easyRNAseq
Where from, where to?

Nicolas Delhomme
12th Dec. 2012, Zurich
Outline

• From where it comes
  – history and rationale

• To where it stands
  – current workflow and caveats

• And where it should be
  – current and planned changes
• Started at the EMBL (European Molecular Biology Laboratory), HD, in 2008
  – The data: a large transcriptome dataset (>30 RNA-Seq samples, fruitfly)
  – Not much (pre-)processing available.
  – R/Bioc capabilities of handling large dataset and performing statistical analyses.
  – At the same time, Bioconductor started to develop packages specifically for NGS data.
4 years later

• Bioc 2.11
  – more than 1,150 packages (550+ soft., 600+ annot.)

  – among which the Bioconductor “core” packages for NGS: 8 of them that leveraged R “base” to work conveniently with “biological” objects

  – but as well, 50+ contributed packages just for NGS analyses, and 18 RNA-Seq specific
As for 3 years ago, the rationale remains:

– to simplify (to ease) my pre-processing because analyses is what’s the most gratifying!

– so I wanted a function to which I would give
  • aligned reads
  • chromosomic
  • genic annotation

– and from which I would get a count table.
Challenges

• Computationally relevant

<table>
<thead>
<tr>
<th>Data volume</th>
<th>disk space and computing capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>complex architecture</td>
</tr>
</tbody>
</table>

• Biologically relevant

<table>
<thead>
<tr>
<th>Reference selection</th>
<th>Genome, transcriptome?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference validity</td>
<td>e.g. different strains</td>
</tr>
<tr>
<td>Alignment policies</td>
<td>quality trimming, gapped alignments, multimapping</td>
</tr>
<tr>
<td>Replicates</td>
<td>etc...</td>
</tr>
</tbody>
</table>
easyRNASeq (for me)

• An R package
  – to ease various RNA-Seq analyses (flexibility)
  – that wraps and combines the functionalities of many R packages
  – all encapsulated into a single function call: easyRNAseq

– additional functions
  • e.g. de-multiplex data
easyRNASeq (for the user)

- The same, but as less a black box as possible!

- A non-negligible part of the computation time is spent on assessing the validity of the user input
Where it stands

- annotation source (e.g. gff, mart DB)
  - genomeIntervals, biomaRt
    - annotation set (e.g. exon, gene)
      - genic annotation iRanges, GenomicRanges
        - read count summarization iRanges, GenomicRanges
          - count table (default) or an easyRNASeq, DESeq or edgeR object

- reads file (e.g. export, BAM)
  - ShortRead, Rsamtools
    - aligned read information
      - read coverage iRanges, GenomicRanges
  - BSgenome
    - chromosome sizes

+ SummarizedExperiment
An example call

```r
library(easyRNASeq)

count.table <- easyRNASeq(
  filesDirectory=system.file("extdata",package="RnaSeqTutorial"),
  pattern="[A,C,T,G]{6}\bam$",
  organism="Dmelanogaster",
  annotationMethod="rda",
  annotationFile=system.file("data","gAnnot.rda",package="RnaSeqTutorial"),
  count="genes",
  summarization="geneModels")
)
Checking arguments...
Fetching annotations...
Computing gene models...
Summarizing counts...
Processing ACACTG.bam
Updating the read length information.
The reads are of 30 bp.
Processing ACTAGC.bam
Updating the read length information.
The reads are of 30 bp.
Processing ATGGCT.bam
“Easy”: count the number of reads that aligned to a gene, exon, splice-junction...

What about multi-mapping reads?
And what about isoform levels?

![Diagram showing read mapping to genome with annotated transcripts labeled g1 t1 and g1 t2.](image-url)
Pitfalls

- annotation source (e.g. gff, mart DB)
- reads file (e.g. export, BAM)

- genomeIntervals, biomaRt
- ShortRead, Rsamtools
- BSgenome

- annotation set (e.g. exon, gene)
- aligned read information
- chromosome sizes

- genic annotation
  - IRanges, GenomicRanges
- read coverage
  - IRanges, GenomicRanges

- read count summarization
  - IRanges, GenomicRanges

- count table (default) or an easyRNASeq, DESeq or edgeR object
Warning messages:
1: In easyRNASeq(filesDirectory = system.file("extdata", package = "RnaSeqTutorial")), :
   You enforce UCSC chromosome conventions, however the provided annotation is not compliant. Correcting it.
2: In easyRNASeq(filesDirectory = system.file("extdata", package = "RnaSeqTutorial")), :
   There are 2238 synthetic exons as determined from your annotation that overlap! This implies that some reads will be counted more than once! Is that really what you want?
Pitfalls

• Mind the alignment specificities

• Mind the “reference” used

• Mind the correspondence of the names between
  – the annotation
  – the read alignments

  – Most frequent subject of users’ email.

• Many of these are controlled for (which is why easyRNAseq emits so many warnings)
BioC package for summarizing expression

- GenomicRanges (summarizeOverlaps())
- easyRNAsSeq (easyRNAsSeq())
- Rsubread (featureCounts())
- ArrayExpressHTS
- ...
htseq-count¹ (Simon Anders) used as a reference

The figure is a courtesy of Mark Robinson
¹ (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html)
Use cases

• Evident use case:
  – RNA-Seq differential expression

• useful as well
  – for further QA
  – for other experimental analyses
    • short RNA
    • Tag-Seq
    • CaGE-Seq
- **RPKM**
  - Reads Per feature Kb per Million reads in the library

- **DESeq**
  - based on a Negative Binomial
  - fit a model to correct for the library sizes

- **edgeR**
  - based on a Negative Binomial
  - use a trimmed mean of M-values to correct for the library sizes
• Automatic if the user requires a DESeq or edgeR ready to use object.
Summary

• Bioc core packages have been consolidated

• but easyRNAseq still needs to mature
  – integrate most recent technologies (multi-mapping)
  – annotation processing
  – vignette pruning and adding an FAQ section
Where to go

- Rsamtools: stream BAM file to minimize mem. footprint.

- SummarizedExperiment as a standard
  – add coercion from/to edgeR - DESeq

- Plugin additional methods (cqn, DEXSeq,...)

- Use BiocParallel instead of parallel
Acknowledgments

• For the content of that talk
  – Angela Gonçalves, Mark Robinson

• For the invitation
  – Mark Robinson

• For precious comments and help
  – The whole Bioconductor “core” group (present and past), Wolfgang Huber, Simon Anders, Charles Girardot, Stefan Bonn

• For feedback
  – Numerous Bioc users: Tim Triche Jr., Wade Davis, Richard Friedman

• You, for your attention