6.12, edgeR 3.18.0, evaluate 0.10, fail 1.3, foreign 0.8-67, genefilter 1.58.0, genepattern 1.54.0, graph 1.54.0, grid 3.4.0, gridExtra 2.2.1, gtable 0.2.0, highr 0.6, htmlTable 1.9, htmltools 0.3.5, htmlwidgets 0.8, hwriter 1.3.2, labeling 0.3, lattice 0.20-35, latticeExtra 0.6-28, lazyeval 0.2.0, limma 3.32.0, locfit 1.5-9.1, magrittr 1.5, memoise 1.1.0, munsell 0.4.3, nlm 3.1-131, nnet 7.3-12, pheatmap 1.0.8, plyr 1.8.4, rjson 0.2.15, rmarkdown 1.1.4, rpact 4.1-11, rprojroot 1.2, rtracklayer 1.36.0, scales 0.4.1, sendmailR 1.2-1, splines 3.4.0, stringr 1.1.5, stringr 1.2.0, survival 2.41-3, tibble 1.3.0, tools 3.4.0, xtable 1.8-2, yaml 2.1.14, zlibbioc 1.22.0

9 Funding

This project was supported by funds from the National Institutes of Health (NIH).

10 References


