Measuring transcriptomes with RNA-Seq

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2017
Anthony Gitter
gitter@biostat.wisc.edu

These slides, excluding third-party material, are licensed under CC BY-NC 4.0 by Colin Dewey, Mark Craven, and Anthony Gitter
Overview

- RNA-Seq technology

- The RNA-Seq quantification problem

- Generative probabilistic models and Expectation-Maximization for the quantification task
Goals for lecture

• What is RNA-Seq?

• How is RNA-Seq used to measure the abundances of RNAs within cells?

• What probabilistic models and algorithms are used for analyzing RNA-Seq?
Measuring transcription the old way: microarrays

- Each spot has “probes” for a certain gene

- Probe: a DNA sequence complementary to a certain gene

- Relies on complementary hybridization

- Intensity/color of light from each spot is measurement of the number of transcripts for a certain gene in a sample

- Requires knowledge of gene sequences
Advantages of RNA-Seq over microarrays

- No reference sequence needed
  - With microarrays, limited to the probes on the chip
- Low background noise
- Large dynamic range
  - $10^5$ compared to $10^2$ for microarrays
- High technical reproducibility
- Identify novel transcripts and splicing events
RNA-Seq technology

- Leverages rapidly advancing sequencing technology (e.g., Illumina)

- Transcriptome analog to whole genome shotgun sequencing

- Two key differences from genome sequencing:
  1. Transcripts sequenced at different levels of coverage - expression levels
  2. Sequences already known (in many cases) - coverage is measurement
A generic RNA-Seq protocol

Sample RNA → RNA fragments → reverse transcription + amplification → cDNA fragments → sequencing machine → reads

- CCTTCNCACTTCGTTTCCCAC
- TTTTNCAAGAGTTTTTTCTTG
- GAACANTCCAACGCTTGGTGA
- GGAAANAAGACCCTGTTGAGC
- CCCGGNGATCCGCTGGGACAA
- GCAGCATATTGATAGATAACT
- CATCTAGCATCGCGTTGCGTT
- CCCGCGCGCTTAGGCTACTCG
- TCACACATCTCTAGCTAGCAT
- CATGCTAGCTATGCCTATCTA
- CCTTCNCACTTCGTTTCCCAC
- TTTTNCAAGAGTTTTTTCTTG
- GAACANTCCAACGCTTGGTGA
- GGAAANAAGACCCTGTTGAGC
- CCCGGNGATCCGCTGGGACAA
- GCAGCATATTGATAGATAACT
- CATCTAGCATCGCGTTGCGTT
- CCCGCGCGCTTAGGCTACTCG
- TCACACATCTCTAGCTAGCAT
- CATGCTAGCTATGCCTATCTA
RNA-Seq data: FASTQ format

name          sequence          read
----------------------          ------------------------
paired-end reads

read1
down

1 Illumina HiSeq
2500 lane

~150 million reads
Tasks with RNA-Seq data

• Assembly:
  • Given: RNA-Seq reads (and possibly a genome sequence)
  • Do: Reconstruct full-length transcript sequences from the reads

• Quantification (our focus):
  • Given: RNA-Seq reads and transcript sequences
  • Do: Estimate the relative abundances of transcripts ("gene expression")

• Differential expression:
  • Given: RNA-Seq reads from two different samples and transcript sequences
  • Do: Predict which transcripts have different abundances between two samples
RNA-Seq is a *relative* abundance measurement technology

- RNA-Seq gives you reads from the ends of a random **sample** of fragments in your library

- Without additional data this only gives information about **relative** abundances

- Additional information, such as levels of “spike-in” transcripts, are needed for absolute measurements
### Issues with relative abundance measures

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample 1 absolute abundance</th>
<th>Sample 1 relative abundance</th>
<th>Sample 2 absolute abundance</th>
<th>Sample 2 relative abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>10%</td>
<td>20</td>
<td>5%</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>10%</td>
<td>20</td>
<td>5%</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>10%</td>
<td>20</td>
<td>5%</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>10%</td>
<td>20</td>
<td>5%</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>10%</td>
<td>20</td>
<td>5%</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>50%</td>
<td>300</td>
<td>75%</td>
</tr>
</tbody>
</table>

- Changes in absolute expression of high expressors is a major factor
- Normalization is required for comparing samples in these situations
The basics of quantification with RNA-Seq data

• For simplicity, suppose reads are of length one (typically they are > 35 bases)

transcripts          reads
1. 200               100 A
2. 60                60 C
3. 80                40 G

• What relative abundances would you estimate for these genes?
**Length dependence**

- Probability of a read coming from a transcript \( \propto \) relative abundance \( \times \) length

\[
\hat{f}_1 \propto \frac{100}{200} = \frac{1}{400}
\]

transcript 1 relative abundance

\[
\text{probability of read from transcript 1} = \frac{\text{transcript 1 reads}}{\text{total reads}}
\]

<table>
<thead>
<tr>
<th>transcripts</th>
<th>reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100 A</td>
</tr>
<tr>
<td>2</td>
<td>60 C</td>
</tr>
<tr>
<td>3</td>
<td>40 G</td>
</tr>
</tbody>
</table>
Length dependence

- Probability of a read coming from a transcript $\propto$ relative abundance $\times$ length

<table>
<thead>
<tr>
<th>transcripts</th>
<th>reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 200</td>
<td>100 A</td>
</tr>
<tr>
<td>2 60</td>
<td>60 C</td>
</tr>
<tr>
<td>3 80</td>
<td>40 G</td>
</tr>
</tbody>
</table>

\[
\hat{f}_1 \propto \frac{100}{200} = \frac{1}{400} \\
\hat{f}_2 \propto \frac{60}{200} = \frac{1}{200} \\
\hat{f}_3 \propto \frac{40}{200} = \frac{1}{400}
\]

\[
\hat{f}_1 = 0.25 \\
\hat{f}_2 = 0.5 \\
\hat{f}_3 = 0.25
\]
The basics of quantification from RNA-Seq data

• Basic assumption:

$$\theta_i = P(\text{read from transcript } i) = Z^{-1} \tau_i \ell'_i$$

expression level (relative abundance)

length

• Normalization factor is the mean length of expressed transcripts

$$Z = \sum_i \tau_i \ell'_i$$
The basics of quantification from RNA-Seq data

- Estimate the probability of reads being generated from a given transcript by counting the number of reads that align to that transcript

\[ \hat{\theta}_i = \frac{c_i}{N} \]

# reads mapping to transcript i

total # of mappable reads

- Convert to expression levels by normalizing by transcript length

\[ \hat{T}_i \propto \frac{\hat{\theta}_i}{\ell_i'} \]
The basics of quantification from RNA-Seq data

- Basic quantification algorithm
  - Align reads against a set of reference transcript sequences
  - Count the number of reads aligning to each transcript
  - Convert read counts into relative expression levels
Counts to expression levels

- RPKM - Reads Per Kilobase per Million mapped reads

\[
\text{RPKM for gene } i = 10^9 \times \frac{c_i}{\ell_i' N}
\]

- TPM - Transcripts Per Million

\[
(\text{estimate of}) \quad \text{TPM for isoform } i = 10^6 \times Z \times \frac{c_i}{\ell_i' N}
\]

- Prefer TPM to RPKM because of normalization factor

- TPM is a technology-independent measure (simply a fraction)
What if reads do not uniquely map to transcripts?

• The approach described assumes that every read can be uniquely aligned to a single transcript

• This is generally not the case
  
  • Some genes have similar sequences - gene families, repetitive sequences
  
  • Alternative splice forms of a gene share a significant fraction of sequence
Central dogma of molecular biology

Double-stranded genomic DNA template

1. Transcription initiation
2. Translation initiation codon (ATG)
3. Translation terminator codons (TAA, TAG, TGA)
4. Poly (A) addition sequence (AATAAA)

Promoter region (RNA pol II binding)

5’ UTR

Exon 1 → Intron 1 → Exon 2

Single-stranded pre-mRNA (nuclear RNA)

1. Transcription and polyadenylation
2. Branch site
3. 3’SS
4. 5’SS
5. ESE
6. ESS
7. ISE
8. ISS

M^7GpppAC (CAP)

DA

A

AAA (n) A OH

Mature mRNA

RNA processing

1. Export to cytoplasm and translation
2. Folding, posttranslational modification, subcellular localization, etc.

Protein (amino acid sequence)

H2N → COOH

Griffith et al. PLoS Computational Biology 2015
Alternative splicing

pre-mRNA

splicing

alternatively spliced mRNAs

translation

protein isoforms
## Multi-mapping reads in RNA-Seq

<table>
<thead>
<tr>
<th>Species</th>
<th>Read length</th>
<th>% multi-mapping reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse</td>
<td>25</td>
<td>17%</td>
</tr>
<tr>
<td>Mouse</td>
<td>75</td>
<td>10%</td>
</tr>
<tr>
<td>Maize</td>
<td>25</td>
<td>52%</td>
</tr>
<tr>
<td>Axolotl</td>
<td>76</td>
<td>23%</td>
</tr>
</tbody>
</table>

- Throwing away multi-mapping reads leads to
  1. Loss of information
  2. Potentially biased estimates of abundance
Distributions of alignment counts
What if reads do not uniquely map to transcripts?

- Multiread: a read that could have been derived from multiple transcripts

<table>
<thead>
<tr>
<th>transcripts</th>
<th>reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
</tr>
</tbody>
</table>

- How would you estimate the relative abundances for these transcripts?
Some options for handling multireads

• Discard multireads, estimate based on uniquely mapping reads only

• Discard multireads, but use “unique length” of each transcript in calculations

• “Rescue” multireads by allocating (fractions of) them to the transcripts

• Three step algorithm

  1. Estimate abundances based on uniquely mapping reads only

  2. For each multiread, divide it between the transcripts to which it maps, proportionally to their abundances estimated in the first step

  3. Recompute abundances based on updated counts for each transcript
Rescue method example - Step 1

transcripts

reads

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>

\[
\hat{f}_{\text{unique}} = \frac{90}{200} + \frac{40}{60} + \frac{40}{80} = 0.278
\]

\[
\hat{f}_{\text{unique}} = 0.412
\]

\[
\hat{f}_{\text{unique}} = 0.309
\]
Rescue method example - Step 2

<table>
<thead>
<tr>
<th>transcripts</th>
<th>reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 200</td>
<td>90 A</td>
</tr>
<tr>
<td>2 60</td>
<td>40 C</td>
</tr>
<tr>
<td>3 80</td>
<td>40 G</td>
</tr>
<tr>
<td></td>
<td>30 T</td>
</tr>
</tbody>
</table>

Step 2

\[
c_{1}^{\text{rescue}} = 90 + 30 \times \frac{0.278}{0.278 + 0.412} = 102.1
\]

\[
c_{2}^{\text{rescue}} = 40 + 30 \times \frac{0.412}{0.278 + 0.412} = 57.9
\]

\[
c_{3}^{\text{rescue}} = 40 + 0 = 40
\]
Rescue method example - Step 3

transcripts

\[ \begin{align*}
\text{Step 3} \\
\hat{f}_{\text{rescue}}^1 &= \frac{102.1}{200} + \frac{57.9}{60} + \frac{40}{80} = 0.258 \\
\hat{f}_{\text{rescue}}^2 &= \frac{57.9}{60} + \frac{40}{80} = 0.488 \\
\hat{f}_{\text{rescue}}^3 &= \frac{40}{80} = 0.253
\end{align*} \]

reads

90 A
40 C
40 G
30 T
An observation about the rescue method

• Note that at the end of the rescue algorithm, we have an updated set of abundance estimates

• These new estimates could be used to reallocate the multireads

• And then we could update our abundance estimates once again

• And repeat!

• This is the intuition behind the statistical approach to this problem
RSEM (RNA-\textit{Seq} by \textit{Expectation-Maximization}) - a generative probabilistic model

- Simplified view of the model
  - Grey – observed variable
  - White – latent (unobserved) variables

![Diagram of RSEM model]

- Number of reads
- Start position
- Transcript probabilities (expression levels)
- Read sequence
- Orientation
- Transcript
- \( \theta \)
RSEM - a generative probabilistic model

transcript probabilities (expression levels) → \( \theta \)

number of reads → \( N \)

transcript → \( G_n \)

fragment length → \( F_n \)

start position → \( S_n \)

read length → \( L^1_n \)

orientation → \( O_n \)

quality scores → \( Q^1_n \)

read sequence → \( R^1_n \)

paired read

\[
P(g, f, s, o, \ell, q, r|\theta) = \prod_{n=1}^{N} \left[ P(g_n|\theta)P(f_n|g_n)P(s_n|f_n, g_n)P(o_n|g_n)P(q_n)P(\ell_n|f_n)P(r_n|g_n, f_n, s_n, o_n, \ell_n, q_n) \right]
\]
Quantification as maximum likelihood inference

- Observed data likelihood

\[ P(r, \ell, q | \theta) = \prod_{n=1}^{N} \sum_{i=0}^{M} \sum_{j=0}^{L_i} \sum_{k=0}^{L_i} \sum_{\sigma=0}^{1} P(R_n = r_n, L_n = \ell_n, Q_n = q_n, S_n = j, F_n = k, O_n = \sigma | G_n = i) \]

- Likelihood function is concave with respect to \( \theta \)
  - Has a global maximum (or global maxima)

- Expectation-Maximization for optimization

“RNA-Seq gene expression estimation with read mapping uncertainty”
Li, B., Ruotti, V., Stewart, R., Thomson, J., Dewey, C.
Bioinformatics, 2010
Approximate inference with read alignments

\[
P(r, \ell, q|\theta) = \prod_{n=1}^{N} \sum_{i=0}^{M} \sum_{j=0}^{L_i} \sum_{k=0}^{L_i} \sum_{o=0}^{1} P(R_n = r_n, L_n = \ell_n, Q_n = q_n, S_n = j, F_n = k, O_n = o | G_n = i)
\]

- Full likelihood computation requires \(O(NML^2)\) time

  - \(N\) (number of reads) \(\sim 10^7\)
  
  - \(M\) (number of transcripts) \(\sim 10^4\)
  
  - \(L\) (average transcript length) \(\sim 10^3\)

- Approximate by alignment

\[
P(r, \ell, q|\theta) = \prod_{n=1}^{N} \sum_{(i,j,k,o) \in \pi_n^x} \theta_i P(R_n = r_n, L_n = \ell_n, Q_n = q_n, Z_{nijk0} = 1 | G_n = i)
\]

  all local alignments of read \(n\) with at most \(x\) mismatches
EM Algorithm

• Expectation-Maximization for RNA-Seq
  
  • E-step: Compute expected read counts given current expression levels
  
  • M-step: Compute expression values maximizing likelihood given expected read counts
  
  • Rescue algorithm ≈ 1 iteration of EM
HMM Interpretation

hidden: read start positions
observed: read sequences

Learning parameters: Baum-Welch Algorithm (EM for HMMs)
Approximation: Only consider a subset of paths for each read
Probabilistically-weighted alignments
Expected read count visualization
Improved accuracy over unique and rescue

Gene-level expression estimation
Improving accuracy on repetitive genomes: maize

Gene-level expression estimation
RNA-Seq and RSEM summary

• **RNA-Seq** is the preferred technology for transcriptome analysis in most settings

• The major challenge in analyzing RNA-Seq data: the reads are much shorter than the transcripts from which they are derived

• Tasks with RNA-Seq data thus require handling **hidden** information: which gene/isoform gave rise to a given read

• The **Expectation-Maximization** algorithm is extremely powerful in these situations
Recent developments in RNA-Seq

• Long read sequences: PacBio and Oxford Nanopore

• Single-cell RNA-Seq: review
  • Observe heterogeneity of cell populations
  • Model technical artifacts (e.g. artificial 0 counts)
  • Detect sub-populations
  • Predict pseudotime through dynamic processes
  • Detect gene-gene relationships

• Alignment-free quantification:
  • Kallisto
  • Salmon
Public sources of RNA-Seq data

  - Both microarray and sequencing data
  - All sequencing data (not necessarily RNA-Seq)
- ArrayExpress: [https://www.ebi.ac.uk/arrayexpress/](https://www.ebi.ac.uk/arrayexpress/)
  - European version of GEO
- All of these have links between them