General considerations for RNA-seq quantification for differential expression and or splice variant assessment
What are the goals of your research? Why did you generate all of the RNAseq data in the first place.

- RNA-seq is generated for a number of reasons
- Transcriptome assembly (& SNP discovery)
- Transcript discovery (variants for Transcription start site, alternative splicing, etc..)
- Quantification of (alternative transcripts)
- Differential expression analysis across treatments.
What was once thought to be separate goals are now clearly recognized as intertwined.

- Early work for RNA-seq tried to “mirror” the type of gene level analysis used in microarrays.

- However, RNA-seq has demonstrated how important it is to take into account alternative transcripts, even when attempting to get “gene level” measures.
How do we put together a useful pipeline for RNAseq

• What are the steps we need to consider?
How do we put together a useful pipeline for RNAseq

• What are the steps we need to consider?
• Genome/transcriptome assembly.
• Mapping reads to genome/transcriptome.
• Deal with alternative transcripts (new transcriptome)?
• Remap & count reads.
• Differential expression.
The “tuxedo” protocol for RNA-seq

Trapnell et al 2012
Pipelines for RNA-seq (geared towards splicing)

Vijay et al. 2012
How should we map reads

- Do we want to map to a reference genome (with a “splice aware” aligner)?
- Or do we want to map to a transcriptome directly.
Mapping to a transcriptome

• What might be the downside to mapping to the transcriptome?

• unspliced read aligners are useful against a transcript (or cDNA) database, such as that generated for a de novo transcriptome.

• For this BW is faster than seed based approaches (shrimp & stampy), but the latter may be preferred if mapping to "distant" transcriptomes.
Mapping to the genome

• How do we deal with alternative transcripts or paralogs during mapping?

"splicing aware" aligners:
  – Exon First: (tophat, MapSplice, SpliceMap) Fig1A Garber
  – Step 1 - map reads to genome
  – Step 2 - unmapped reads are split, and aligned.

• Seed & extend (Fig1B Garber) (GSNAP, QPALMA)
  – kmers from reads are mapped (the seeds), and then extended
a  Exon-first approach

Exon 1  Exon 2  RNA

Exon read mapping

Spliced read mapping

Seed-extend approach

Exon 1  Exon 2  RNA

Seed matching

k-mer seeds

Seed extend

Potential limitations of exon-first approaches

Gene

Pseudogene

Garber et al. 2011
Merging all transcripts?

Trapnell et al 2012.
Counting

• One of the most difficult issues has been how to count reads.

• What are some of the issues that we need to account for during counting of reads?
Counting
Counting

• We are interested in transcript abundance.
• But we need to take into account a number of things.
  • How many reads in the sample.
  • Length of transcripts
  • GC content and sequencing bias
Counting

- RPKM (reads aligned per kilobase of exon per million reads mapped) – Mortazavi et al 2008
- FPKM (fragments per kilobase of exon per million fragments mapped). Same idea for paired end sequencing.
Accounting for multiple isoforms.

- Only count reads that map uniquely to an isoform (Alexa-Seq). Can be very problematic, when isoforms do not have unique exons.

- So called "isoform-expression" methods (cufflinks, MISO) model the uncertainty parametrically (often using MLE). The model with the best mix of isoforms that models the data (highest joint probability) is the best estimate. How this is handled differs a great deal by the different
What does this tell us?

These cells are biological replicates (diagonals)

These cells are for different software

Nookaew et al 2102 NAR
Differentially expressed genes based on software for quantification

Differentially expressed genes based on software for mapping

Nookaew et al 2102 NAR
Seqanswer or blog postings of use

Problems with cufflink and cuffdiff?
Reproducibility...

- http://www.biostars.org/p/6317/
Counting reads

• Htseq (python library) works with DEseq
Differential expression

- EDGE-R
- EBseq (RSEM/EBseq)
- RSEM (http://deweylab.biostat.wisc.edu/rsem/)
- eXpress (http://bio.math.berkeley.edu/eXpress/overview.html)
- Beers simulation pipeline (http://www.cbil.upenn.edu/BEERS/)
- DEXseq (http://bioconductor.org/packages/release/bioc/html/DEXSeq.html)
Example workflows
