Analysis of RNA-Seq Data with R/Bioconductor

MCBIOS-2015 Workshop

Thomas Girke

March 12, 2015
Overview

RNA-Seq Analysis
  Quality Report
  Aligning Short Reads
  Counting Reads per Feature
  DEG Analysis
  GO Analysis
  View Results in IGV & ggbio
  Differential Exon Usage

References
Outline

Overview

RNA-Seq Analysis
- Quality Report
- Aligning Short Reads
- Counting Reads per Feature
- DEG Analysis
- GO Analysis
- View Results in IGV & ggbio
- Differential Exon Usage

References
RNA-Seq Technology

1. mRNA Isolation

2. Illumina Sequencing

3. Align Sequences against Genome

4. Generate Sequence Counts for all Genes in Genome

Sample 1
Gene A: 30
Gene B: 10
30/10 = 3 fold change

Sample 2
Gene A: 10
Gene B: 5
10/5 = 2 fold change
Analysis Workflow of RNA-Seq Gene Expression Data

1. Alignment of RNA reads to reference
   - Reference can be genome or transcriptome.

2. Count reads overlapping with annotation features of interest
   - Most common: counts for exonic gene regions, but many viable alternatives exist here: counts per exons, genes, introns, etc.

3. Normalization
   - Main adjustment for sequencing depth and compositional bias.

4. Identification of Differentially Expressed Genes (DEGs)
   - Identification of genes with significant expression differences.
   - Identification of expressed genes possible for strongly expressed ones.

5. Specialty applications
   - Splice variant discovery (semi-quantitative), gene discovery, antisense expressions, etc.

6. Cluster Analysis
   - Identification of genes with similar expression profiles across many samples.

7. Enrichment Analysis of Functional Annotations
   - Gene ontology analysis of obtained gene sets from steps 5-6.
Important Aspects in RNA-Seq Analysis

- Alignment reference
  - Genome
  - Transcript models
  - Both
- How to quantify expression?
  - Read count per range
  - Coverage statistics per range
- What features?
  - Genes, transcript models, exons
- Alternative splicing
  - Often restricted to splice junction analysis
  - Objective: discovery vs. quantification
In NGS we usually want to find the **origin of reads** (NG sequences) in a reference genome or transcriptome. Thus, we are mostly interested in finding the best scoring or multiple best scoring locations for each read, but not lower scoring alternative solutions as in paralog/ortholog search applications.

**Ambiguous mappings** should be removed, because there is no evidence for their origin. However, for certain applications one needs to include them, e.g. when mapping RNA-Seq reads against transcript sequences instead of genome.
Short Read Aligner for RNA-Seq

No special requirements for alignments with low number of variants

- ChIP-Seq
- RNA-Seq (if mapping against transcriptome or intron-less genome)
- Bis-Seq (with injected reference)
- ...

Variant tolerant aligners to account for mismatches and indels

- VAR-Seq
- Bis-Seq (without injected reference)
- ...

Splice tolerant aligner to account for introns

- RNA-Seq (if mapping against genome with introns)
SAM is a tab-delimited alignment format consisting of a header section (lines starting with @) and an alignment section with 12 columns. BAM is the compressed, indexed and binary version of this format.

The below sample alignment contains the following features: (1) bases in lower cases are clipped from the alignment; (2) read r001/1 and r001/2 constitute a read pair; (3) r003 is a chimeric read; (4) r004 represents a split alignment.

Coor 12345678901234 56789012345678901234567890123456789012345
ref AGCATGGTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1 TTAGATAAAGGATA*CTG
+r002 aaaaGATAA*GGATA
+r003 gcctAGGCTAA
+r004 ATAGCT... ........... TCAGC
-r003 ttagctTAGGC
-r001/2 CAGCGGCAT

↓ SAM Format

r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACGT *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,−,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+5S6M,30,1;
r001 83 ref 37 30 9M = 7 −39 CAGCGGCAT * NM:i:1

For details see the SAM Format Specification [Link]
Normalization Required

Log ratio distributions (a and b) and MA plot (c) for two tissue samples (from Robinson and Oshlack, 2010).
Be Careful with RPKM/FPKM Values

RPKM Concept (FPKM is paired-end version of it)

- RPKM (FPKM): reads (fragments) per kp per million mapped reads
- The more we sequence, the more reads we expect from each gene. This is the most relevant correction of this method.
- Longer transcript are expected to generate more reads. The latter is only relevant for comparisons among different genes which we rarely perform!
- RPKM/FPKM are not suitable for statistical testing. Why? Consider the following example: in two libraries, each with one million reads, gene X may have 10 reads for treatment A and 5 reads for treatment B, while it is 100x as many after sequencing 100 millions reads from each library. In the latter case we can be much more confident that there is a true difference between the two treatments than in the first one. However, the RPKM values would be the same for both scenarios.
- Thus, RPKM/FPKM are useful for reporting expression values, but not for statistical testing!
Trimmed Mean of M Values (TMM) by Robinson and Oshlack (2010)

- Many normalization RNA-Seq normalization methods perform poorly on samples with extreme composition bias. For instance, in one sample a large number of reads comes from rRNAs while in another they have been removed more efficiently. Most scaling based methods, including RPKM and CPM, will underestimate the expression of weaker expressed genes in the presence of extremely abundant mRNAs (less sequencing real estate available for them). The TMM methods tries to correct this bias.

- Method implemented in edgeR library (Robinson et al., 2010).
Analysis of Differentially Expressed Genes (DEGs)

- Data is discrete, positively skewed
  ⇒ no (log-)normal model
- Small numbers of replicates
  ⇒ no rank based or permutation methods
- Sequencing depth (coverage) varies among samples
  ⇒ normalization
DEG Analysis Methods

Requirements

- One would like to perform a t-test or something similar for each gene.
- t-test assumes normal distribution and no mean-variance dependence. Both are not appropriate assumptions for RNA-Seq data.
- Variance estimation and rank-order statistics is difficult on small sample numbers.

Statistical Testing

- Poisson distribution (initially used but not very common anymore)
- Most statistical methods for RNA-Seq DEG analysis use negative binomial distribution along with modified statistical tests based on that.
- The multiple testing issue is very similar as in microarray data analysis. Thus, most tools provide False Discovery Rates (FDRs), which are derived from p-values corrected for multiple testing using the Benjamini-Hochberg method.
- For variance estimation most methods borrow information across genes
Software for RNA-Seq DEG Analysis

- edgeR (Robinson et al., 2010)
- DESeq/DESeq2 (Anders and Huber, 2010)
- DEXSeq (Anders et al., 2012)
- limmaVoom
- Cuffdiff/Cuffdiff2 (Trapnell et al., 2013)
- PoissonSeq
- baySeq
- ...

Analysis of RNA-Seq Data with R/Bioconductor
Packages for RNA-Seq Analysis in R

- GenomicRanges Link: high-level infrastructure for range data
- Rsamtools Link: BAM support
- rtracklayer Link: Import/export of range and annotation data, interface to online genome browsers, etc.
- DESeq Link: RNA-Seq DEG analysis
- DESeq2 Link: RNA-Seq DEG analysis
- edgeR Link: RNA-Seq DEG analysis
- DEXSeq Link: RNA-Seq Exon analysis
- QuasR Link: RNA-Seq workflows
- systemPipeR Link: NGS workflows and reports
Outline

Overview

RNA-Seq Analysis
- Quality Report
- Aligning Short Reads
- Counting Reads per Feature
- DEG Analysis
- GO Analysis
- View Results in IGV & ggbio
- Differential Exon Usage

References
Data Sets and Experimental Variables

To make the following sample code work, please follow these instructions:

- Download and unpack the sample data for this practical.
- Direct your R session to the resulting Rrnaseq directory. It contains 18 slimmed down FASTQ files (SRP010938) from A. thaliana (Howard et al., 2013). To minimize processing time, each FASTQ file has been subsetted to 90,000-100,000 randomly sampled reads that map to the first 100,000 nucleotides of each chromosome. The corresponding reference genome sequence (FASTA) and its GFF annotation files have been truncated accordingly.
- Start the analysis by opening in your R session the Rrnaseq.R script which contains the code shown in this slide show in pure text format.

The FASTQ files are organized in the provided targets.txt file. This is the only file in this analysis workflow that needs to be generated manually, e.g. in a spreadsheet program. To import targets.txt, we run the following commands from R:

```r
> # download.file("http://biocluster.ucr.edu/~tgirke/HTML_Presentations/Manuals/MCBIOS2015/Rrnaseq.zip", "Rrnaseq.zip")
> # unzip("Rrnaseq.zip")
> # setwd("Rrnaseq")
> library(systemPipeR)
> args <- systemArgs(sysma="tophat.param", mytargets="targets.txt")
> targetsin(args)[1:3,]
```

<table>
<thead>
<tr>
<th>FileName</th>
<th>SampleName</th>
<th>Factor</th>
<th>SampleLong</th>
<th>Experiment</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>./data/SRR446027_1.fastq</td>
<td>M1A</td>
<td>M1</td>
<td>Mock.1h.A</td>
<td>1</td>
<td>23-Mar-2012</td>
</tr>
<tr>
<td>./data/SRR446028_1.fastq</td>
<td>M1B</td>
<td>M1</td>
<td>Mock.1h.B</td>
<td>1</td>
<td>23-Mar-2012</td>
</tr>
<tr>
<td>./data/SRR446029_1.fastq</td>
<td>A1A</td>
<td>A1</td>
<td>Avr.1h.A</td>
<td>1</td>
<td>23-Mar-2012</td>
</tr>
</tbody>
</table>
Outline

Overview

RNA-Seq Analysis
Quality Report
Aligning Short Reads
Counting Reads per Feature
DEG Analysis
GO Analysis
View Results in IGV & ggbio
Differential Exon Usage

References
Quality Reports

The following shows how to create read quality reports with the `seeFastq` function from `systemPipeR`.

```r
> fqlist <- seeFastq(fastq=infile1(args), batchsize=10000, klength=8)
> pdf("./results/fastqReport.pdf", height=18, width=4*length(fqlist))
> seeFastqPlot(fqlist); dev.off()
```

**Figure:** QC report for 18 FASTQ files.
Outline

Overview

**RNA-Seq Analysis**
- Quality Report
- **Aligning Short Reads**
- Counting Reads per Feature
- DEG Analysis
- GO Analysis
- View Results in IGV & ggbio
- Differential Exon Usage

References
**Align Reads with Rsubread**

*Rsubread* is an R/Bioc package that implements an extremely fast aligner for RNA-Seq data. It is currently only available for OS X and Linux, but not for Windows.

(1) Index reference genome

```r
> library(Rsubread); library(systemPipeR)
> args <- systemArgs(sysma="rsubread.param", mytargets="targets.txt")
> buildindex(basename=reference(args), reference=reference(args)) # Build indexed reference genome
```

(2) Align all FASTQ files with *Rsubread* in loop. Includes generation of indexed BAM files.

```r
> align(index=reference(args), readfile1=infile1(args), input_format="FASTQ",
+       output_file=outfile1(args), output_format="SAM", nthreads=8, indels=1, TH1=2)
> for(i in seq(along=outfile1(args))) asBam(file=outfile1(args)[i], destination=gsub(".sam", "", outfile1(args)[i]), overwrite=TRUE, indexDestination=TRUE)
> unlink(outfile1(args)); unlink(paste0(outfile1(args), ".indel"))
```
Align Reads with **Tophat2**

systemPipeR NGS workflow and report generation environment that can run command-line software on local computers and compute clusters. Note: this step requires the command-line tools tophat2/bowtie2 [Link](#).

1. **Index reference genome**
   ```r
   > library(systemPipeR)
   > args <- systemArgs(sysma="tophat.param", mytargets="targets.txt")
   > moduleload(modules(args)) # Skip if a module system is not available
   > system("bowtie2-build ./data/tair10.fasta ./data/tair10.fasta")
   ```

2. **Align all FASTQ files with Bowtie2/Tophat2 on a single computer. Includes generation of indexed BAM files.**
   ```r
   > bampaths <- runCommandline(args=args)
   Missing alignment results (bam files): 0
   Existing alignment results (bam files): 18
   > bampaths
   ```

<table>
<thead>
<tr>
<th>M1A</th>
<th>M1B</th>
<th>A1A</th>
<th>A1B</th>
<th>V1A</th>
<th>V1B</th>
<th>M6A</th>
<th>M6B</th>
<th>A6A</th>
<th>A6B</th>
<th>V6A</th>
<th>V6B</th>
<th>M12A</th>
<th>M12B</th>
<th>A12A</th>
<th>A12B</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
<td>&quot;TRUE&quot;</td>
</tr>
</tbody>
</table>

3. **Alternatively, align all FASTQ files with Bowtie2/Tophat2 on many compute nodes in parallel. The following submits to Torque scheduler 18 processes each with 4 CPU cores.**
   ```r
   > 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)
   > showStatus(reg)
   ```
Alignment Summary

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

```r
> (read_statsDF <- alignStats(args=args))[1:8,]

<table>
<thead>
<tr>
<th>FileNName</th>
<th>Nreads</th>
<th>Nalign</th>
<th>Perc_Aligned</th>
<th>Nalign_Primary</th>
<th>Perc_Aligned_Primary</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1A</td>
<td>M1A</td>
<td>96459</td>
<td>89376</td>
<td>92.65698</td>
<td>92.65698</td>
</tr>
<tr>
<td>M1B</td>
<td>M1B</td>
<td>98742</td>
<td>86014</td>
<td>87.10984</td>
<td>87.10984</td>
</tr>
<tr>
<td>A1A</td>
<td>A1A</td>
<td>94935</td>
<td>88360</td>
<td>93.07421</td>
<td>93.07421</td>
</tr>
<tr>
<td>A1B</td>
<td>A1B</td>
<td>94427</td>
<td>83172</td>
<td>88.08074</td>
<td>88.08074</td>
</tr>
<tr>
<td>V1A</td>
<td>V1A</td>
<td>99366</td>
<td>80869</td>
<td>81.38498</td>
<td>81.38498</td>
</tr>
<tr>
<td>V1B</td>
<td>V1B</td>
<td>97771</td>
<td>93637</td>
<td>95.77175</td>
<td>95.77175</td>
</tr>
<tr>
<td>M6A</td>
<td>M6A</td>
<td>98617</td>
<td>92917</td>
<td>94.22006</td>
<td>94.22006</td>
</tr>
<tr>
<td>M6B</td>
<td>M6B</td>
<td>90452</td>
<td>80074</td>
<td>88.52651</td>
<td>88.52651</td>
</tr>
</tbody>
</table>
```

```r
> write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE,
```
```r
```

Analysis of RNA-Seq Data with R/Bioconductor
Outline

Overview

RNA-Seq Analysis
  Quality Report
  Aligning Short Reads
  Counting Reads per Feature
  DEG Analysis
  GO Analysis
  View Results in IGV & ggbio
  Differential Exon Usage

References
Storing annotation ranges in *TranscriptDb* databases makes many operations more robust and convenient.

```r
> library(GenomicFeatures)
> txdb <- makeTranscriptDbFromGFF(file="data/tair10.gff",
+ format="gff3",
+ dataSource="TAIR",
+ species="Arabidopsis thaliana")
> saveDb(txdb, file="./data/tair10.sqlite")
> txdb <- loadDb("./data/tair10.sqlite")
> eByg <- exonsBy(txdb, by="gene")
```
The `summarizeOverlaps` function from the GenomicRanges package is easier to use, it provides more options and it is much more memory efficient. See here [Link] for details.

```r
> txdb <- loadDb("./data/tair10.sqlite")
> eByg <- exonsBy(txdb, by=c("gene"))
> bfl <- BamFileList(outpaths(args), yieldSize=50000, index=character())
> counteByg <- summarizeOverlaps(eByg, bfl, mode="Union", ignore.strand=TRUE, inter.feature=FALSE)
> countDFeByg <- assays(counteByg)$counts
> countDFeByg[1:4,1:12]

     M1A  M1B  A1A  A1B  V1A  V1B  M6A  M6B  A6A  A6B  V6A  V6B
AT1G01010  28  128   99   87  183  118   22   19   77   23  148  201
AT1G01020   12   47   35   50   49   41    9   11    9    3   42   71
AT1G01030   19   51   36   33   47   78    5    8    4    4   18   73
AT1G01040  98  354  259  345  298  350   82   81  128   34  359  530

> write.table(countDFeByg, "results/countDFeByg.xls", col.names=NA, quote=FALSE, sep=
```
Simple RPKM Normalization

RPKM: reads per kilobase of exon model per million mapped reads

```r
> rpkmDFeByg <- apply(countDFeByg, 2, function(x) returnRPKM(counts=x, ranges=eByg))
> write.table(rpkmDFeByg, "results/rpkmDFeByg.xls", col.names=NA, quote=FALSE, sep="	")
> rpkmDFeByg[1:4,1:7]
```

<table>
<thead>
<tr>
<th></th>
<th>M1A</th>
<th>M1B</th>
<th>A1A</th>
<th>A1B</th>
<th>V1A</th>
<th>V1B</th>
<th>M6A</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G01010</td>
<td>2424.745</td>
<td>6201.798</td>
<td>5431.999</td>
<td>4225.653</td>
<td>11608.558</td>
<td>5482.332</td>
<td>1315.552</td>
</tr>
<tr>
<td>AT1G01020</td>
<td>988.799</td>
<td>2166.827</td>
<td>1827.306</td>
<td>2310.805</td>
<td>2957.618</td>
<td>1812.533</td>
<td>512.090</td>
</tr>
<tr>
<td>AT1G01030</td>
<td>1457.938</td>
<td>2189.552</td>
<td>1750.267</td>
<td>1420.254</td>
<td>2641.816</td>
<td>3211.111</td>
<td>264.931</td>
</tr>
<tr>
<td>AT1G01040</td>
<td>2290.597</td>
<td>4629.408</td>
<td>3835.648</td>
<td>4522.808</td>
<td>5102.206</td>
<td>4389.003</td>
<td>1323.469</td>
</tr>
</tbody>
</table>
Reproducibility Check by Sample-Wise Clustering

QC check of the sample reproducibility by computing a correlating matrix and plotting it as a tree. Note: the plotMDS function from edgeR is a more robust method for this task.

```r
> library(ape)
> d <- cor(rpkmDFeByg, method="spearman")
> hc <- hclust(dist(1-d))
> plot.phylo(as.phylo(hc), type="p", edge.col=4, edge.width=3, show.node.label=TRUE, no.margin=TRUE)
```
Sample-Wise Clustering with \texttt{\textit{rlog}} Values

\begin{verbatim}
> library(DESeq2)
> 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))
> plot.phylo(as.phylo(hc), type="p", edge.col=4, edge.width=3, show.node.label=TRUE, no.margin=TRUE)
\end{verbatim}
Exercise 1: Strand-specific Read Counting

Task 1  Count reads mapping to both strands and also those mapping only to the sense strand. Discuss differences. Why is this analysis meaningless for the provided non-strand-specific RNA-Seq samples?

Task 2  Identify all genes where more than 75% of the reads in at least 4 samples map to the sense strand.

Task 3  Plot the result of the most pronounced strand-selective cases with `ggbio`.
Outline

Overview

RNA-Seq Analysis
- Quality Report
- Aligning Short Reads
- Counting Reads per Feature

DEG Analysis
- GO Analysis
- View Results in IGV & ggbio
- Differential Exon Usage

References
DEGs for Comparisons in targets File Using `edgeR`

```r
> countDF <- read.table("./results/countDFeByg.xls")
> cmp <- readComp(args, format="matrix", delim="-")
> edgeDF <- run_edgeR(countDF=countDF, targets=targetsIn(args), cmp=cmp[[1]], independent=FALSE, mdsplot="")

Disp = 0.17287 , BCV = 0.4158

> write.table(edgeDF, "/results/edgeRcomp.xls", quote=FALSE, sep="\t", col.names = NA)
> DEG_list <- filterDEGs(degDF=edgeDF, filter=c(Fold=2, FDR=20))
```

### DEG Counts (Fold: 2 & FDR: 20%)

![Bar chart showing DEG counts for different comparisons.]

**Type**
- Down
- Up

**Comparisons**
- M6–V6
- M6–A6
- M12–V12
- M12–A12
- M1–V1
- M1–A1
- A6–V6
- A12–V12
- A1–V1
The function `overLapper` can compute Venn intersects for any number of sample sets and `vennPlot` can plot 2-5 way Venn diagrams. A useful feature is the possibility to combine the counts from several Venn comparisons in a single plot.

```r
> vennsetup <- overLapper(DEG_list$Up[6:9], type="vennsets")
> vennsetdown <- overLapper(DEG_list$Down[6:9], type="vennsets")
> vennPlot(list(vennsetup, vennsetdown), mymain="", mysub="", colmode=2, ccol=c("blue", "red"))
```
Heatmap of DEGs

Note: the following plots the scaled expression values (here RPKMs) in form of a heatmap.

```r
> library(lattice); library(gplots)
> rpkmDFeByg <- read.table("./results/rpkmDFeByg.xls")
> y <- rpkmDFeByg[unlist(DEG_list$UporDown),]
> colnames(y) <- targetsin(args)$Factor
> y <- t(scale(t(as.matrix(y))))
> y <- y[order(y[,1]),]
> levelplot(t(y), height=0.2, col.regions=colorpanel(40, "darkblue", "yellow", "white"), main="Expression Values (DEG Filter: FDR 1%, FC > 2)")
```

Expression Values (DEG Filter: FDR 1%, FC > 2)
DEGs for Comparisons in targets File Using \textit{DESeq2}

```r
> countDF <- read.table("./results/countDFeByg.xls")
> cmp <- readComp(args, format="matrix", delim="-")
> degseqDF <- run\_DESeq2(countDF=countDF, targets=targetsin(args), cmp=cmp[[1]], independent=FALSE)
> write.table(edgeDF, "./results/DESeq2comp.xls", quote=FALSE, sep="\t", col.names = NA)
> DEG\_list2 <- filterDEGs(degDF=degseqDF, filter=c(Fold=2, FDR=10))
```

### DEG Counts (Fold: 2 & FDR: 10%)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1−V1</td>
<td>0.0</td>
</tr>
<tr>
<td>A12−V12</td>
<td>2.5</td>
</tr>
<tr>
<td>A6−V6</td>
<td>5.0</td>
</tr>
<tr>
<td>M1−A1</td>
<td>7.5</td>
</tr>
<tr>
<td>M1−V1</td>
<td>10.0</td>
</tr>
<tr>
<td>M12−A12</td>
<td>0.0</td>
</tr>
<tr>
<td>M12−V12</td>
<td>2.5</td>
</tr>
<tr>
<td>M6−A6</td>
<td>5.0</td>
</tr>
<tr>
<td>M6−V6</td>
<td>7.5</td>
</tr>
</tbody>
</table>

- **Type**
  - **Down**
  - **Up**
Outline

Overview

RNA-Seq Analysis
  Quality Report
  Aligning Short Reads
  Counting Reads per Feature
  DEG Analysis

GO Analysis
  View Results in IGV & ggbio
  Differential Exon Usage

References
The following shows how to obtain gene-to-GO mappings from bioMaRt. This is relatively slow, but it needs to be done only once.

```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])
> dir.create("./data/GO")
> write.table(go, "data/GO/GOannotationsBiomart_mod.txt", quote=FALSE, row.names=FALSE, col.names=FALSE, sep="\t")
> 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")
```
The Batch enrichment analysis of many gene sets is performed with the `GOCluster_Report` function. When `method="all"`, it returns all GO terms passing the p-value cutoff specified under the `cutoff` arguments. When `method="slim"`, it returns only the GO terms specified under the `myslimv` argument. The given example shows how one can obtain such a GO slim vector from BioMart for a specific organism.

```r
code
> 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[sapply(DEGlist, length) > 0]
> BatchResult <- GOCluster_Report(catdb=catdb, setlist=DEGlist, method="all", id_type="gene", CLSZ=2, cutoff=0.9)
> write.table(BatchResult, "/results/GO_BatchResult.xls", quote=FALSE, sep="\t", row.names = FALSE)
> library("biomaRt"); m <- useMart("ensembl", dataset="mmulatta_gene_ensembl")
> 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)
> write.table(BatchResultslim, "/results/GO_BatchResultslim.xls", quote=FALSE, sep="\t", row.names = FALSE)
```
Batch GO Term Enrichment Analysis (Part III)

Plot batch GO term results

> goBarplot(BatchResultslim, gocat="MF")

Figure: GO Slim Barplot for MF Ontology.

Analysis of RNA-Seq Data with R/Bioconductor
RNA-Seq Analysis
GO Analysis
Outline

Overview

RNA-Seq Analysis
  Quality Report
  Aligning Short Reads
  Counting Reads per Feature
  DEG Analysis
  GO Analysis

View Results in IGV & ggbio

Differential Exon Usage

References
View results in IGV

- Download and open IGV [Link]
- Select in menu in top left corner A. thaliana (TAIR10)
- Upload the following indexed/sorted Bam files with File -> Load from File...
  - ./data/SRR446039_1.fastq.tophat/accepted_hits.bam
  - ./data/SRR446040_1.fastq.tophat/accepted_hits.bam
  - ./data/SRR446041_1.fastq.tophat/accepted_hits.bam
  - ./data/SRR446042_1.fastq.tophat/accepted_hits.bam

- To view area of interest, enter its coordinates Chr1:45,296-47,019 in position menu on top.
Controlling IGV from R

Create previous IGV session with required tracks automatically, and direct it to a specific position, here Chr1:45,296-47,019.

```r
> library(SRAdb)
> startIGV("lm")
> sock <- IGVsocket()
> session <- IGVsession(files=outpaths(args)[c("M12A", "M12B", "A12A", "A12B")],
                 +   sessionFile="session.xml",
                 +   genome="A. thaliana (TAIR10)"
> IGVload(sock, session)
> IGVgoto(sock, 'Chr1:45296-47019')
```
Generate Similar View with *ggbio* Programatically

```r
> library(ggbio)
> M12A <- readGAlignmentsFromBam(outpaths(args)["M12A"], use.names=TRUE, param=ScanBamParam(which=GRanges("Chr1", IRanges(45296, 47019))))
> A12A <- readGAlignmentsFromBam(outpaths(args)["A12A"], use.names=TRUE, param=ScanBamParam(which=GRanges("Chr1", IRanges(45296, 47019))))
> p1 <- autoplot(M12A, geom = "rect", aes(color = strand, fill = strand))
> p2 <- autoplot(A12A, geom = "rect", aes(color = strand, fill = strand))
> p3 <- autoplot(txdb, which=GRanges("Chr1", IRanges(45296, 47019)), names.expr = "gene_id")
> tracks(M12A=p1, A12A=p2, Transcripts=p3, heights = c(0.3, 0.3, 0.4)) + ylab("")
```
Outline

Overview

RNA-Seq Analysis
  Quality Report
  Aligning Short Reads
  Counting Reads per Feature
  DEG Analysis
  GO Analysis
  View Results in IGV & ggbio

Differential Exon Usage

References
### Analysis of Differential Exon Usage with DEXSeq

**Number of reads overlapping gene ranges**

```r
> library(DEXSeq)
> exonicParts <- disjointExons(txdb, aggregateGenes=FALSE)
> bamlst <- BamFileList(outpaths(args)[c("M12A", "M12B", "A12A", "A12B")], index=character(), yieldSize=100000, obeyQname=TRUE)
> SE <- summarizeOverlaps(exonicParts, bamlst, mode="Union", singleEnd=TRUE, ignore.strand=TRUE, inter.feature=FALSE)
> colData(SE)$condition <- colData$condition
> dx <- DEXSeqDataSetFromSE(SE, design= ~ sample + exon + condition:exon)
> featureCounts(dx)[1:2,] # Counts for individual exons
```

<table>
<thead>
<tr>
<th></th>
<th>M12A</th>
<th>M12B</th>
<th>A12A</th>
<th>A12B</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G01010:E001</td>
<td>9</td>
<td>12</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>AT1G01010:E002</td>
<td>12</td>
<td>15</td>
<td>12</td>
<td>11</td>
</tr>
</tbody>
</table>

```r
> assays(dx)$counts[1:2,] # Counts for individual exons plus for all remaining exons of a gene
```

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G01010:E001</td>
<td>9</td>
<td>12</td>
<td>14</td>
<td>6</td>
<td>59</td>
<td>71</td>
<td>67</td>
<td>37</td>
</tr>
<tr>
<td>AT1G01010:E002</td>
<td>12</td>
<td>15</td>
<td>12</td>
<td>11</td>
<td>56</td>
<td>68</td>
<td>69</td>
<td>32</td>
</tr>
</tbody>
</table>

```r
> rowData(dx)[1:4,] # Exon ranges
> colData(dx) # Sample data
> write.table(featureCounts(dx), ".results/countDFdex", quote=FALSE, sep="\t", col.names = NA)
```
## Identify genes with differential exon usage

```r
## Performs normalization
>dxd <- estimateSizeFactors(dxd)

## Evaluate variance of the data by estimating dispersion using Cox-Reid (CR) likelihood estimation
>dxd <- estimateDispersions(dxd)

## Performs Chi-squared test on each exon and Benjmini-Hochberg p-value adjustment for multiple testing
>dxd <- testForDEU(dxd)

## Estimates fold changes of exons
>dxd <- estimateExonFoldChanges(dxd, fitExpToVar="condition")

## Obtain results as DataFrame
>dxr1 <- DEXSeqResults(dxd)

## Column descriptions
> col_descr <- elementMetadata(dxr1)$description

## Count number of genes with differential exon usage
> dxr1[is.na(dxr1$padj), "padj"] <- 1
> table(tapply(dxr1$padj < 0.2, dxr1$groupID, any))

### FALSE	TRUE
115	1

---

### DEU sample

```r
> dxr1[dxr1$groupID=="AT4G00050",][1:4, c(1:2,7:10)]
```

<table>
<thead>
<tr>
<th>groupID</th>
<th>featureID</th>
<th>padj</th>
<th>A12</th>
<th>M12</th>
<th>log2fold_A12_M12</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT4G00050:E001</td>
<td>AT4G00050</td>
<td>6.275199</td>
<td>6.688707</td>
<td>-0.09206611</td>
<td></td>
</tr>
<tr>
<td>AT4G00050:E002</td>
<td>AT4G00050</td>
<td>12.137741</td>
<td>10.860877</td>
<td>0.16035925</td>
<td></td>
</tr>
<tr>
<td>AT4G00050:E003</td>
<td>AT4G00050</td>
<td>4.998105</td>
<td>5.463383</td>
<td>-0.12841322</td>
<td></td>
</tr>
<tr>
<td>AT4G00050:E004</td>
<td>AT4G00050</td>
<td>4.793227</td>
<td>5.288418</td>
<td>-0.14183906</td>
<td></td>
</tr>
</tbody>
</table>
Sample plot showing fitted expression of exons

```r
> plotDEXSeq(dxr1, "AT4G00050", displayTranscripts=TRUE, legend=TRUE, cex.axis=1.2, cex=1.3, lwd=2 )
```

Generate many plots and write them to results directory

```r
> mygeneIDs <- unique(as.character(na.omit(dxr1[dxr1$groupID %in% unique(dxr1$groupID),][,"groupID"])))
> DEXSeqHTML(dxr1, genes=mygeneIDs[1:10], path="results", file="DEU.html")
```
**systemPipeR: Run Entire RNA-Seq Workflow and Generate Analysis Report**

- *systemPipeR* is useful for building end-to-end analysis pipelines with automated report generation for NGS applications such as RNA-Seq and many others.
- It provides support for running command-line software, such as NGS aligners, on both single machines or compute clusters. This includes both interactive job submissions or batch submissions to queuing systems of clusters.
- To generate the report for the data sets and analysis steps demonstrated in this tutorial, open the file *systemPipeR.Rnw* in RStudio’s code editor and then click the Compile PDF button. This will run the entire analysis and generate the corresponding RNA-Seq analysis report (PDF format) along with a bibliography of citations included in the text.
- Alternatively, one can achieve the same result by running the following commands from the command-line:
  ```
echo 'Sweave("systemPipeR.Rnw")' | R --slave # Runs R code
echo 'Stangle("systemPipeR.Rnw")' | R --slave # Extracts R code
pdflatex systemPipeR.tex; bibtex systemPipeR; pdflatex systemPipeR.tex # Compiles PDF
  ```
- Note: for time reasons, not all code chunks are evaluated (change `eval=FALSE` to `eval=TRUE`) when the report is generated.
- A sample report can be viewed here *systemPipeR.pdf*.
- To efficiently customize these reports, users want to learn how to use *LaTeX/Sweave* and/or *knitr*. 
> sessionInfo()

R version 3.1.2 (2014-10-31)
Platform: x86_64-unknown-linux-gnu (64-bit)
locale:
[1] C

attached base packages:
[1] parallel stats4 stats graphics utils datasets grDevices methods base

other attached packages:
[1] DEXSeq_1.12.1 ggbio_1.14.0 ggplot2_1.0.0 gplots_2.16.0 lattice_0.20-29 DESeq2_1.6.3 RcppArmadillo_0.4.600.4.0 Rcpp_0.11.4
[9] ape_3.2 GenomicFeatures_1.18.3 systemPipeR_1.0.11 AnnotationDbi_1.28.1 Biobase_2.26.0 RSQLite_1.0.0 DBI_0.3.1 ShortRead_1.24.0
[17] GenomicAlignments_1.2.1 BiocParallel_1.0.2 Rsamtools_1.18.2 BiocGenerics_0.12.1
[25] S4Vectors_0.4.0

loaded via a namespace (and not attached):
[1] AnnotationForge_1.8.2 BBmisc_1.9 BSgenome_1.34.1 BatchJobs_1.5 Catego
[9] G0stats_2.32.0 GSEABase_1.28.0 Hmisc_3.14-6 KernSmooth_2.23-13 MASS
[17] RColorBrewer_1.1-2 RCurl_1.95-4.5 VariantAnnotation_1.12.9 XML_3.98-1.1 acepa
[25] biovizBase_1.14.1 bitops_1.0-6 brew_1.0-6 caTools_1.17.1 check
[33] dichromat_2.0-0 digest_0.6.8 edgeR_3.8.5 fail_1.2 forensi
[41] geneplotter_1.44.0 graph_1.44.1 grid_3.1.2 gridExtra_0.9.1 gtable
[49] labeling_0.3 latticeExtra_0.6-26 limma_3.22.4 locfit_1.5-9.1 munsell
[57] plyr_1.8.1 proto_0.3-10 reshape_0.8.5 reshape2_1.4.1 rjson
[65] sendmailR_1.2-1 splines_3.1.2 statmod_1.4.20 stringr_0.6.2 survival
Outline

Overview

RNA-Seq Analysis
  Quality Report
  Aligning Short Reads
  Counting Reads per Feature
  DEG Analysis
  GO Analysis
  View Results in IGV & ggbio
  Differential Exon Usage

References
URL http://www.hubmed.org/display.cgi?uids=20979621

URL http://www.hubmed.org/display.cgi?uids=22722343

URL http://dx.doi.org/10.1371/journal.pone.0074183

URL http://www.hubmed.org/display.cgi?uids=19910308

URL http://www.hubmed.org/display.cgi?uids=20196867

URL http://www.hubmed.org/display.cgi?uids=23222703