RNA Sequencing

Kasper D. Hansen (khansen@jhsph.edu)
140.668 SPECIAL TOPICS IN GENOMICS
Background
Much excitement over RNA-Sequencing
Measuring Gene Expression

1. Hybridize to array
2. Sequence and count
3. Measure intensity
Data from *D. melanogaster*
Why the excitement?

Technical replicates
$R^2 = 0.96$

Sensitivity and dynamic range
$R^2 = 0.99$

Mortazavi (Nat. Methods 2008)

A great paper!

(Other compelling reasons as well)
Mapping
Mapping transcripts

- Genome
- Transcript
- Length in genome space
- Paired-end reads
The basic approaches

(a) De novo assembly of the transcriptome

Highly expressed gene
Lowly expressed gene

Read coverage must be high enough to build EST contigs (solid bar)

(b) Map onto the genome

Read mapper must support splitting reads to record splices

(c) Map onto the genome and splice junctions

Splice junctions sequences from either annotations or inferred

From Pepke (2009 Nat Methods)
Mapping reads to the transcriptome

- Genome
- Transcriptome
- $2^\text{Genome}$

Illustration idea from Lior Patcher

Well established

Reads
Junction reads

Genomic Reads

Splice Junction Reads

Genomic Reads

Splice Junction Reads

Image from Brenton Gravely
Junction reads, zoom

Image from Brenton Gravely
FP rates for junctions: mapping strategy

Strategies for judging trustworthiness of a junction given the set of reads covering it, from Brooks 2010

### Table: Overhang Positions

<table>
<thead>
<tr>
<th>Offset</th>
<th>5' Exon</th>
<th>3' Exon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Bar Graph: Annotated Junctions (n=58,212)

- 1+ Offsets: 40%
- 2+ Offsets: 30%
- 3+ Offsets: 20%
- 4+ Offsets: 10%
- 5+ Offsets: 0%

### Bar Graph: Randomly Generated Junctions (n=5,409,600)

- 1+ Offsets: 0.070%
- 2+ Offsets: 0.053%
- 3+ Offsets: 0.035%
- 4+ Offsets: 0.018%
- 5+ Offsets: 0%

Supplemental Figure 3.

Analysis of optimal overhang and mismatch for splice junction alignments. (A) Distribution of Overhang Positions. A histogram of the number of uniquely aligned reads across all annotated junctions is shown. An even distribution of read alignments across all base positions occurs if at least a 6nt overhang is enforced. (B) Distinguishing true junctions from false positive alignments. To reduce the number of false positive junctions, as determined by randomly generated junctions, a total of 3 alignment start positions (offsets) were required to consider a junction to be truly present.
Mapping - conclusions

Mapping to transcript space is not easy.

But essential for really understanding alternative splicing.


Bias
Data from *D. melanogaster*

Image from Brenton Gravely
Base effect - single sample

14.1m reads

proportion of total reads in region

position
Base effect - multiple samples

Proportion of total reads in region vs position for different samples.

- wt
- rrp
- ski

Position range: 369860 - 369940

Proportion range: 0.00 - 0.04
Base effect - different study (and prep)
Base effect

Reproducible base effect - like probe affinities in microarrays. Seems to be prep dependent.

Creates issues for comparing different regions in the genome.

Less of an issue for comparing the same region across samples?
Looking for explanation for position-bias

Cycle, also 5’ to 3’
Believed that higher cycle has lower quality

Proportions of the four nucleotides at cycle 24
Proportions of the 4096 hexamers starting at cycle 1

May contain ‘N’s
Supplementary Figure S1. Nucleotide frequency vs. position for stringently mapped reads. For each experiment, mapped reads were extended upstream of the 5' start position, such that the first position of the actual read is and positions to are obtained from the mapping to the genome. The first hexamer of the read is shaded dark grey. The experiments under "RNA-Seq" were all conducted using random hexamer priming, in most cases preceded by fragmentation of the RNA (two experiments did not use fragmentation). The "DNA-Seq" experiments were done using the standard DNA sequencing protocol and consist of a variety of resequencing and ChIP"Seq experiments. The "RNA-Seq, other protocols" experiments were conducted using a variety of protocols: oligo"dT priming followed by fragmentation using nebulization, sonication, and DNase I, as well as random hexamer priming followed by DNase fragmentation, see also Supplementary Table S4.

6 experiments
5 Labs (including Illumina at Hayward)
4 organisms (yeast, fly, mouse, human)

All used the standard protocol, expect 1 experiment that did not employ fragmentation (old protocol).

Uniquely, perfectly mapped reads.
Nucleotide frequency vs. position

Supplementary Figure S1. Nucleotide frequency vs. position for stringently mapped reads. For each experiment, mapped reads were extended upstream of the 5' start position, such that the first position of the actual read is 1 and positions to (10 are obtained from the mapping to the genome. The first hexamer of the read is shaded dark grey. The experiments under “RNA-Seq” were all conducted using random hexamer priming, in most cases preceded by fragmentation of the RNA (two experiments did not use fragmentation). The “DNA-Seq” experiments were done using the standard DNA sequencing protocol and consist of a variety of resequencing and ChIP-Seq experiments. The “RNA-Seq, other protocols” experiments were conducted using a variety of protocols: oligo-dT priming followed by fragmentation using nebulization, sonication, and DNase I, as well as random hexamer priming followed by DNase fragmentation, see also Supplementary Table S4.
Bias - conclusions

There is clear bias in RNA-Seq

But how much, why and how to deal with it are still open problems.


Coverage
Data from *D. melanogaster*

Image from Brenton Gravely
Detection in *Drosophila*

- untreated - 32.9m reads
- brr2 - 15.0m reads
- pasilla - 28.3m reads

1 read in 200bp, 10m reads

10 reads in 200bp, 10m reads
Differential Expression
Lessons from MA: Biological variation

Pooled samples: Technical variation

Individual samples: Biological (and Technical) variation

From Kendziorski et al (2005) PNAS

Microarray Data: Mouse
Same samples hybridized individually and as a pool
What is our measure of “gene expression”?

With microarrays, we have predefined features which collect measurements for a given gene.

Here we need to ask:

- What is a gene?
- What is the expression level of the gene, given our observed data?
Counting reads that overlap gene annotation

The answers to these questions depend on the purpose of our experiment. For simplicity, imagine we want to compare gene expression as in a microarray study. We will:

- Use known gene boundaries to determine a “gene region”.
- Measure gene expression as the count of reads that start in this region.

<table>
<thead>
<tr>
<th></th>
<th>mut_1</th>
<th>mut_2</th>
<th>wt_1</th>
<th>wt_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>YHR055C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YPR161C</td>
<td>38</td>
<td>39</td>
<td>35</td>
<td>34</td>
</tr>
<tr>
<td>YOL138C</td>
<td>31</td>
<td>33</td>
<td>40</td>
<td>26</td>
</tr>
<tr>
<td>YDR395W</td>
<td>55</td>
<td>52</td>
<td>47</td>
<td>47</td>
</tr>
<tr>
<td>YGR129W</td>
<td>29</td>
<td>26</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>YPR165W</td>
<td>189</td>
<td>180</td>
<td>151</td>
<td>180</td>
</tr>
</tbody>
</table>
Representing genes

Union-Intersection representation

Union representation
Making samples comparable

First thought - adjust for total lane counts

RPKM: reads per kilobase per million reads mapped; adjusts for gene length and total sequencing depth, but maybe not in a good way.

More advanced normalization to adjust for distributional differences:

- Upper quartile (Bullard et al, 2010)
- Trimmed mean of M-values (TMM, Robinson and Oshlack, 2010)
- Robust scaling to pseudo-reference sample (Anders and Huber, 2010)
Why we need normalization

Other datasets

The global shift in log-fold-change caused by RNA composition differences occurs at varying degrees in other RNA-seq datasets. For example, an M versus A plot for the Cloonan et al. [12] dataset (Figure S3 in Additional file 1) gives an estimated TMM scaling factor of 1.04 between the two samples (embryoid bodies versus embryonic stem cells), sequenced on the SOLiD™ system. The M versus A plot for this dataset also highlights an interesting set of genes that have lower overall expression, but higher in embryoid bodies. This explains the positive shift in log-fold-changes for the remaining genes. The TMM scale factor appears close to the median log-fold-changes amongst a set of approximately 500 mouse housekeeping genes (from [17]). As another example, the Li et al. [18] dataset, using the illumina 1G Genome Analyzer, exhibits a shift in the overall distribution of log-fold-changes and gives a TMM scaling factor of 0.904 (Figure S4 in Additional file 1). However, there are sequencing-based datasets that have quite similar RNA outputs and may not need a significant adjustment. For example, the small-RNA-seq data from Kuchenbauer et al. [19] exhibit only modest bias in the log-fold-changes (Figure S5 in Additional file 1). Spike-in controls have the potential to be used for normalization. In this scenario, small but known amounts of RNA from a foreign organism are added to each sample at a specified concentration. In order to use spike-in controls for normalization, the ratio of the concentration of the spike to the sample must be kept constant throughout the experiment. In practice, this is difficult to achieve and small variations will lead to biased estimation of the normalization factor. For example, using the spiked-in DNA from the Mortazavi et al. data set [11] would lead to unrealistic normalization factor estimates (Figure S6 in Additional file 1). As with Figure 1 Normalization is required for RNA-seq data. Data from [6] comparing log ratios of (a) technical replicates and (b) liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of the housekeeping genes. (c) An M versus A plot comparing liver and kidney shows a clear offset from zero. Green points indicate 545 housekeeping genes, while the green line signifies the median log-ratio of the housekeeping genes. The red line shows the estimated TMM normalization factor. The smear of orange points highlights the genes that were observed in only one of the liver or kidney tissues. The black arrow highlights the set of prominent genes that are largely attributable for the overall bias in log-fold-changes.

Table 1 Number of genes called differentially expressed between liver and kidney at a false discovery rate <0.001 using different normalization methods

<table>
<thead>
<tr>
<th>Library size</th>
<th>TMM normalization</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher in liver</td>
<td>2,355</td>
<td>2,355</td>
</tr>
<tr>
<td>Higher in kidney</td>
<td>8,332</td>
<td>4,935</td>
</tr>
<tr>
<td>Total</td>
<td>10,867</td>
<td>7,290</td>
</tr>
</tbody>
</table>

House keeping genes (545)

<table>
<thead>
<tr>
<th>Library size</th>
<th>TMM normalization</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher in liver</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Higher in kidney</td>
<td>376</td>
<td>220</td>
</tr>
<tr>
<td>Total</td>
<td>421</td>
<td>265</td>
</tr>
</tbody>
</table>

TMM, trimmed mean of M values.

Robinson and Oshlack, 2010
RPKMs can be bad

Fold change estimates using RNA-Seq compared with qRT-PCR

Points are RPKMs for roughly 1000 genes.

The sequence data has no biological reps (but this is commercial RNA).

Bias based on gene length

(a) Full-length UI genes
(b) 250-bp UI gene regions

Bad for interpretation
Comparing replicate samples

Sources of variation

Lane/Flowcell
Library prep
Extraction
Biological variation

Systematic differences

?: Is absolute quantification possible

Poisson model

good fit

less good fit

mut

observed quantiles

theoretical quantiles
Biological variation is (still) real

Hansen, Irizarry, Leek. Nat Biotech in press
Differential Expression, Conclusion

Still need normalization

Methods for DE are appearing


