Computational approaches to the analysis of RNA-seq data

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB
Next-generation gap

http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.f.268.html
The evolution of transcriptomics

**Hybridization-based**

1995 P. Brown, et. al.
Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)

RNA-seq is still a technology under active development
How RNA-seq works

Sample preparation

Next generation sequencing (NGS)

Data analysis:
- Mapping reads
- Visualization (Gbrowse)
- De novo assembly
- Quantification

Figure from Wang et. al, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genetics 10, 57-63, 2009).
RNA-seq vs. microarray

- RNA-seq can be used to characterize novel transcripts and splicing variants as well as to profile the expression levels of known transcripts (but hybridization-based techniques are limited to detect transcripts corresponding to known genomic sequences)

- RNA-seq has **higher resolution** than whole genome tiling array analysis
  - In principle, mRNA can achieve single-base resolution, where the resolution of tiling array depends on the density of probes

- RNA-seq can apply the same experimental protocol to various purposes, whereas specialized arrays need to be designed in these cases
  - Detecting single nucleotide polymorphisms (needs SNP array otherwise)
  - Mapping exon junctions (needs junction array otherwise)
  - Detecting gene fusions (needs gene fusion array otherwise)
  - Next-generation sequencing (NGS) technologies are now challenging microarrays as the tool of choice for genome analysis.
Advantages of RNA-Seq compared with other transcriptomics methods

<table>
<thead>
<tr>
<th>Technology specifications</th>
<th>Tiling microarray</th>
<th>cDNA or EST sequencing</th>
<th>RNA-Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principle</td>
<td>Hybridization</td>
<td>Sanger sequencing</td>
<td>High-throughput sequencing</td>
</tr>
<tr>
<td>Resolution</td>
<td>From several to 100 bp</td>
<td>Single base</td>
<td>Single base</td>
</tr>
<tr>
<td>Throughput</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Reliance on genomic sequence</td>
<td>Yes</td>
<td>No</td>
<td>In some cases</td>
</tr>
<tr>
<td>Background noise</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Application</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simultaneously map transcribed regions and gene expression</td>
<td>Yes</td>
<td>Limited for gene expression</td>
<td>Yes</td>
</tr>
<tr>
<td>Dynamic range to quantify gene expression level</td>
<td>Up to a few-hundredfold</td>
<td>Not practical</td>
<td>&gt;8,000-fold</td>
</tr>
<tr>
<td>Ability to distinguish different isoforms</td>
<td>Limited</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ability to distinguish allelic expression</td>
<td>Limited</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Practical issues</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Required amount of RNA</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Cost for mapping transcriptomes of large genomes</td>
<td>High</td>
<td>High</td>
<td>Relatively low</td>
</tr>
</tbody>
</table>
RNA-seq and microarray agree fairly well only for genes with medium levels of expression. Saccharomyces cerevisiae cells grown in nutrient-rich media. Correlation is very low for genes with either low or high expression levels.
Challenges for RNA-Seq: library construction

Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3' end of the transcript. RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5' and 3' ends.

A specific yeast gene, SES1 (seryl-tRNA synthetase)
Library construction

- Unlike small RNAs (microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), short interfering RNAs (siRNAs) and many others), which can be directly sequenced after adaptor ligation, larger RNA molecules must be fragmented into smaller pieces (200–500 bp) to be compatible with most deep-sequencing technologies.

- Common fragmentation methods include RNA fragmentation (RNA hydrolysis or nebulization) and cDNA fragmentation (DNase I treatment or sonication).

- Each of these methods creates a different bias in the outcome.
Library construction

- **PCR artefacts**
  - Many short reads that are identical to each other can be obtained from cDNA libraries that have been amplified. These could be a genuine reflection of abundant RNA species, or they could be PCR artefacts.
  - Use replicates

- **Whether or not to prepare strand-specific libraries**
  - Strand-specific libraries are valuable for transcriptome annotation, especially for regions with overlapping transcription from opposite direction.
  - Strand-specific libraries are currently laborious to produce because they require many steps or direct RNA–RNA ligation, which is inefficient.
Data analysis for mRNA-seq: key steps

• Mapping reads to the *reference* genome
  – Many mapping tools have been developed—speed & sensitivity
• Quantifying the known genes
• Prediction of novel transcripts
  – Assembly of short reads: comparative vs. *de novo*
• Quantifying splicing variants
Task 1: Short read mapping

- **Input:**
  - A reference genome
  - A collection of many short reads
  - User-specified parameters

- **Output:**
  - One or more genomic coordinates for each read

- In practice, only 70-75% of reads successfully map to the reference genome. Why?
Multiple mapping

- A single read may occur more than once in the reference genome.
- The user may choose to ignore reads that appear more than $n$ times.
- As $n$ gets large, you get more data, but also more noise in the data.
An observed read may not exactly match any position in the reference genome.

Sometimes, the read *almost* matches one or more positions.

Such mismatches may represent a SNP (single-nucleotide polymorphism, see [wikipedia](https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism)) or an sequencing error.

The user can specify the maximum number of mismatches, or a phred-style quality score threshold.

As the number of allowed mismatches goes up, the number of mapped reads increases, but so does the number of incorrectly mapped reads.
Task 2: Quantification of known transcripts

• The expression levels of known transcripts (exon model) are measured by the number of reads per kilobase of transcript per million mapped reads (RPKM)

Lander-Waterman model: foundation of gene expression profiling using mRNA-seq

- Consider a shotgun sequencing process in which \( N \) reads (of length \( T \)) are sampled from a pool of transcripts of total length \( G \). The number of reads mapped to a transcript (with length \( g \) and \( n \) copies) follows a Poisson distribution with mean \( \alpha \sim n(g-T)/G \times NT/G \).

- The variation of the number of reads can also be computed based on this model.

- For details see: Sharon et. al, RECOMB 2009.
Task 3: Prediction of novel transcripts: assembling the short reads

- Comparative assembly: using an annotated reference genome
  - Searching for junctions of annotated exons
  - Assembling splice variants
  - Finding paired-end (PED) links between exons

- De novo assembly: without using a reference genome
  - Splicing graph approach: Heber, et. al. ISMB 2002 for EST assembly
  - K-mer based approach, working efficiently for short reads (e.g. Velvet, ALLPATH and EULER-ESR)
  - Compact representation for alternative splicing variants
  - Can also be adapted for comparative assembly & assembly with PED reads
Alternative Splicing (AS)

35% - 60% of human genes show AS
- some genes have a huge number of isoforms
  \( (slo > 500, \text{neurexin} > 1000, DSCAM > 38000) \)
How to assemble multiple alternative spliced transcripts?

In the presence of AS, conventional assembly may be erroneous, ambiguous, or truncated.
Splice graph approach

Replace the problem of finding a list of consensus sequences with *Graph Reconstruction Problem*:
Given an set of expressed sequence, find a minimal graph (*splicing graph*) representing *all* transcripts as paths.

Heber, et. al. ISMB 2002
Splicing graphs

ADSL gene

transcripts

splicing graph

collapsed graph
Splicing graph construction

If a reference genome is used:
- Map reads to the reference genome (short read aligner)
- Check alignment (splice sites, quality)
- Connect consecutive positions
- Build splicing graph

![Splicing graph diagram](Image)
Splicing graph construction

If a reference genome sequence is not used:

- Break sequences into $k$-mers (20-mers).
- Build graph using $k$-mers as vertices, connect them iff they occur consecutively in a sequence [Pevzner et al., 2001].

Example (3-mers):

Sequences: CTCGATGAC, CTCGGAC
Vertices: $\{\text{CTC, TCG, CGA, GAT, ATG, TGA, GAC, CGG, GGA}\}$

splicing graph

collapsed splicing graph
An edge in the splicing graph, called a *block*, represents a maximal sequence of adjacent exons or exon fragments that always appear together in a given set of splicing variants. Therefore, variants can be represented by sequence of blocks, e.g. \{ABCD, C, AD\}.

Vertices \(s\) and \(f\) are included into graph, and are linked to the 5’ and 3’ of each variant, respectively. Each splicing variant corresponds to a directed path that goes from \(s\) to \(t\). But note that some paths in the splicing graph do not correspond to real variants, e.g. \{ABC, CD\}.

Lacroix, et. al. WABI, 2009
Quantifying splicing variants

(Transcriptome reconstruction problem) Given a set of candidate splicing variants $S = \{x_1, \ldots, x_k\}$, and a set of constraints $\{C_1, \ldots, C_k\}$, each indicating the total abundance of a subset of variants in $S$, assign the most possible abundance to each variant such at all constraints are satisfied?

In practice, a constraint $C_j$, will reflect the abundance of a block *junction* $j$, measured by mRNA-seq reads mapped to the two linked blocks.
Quantifying all splicing variants

For example, in the case of previous slide, constraints can be generated for junctions of blocks \{sA=A, AB, BC, CD, sC=Cf=C, Df=D, AD\}.

Specifically, for all potential variants in the graph \{ABCD, ABC, AD, C, CD\}, we have independent constraints of,

\[
\begin{align*}
ABC + C &= C_{Cf}; \\
C + CD &= C_{sC}; \\
ABCD + ABC &= C_{AB}; \\
AD &= C_{AD}
\end{align*}
\]
Quantifying *known* splicing variants

In we known the variants in advance, the same approach can be applied. For example, assuming three *known* variants in the graph are \{ABCD, AD, C\}, we have (much simpler) independent constraints of,

\[
C = C_{Cf}; \quad \text{ABCD} = C_{AB}; \\
AD = C_{AD}
\]
Integrating mapped PED reads

- Given a splicing graph and a set of observed PED reads mapped to block junctions, we first identify all paths (corresponding to potential splicing variants) containing *only* observed pairs of blocks.

- The resulting variants can be used for quantification as indicated in the previous slides.

Lacroix, et. al. WABI, 2009
The previous model assumes all observed abundance values in the constraints are *accurate*. In practice, however, it may not be true due to the noise in the mRNA-seq data. To handle this, a maximum likelihood (ML) model was introduced.

For a gene $g$, suppose it has $m$ exons with lengths $L = [l_1, l_2, \ldots, l_m]$ and $n$ variants with abundances $\Theta = [\theta_1, \theta_2, \ldots, \theta_n]$. Consider a set of constraints in which each constraint $k$ has an observed abundance $x_k$ reflecting the *approximate* total abundance of a subset of variants (denoted by $c_{ki}$, $c_{ki}=1$ if variant $i$ is included; $c_{ki}=0$ otherwise).
ML model

- Based on the previous notation, the expected abundance in constraint \( k \) is \( \lambda_k = \Sigma_i c_{ki} \theta_i \).
- The likelihood of observing an abundance \( \lambda_k \) for constraint \( k \) can be computed by Lander-Waterman’s Poisson model:
  \[
  L(\Theta \mid X) = \prod_k e^{-\lambda_k} \cdot \frac{\lambda_k^{x_k}}{x_k!}
  \]
- The MLE of the abundances of known variants can be computed by maximizing the likelihood function.
- The uncertainty of the MLE can be estimated using an important sampling (IP) on the joint distribution.

Jiang and Wong, Bioinformatics, 2009
Other applications of mRNA-seq: gene fusion

Following the alignment of the short m-RNA reads to a reference genome, most reads will fall within a single exon, and a smaller but still large set would be expected to map to known exon-exon junctions. The remaining unmapped short reads can then be further analyzed to determine whether they match an exon-exon junction where the exons come from different genes. An alternative approach is using pair-end reads, when potentially a large number of paired reads would map each end to a different exon, giving better coverage of these events. Nonetheless, the end result consists of multiple and potentially novel combinations genes providing an ideal starting point for further validation.

Acknowledgement: Wiki – mRNA-seq
New discoveries from mRNA-Seq

- Novel features of eukaryotic gene organization
  - Many yeast genes were found to overlap at their 3' ends; 808 pairs of ORFs overlap at their 3' ends (Science 320, 1344–1349, 2008)
  - Antisense expression (may be involved in gene regulation) is enriched in the 3' exons of mouse transcripts (Nature Methods 5, 613–619, 2008)

- Novel splicing forms
  - 31,618 known splicing events were confirmed (11% of all known splicing events) and 379 novel splicing events were discovered in human (Biotechniques 45, 81–94, 2008)

- Novel transcripts
Deeper insights into the complex transcriptomes

Deep sequencing of small RNAs (<200 nucleotides) from human HeLa and HepG2 cells

Individual non-coding RNAs dwarf the number of protein-coding genes, and include classes that are well understood as well as classes for which the nature, extent and functional roles are obscure

A proposed model for the metabolism of genic transcripts into a diversity of long and short RNAs.

Nature 457, 1028-1032, 2009
Another paper of deep sequencing of human transcriptome

- 50% of the reads mapped to unique genomic locations, of which 80% corresponded to known exons.
- 66% of the polyadenylated transcriptome mapped to known genes and 34% to nonannotated genomic regions.
- On the basis of known transcripts, RNA-Seq can detect 25% more genes than can microarrays.
- A global survey of messenger RNA splicing events identified 94,241 splice junctions (4096 of which were previously unidentified) and showed that exon skipping is the most prevalent form of alternative splicing.

Challenges: mapping reads to reference genome

- Sequencing errors and polymorphisms
- It is more difficult to map reads that span splice junctions (for complex transcriptomes), due to the presence of extensive alternative splicing and trans-splicing.
- Repetitive sequences: a significant portion of sequence reads match multiple locations in the genome
  - Obtaining longer sequence reads, or paired-end sequencing strategy, should help alleviate the multi-matching problem.
Challenges: novel splicing variant & quantification

- Discovery of novel splicing variants
  - Reconstruction of complete splice forms
  - Reliability: assignment of a P-value
- Quantifying the expression levels of recently duplicated genes
  - Equivalent to the problem of quantifying splicing variants, but harder because many isoforms need to be consider simultaneously
Readings


- Go with microarray or RNA-seq?
  - Advantages of next-generation sequencing versus the microarray in epigenetic research,
  - Briefings in Functional Genomics and Proteomics 2009 8(3):174-183

- Full-length transcriptome assembly from RNA-Seq data without a reference genome

- Next-generation transcriptome assembly
  - Nature Reviews Genetics 12, 671-682
  - This Review summarizes the recent developments in transcriptome assembly approaches — reference-based, de novo and combined strategies
What’s next

- Wednesday: indexing algorithms behind the mapping tools
- This week’s lab: BWA & Bowtie2
- Next week: group presentations

If you are interested:
  - Analyzing RNA-Seq from Heart and Liver Tissues using Galaxy
  - [http://biowhat.ucsd.edu/sdcsb/](http://biowhat.ucsd.edu/sdcsb/)