Counting reads for RNA-seq

Martin T. Morgan mtmorgan@fhcrc.org
Fred Hutchinson Cancer Research Center
Seattle, WA, USA

25 August 2014

Varieties of RNA-seq

1. Known gene differential expression
    - Genes, *DESeq2*, *edgeR*
    - Transcripts
    - Exons, *DEXSeq*
2. Novel transcripts

RNA-seq work flow

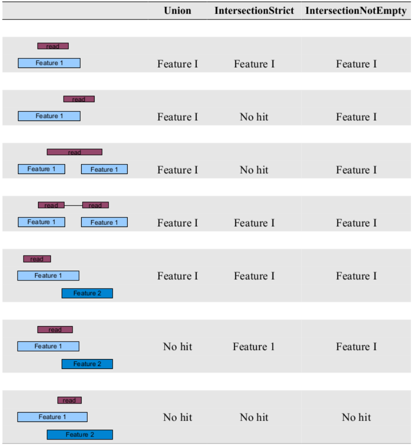
1. Experimental design – keep it simple; replicate.
2. Wet-lab preparation – covariates & opportunities for ‘batch’ effects
3. Sequencing – paired-end valuable for transcript-level inference
4. Alignment – typically to whole genome; requires aligner capable of gapped alignments
5. Summary – reads overlapping each gene or region of interest
6. Analysis – linear model (e.g., t-test) fit to each region of interest; ‘top table’ of differentially expressed genes
7. Comprehension – annotation of differentially expressed regions, gene set enrichment, comparison to other studies, integration with other data types

RNA-seq summary: counts per region of interest

- Input: BAM files of aligned reads, typically one per sample
- How to count?
  - What is an ‘overlap’?
  - What (*Bioconductor*) software to use?
- Output: region × sample matrix of read counts
- *Not* RPKM or other ‘normalized’ measure

RNA-seq summary: how to count?

Counting modes



	Union	IntersectionStrict	IntersectionNotEmpty
read Feature 1	Feature I	Feature I	Feature I
read Feature 1	Feature I	No hit	Feature I
read Feature 1 Feature 1	Feature I	No hit	Feature I
read read Feature 1 Feature 1	Feature I	Feature I	Feature I
read Feature 1 Feature 2	Feature I	Feature I	Feature I
read Feature 1 Feature 2	No hit	Feature 1	Feature I
read Feature 1 Feature 2	No hit	No hit	No hit

* Picture reproduced from HTSeq web site :
http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

Counting in *Bioconductor*

- *GenomicAlignments* `summarizeOverlaps()` – standard adn customized counting modes
- *Rsubread* `featureCounts()` – fast; Linux and Mac only