Overview of (some) R/Bioconductor packages for short read analysis

Application of high throughput genomics tools in nutrition

Nutrigenomics

Understanding of how nutrition influences metabolic pathways, how this regulation is disturbed in diet-related diseases

Philip de Groot (Philip.deGroot@wur.nl)
NBIC Advanced RNA-SEQ course 15-17 October 2012
**Transcriptomics:** The global study of gene expression at the RNA level

**Microarray:**
- Measures RNA abundances by exploiting complementary hybridization with the fixed number of genes on the array chip

**RNA-seq:**
- Refers to the use of high-throughput sequencing technologies to sequence cDNA in order to get information about a sample's RNA content
- More genes can be studied in the samples because it is not only limited to annotated genes
RNA-SEQ biases

The influence of sequencing depth
the higher sequencing depth, the higher counts

Dependence on gene length
counts are proportional to the transcript length times the mRNA expression level

Differences on the counts distribution among samples
1- **RPKM (Mortazavi et al., 2008):** counts are divided by the transcript length (kb) times the total number of millions of mapped reads.

2- **Upper quartile: (Bullard et al., 2010):** counts are divided by upper-quartile of counts for transcripts with at least one read.

3- **TMM (Robinson and Oshlack, 2010):** Trimmed mean of value.

4- **quantiles:** as in microaary normalization (Irizarry et al. 2003).

5- **FPKM (Trapnell et al., 2010):** Instead of counts, cufflinks software generates FPKM values (Fragments per kilobase of exon per million fragments mapped) to estimate gene expression, which are analogous to RPKM.
R packages in bioconductor:

edgeR (Robinson et al., 2010): exact test based on negative binomial distribution.

DESeq (Anders and Huber, 2010): exact test based on negative binomial distribution.

DEGseq (Wang et al., 2010): MA-plots based methods (MATR and MARS), assuming Normal distribution for M|A

baySeq (Hardcastle et al., 2010): Estimation of the posterior likelihood of differential expression (or more complex hypotheses) via empirical Bayesian methods using posssion or NB distribution.
easyRNASeq: a bioconductor package for processing RNA-Seq data
Nicolas Delhomme; Ismael Padioleau; Eileen E. Furlong; Lars M. Steinmetz
Bioinformatics 2012; doi: 10.1093/bioinformatics/bts477

Figure 1: easyRNAseq Overview. The R packages used for the different steps are emphasized in bold face.
easyRNASeq example:
> library(easyRNASeq)
> library(RnaSeqTutorial)
> library(BSgenome.Dmelanogaster.UCSC.dm3)
> count.table <- easyRNASeq(filesDirectory = system.file(
+ "extdata",
+ package = "RnaSeqTutorial"),
+ pattern = "[A,C,T,G]{6}\\.bam$",
+ readLength = 30L,
+ organism = "Dmelanogaster",
+ chr.sizes = seqlengths(Dmelanogaster),
+ annotationMethod = "rda",
+ annotationFile = system.file(
+ "data",
+ "gAnnot.rda",
+ package = "RnaSeqTutorial"),
+ count = "exons"
+ )
easyRNASeq example:
> head(count.table)

<table>
<thead>
<tr>
<th></th>
<th>ACACTG.bam</th>
<th>ACTAGC.bam</th>
<th>ATGGCT.bam</th>
<th>TTGCXA.bam</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG11023:1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CG11023:2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CG11023:3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CG2671:1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CG2671:2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CG2671:3</td>
<td>13</td>
<td>8</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

> dim(count.table)
[1] 69306 4
Output

The results can be exported in four different formats:

- count table (the default, a n (features) x m (samples) matrix).
- a DESeq [Anders and Huber, 2010] CountDataSet class object. Useful to perform further analyses using the DESeq package.
- an edgeR [Robinson et al., 2010] DGEList class object. Useful to perform further analyses using the edgeR package.
- an RNAseq class object. Useful for performing additional pre-processing without re-loading the reads and annotations.
All these steps are available in Galaxy! CummeRbund is used for data exploration!
CummeRbund assumes that all output files from cuffdiff are in the current working directory. To read these files, populate the 'cuffData.db' database backend, and return the CuffSet pointer object, you can do the following:

```r
> library(cummeRbund)
> cuff<-readCufflinks()

> cuff
CuffSet instance with:
3 samples
400 genes
1203 isoforms
662 TSS
906 CDS
1062 promoters
1986 splicing
990 relCDS
```
> disp<-dispersionPlot(genes(cuff))
The squared coefficient of variation allows visualization of cross-replicate variability between conditions and can be a useful metric in determining data quality.
(a) Density plot of individual conditions. (b) Density plot with replicates=True exposes individual replicate FPKM distributions.
(a) Box plot of FPKM distributions for individual conditions.

(b) Box plot with replicates=TRUE exposes individual replicate FPKM distributions.
MVA plots
> dend<-csDendro(genes(cuff))
> dend.rep<-csDendro(genes(cuff),replicates=T)

(a) Dendrogram of JS distances between conditions.
(b) Dendrogram with replicates=TRUE can identify outlier replicates.
(a) Volcano plots explore the relationship between fold-change and significance.
h<-csHeatmap(myGenes,cluster='both')

(a) Heatmaps provide a convenient way to visualize the expression of entire gene sets at once.
b<-expressionBarplot(myGenes)

(a) A (somewhat crowded) barplot for all genes in a CuffGeneSet object.
myGeneId<"PINK1"; myGene<-getGene(cuff,myGeneId); gl<-expressionPlot(myGene)

(a) Expression plot of a single gene.
(b) Expression plot of a single gene with replicate FPKMs exposed.
gb<-expressionBarplot(myGene)

(a) Expression Barplot of a single gene.

(b) Expression Barplot of a single gene with replicate FPKMs exposed.
mySigMat<-sigMatrix(cuff, level='genes',
alpha=0.05)

(a) Significant features overview matrix. This plot describes the number of significant genes at a 5%FDR for each pairwise interaction tested.
(a) PCA plot for gene-level features

(b) MDS plot for gene-level features
ic<-csCluster(myGenes,k=4)
icp<-csClusterPlot(ic)

(a) PAM clustering with JS distance for a CuffGeneSet.
mySimilar<- findSimilar(cuff,"PINK1",n=20)

(a) Top 20 most similar genes to 'PINK1'.
Two different pipeline to analyse RNA-seq data

1. Reads → TopHat → Mapped reads → Read counts for each gene → Htseq (Python package) → exact test based on negative binomial distribution → EdgeR (R package) → Expression plots

2. Reads → TopHat → Mapped reads → Read counts for each gene → Htseq (Python package) → exact test based on negative binomial distribution → EdgeR (R package) → Expression plots

Philip de Groot (Philip.deGroot@wur.nl)
NBIC Advanced RNA-SEQ course 15-17 October 2012
## Bioconductor packages #1

Table 2: Selected Bioconductor packages for high-throughput sequence analysis.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Packages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data representation</td>
<td><code>IRanges</code>, <code>GenomicRanges</code>, <code>GenomicFeatures</code>, <code>Biostrings</code>, <code>BSgenome</code>, <code>girafe</code></td>
</tr>
<tr>
<td>Input / output</td>
<td><code>ShortRead</code> (fastq), <code>Rsamtools</code> (bam), <code>rtracklayer</code> (gff, wig, bed), <code>VariantAnnotation</code> (vcf), <code>R453Plus1Toolbox</code> (454)</td>
</tr>
<tr>
<td>Annotation</td>
<td><code>GenomicFeatures</code>, <code>ChIPpeakAnno</code>, <code>VariantAnnotation</code></td>
</tr>
<tr>
<td>Alignment</td>
<td><code>Rsubread</code>, <code>Biostrings</code></td>
</tr>
<tr>
<td>Visualization</td>
<td><code>ggbio</code>, <code>Gviz</code></td>
</tr>
<tr>
<td>Quality assessment</td>
<td><code>qrqc</code>, <code>seqbias</code>, <code>ReQON</code>, <code>htSeqTools</code>, <code>TEQC</code>, <code>Rolexa</code>, <code>ShortRead</code></td>
</tr>
<tr>
<td>RNA-seq</td>
<td><code>BitSeq</code>, <code>cqn</code>, <code>cummeRbund</code>, <code>DESeq</code>, <code>DEXSeq</code>, <code>EDASeq</code>, <code>edgeR</code>, <code>gage</code>, <code>goseq</code>, <code>iASEq</code>, <code>tweeDEseq</code></td>
</tr>
<tr>
<td>ChIP-seq, etc.</td>
<td><code>BayesPeak</code>, <code>baySeq</code>, <code>ChIPpeakAnno</code>, <code>chipseq</code>, <code>ChIPseqR</code>, <code>ChIPsim</code>, <code>CSAR</code>, <code>DiffBind</code>, <code>MEDIPS</code>, <code>mosaics</code>, <code>NarrowPeaks</code>, <code>nucleR</code>, <code>PICS</code>, <code>PING</code>, <code>REDseq</code>, <code>Reptools</code>, <code>TSSi</code></td>
</tr>
<tr>
<td>Motifs</td>
<td><code>BCRANK</code>, <code>cosmo</code>, <code>cosmoGUI</code>, <code>MotIV</code>, <code>seqLogo</code>, <code>rGADEM</code></td>
</tr>
<tr>
<td>3C, etc.</td>
<td><code>HiTC</code>, <code>r3Cseq</code></td>
</tr>
<tr>
<td>Copy number</td>
<td><code>cn.mops</code>, <code>CNAnorm</code>, <code>exomeCopy</code>, <code>segmentSeq</code></td>
</tr>
<tr>
<td>Microbiome</td>
<td><code>phyloseq</code>, <code>DirichletMultinomial</code>, <code>clstutils</code>, <code>manta</code>, <code>mcaGUI</code></td>
</tr>
<tr>
<td>Work flows</td>
<td><code>ArrayExpressHTS</code>, <code>Genominator</code>, <code>easyRNASeq</code>, <code>oneChannelGUI</code>, <code>rnaSeqMap</code></td>
</tr>
<tr>
<td>Database</td>
<td><code>SRAdb</code></td>
</tr>
</tbody>
</table>
### Table 4: Selected Bioconductor packages for representing and manipulating ranges, strings, and other data structures.

<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IRanges</strong></td>
<td>Defines important classes (e.g., <em>IRanges, Rle</em>) and methods (e.g., <em>findOverlaps, countOverlaps</em>) for representing and manipulating ranges of consecutive values. Also introduces <em>DataFrame, SimpleList</em> and other classes tailored to representing very large data.</td>
</tr>
<tr>
<td><strong>GenomicRanges</strong></td>
<td>Range-based classes tailored to sequence representation (e.g., <em>GRanges, GRangesList</em>), with information about strand and sequence name.</td>
</tr>
<tr>
<td><strong>GenomicFeatures</strong></td>
<td>Foundation for manipulating data bases of genomic ranges, e.g., representing coordinates and organization of exons and transcripts of known genes.</td>
</tr>
<tr>
<td><strong>Biostrings</strong></td>
<td>Classes (e.g., <em>DNAStringSet</em>) and methods (e.g., <em>alphabetFrequency, pairwiseAlignment</em>) for representing and manipulating DNA and other biological sequences.</td>
</tr>
<tr>
<td><strong>BSgenome</strong></td>
<td>Representation and manipulation of large (e.g., whole-genome) sequences.</td>
</tr>
</tbody>
</table>
Table 6: Selected *Bioconductor* packages for sequence reads and alignments.

<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ShortRead</em></td>
<td>Defines the <em>ShortRead</em> class and functions for manipulating <em>fastq</em> files; these classes rely heavily on <em>Biostrings</em>.</td>
</tr>
<tr>
<td><em>GenomicRanges</em></td>
<td><em>GappedAlignments</em> and <em>GappedAlignmentPairs</em> store single- and paired-end aligned reads.</td>
</tr>
<tr>
<td><em>Rsamtools</em></td>
<td>Provides access to BAM alignment and other large sequence-related files.</td>
</tr>
<tr>
<td><em>rtracklayer</em></td>
<td>Input and output of <em>bed</em>, <em>wig</em> and similar files.</td>
</tr>
</tbody>
</table>
Table 8: Selected *Bioconductor* packages for RNA-seq analysis.

<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EDASeq</strong></td>
<td>Exploratory analysis and QA; also <em>qrqc, ShortRead</em>.</td>
</tr>
<tr>
<td><strong>edgeR, DESeq</strong></td>
<td>Generalized Linear Models using negative binomial error.</td>
</tr>
<tr>
<td><strong>DEXSeq</strong></td>
<td>Exon-level differential representation.</td>
</tr>
<tr>
<td><strong>goseq</strong></td>
<td>Gene set enrichment tailored to RNAseq count data; also <em>limma</em>’s roast or camera after transformation with <em>voom</em>.</td>
</tr>
<tr>
<td><strong>easyRNASeq</strong></td>
<td>Workflow; also <em>ArrayExpressHTS, rnaSeqMap, oneChannelGUI</em>.</td>
</tr>
<tr>
<td><strong>Rsubread</strong></td>
<td>Alignment (Linux only); also <em>Biostrings matchPDict</em> for special-purpose alignments.</td>
</tr>
</tbody>
</table>
Questions?

http://www.r-consultancy.com