NGS2014 Monday & Tuesday

• Today, Tomorrow and maybe part of Wed we are going to get more in depth on aspects of RNAseq.
• I will be giving a lecture on some additional aspects of RNAseq (M&T).
• Matt will go over in more detail Transcriptome assembly (and will discuss SOAP-deNovo Trans as well as Trinity).
• Chris, Matt and Meg will help lead you through a “choose your own adventure” of RNAseq analysis.
Tomorrow I will give a short lecture on some of the basics of performing differential expression analysis with counts from RNAseq data.

Meg and Matt may discuss or do a short tutorial on aspects of annotating and "cleaning" your transcriptomes.

We will also do a tutorial comparing some of the results of your "counting" from today.
General considerations for RNA-seq quantification for differential expression

or how to count.
First some apologies (I am moving back to Canada, and I need to practice this).

• From all I have seen and heard, Meg has already covered a number of important aspects of this, so there will be some repetition.

• I am purposefully repeating some of the same highlights in a few places, because it truly bears repeating!
Biological replicates Not technical ones.

• Meg already went over this, but there is little purpose in using technical replication from a given biological sample UNLESS part of your question revolves around it.

• Focus on biological variability. While you are confounding some sources of technical and biological variability, we already know a lot about the former, and little about the latter.
Blocking

Blocks in experimental design represent some factor (usually something not of major interest) that can strongly influence your outcomes. More importantly it is a factor which you can use to group other factors that you are interested in.

For instance in agriculture there is often plot to plot variation. You may not be interested in the plot themselves but in the variety of crops you are growing.

But what would happen if you grew all of strain 1 on plot 1 and all of strain 2 on plot 2?

Whiteboard.

These plots would represent blocking levels.
Blocking

Sampling

Replication

Blocking

Randomization

In genomic studies the major blocking levels are often the slide/chip for microarrays (i.e. two samples /slide for 2 color arrays, 16 arrays/slide for Illumina arrays).

For GAII/HiSeq RNA-seq data the major blocking effect is the flow cell itself, or lanes within the flow cell.

Auer and Doerge 2010
Blocking
Incorporating lanes as a blocking effect

Sampling
Replication
Blocking
Randomization

Auer and Doerge 2010
Blocking designs

Sampling

Replication

Blocking

Randomization

Balanced for treatments across flow cells... Randomized for location

Balanced Incomplete Blocking Design (BIBD)

Let’s dissect these subscripts.

Auer and Doerge 2010
What is your research question?:

What are the goals of your research?

Why did you generate all of the RNAseq data in the first place?
What can you use RNAseq data for?
Using RNAseq

• Transcriptome assembly.
• Improving genome assembly/annotation.
• SNP discovery (large genomes)
• Transcript discovery (variants for Transcription start site, alternative splicing, etc..)
• Quantification of (alternative transcripts)
• Differential expression analysis across treatments.
Using RNAseq

• Transcriptome assembly.
• Improving genome assembly/annotation.
• SNP discovery (large genomes)
• Transcript discovery (variants for Transcription start site, alternative splicing, etc..)

• Differential expression analysis...
Using RNAseq: differential expression

- Differential expression of what?
- Differential expression at the level of “genes”
- Allele specific expression
- Quantification of alternative transcripts
Your primary goals of experiment should guide how you perform your experiment.

- The exact details (# biological samples, sample depth, read_length, strand specificity) of how you perform your experiment needs to be guided by your primary goal.

- Unless you have all the $$, no single design can capture all of the variability.
Your goals matter

• For instance: If your primary interest in discovery of new transcripts, sampling deeply within a sample is probably best.

• For differential expression analyses, you will almost never have the ability to perform Differential expression analysis on very rare transcripts, so it is rarely useful to generate more than 15-20 million read pairs (see Meg’s slides).
Are singleEnded reads ever useful?

• In my experience (plants and animals), almost never.
• My primary organism (Drosophila melanogaster) is one of the best annotated and experimentally validated genomes.
• Even still, we get surprising ambiguity for reads 75bp and shorter, which mostly goes away with PE.
• Hopefully less of a problem now (as most people are doing 100 -150 bp+).
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
The point...

- There is no single “best” way forward yet. It is probably best to try several pipelines and think carefully about each of the steps.
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.

• What is preferable, to generate a de novo transcriptome or map to a “closely” related species?
And before we map reads...

• How should we filter (based on quality) reads (if at all)?

• What are some of the considerations (Matt...?)
Mapping to a transcriptome

• What are the downsides to mapping to a transcriptome?
Mapping to a 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
Garber et al. 2011
The variation in the mapping step (at least with a reference genome) seems to have modest effects.

RNA-seq gene profiling - a systematic empirical comparison
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

Nookaew et al 2102 NAR
Which to use

• If a (close to?) perfect match transcriptome assembly is available for mapping. Burrows-wheeler based aligners can be much faster than seed based methods (upto 15x faster)

• BW based aligners have reduced performance once mismatches are considered.
  – Exponential decrease in performance with each additional mismatch (iteratively performs perfect searches).
  – Seed methods may be more sensitive when mapping to transcriptomes of distantly related species (or high polymorphism rates).

From Garber et al. 2011
How could mapping reads (whether to a reference genome or transcriptome) influence our downstream counts?
How could mapping reads (whether to a reference genome or transcriptome) influence our downstream counts?
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

• What are we trying to count?
• Gene level measure (eXpress, corset, RSEM, HTSeq,...).
• Exon level (HTSeq, ???)
• Transcript level (HTSeq, Cufflinks, ....)
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
  - (how many transcripts)
Old ways of counting 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.
- Transcripts per million (we will come to that).
None of these measures are great for differential expression analysis.

• For appropriate differential expression analysis (as with all statistical modeling), keeping all of the data is better.

• So having counts of mapped reads, along with information like GC content, transcript length, total # reads is far more useful.

• We will discuss this tomorrow.
Accounting for multiple isoforms (when counting alternative transcripts).

- Only count reads that map uniquely to an isoform (Alexa-Seq, HTSeq). 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.
Figure a: Illustration of different exon models.

Figure b: Table showing log fold-change calculations for different conditions.

<table>
<thead>
<tr>
<th>Condition A</th>
<th>Condition B</th>
<th>Log fold-change (union count)</th>
<th>Log fold-change (intersect count)</th>
<th>Log fold-change (true expression)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\log_2 \left( \frac{10}{10} \right) = 0$</td>
<td>$\log_2 \left( \frac{7}{1} \right) = 0.19$</td>
<td>$\log_2 \left( \frac{8}{6 + \frac{6}{2L}} \right) = 0.32$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\log_2 \left( \frac{8}{8} \right) = -0.41$</td>
<td>$\log_2 \left( \frac{5}{5} \right) = 0$</td>
<td>$\log_2 \left( \frac{6/1}{5/L} \right) = 0.53$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\log_2 \left( \frac{5}{10} \right) = -1$</td>
<td>$\log_2 \left( \frac{4}{5} \right) = -0.1$</td>
<td>$\log_2 \left( \frac{5/10}{2L} \right) = 0$</td>
</tr>
</tbody>
</table>

Trapnell et al 2013
However...

• There has been a great deal of discussion and disagreement about this (see seqanswer forums, and “discussions” between Simon Anders and Lior Patcher).

• Fundamentally there are numerous cases where both methods fail. So take care.
Seqanswer or blog postings of use

Problems with cufflink and cuffdiff? Reproducibility...

- http://www.biostars.org/p/6317/
Counting at the “gene” or exon level may be simpler (at least initially).

- i.e. all mapped reads for transcripts associated with a particular “gene” get counted (HTSeq, corset, eXpress, RSEM (?)).
Counting reads

• Htseq (python library) works with Deseq.

• In our experience this is both easy (ish) to use and counting in a sensible manner.

• I remain very confused about getting “counts” out of both RSEM and Cufflinks...
Differential expression

- **EDGE-R**
- **EBseq (RSEM/EBseq)**
- **RSEM** ([http://deweylab.biostat.wisc.edu/rsem/](http://deweylab.biostat.wisc.edu/rsem/))
- **eXpress** ([http://bio.math.berkeley.edu/eXpress/overview.html](http://bio.math.berkeley.edu/eXpress/overview.html))
- **Beers simulation pipeline** ([http://www.cbil.upenn.edu/BEERS/](http://www.cbil.upenn.edu/BEERS/))
- **Limma (voom)**
Example workflows
