Read Counting in RNA-seq

Hervé Pagès
hpages@fhcrc.org

Fred Hutchinson Cancer Research Center
Seattle, WA, USA

21 January 2014

Outline

Introduction

Counting the reads with Bioconductor

The 2 types of applications of RNA-seq

Discovery

- find new transcripts
- find transcript boundaries
- find splice junctions

Comparison Given samples from different experimental conditions, find effects of the treatment on

- gene expression strengths (a.k.a. “differential analysis at the gene level”)
- isoform abundance ratios

Workflow of a differential analysis of RNA-Seq data

- Start with: Short-read sequences with qualities (FASTQ files)
- Align to a reference genome ==> SAM files
- Count reads per gene or exon (based on a gene model) => matrix of counts
- Statistical analysis on the counts (fold-changes, p values, etc...)
- Downstream analyses (gene set enrichment analysis, nearest peak to a differentially expressed gene, etc...)

Alignment

Typically done with a stand-alone software.
For RNA-Seq, we need a splice-aware aligner:

- TopHat2
- GSNAP
- etc...

Counting reads per gene

- Count each read at most once.
- Discard a read if
    - it cannot be uniquely mapped
    - its alignment overlaps with several genes
    - the alignment quality score is bad
    - (for paired-end reads) the mates do not map to the same gene

Outline

Introduction

Counting the reads with Bioconductor

Reading BAM files

TODO...

Chosing and loading a gene model

TODO...

Using `summarizeOverlaps`

TODO...

Basic manipulation of a *SummarizedExperiment* object

TODO...