Estimating isoform-specific gene expression from RNA-Seq data

Hui Jiang
Stanford University
Outline

• Biological Motivation
  – Alternative Splicing

• Experimental and Technological Background
  – RNA-Seq
  – Ultra High Throughput Sequencing
  – Data Preprocessing - Sequence Mapping

• Statistical Modeling (Joint work with Wing Wong)
  – Joint Poisson Model
  – Solving Model using EM Algorithm
  – Bayesian Inference

• Further Study (Joint work with Julia Salzman and Wing Wong)
  – Paired-End RNA-Seq
  – Reduce Computation - Minimal Sufficiency
  – Model Comparison
Central dogma of molecular biology

Gene Expression

Genome

Gene

Transcription

pre-mRNA

Splicing

Alternative Splicing

mRNA

Translation

Protein

Nucleotide

Base Pair
92–94% of human genes undergo alternative splicing, 86% with a minor isoform frequency of 15% or more

Functional Consequences of alternative splicing (Isoforms)

- Protein Localization
- Protein Function
- mRNA Degradation (microRNAs)

**NMDAR1**  N-methyl-D-aspartate (NMDA) receptor R1 subunit

Exon 5 encodes the N1 Cassette in the extracellular domain and affects receptor sensitivity to agonists.

Exons 21 and 22a encode the C1 and C2 cassettes in the intracellular domain, which control phosphorylation by PKC, trafficking from the ER to the plasma membrane, and other functions.
Ezh2  Histone-lysine N-methyltransferase, an important B-cell gene

Splicing of Ezh2 is on a fine scale
Outline

• Biological Motivation
  – Alternative Splicing

• Experimental and Technological Background
  – RNA-Seq
  – Ultra High Throughput Sequencing
  – Data Preprocessing - Sequence Mapping

• Statistical Modeling (Joint work with Wing Wong)
  – Joint Poisson Model
  – Solving Model using EM Algorithm
  – Bayesian Inference

• Further Study (Joint work with Julia Salzman and Wing Wong)
  – Paired-End RNA-Seq
  – Reduce Computation - Minimal Sufficiency
  – Model Comparison
RNA-Seq

mRNA  

RNA fragments

or

cDNA

EST library with adaptors

ATCACAGTGGGACTCCATAAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAACAAAA
GGACAGAGTCCCCAGCGGGCTGAAGG
ATGAAACATTTAAAGTCAAAACAATATGAA

Short sequence reads

Sequencing

• Sequencing as a primary tool in molecular biology
  – Human genome project (1990-2003, $3B)
  – 1000 genomes project (2008-, $50M)

• A variety of sequencing techniques
  – Sanger sequencing (1975)
    • 10 million bases per machine per week (5-day week)
  – 454 sequencing
    • 1 million reads 400 bases/read per run
  – Illumina sequencing
    • 50 million reads 50 bp/read
  – ABI SOLiD systems
    • similar to Illumina

Next generation sequencing
Second generation sequencing
Massively parallel sequencing
Ultra high-throughput sequencing
Next-generation sequencing

What can next-generation sequencing do

Sequencing platforms

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing Chemistry</td>
<td>Pyrosequencing</td>
<td>Polymerase-based sequence-by-synthesis</td>
<td>Ligation-based sequencing</td>
</tr>
<tr>
<td>Amplification approach</td>
<td>Emulsion PCR</td>
<td>Bridge amplification</td>
<td>Emulsion PCR</td>
</tr>
<tr>
<td>Paired end separation</td>
<td>3 kb</td>
<td>200 bp</td>
<td>3 kb</td>
</tr>
<tr>
<td>Mb per run</td>
<td>100 Mb</td>
<td>1300 Mb</td>
<td>3000 Mb</td>
</tr>
<tr>
<td>Time per paired end run</td>
<td>7 hours</td>
<td>4 days</td>
<td>5 days</td>
</tr>
<tr>
<td>Read length (update)</td>
<td>250 bp (400)</td>
<td>35, 75 and 100 bp</td>
<td>35 and 50 bp</td>
</tr>
<tr>
<td>Cost per run</td>
<td>$ 8,438 USD</td>
<td>$ 8,950 USD</td>
<td>$ 17,447 USD</td>
</tr>
<tr>
<td>Cost per Mb</td>
<td>$ 84.39 USD</td>
<td>$ 5.97 USD</td>
<td>$ 5.81 USD</td>
</tr>
</tbody>
</table>

Table 1. Comparing metrics and performance of next-generation DNA sequencers [9]
How does the data look like?

AAAAATCTCTTCTGAAACCATTCTACAGAAAATGC
AAAAAAAGAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AACAGACCTAATCGCTCATTTGCTATATTCTT
AACCCGACCTGCGACTCCTTGACGTGAC
ATGTTAGGTTGTACGTTAGAATCTCTATTAT
ATTGCCAGAAAGTACTCTGAGCTATCTAGAT
ATCCCGATCCCGTTACAGGTCACAGTGAATTTA
ACCCACCAACAATGAATACTAATCAAACACCTC
ATGGGGGAATATTGCAATTATATGAAAGGTTA
ATGTTTAAAGCTCCTTTTTAAAACATATATTT
ATATATCTCTTTCCCTCTCAACTCTTCTCTTC
AGGGGACTACTCCACCCCTGAGCTCCCTCGTAG
AAAAGATATAATATATATATATATATATTACAT
AGTCCACCTCGACCTGGTCGCTCTGCTCGAGA
ATTTGGTGAGTATTAAAGAGAGTAGCAT
GGTCTGTTTGTGGTACGCCGCTCTTTCTTTTT
ATGGAAAGAAGCTTTCTCTATAATGTAAATA
AGGGACTGACGTCGTGGGGCCTATGGTTTATTAG
AGAAAAATTTAAAAATCTTTGAAAGAAGAAGAG
AAGGGGATTATAGGTTCTCGGGGCAAATT
AGACACCCATCAAACCTGGAGGTAGATATAG
AGATGTCGGTCCAGGATCTACAGGACCTT
ATTGGGTGTTGGCTATCCCACCCCGTTACAA
CGGGGATAAGTGTGGTTTCGAAAGAGATAAAAA
Sequence Mapping

- Find where the reads come from
- An essential step for sequencing analysis
- Efficiency is critical
  - Gigabytes of data can be generated in a single run
  - It takes months to map them if using tools like BLAST
What does sequence mapping do?

Find all the matches for a read in the genome

A DNA Sequence: ACATAGGATCATGAAGTACCCATATCTAGTG
reads: AGGA, CATC, ATAT, TTTG, GTGT

Matched Results: ACATAGGATCATGAAGTACCCATATCTAGTGGG
Perfect Match: AGGA
1bp Mismatch: CATC CATC CATG
## Sequence mapping softwares

<table>
<thead>
<tr>
<th>Software</th>
<th>Max. mismatches</th>
<th>Insdel</th>
<th>Max. read length</th>
<th>Report All matches</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELAND</td>
<td>2</td>
<td>N</td>
<td>32</td>
<td>N</td>
<td>A. Cox, Illumina</td>
</tr>
<tr>
<td>RMAP</td>
<td>4+</td>
<td>N</td>
<td>64</td>
<td>N</td>
<td>A. D. Smich, <em>et al</em> (Feb, 2008)</td>
</tr>
<tr>
<td>SOAP</td>
<td>2</td>
<td>1 gap</td>
<td>60</td>
<td>N</td>
<td>R. Li, <em>et al</em> (Mar, 2008)</td>
</tr>
<tr>
<td>Bowtie</td>
<td>3</td>
<td>N</td>
<td>200+</td>
<td>Y</td>
<td>B. Langmead, <em>et al</em> (Mar, 2009)</td>
</tr>
</tbody>
</table>
Outline

• Biological Motivation
  – Alternative Splicing

• Experimental and Technological Background
  – RNA-Seq
  – Ultra High Throughput Sequencing
  – Data Preprocessing - Sequence Mapping

• Statistical Modeling (Joint work with Wing Wong)
  – Joint Poisson Model
  – Solving Model using EM Algorithm
  – Bayesian Inference

• Further Study (Joint work with Julia Salzman and Wing Wong)
  – Paired-End RNA-Seq
  – Reduce Computation - Minimal Sufficiency
  – Model Comparison
How to quantify gene expression?

Experiment 1 (total 10M reads)

Chr1

Gene 1

Gene 2

Gene 3

Experiment 2 (total 20M reads)

Chr1

Gene 1

Gene 2

Gene 3
Gene level expression index

- More reads mapped to gene if transcript is long
- More reads mapped to gene if sequencing is deep
- Expression index
  
  Let \( l = \) size of transcript in kb

  \( N = \) total \# of mappable reads

  then the gene expression index is

  \[
  \text{RPKM} = \frac{\text{(# reads mapped to gene)}}{(l \times N)}
  \]

  “reads per kb per million reads”

  1 RKPM \( \sim \) 0.3 to 1 transcript per cell

Gene expression

Experiment 1 (total 10M reads)

Chr1  
10 RPKM 6 RPKM 4 RPKM

Experiment 2 (total 20M reads)

Chr1  
2.5 RPKM 0 RPKM 6 RPKM

RPKM: Reads Per Kilo base-pair per Million mapped reads

Isoform abundance estimation

Isoform 1
Isoform 2
Isoform 3

Length
$L_1$ $L_2$ $L_3$ $L_4$ $L_5$

#reads
$n_1$ $n_2$ $n_3$ $n_4$ $n_5$

abundance
$\theta_1$ $\theta_2$ $\theta_3$
Our approach

• Assume each read is sampled uniformly along the length of each transcript in the sample, and that longer transcripts are proportionally more likely to be sampled.

• Under this model, $n_1, n_2, ..$ are independent Poisson variables with parameters $\lambda_1, \lambda_2, ..$, where

$$\lambda_j = wL_j \left( \delta_{1j} \theta_1 + \delta_{2j} \theta_2 + \ldots + \delta_{ij} \theta_i \right)$$

$$\delta_{ij} = 1_{\{\text{isoform } i \text{ contains exon } j\}}$$

• Draw inference on $\theta_1, \theta_2, ..$ from the likelihood or the posterior distribution
Joint Poisson Model

• In general, assuming Poisson R.V.s, Let $a_{ij} = w L_j 1 \{ \text{isoform } i \text{ contains exon } j \}$, MLE gives us

$$
\hat{\Theta}_{ML} = \arg \max_{\Theta} \sum_j \left( n_j \log \sum_i a_{ij} \theta_i \right) - \sum_j \left( \sum_i a_{ij} \theta_i \right)
$$

• Constraint optimization: $\theta_1, \theta_2, \ldots$ are nonnegative

• Difficult to solve analytically
Concavity of log-likelihood

• Let $a_{ij} = wL_j 1\{\text{isoform } i \text{ contains exon } j\}$

$$f = \sum_j \left( n_j \log \sum_i a_{ij} \theta_i \right) - \sum_j \left( \sum_i a_{ij} \theta_i \right)$$

• Gradient

$$\frac{\partial f}{\partial \theta_k} = \sum_j a_{kj} \left( \frac{n_j}{\sum_i a_{ij} \theta_i} - 1 \right)$$

• second order derivatives

$$\frac{\partial^2 f}{\partial \theta_k \partial \theta_l} = -\sum_j a_{kj} a_{lj} \left( \frac{n_j}{\left(\sum_i a_{ij} \theta_i\right)^2} \right)$$
Concavity of log-likelihood

- Hessian matrix \[ Hf = \frac{\partial^2 f}{\partial \Theta^2} = -ADA^T \]

- Where \( A = \{a_{ij}\} \), \( D \) is a diagonal matrix, with \[ D_{jj} = n_j \left( \sum_i a_{ij} \theta_i \right)^2 \geq 0 \]

- Hessian is negative semidefinite, therefore \( f \) is concave.
- The constraints define a convex subset.
- It suffices to find local maximum.
Solving convex optimization

• Specialized convex optimization methods
  – Interior-point method
  – Cutting-plane method
  – Ellipsoid method
  – Subgradient method

• General optimization methods
  – Hill-climbing method
  – Steepest descent method
  – Newton–Raphson method
  – Conjugate gradient method

• We use the Expectation-Maximization algorithm
The EM algorithm

• Define unobserved data $n_{ij} =$ number of reads in exon $j$ generated by isoform $i$

• Log-likelihood function of the complete data

$$f = \sum_j \sum_i n_{ij} \log(a_{ij} \theta_i) - a_{ij} \theta_i$$
The EM algorithm (cont.)

- E-Step
  \[ E(n_{ij} \mid n_j, \Theta) = \frac{n_j \theta_i a_{ij}}{\sum_i \theta_i a_{ij}} \]

- M-Step
  \[ \Theta^{(n+1)} = \arg \max E[f(X \mid \Theta) \mid Y, \Theta^{(n)}] \]
  \[ \Rightarrow \theta_i^{(n+1)} = \frac{j n_{ij}}{\sum_j a_{ij}} \]

- The nonnegative constraints are automatically satisfied
Isoform expression (Wold data)

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Isoform 1</th>
<th>Isoform 2</th>
<th>Isoform 3</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain</td>
<td>5.05</td>
<td>0.42</td>
<td>0</td>
<td>5.47</td>
</tr>
<tr>
<td>Muscle</td>
<td>1.91</td>
<td>238.67</td>
<td>14.89</td>
<td>255.47</td>
</tr>
</tbody>
</table>

Statistical inference

• Multiple isoforms
  – Correlated expression
  – Asymptotics of the MLE
    $$\hat{\theta} \sim N(\theta, I(\theta)^{-1})$$
  – Fisher information matrix

$$I_{jk} = \text{Cov}\left(\frac{\partial}{\partial \theta_j} \log f(X; \theta), \frac{\partial}{\partial \theta_k} \log f(X; \theta)\right) = -E_{\theta}\left[\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X; \theta)\right]$$
Statistical inference

• **Difficulty**: When some isoform(s) are not or lowly expressed, MLE lies on or close to the boundary of the parameter space. Also, Fisher Information may become singular.

• **Our approach**: Use importance sampling to draw from the posterior, starting with a proposal density related to the asymptotic distribution.

• Summarize marginal inferences for single or pairs of isoform expressions.
Importance sampling:

Draw $x$ from $q(x)$. $q(\cdot)$ is a Gaussian density with mean=MLE, inverse covariance = a modified Fisher information matrix.

Compute importance weight $w(x) = p(x)/q(x)$ where $p(x) = \text{likelihood } (x) * \text{prior density } (x)$.

For any function $g(\cdot)$, estimate posterior mean of $g(X)$ by

$\sum x g(x)w(x) / \sum x w(x)$

To estimate posterior probabilities or histograms, take $g(\cdot)$ to be suitable Indicator functions.
95% probability interval (Wold data)

Figure 3.5: Visualization of RNA-Seq reads falling into mouse gene Dbi in brain tissue sample in CisGenome Browser.

<table>
<thead>
<tr>
<th>Isoform</th>
<th>Expression</th>
<th>95% probability interval</th>
<th>Relative 95% probability interval length</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>3.87</td>
<td>(2.22, 6.76)</td>
<td>117%</td>
</tr>
<tr>
<td>(2)</td>
<td>580.68</td>
<td>(559.52, 601.86)</td>
<td>7%</td>
</tr>
</tbody>
</table>

Gene level expression

- Gene expression is obtained by summing isoform expressions
  \[ \hat{g} = \sum_{i} \hat{t}_i \]

- Marginal posterior for \( g \) can be obtained from that of \( t \)

- In many cases we may have tight inference for the gene level expression but yet have great uncertainty about the expression for individual isoforms
More certain for gene than isoforms
(Wold data)

Figure 3.6: Visualization of RNA-Seq reads falling into mouse gene CLK1 in brain tissue sample in CisGenome Browser.

<table>
<thead>
<tr>
<th></th>
<th>Expression</th>
<th>95% probability interval</th>
<th>Relative 95% probability interval length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoform 1</td>
<td>0.48</td>
<td>(0.05, 3.01)</td>
<td>617%</td>
</tr>
<tr>
<td>Isoform 2</td>
<td>6.60</td>
<td>(4.20, 7.28)</td>
<td>47%</td>
</tr>
<tr>
<td>Gene</td>
<td>7.09</td>
<td>(6.52, 7.84)</td>
<td>19%</td>
</tr>
</tbody>
</table>

Marginal inference

Another example

<table>
<thead>
<tr>
<th>Isoform</th>
<th>Expression</th>
<th>95% interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoform 1 (upper)</td>
<td>2.1</td>
<td>(0.05, 7.76)</td>
</tr>
<tr>
<td>Isoform 2 (middle)</td>
<td>35</td>
<td>(4, 71)</td>
</tr>
<tr>
<td>Isoform 3 (lower)</td>
<td>350</td>
<td>(316, 379)</td>
</tr>
<tr>
<td>Gene level</td>
<td>387</td>
<td>(371, 402)</td>
</tr>
</tbody>
</table>
Another example
Outline

• Biological Motivation
  – Alternative Splicing

• Experimental and Technological Background
  – RNA-Seq
  – Ultra High Throughput Sequencing
  – Data Preprocessing - Sequence Mapping

• Statistical Modeling (Joint work with Wing Wong)
  – Joint Poisson Model
  – Solving Model using EM Algorithm
  – Bayesian Inference

• Further Study (Joint work with Julia Salzaman and Wing Wong)
  – Paired-End RNA-Seq
  – Reduce Computation - Minimal Sufficiency
  – Model Comparison
Paired-end RNA-Seq
Paired-end RNA-Seq provides more information for isoform expression
The model

\[ f_\theta(n_1, n_2, \ldots, n_J) = \prod_{j=1}^{J} \frac{(\theta \cdot a_j)^{n_j} e^{-\theta \cdot a_j}}{n_j!}, \quad \text{a curved exponential family} \]

\[ \begin{align*}
\theta_i & \quad \text{The abundance of transcript type } i, \; i = 1, \ldots, I. \\
\theta & \quad \text{The isoform abundance vector } [\theta_1, \theta_2, \ldots, \theta_I]. \\
n_{i,j} & \quad \text{The number of reads } s_j \text{ that are generated from transcripts } i. \\
n_j & \quad \text{The number of read } s_j \text{ that are generated from all the transcripts.} \\
a_{i,j} & \quad \text{Up to proportionality, the sampling rate of } n_{i,j}, \text{ i.e., the rate that read } s_j \text{ is generated from each individual transcript } i. \\
\theta \cdot a_j & \quad \text{The sampling rate of } n_j, \text{ i.e., the rate that read } s_j \text{ is generated from all the transcripts.} \\
a_j & \quad \text{The sampling rate vector } [a_{1,j}, a_{2,j}, \ldots, a_{I,j}] \text{ for read } s_j. 
\end{align*} \]
Sampling rates

<table>
<thead>
<tr>
<th></th>
<th>Read $s_1$</th>
<th>Read $s_2$</th>
<th>...</th>
<th>Read $s_J$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcript 1</td>
<td>$a_{1,1}$</td>
<td>$a_{1,2}$</td>
<td>...</td>
<td>$a_{1,J}$</td>
</tr>
<tr>
<td>Transcript 2</td>
<td>$a_{2,1}$</td>
<td>$a_{2,2}$</td>
<td>...</td>
<td>$a_{2,J}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Transcript $I$</td>
<td>$a_{I,1}$</td>
<td>$a_{I,2}$</td>
<td>...</td>
<td>$a_{I,J}$</td>
</tr>
</tbody>
</table>

\[
\frac{a_{i_1,j_1}}{a_{i_2,j_2}} = \frac{\Pr(\text{read } s_{j_1} \text{ observed after process one copy of transcript } i_1)}{\Pr(\text{read } s_{j_2} \text{ observed after process one copy of transcript } i_2)}
\]

If transcript $i$ cannot generate read $s_j$ and so $a_{i,j} = 0$. 
Model choice and analysis: $a_{i,j}$

- Uniform sampling model
  
  $a_{i,j} = w$, if transcript $i$ can generate read $j$
  
  $a_{i,j} = 0$, otherwise

- Insert length model
  
  $a_{i,j} = q(l_{i,j})w$, if transcript $i$ can generate read $j$
  
  $a_{i,j} = 0$, otherwise

$w$ is the total number of mapped reads, serve as a normalization factor, making our model compatible to the RPKM model.
\( q(l) \)
Uniform assumption
Computational issues

• Equivalent optimization problem to the MLE
  \[ \max_{\theta \in \mathbb{R}^i, +} \ n^T \log(A^T \theta) - \| A^T \theta \|_1 \]

• Good news: convex optimization
• Bad news: \( A \) is \( I \times J \), \( J > 2000 \) for uniform model, \( J > O(10^5) \) for insert-length model
• Solution: Minimal Sufficient Statistics (MSS)
Interpretation of MSS as Collapsing Reads gives interpretation
\[ n_1 \sim \text{Po}(a_{1,1} \theta_1 + a_{2,1} \theta_2) \quad \text{and} \quad n_2 \sim \text{Po}(a_{1,2} \theta_1 + a_{2,2} \theta_2) \]
\[ \Rightarrow n_1 + n_2 \sim \text{Po}((a_{1,1} + a_{1,2}) \theta_1 + (a_{2,1} + a_{2,2}) \theta_2) \]

\( n_1 + n_2 \) is a sufficient statistic for \( \theta \) iff \( \frac{a_{1,1}}{a_{2,1}} = \frac{a_{1,2}}{a_{2,2}} \), and is said to be a "collapsing"
Collapse the reads

<table>
<thead>
<tr>
<th></th>
<th>Read $s_1$</th>
<th>Read $s_2$</th>
<th>...</th>
<th>Read $s_J$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcript 1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Transcript 2</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Transcript $I$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

Collapse these reads
Form a “category” of reads
Maximal collapsing

Maximal Collapsing Statistics: $n^{(1)}, n^{(2)}, \ldots, n^{(C)}$

- Read counts $j, k$ collapsed (i.e. summed) if their relative sampling rates from each isoform are proportional, i.e.

$$\frac{\langle a_j, a_k \rangle}{\|a_j\|_2 \|a_k\|_2} = 1.$$

producing counts $n^{(1)}, n^{(2)}, \ldots, n^{(C)}$.

- If counts $n^{(1)}, n^{(2)}, \ldots, n^{(C)}$ cannot be collapsed further, they are a maximal collapsing
Statistical properties of maximal collapsing

• Maximal collapsing is unique.

• Any collapsing is a sufficient statistic.

• Maximal collapsing is a minimal sufficient statistic.
Computational Properties of Maximal Collapsing

For Insert Length Model:

• Computational time before collapsing is hours per gene: over 10,000 categories.
• Computational time to perform the collapsing is $J \log J$.
• Collapsed problem can be solved in seconds per gene: several hundred categories.
A toy example

- **Before collapsing**
  \[
  A = \begin{pmatrix}
  1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
  1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\
  \end{pmatrix}
  \quad \Rightarrow \theta = [0.5, 0.25]
  \]
  \[
  n = [1, 1, 0, 1, 1, 0, 1, 0, 1, 1]
  \]

- **After collapsing**
  \[
  A = \begin{pmatrix}
  4 & 2 & 4 \\
  4 & 0 & 4 \\
  \end{pmatrix}
  \quad \Rightarrow \theta = [0.5, 0.25]
  \]
  \[
  n = [3, 1, 3]
  \]
A real example
A real example (cont.)

- Unpaired reads

<table>
<thead>
<tr>
<th>Category ID</th>
<th>Sampling rate vector</th>
<th>Read count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[4242n, 4242n]</td>
<td>216</td>
</tr>
<tr>
<td>2</td>
<td>[296n, 0]</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>[0, 62n]</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ \text{MLE} = [15.47, 2.70], \ \text{95\% intervals (Bayesian)}: (7.89, 18.81) \text{ and } (0.25, 10.83) \]

- Paired reads

<table>
<thead>
<tr>
<th>Category ID</th>
<th>Sampling rate vector</th>
<th>Read count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[1681.82n, 1681.82n]</td>
<td>95</td>
</tr>
<tr>
<td>2</td>
<td>[294.60n, 0]</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>[0, 245.80n]</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>138</td>
<td>[0.0057n, 0.0018n]</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ \text{MLE} = [16.73, 3.43], \ \text{95\% intervals (Bayesian)}: (11.22, 21.02) \text{ and } (1.03, 9.29) \]
Biological validation

Insert Length Model: $[\hat{\theta}_1, \hat{\theta}_2] = [15.47, 2.70]$
Uniform Sampling Length Model: $[\hat{\theta}_1, \hat{\theta}_2] = [16.73, 3.43]$

Ladder: 100 (5), 200(10), 400(20), 800(40), 1200(60), 2000(100)
Predicted lengths: 263 and 379
Another example

Single reads: \( \hat{\theta}_1 = 12.48, \hat{\theta}_2 = 7.35 \)

Insert length model: \( \hat{\theta}_1 = 22.16, \hat{\theta}_2 = 0 \)
Genome-wide model comparison
Some statistics (genes with two isoforms, expression between 5 and 100)
Fisher Information Larger in Insert Length vs. Uniform Sampling Models

- Single reads of 30bp vs. paired reads 30bp with uniform insert length $200 \pm 20$bp
- Fisher information: $I_S / I_P = 0.27$
Simulation study

- Exons of lengths 500bp, 50bp and 500bp
- Reads distributed according to Gaussian and uniform distributions

Relative $L^2$ loss:

$$\frac{||\theta - \hat{\theta}||_2}{||\theta||_2} = \sqrt{(\theta_1 - \hat{\theta}_1)^2 + (\theta_2 - \hat{\theta}_2)^2} = \sqrt{\frac{(\frac{1}{2} - \hat{\theta}_1)^2 + (\frac{1}{2} - \hat{\theta}_2)^2}{\sqrt{2}/2}}$$
Simulation Results

Figure: y axis: Relative Error vs. x axis: Read Number. (a) paired 30bp with Gaussian insert size; (b) 30 bp with uniform insert size.
Summary

• RNA-Seq for accurate measurement on transcriptomes
• Simple statistical model for isoform-specific gene expression estimation
• EM algorithm for solving the corresponding convex optimization problem
• Bayesian approach for statistical inference
• Minimal sufficiency for reducing computation
• Paired-end RNA-Seq for improving estimation accuracy