Bioinformatics Pipelines for RNA-Seq Data Analysis

Ion Măndoiu and Sahar Al Seesi

Computer Science & Engineering Department
University of Connecticut

BIBM 2011 Tutorial
Outline

• Background
• RNA-Seq read mapping
• Variant detection and genotyping from RNA-Seq reads
• Transcriptome quantification using RNA-Seq
• Implementing RNA-Seq analysis pipelines using Galaxy
• Novel transcript reconstruction
• Conclusions
Outline

• Background
  – NGS technologies

• RNA-Seq read mapping

•Variant detection and genotyping from RNA-Seq reads

• Transcriptome quantification using RNA-Seq

• Implementing RNA-Seq analysis pipelines using Galaxy

• Novel transcript reconstruction

• Conclusions
Cost of DNA Sequencing

Cost of genome sequencing compared with Moore’s law for computers

Cost of computing (Moore’s law)

$ per million DNA bases

Log scale

100,000
10,000
1,000
100
10
1.0
0.1

1999 2002 04 06 08 10

http://www.economist.com/node/16349358
2\textsuperscript{nd} Gen. Sequencing: Illumina

1. PREPARE GENOMIC DNA SAMPLE
   - Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE
   - Bind single-stranded fragments randomly to the interior surface of the flow cell channels.

3. BRIDGE AMPLIFICATION
   - Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE-STRANDED
   - The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES
   - Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION
   - Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.
2nd Gen. Sequencing: Illumina

**7. DETERMINE FIRST BASE**

The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

**8. IMAGE FIRST BASE**

After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

**9. DETERMINE SECOND BASE**

The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

**10. IMAGE SECOND CHEMISTRY CYCLE**

After laser excitation, the image is captured as before, and the identity of the second base is recorded.

**11. SEQUENCING OVER MULTIPLE CHEMISTRY CYCLES**

The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.
2nd Gen. Sequencing: SOLiD

- Emulsion PCR used to perform single molecule amplification of pooled library onto magnetic beads
2\textsuperscript{nd} Gen. Sequencing: SOLiD

\textbf{SOLiD}\textsuperscript{TM} substrate

Glass slide

Di base probes

Template

2nd base

1st base

Cleavage site
2nd Gen. Sequencing: SOLiD

1. Prime and ligate

Primer round 1 + Ligase

Universal seq primer (n)

Template sequence

2. Image

Excite Fluorescence

3. Cap unextended strands

Phosphatase

4. Cleave off fluor

Cleavage agent

5. Repeat steps 1–4 to extend sequence

Ligation cycle 1 2 3 4 5 6 7 ... (n cycles)

6. Primer reset

Universal seq primer (n-1)

1. Melt off extended sequence

2. Primer reset

7. Repeat steps 1–5 with new primer

Primer round 2

1 base shift

Universal seq primer (n-1)
8. Repeat Reset with \( n-2 \), \( n-3 \), \( n-4 \) primers

| Primer round | Read position | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
| 1            | Universal seq primer (n) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 2            | Universal seq primer (n-1) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 3            | Universal seq primer (n-2) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4            | Universal seq primer (n-3) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 5            | Universal seq primer (n-4) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

- Indicates positions of interrogation

Ligation cycle: 1, 2, 3, 4, 5, 6, 7
2nd Gen. Sequencing: SOLiD

Possible dinucleotides encoded by each color

1st base
A
C
G
T

2nd base
A
C
G
T

Template sequence
AT
AC
AA
GA
CG
CA
CC
TC
GC
GT
GG
AG
TA
TG
TT
CT

Double interrogation

With 2 base encoding each base is defined twice

Decoding

Color space sequence
TA
AC
AA
GA
GC
CA
CC
TC
CG
GT
GG
AG
AT
TG
TT
CT

Possible dinucleotides

AT
TG
GG
GA

Decoded sequence

Base space sequence
A
T
G
G
A
2nd Gen. Sequencing: 454

**Roche (454) GSFLX Workflow:**

**Library construction**

**Emulsion PCR**

**PTP loading**

**Pyrosequencing reaction**

*Figure 1. 454 Workflow: library construction ligates 454-specific adapters to DNA fragments and couples amplification beads with DNA in an emulsion PCR to amplify fragments before sequencing. The beads are loaded into the picoliter plate (PTP). The bottom panel illustrates the pyrosequencing reaction that occurs on nucleotide incorporation to report sequencing by synthesis.*
• High-density array of micro-machined wells
• Each well holds a different clonally amplified DNA template generated by emulsion PCR
• Beneath the wells is an ion-sensitive layer and beneath that a proprietary Ion sensor
• The sequencer sequentially floods the chip with one nucleotide after another (natural nucleotides)
• If currently flooded nucleotide complements next base on template, a voltage change is recorded
3\textsuperscript{rd} Gen. Sequencing

PacBio SMRT

Nanopore sequencing
## Cost/Performance Comparison

[Glenn 2011]

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Run time$^a$</th>
<th>Millions of reads/ run</th>
<th>Bases/read$^b$</th>
<th>Yield Mb/run</th>
<th>Reagent cost/ run$^c$</th>
<th>Reagent cost/Mb</th>
<th>Minimum unit cost (% run)$^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3730xl (capillary)</td>
<td>2 h</td>
<td>0.000096</td>
<td>650</td>
<td>0.06</td>
<td>$96</td>
<td>$1500</td>
<td>$6 (1%)</td>
</tr>
<tr>
<td>Ion Torrent – ‘314’chip</td>
<td>2 h</td>
<td>0.10</td>
<td>100</td>
<td>&gt;10</td>
<td>$500</td>
<td>&lt;$50</td>
<td>~$750 (100%)</td>
</tr>
<tr>
<td>454 GS Jr. Titanium</td>
<td>10 h</td>
<td>0.10</td>
<td>400</td>
<td>50</td>
<td>$1100</td>
<td>$22</td>
<td>$1500 (100%)</td>
</tr>
<tr>
<td>Starlight*</td>
<td>†</td>
<td>~0.01</td>
<td>&gt;1000</td>
<td>†</td>
<td>†</td>
<td>†</td>
<td>†</td>
</tr>
<tr>
<td>PacBio RS</td>
<td>0.5–2 h</td>
<td>0.01</td>
<td>860–1100</td>
<td>5–10</td>
<td>$110–900</td>
<td>$11–180</td>
<td>†</td>
</tr>
<tr>
<td>454 FLX Titanium</td>
<td>10 h</td>
<td>1</td>
<td>400</td>
<td>500</td>
<td>$6200</td>
<td>$12.4</td>
<td>$2000 (10%)</td>
</tr>
<tr>
<td>454 FLX+e</td>
<td>18–20 h</td>
<td>1</td>
<td>700</td>
<td>900</td>
<td>$6200</td>
<td>$7</td>
<td>$2000 (10%)</td>
</tr>
<tr>
<td>Ion Torrent – ‘316’chip*</td>
<td>2 h</td>
<td>1</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td>$750</td>
<td>&lt;$7.5</td>
<td>~$1000 (100%)</td>
</tr>
<tr>
<td>Helicos$^f$</td>
<td>N/A</td>
<td>800</td>
<td>35</td>
<td>28 000</td>
<td>N/A</td>
<td>NA</td>
<td>$1100 (2%)</td>
</tr>
<tr>
<td>Ion Torrent – ‘318’chip*</td>
<td>2 h</td>
<td>4–8</td>
<td>&gt;100</td>
<td>&gt;1000</td>
<td>~$925</td>
<td>~$0.93</td>
<td>~$1200 (100%)</td>
</tr>
<tr>
<td>Illumina MiSeq$^g$</td>
<td>26 h</td>
<td>3.4</td>
<td>150 + 150</td>
<td>1020</td>
<td>$750</td>
<td>$0.74</td>
<td>~$1000 (100%)</td>
</tr>
<tr>
<td>Illumina iScanSQ</td>
<td>8 days</td>
<td>250</td>
<td>100 + 100</td>
<td>50 000</td>
<td>$10 220</td>
<td>$0.20</td>
<td>$3000 (14%)</td>
</tr>
<tr>
<td>Illumina GAIIx</td>
<td>14 days</td>
<td>320</td>
<td>150 + 150</td>
<td>96 000</td>
<td>$11 524</td>
<td>$0.12</td>
<td>$3200 (14%)</td>
</tr>
<tr>
<td>SOLiD – 4</td>
<td>12 days</td>
<td>&gt;840$^5$</td>
<td>50 + 35</td>
<td>71 400</td>
<td>$8128</td>
<td>&lt;$0.11</td>
<td>$2500 (12%)</td>
</tr>
<tr>
<td>Illumina HiSeq 1000</td>
<td>8 days</td>
<td>500</td>
<td>100 + 100</td>
<td>100 000</td>
<td>$10 220</td>
<td>$0.10</td>
<td>$3000 (12%)</td>
</tr>
<tr>
<td>Illumina HiSeq 2000</td>
<td>8 days</td>
<td>1000</td>
<td>100 + 100</td>
<td>200 000</td>
<td>$20 120$</td>
<td>$0.10</td>
<td>$3000 (6%)</td>
</tr>
<tr>
<td>SOLiD – 5500 (PI)*</td>
<td>8 days</td>
<td>&gt;700$^5$</td>
<td>75 + 35</td>
<td>77 000</td>
<td>$6101</td>
<td>&lt;$0.08</td>
<td>$2000 (12%)</td>
</tr>
<tr>
<td>SOLiD – 5500xl (4hq)$^e$</td>
<td>8 days</td>
<td>&gt;1410$^8$</td>
<td>75 + 35</td>
<td>155 100</td>
<td>$10 503$</td>
<td>&lt;$0.07</td>
<td>$2000 (12%)</td>
</tr>
<tr>
<td>Illumina HiSeq 2000 – v3$^l$</td>
<td>10 days</td>
<td>≤3000</td>
<td>100 + 100</td>
<td>≤600 000</td>
<td>$23 470$</td>
<td>≥$0.04</td>
<td>~$3500 (6%)</td>
</tr>
</tbody>
</table>
A transformative technology

- Re-sequencing
- De novo genome sequencing
- RNA-Seq
- Non-coding RNAs
- Structural variation
- ChIP-Seq
- Methyl-Seq
- Metagenomics
- Viral quasispecies
- Shape-Seq
- ... many more biological measurements “reduced” to NGS sequencing
Outline

• Background
• **RNA-Seq read mapping**
  – Mapping strategies
  – Merging read alignments
• Variant detection and genotyping from RNA-Seq reads
• Transcriptome quantification using RNA-Seq
• Implementing RNA-Seq analysis pipelines using Galaxy
• Novel transcript reconstruction
• Conclusions
Mapping RNA-Seq Reads

pre-mRNA
Exon
Intron
mRNA
Short reads

Short read is split by intron when aligning to reference Genome

Mapping Strategies for RNA-Seq reads

- **Short reads (Illumina, SOLiD)**
  - Ungapped mapping (with mismatches) on genome
    - Leverages existing tools: bowtie, BWA, ...
    - Cannot align reads spanning exon-junctions
  - Mapping on transcript libraries
    - Cannot align reads from un-annotated transcripts
  - Mapping on exon-exon junction libraries
    - Cannot align reads overlapping un-annotated exons
  - Spliced alignment on the genome
    - Similar to classic EST alignment problem, but harder due to short read length and large number of reads

- Hybrid approaches

- **Long read mapping (454, ION Torrent)**
  - Local alignment (Smith-Waterman) to the genome
    - Handles indel errors characteristic of current long read technologies
Spliced Read Alignment with Tophat

Hybrid Approach Based on Merging Alignments

mRNA reads

Transcript Library Mapping

Transcript mapped reads

Read Merging

Genome Mapping

Genome mapped reads

Mapped reads
Alignment Merging for Short Reads

<table>
<thead>
<tr>
<th>Genome</th>
<th>Transcripts</th>
<th>Agree?</th>
<th>Hard Merge</th>
<th>Soft Merge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique</td>
<td>Unique</td>
<td>Yes</td>
<td>Keep</td>
<td>Keep</td>
</tr>
<tr>
<td>Unique</td>
<td>Unique</td>
<td>No</td>
<td>Throw</td>
<td>Throw</td>
</tr>
<tr>
<td>Unique</td>
<td>Multiple</td>
<td>No</td>
<td>Throw</td>
<td>Keep</td>
</tr>
<tr>
<td>Unique</td>
<td>Not Mapped</td>
<td>No</td>
<td>Keep</td>
<td>Keep</td>
</tr>
<tr>
<td>Multiple</td>
<td>Unique</td>
<td>No</td>
<td>Throw</td>
<td>Keep</td>
</tr>
<tr>
<td>Multiple</td>
<td>Multiple</td>
<td>No</td>
<td>Throw</td>
<td>Throw</td>
</tr>
<tr>
<td>Multiple</td>
<td>Not Mapped</td>
<td>No</td>
<td>Throw</td>
<td>Throw</td>
</tr>
<tr>
<td>Not mapped</td>
<td>Unique</td>
<td>No</td>
<td>Keep</td>
<td>Keep</td>
</tr>
<tr>
<td>Not mapped</td>
<td>Multiple</td>
<td>No</td>
<td>Throw</td>
<td>Throw</td>
</tr>
<tr>
<td>Not mapped</td>
<td>Not Mapped</td>
<td>Yes</td>
<td>Throw</td>
<td>Throw</td>
</tr>
</tbody>
</table>
Merging Of Local Alignments

1. Convert transcriptome alignment to genomic coordinates, concatenate converted SAM with genome SAM (adding pairing info to read IDs for paired data), and sort by (modified) read ID

2. Filter local alignments using “hard merge”, as follows

   For each read r, let G be an array of 3-tuples (chr, s, p) storing the genomic location for each base of r, where chr is the chromosome, s is the strand and p is the position to which r is mapped. Let S be an array storing the ambiguity status for each base of r.
   a. For each base b of r
      \[ G[b] \leftarrow \text{undefined}; S[b] \leftarrow \text{undefined} \]
   b. For each alignment a of r (first pass over r’s alignments)
      If a has fewer than t1 bases then discard a
      Else
      i. For each read base b included in a, use the SAM info to compute the genomic location g (consisting of chr, s, and p) to which b is mapped according to a
      ii. If G[b] = undefined then G[b] \leftarrow g; S[b] \leftarrow \text{non-ambiguous}
          Else
          If G[b] \neq g then S[b] \leftarrow \text{ambiguous}
          Endif
          Endif
      Endif
   c. For each alignment a of r (second pass over r’s alignments)
      i. If fewer than t2 bases of a are non-ambiguous in r according to S, mark these bases as ambiguous in S
   d. If there are t3 or more non-ambiguous bases of r according to S,
      If locations for different bases are in different chromosomes, different strands or are not monotonically increasing/decreasing along the read then all alignments of r are discarded
      Else generate local alignments for all stretches of non-ambiguous bases of r.
      These alignments are based on the (unique) genomic locations saved for each non-ambiguous base in G
      Endif
   Endif
Outline

- Background
- RNA-Seq read mapping
- Variant detection and genotyping from RNA-Seq reads
  - SNVQ algorithm
  - Experimental results
- Transcriptome quantification using RNA-Seq
- Implementing RNA-Seq analysis pipelines using Galaxy
- Novel transcript reconstruction
- Conclusions
Motivation

- RNA-Seq is much less expensive than genome sequencing
- **Can sequence variants be discovered reliably from RNA-Seq data?**
  - SNVQ: novel Bayesian model for SNV discovery and genotyping from RNA-Seq data [Duitama et al., ICCABS 2011]
  - Particularly appropriate when interest is in expressed mutations (cancer immunotherapy)
SNP Calling from Genomic DNA Reads

Read sequences & quality scores

@HWI-EAS299_2:2:1:1536:631
GGGATGTCAGGATTCAAGACGAGCTTGGGATGAG
+HWI-EAS299_2:2:1:1536:631
GGGATGTCAGGATTCAAGACGAGCTTGGGATGAG
@HWI-EAS299_2:2:1:771:771
ATTACACACTTCAGAGGTTGGAGTACTC
+HWI-EAS299_2:2:1:771:771
ATTACACACTTCAGAGGTTGGAGTACTC

Reference genome sequence

>ref|NT_082868.6|Mm19_82865_37:1-3688105 Mus musculus chromosome 19 genomic contig, strain C57BL/6J
GATCATACTCCTCATGCTGACATTCTGGTGATGTAT
ATCTGGGAGAGCTATGAGGATGGGAATTGCTA
ACTTTCCCTCTCTATAGGTTAGCTAATGCATAATGCAAA
ATATTTCCAGCGTTTTTTACTAGAGATAAAG
AACTGGGACTTGCTTATTTACCTTTTAGAAGACAGATTC
AGGCTCTGCAAGAGAATATGCAGTTCTCTCAT
ACAGGGAAAGCCTGCTTTTCTACTAAAGCTTGTTGCTAAT

Read Mapping

SNP calling

1 4764558 G T 2 1
1 4767621 C A 2 1
1 4767623 T A 2 1
1 4767633 T A 2 1
1 4767643 A C 4 2
1 4767656 T C 7 1
SNV Detection and Genotyping

Locus i

Reference

| GCGGCCAGCCGGCTTCTGTCGGCCAGCAGCCAGGAATCTGGAAACAATGGCTACAGCGTGC |
| GCGGCCAGCGGCCCTTCTGTCGGCCAGCCCGAGCAG |
| CGCGGCCAGCGCCCGCTTCTGTCGGCCAGCGCCAGG |
| GCCAGCCGGGCTTCTGTCGGCCAGCGCGAGAATCT |
| GCCGGCTTCTGTCGGCCAGCCAGGAATCTGGAA |
| CTTCTGTCGGCCAGCGCGAGAATCTGGAAACAAT |
| GCGGCCAGCAGCCAGGAATCTGGAAACAATGGCTACAGCG |
| CCAGCAGCCAGGAATCTGGAAACAATGGCTACAGCG |
| CAAGCAGCCAGGAATCTGGAAACAATGGCTACAGCG |
| GCAGCCAGGAATCTGGAAACAATGGCTACAGCG |

\[ r(i) : \text{Base call of read } r \text{ at locus } i \]
\[ \varepsilon_{r(i)} : \text{Probability of error reading base call } r(i) \]
\[ G_i : \text{Genotype at locus } i \]
SNV Detection and Genotyping

• Use Bayes rule to calculate posterior probabilities and pick the genotype with the largest one

\[
P(G_i|R_i) = \frac{P(R_i|G_i)P(G_i)}{P(R_i)}
\]

\[
P(G_i = H_i H_i') = \begin{cases} 
\frac{1-h}{4}, & \text{if } H_i = H_i' \\
\frac{h}{6}, & \text{otherwise}
\end{cases}
\]
Current Models

• **Maq:**
  – Keep just the alleles with the two largest counts
  – \( \Pr (R_i \mid G_i=H_iH_i) \) is the probability of observing \( k \) alleles \( r(i) \) different than \( H_i \)
  – \( \Pr (R_i \mid G_i=H_iH'_i) \) is approximated as a binomial with \( p=0.5 \)

• **SOAPsnp**
  – \( \Pr (r_i \mid G_i=H_iH'_i) \) is the average of \( \Pr(r_i\mid H_i) \) and \( \Pr(r_i\mid G_i=H'_i) \)
  – A rank test on the quality scores of the allele calls is used to confirm heterozygocity
SNVQ Model

• Calculate conditional probabilities by multiplying contributions of individual reads

\[
P(R_i|G_i) = \prod_{r \in R_i} P(r|G_i)
\]

\[
P(r|G_i = H_i H'_i) = \begin{cases} 
1 - \varepsilon_{r(i)}, & \text{if } H_i = H'_i = r(i) \\
\frac{\varepsilon_{r(i)}}{3}, & \text{if } H_i \neq r(i) \land H'_i \neq r(i) \\
\frac{1}{2} - \frac{\varepsilon_{r(i)}}{3}, & \text{otherwise}
\end{cases}
\]
Experimental Setup

• 113 million 32bp Illumina mRNA reads generated from blood cell tissue of Hapmap individual NA12878 (NCBI SRA database accession numbers SRX000565 and SRX000566)
  – We tested genotype calling using as gold standard 3.4 million SNPs with known genotypes for NA12878 available in the database of the Hapmap project
  – **True positive**: called variant for which Hapmap genotype coincides
  – **False positive**: called variant for which Hapmap genotype does not coincide
Comparison of Mapping Strategies

![Graph comparing mapping strategies](image-url)
Comparison of Variant Calling Strategies

![Graph showing comparison of variant calling strategies with True Positives on the y-axis and False Positives on the x-axis. The graph compares SNVQ, SOAPsnp, and Maq.]
Data Filtering

• Allow just x reads per start locus to eliminate PCR amplification artifacts

• [Chepelev et al. 2010] algorithm:
  – For each locus groups starting reads with 0, 1 and 2 mismatches
  – Choose at random one read of each group
Comparison of Data Filtering Strategies

The graph shows the comparison of different data filtering strategies. The x-axis represents the number of false positives, while the y-axis represents the number of true positives.

- **None**: No filtering applied.
- **Alignment Trimming**: Alignment trimming is applied.
- **Three Reads Per Start Locus**: Three reads per start locus are considered.
- **One Read Per Start Locus**: Only one read per start locus is considered.

The graph illustrates how each strategy affects the trade-off between true positives and false positives.
Accuracy per RPKM bins

<table>
<thead>
<tr>
<th></th>
<th>SOAPsnp</th>
<th>Maq</th>
<th>SNVQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPKM &lt; 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 &lt; RPKM &lt; 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 &lt; RPKM &lt; 10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 &lt; RPKM &lt; 50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 &lt; RPKM &lt; 100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPKM &gt; 100</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- TPHomoVar
- TPHetero
- FP
- FNHomoVar
- FNHetero
Outline

• Background
• RNA-Seq read mapping
• Variant detection and genotyping from RNA-Seq reads
• Transcriptome quantification using RNA-Seq
  – Background
  – IsoEM algo
  – Experimental results
  – Alternative protocols and inference problems
    • DGE protocol
    • Inference of allele specific expression levels
• Implementing RNA-Seq analysis pipelines using Galaxy
• Novel transcript reconstruction
• Conclusions
Alternative splicing

Simple transcription

Canonical isoform

Alternative transcript initiation

Alternative splicing

Alternative polyadenylation
## Alternative Splicing

### Table 1. Identification of known alternative events in the mouse cerebellum tissue

<table>
<thead>
<tr>
<th>Alt event type</th>
<th>Reference set</th>
<th>Both isoforms expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P0</td>
<td>P5</td>
</tr>
<tr>
<td><strong>Transcriptional events</strong> [No. of events (No. of genes)]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFE</td>
<td>26147 (9334)</td>
<td>8873 (3606)</td>
</tr>
<tr>
<td>ALE</td>
<td>22469 (8610)</td>
<td>8176 (3447)</td>
</tr>
<tr>
<td><strong>Alternative splicing events</strong> [No. of events (No. of genes)]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon skipping</td>
<td>20589 (7547)</td>
<td>1579 (1192)</td>
</tr>
<tr>
<td>Intron retention</td>
<td>2389 (2144)</td>
<td>1038 (845)</td>
</tr>
<tr>
<td>A5SS</td>
<td>1659 (1425)</td>
<td>251 (236)</td>
</tr>
<tr>
<td>A3SS</td>
<td>3151 (2383)</td>
<td>423 (333)</td>
</tr>
</tbody>
</table>

Using the combined known gene models, we generated a library of all known alternative events in the mouse genome (reference set) and identified the events occurring in each stage, as well as over the course of cerebellum development (overall). If the isoforms are coexpressed in a stage, then it is counted as an alt event in that stage, and in case the alternative isoforms are expressed during different developmental stages, the alt event will be included in the overall category.

---

Pal S. et al., Genome Research, June 2011
Computational Problems

Make cDNA & shatter into fragments

Sequence fragment ends

Map reads

Gene Expression (GE)  Isoform Expression (IE)  Isoform Discovery (ID)
Challenges to accurate estimation of gene expression levels

- Read ambiguity (multireads)

- What is the gene length?
Previous Approaches to GE

• Ignore multireads
• [Mortazavi et al. 08]
  – Fractionally allocate multireads based on unique read estimates
• [Pasaniuc et al. 10]
  – EM algorithm for solving ambiguities
• Gene length: sum of lengths of exons that appear in at least one isoform
  ➔ Underestimates expression levels for genes with 2 or more isoforms [Trapnell et al. 10]
Read Ambiguity in IE
Previous Approaches to IE

- [Jiang&Wong 09]
  - Poisson model + importance sampling, single reads
- [Richard et al. 10]
  - EM Algorithm based on Poisson model, single reads in exons
- [Li et al. 10]
  - EM Algorithm, single reads
- [Feng et al. 10]
  - Convex quadratic program, pairs used only for ID
- [Trapnell et al. 10]
  - Extends Jiang’s model to paired reads
  - Fragment length distribution
IsoEM algorithm
[Nicolae et al. 2011]

• Unified probabilistic model and Expectation-Maximization Algorithm (IsoEM) for IE considering
  – Single and/or paired reads
  – Fragment length distribution
  – Strand information
  – Base quality scores
  – Repeat and hexamer bias correction
Read-isoform compatibility

\[ W_{r,i} = \sum_a O_a Q_a F_a \]

\[
Q_a = \prod_{k=1}^{\lfloor r \rfloor} \left[ (1 - \varepsilon_k)M_{a_k} + \frac{\varepsilon_k}{3} (1 - M_{a_k}) \right]
\]

\[
M_{a_k} = \begin{cases} 
1 & \text{if position } k \text{ of } a \text{ matches genome} \\
0 & \text{otherwise}
\end{cases}
\]

\[
\varepsilon_k = \text{error probability of } k^{\text{th}} \text{ base of } r
\]
Fragment length distribution

- Paired reads
Fragment length distribution

- Single reads
IsoEM pseudocode

assign random values to all $f(i)$

while not converged do
  initialize all $n(j)$ to 0
  for each read $r$ do
    sum = $\sum_{j \in \text{compatible}(r)} w_{r,j} f(j)$
    for each isoform $j$ compatible with $r$ do
      $n(j)$ += $w_{r,j} f(j)$ / sum
    end for
  end for
  $s = \sum_j n(j) / (l(j) - \mu + 1)$
  for each isoform $j$ do
    $f(j) = \frac{n(j) / (l(j) - \mu + 1)}{s}$
  end for
end while
Implementation details

- Collapse identical reads into read classes

Isoforms: i1, i2, i3, i4, i5, i6
Reads: (i3, i4)
LCA(i3, i4)
Implementation details

• Run EM on connected components, in parallel

Graph showing the connection between isoforms i1, i2, i3, i4, i5, and i6 with a chart indicating the number of components and their size (# isoforms).
Simulation setup

- Human genome UCSC known isoforms

- GNFAAtlas2 gene expression levels
  - Uniform/geometric expression of gene isoforms

- Normally distributed fragment lengths
  - Mean 250, std. dev. 25
Accuracy measures

• Error Fraction ($EF_t$)
  – Percentage of isoforms (or genes) with relative error larger than given threshold $t$

• Median Percent Error (MPE)
  – Threshold $t$ for which $EF$ is 50%

• $r^2$
Error fraction curves - isoforms

- 30M single reads of length 25 (simulated)
Error fraction curves - genes

- 30M single reads of length 25 (simulated)
MPE and EF_{15} by gene expression level

- 30M single reads of length 25

<table>
<thead>
<tr>
<th>Expression range</th>
<th># genes</th>
<th>(0, 10^{-6}]</th>
<th>(10^{-6}, 10^{-5}]</th>
<th>(10^{-5}, 10^{-4}]</th>
<th>(10^{-4}, 10^{-3}]</th>
<th>(10^{-3}, 10^{-2}]</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>120</td>
<td>5,610</td>
<td>11,907</td>
<td>1,632</td>
<td>102</td>
<td>19,372</td>
</tr>
<tr>
<td>MPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniq</td>
<td>37.4</td>
<td>43.6</td>
<td>42.7</td>
<td>43.0</td>
<td>48.2</td>
<td>43.0</td>
<td></td>
</tr>
<tr>
<td>Rescue</td>
<td>32.8</td>
<td>28.7</td>
<td>26.0</td>
<td>25.1</td>
<td>28.8</td>
<td>26.7</td>
<td></td>
</tr>
<tr>
<td>GeneEM</td>
<td>30.6</td>
<td>28.2</td>
<td>25.7</td>
<td>25.1</td>
<td>28.0</td>
<td>26.3</td>
<td></td>
</tr>
<tr>
<td>Cufflinks</td>
<td>33.0</td>
<td>21.1</td>
<td>19.0</td>
<td>20.2</td>
<td>40.2</td>
<td>19.7</td>
<td></td>
</tr>
<tr>
<td>RSEM</td>
<td>23.6</td>
<td>11.0</td>
<td>7.2</td>
<td>7.9</td>
<td>11.4</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>IsoEM</td>
<td>18.2</td>
<td>8.4</td>
<td>3.2</td>
<td>2.0</td>
<td>1.9</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>EF_{15}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniq</td>
<td>77.5</td>
<td>82.4</td>
<td>81.7</td>
<td>79.7</td>
<td>82.4</td>
<td>81.7</td>
<td></td>
</tr>
<tr>
<td>Rescue</td>
<td>74.2</td>
<td>74.0</td>
<td>71.6</td>
<td>72.8</td>
<td>76.5</td>
<td>72.4</td>
<td></td>
</tr>
<tr>
<td>GeneEM</td>
<td>72.5</td>
<td>73.8</td>
<td>71.5</td>
<td>73.0</td>
<td>74.5</td>
<td>72.3</td>
<td></td>
</tr>
<tr>
<td>Cufflinks</td>
<td>73.3</td>
<td>64.7</td>
<td>62.3</td>
<td>66.2</td>
<td>82.3</td>
<td>63.5</td>
<td></td>
</tr>
<tr>
<td>RSEM</td>
<td>64.2</td>
<td>37.3</td>
<td>17.4</td>
<td>16.3</td>
<td>41.2</td>
<td>23.5</td>
<td></td>
</tr>
<tr>
<td>IsoEM</td>
<td>57.5</td>
<td><strong>28.1</strong></td>
<td><strong>6.7</strong></td>
<td><strong>6.1</strong></td>
<td><strong>4.9</strong></td>
<td><strong>13.2</strong></td>
<td></td>
</tr>
</tbody>
</table>
Read length effect on IE MPE

- Fixed sequencing throughput (750Mb)
Read length effect on IE $r^2$

- Fixed sequencing throughput (750Mb)
Effect of pairs & strand information

- 75bp reads
Runtime scalability

- Scalability experiments conducted on a Dell PowerEdge R900
  - Four 6-core E7450 Xeon processors at 2.4Ghz, 128Gb of internal memory
MAQC data

RNA samples: UHRR, HBRR

- 6 libraries, 47-92M 35bp reads each [Bullard et al. 10]
- Bases called using both auto and phi X calibration for 2 libraries

qPCR

- Quadruplicate measurements for 832 Ensembl genes [MAQC Consortium 06]
$r^2$ comparison for MAQC samples

![Graph showing $r^2$ comparison for MAQC samples. The x-axis represents Million Mapped Bases ranging from 0 to 2000, and the y-axis represents $r^2$ values ranging from 0.35 to 0.85. The graph includes lines and markers for different sample combinations such as HBRR 1X, IsoEM, HBRR 1A, IsoEM, HBRR 1X, Cufflinks, HBRR 1A, Cufflinks, UHRR 1X, IsoEM, UHRR 1A, IsoEM, UHRR 2, IsoEM, UHRR 3, IsoEM, UHRR 4, IsoEM, UHRR 5, IsoEM, UHRR 2, Cufflinks, UHRR 3, Cufflinks, UHRR 4, Cufflinks, and UHRR 5, Cufflinks.]}
$R^2$ of IsoEM estimates from ION Torrent & Illumina HBR reads

Average $R^2$ for 5 ION Torrent MAQC HBR Runs (avg. 1,559,842 reads)

$R^2$ for combined reads from 5 ION Torrent MAQC HBR Runs (7,799,210 reads)
DGE/SAGE-Seq protocol

1. Gene Expression (GE)
2. Cleave with tagging enzyme
3. Map tags
4. Cleave with anchoring enzyme (AE)
5. Attach primer for tagging enzyme (TE)

Gene Expression (GE)
Inference algorithms for DGE data

- Discard ambiguous tags [Asmann et al. 09, Zaretzki et al. 10]
- Heuristic rescue of some ambiguous tags [Wu et al. 10]
- DGE-EM algorithm [Nicolae & Mandoiu, ISBRA 2011]
  - Uses all tags, including all ambiguous ones
  - Uses quality scores
  - Takes into account partial digest and gene isoforms
Tag formation probability

\[ p(1-p)^{k-1} \quad \cdots \quad p(1-p) \quad p \]

5' \quad k \quad \cdots \quad 2 \quad 1 \quad 3'  

AE site  
MRNA  
Tag formation probability
Tag-isoform compatibility

\[ w_{t,i,j} = Q_a p(1 - p)^{j-1} \]

\[ Q_a = \prod_{k=1}^{\left| r \right|} \left[ (1 - \varepsilon_k) M_{a_k} + \frac{\varepsilon_k}{3} (1 - M_{a_k}) \right] \]

\[ M_{a_k} = \begin{cases} 
1 & \text{if position } k \text{ of } a \text{ matches genome} \\
0 & \text{otherwise}
\end{cases} \]

\[ \varepsilon_k = \text{error probability of } k^{\text{th}} \text{ base of } r \]
DGE-EM algorithm

assign random values to all \( f(i) \)

while not converged

init all \( n(i,j) \) to 0

for each tag \( t \)

\[
s = \sum_{(i,j,w) \in t} wf(i)
\]

for \( (i,j,w) \) in \( t \)

\[
n(i, j)^+ = \frac{wf(i)}{s}
\]

for each isoform \( i \)

\[
N = \sum_{j=1}^{sites(i)} n_{i,j}
\]

\[
f(i) = \frac{N / \left(1 - (1 - p)^{sites(i)}\right)}{}
\]
MAQC data

DGE
  • 9 Illumina libraries, 238M 20bp tags [Asmann et al. 09]
  • Anchoring enzyme DpnII (GATC)

RNA-Seq
  • 6 libraries, 47-92M 35bp reads each [Bullard et al. 10]

qPCR
  • Quadruplicate measurements for 832 Ensembl genes [MAQC Consortium 06]
DGE-EM vs. Uniq on HBRR Library 4

- Median Percent Error
- Million Mapped Tags

- Uniq 0 mismatches
- Uniq 1 mismatch
- Uniq 2 mismatches
- DGE-EM 0 mismatches
- DGE-EM 1 mismatch
- DGE-EM 2 mismatches
Synthetic data

• 1-30M tags, lengths 14-26bp
• UCSC hg19 genome and known isoforms
• Simulated expression levels
  – Gene expression for 5 tissues from the GNFAtlas2
  – Geometric expression for the isoforms of each gene
• Anchoring enzymes from REBASE
  – DpnII (GATC) [Asmann et al. 09]
  – NlaIII (CATG) [Wu et al. 10]
  – CviII (RGCY, R=G or A, Y=C or T)
MPE for 30M 21bp tags

RNA-Seq: 8.3 MPE
DGE vs. RNA-Seq Summary

- RNA-Seq and DGE based estimates have comparable cost-normalized accuracy on MAQC data
  - When using best inference algorithm for each type of data
- Simulations suggest possible DGE protocol improvements
  - Enzymes with degenerate recognition sites (e.g. CviII)
  - Optimizing cutting probability
Allele Specific Expression in F1 Hybrids

• [McManus et al. 10]

Paired-end reads (37bp)
Analysis Pipeline for Allele-Specific Isoform Expression in F1 Hybrids

1. **Parent Genome Sequences**
   - ChrX
   - GAATTCTGTGAAAGCCTGT
   - AGCTATAAAAAATGTTGA
   - GCCATAAATACCATCACTTT
   - GAAGTATTCTGAGACTTGT
   - AGGAAGGTGAAGTAAATA
   - TCTAATATAATTGGATTGTA
   - TGTTTTTGATTATTTTTTGTT
   - AGGCTGTGATGGGCTCAA
   - GTAATTGAAA

2. **Generate Isoform Sequences**

3. **Diploid Transcriptome**

4. **Align to Diploid Transcriptome**

5. **Short Reads**

6. **IsoEM**

7. **Allele Specific Expression Levels**

8. **Allele Specific Read Mapping**

**Reference Transcriptome**

**Generate Isoform Sequences**

**Diploid Transcriptome**

**Align to Diploid Transcriptome**

**IsoEM**

**Allele Specific Expression Levels**

**Short Reads**

**Parent Genome Sequences**
Allele Specific Expression on Drosophila RNA-Seq data from [McManus et al. 10]
Allele Specific Expression for Mouse RNA-Seq Data from [Gregg et al. 2010]

(A) IsoEM

(B) Unique Mapping
General Pipeline for Allele-Specific Isoform Expression

Reference Genome

Reference Transcriptome

Generate Isoform Sequences

Reference Isoform Sequences

Short Reads

Align to Transcriptome And Call SNPs

Read Mapping

IsoEM

Allele Specific Read Mapping

Allele Specific Isoform Expression

Align to Diploid Transcriptome

Diploid Transcriptome

Short Reads

Haplotyping (RefHap)
Outline

• Background
• RNA-Seq read mapping
• Variant detection and genotyping from RNA-Seq reads
• Transcriptome quantification using RNA-Seq
• Implementing RNA-Seq analysis pipelines using Galaxy
  – Running analyses, creating flows, and adding tools in Galaxy
  – Hands on exercise
• Novel transcript reconstruction
• Conclusions
Galaxy

- Web-based platform for bioinformatics analysis
- Aims to facilitate reproducing results
- Provides user friendly interface to many available tools
- Free public server (maintained by PSU)
- Downloadable galaxy instance for installation and expansion (adding tools)
Local Galaxy Instance

• [http://rna1.engr.uconn.edu:7474/](http://rna1.engr.uconn.edu:7474/)

• Lab Tools
  – NGS: IsoEM
  – SNVQ

• Tools available on the PSU server
Adding Tools to Local Galaxy Instances

- Galaxy Wiki for tool configuration syntax
  http://wiki.g2.bx.psu.edu/Admin/Tools/Tool%20Config%20Syntax
Outline

• Background
• RNA-Seq read mapping
• Variant detection and genotyping from RNA-Seq reads
• Transcriptome quantification using RNA-Seq
• Implementing RNA-Seq analysis pipelines using Galaxy
• Novel transcript reconstruction
  – Overview of existing approaches
  – DRUT algorithm
• Conclusions
Existing approaches

• Genome-guided reconstruction (ab initio)
  – Exon identification
  – Genome-guided assembly

• Genome independent (de novo) reconstruction
  – Genome-independent assembly

• Annotation-guided reconstruction
  – Explicitly use existing annotation during assembly
Genome-guided reconstruction

• Scripture (2010), IsoLasso (2011)
  – Reports “all” isoforms

• Cufflinks (2010)
  – Reports a minimal set of isoforms
Genome independent reconstruction

  - Euler/de Bruijn k-mer graph
Max Set vs Min Set

Cufflinks

- **G(V,E)**
  - V – pe reads
  - E – compatible reads

Fragment x₄ in (d) is uncertain, because y₄ and y₅ are incompatible with each other.
**Scripture**

- **Connectivity graph**
  - $V$ – bases
  - $E$ – spliced event

- **Filter isoforms**
  - Coverage (p-value)
  - Insert length
Other statistical assembly

• IsoLasso
  – multivariate regression method – Lasso
    • balance between the maximization of prediction accuracy and the minimization of interpretation

• SLIDE
  – sparse estimation as a modified Lasso
    • limiting the number of discovered isoforms and favoring longer isoforms

Li, W. et al. *RECOMB 2011*, Li J et all Proc Natl Acad Sci. USA 2011
Reconstruction Strategies Comparison

Detection and Reconstruction of Unannotated Transcripts

a) **Map** reads to annotated transcripts (using Bowtie)

b) **VTEM**: Identify overexpressed exons (possibly from unannotated transcripts)

c) **Assemble Transcripts** (e.g., Cufflinks) using reads from “overexpressed” exons and unmapped reads

d) **Output**: annotated transcripts + novel transcripts
**Virtual Transcript Expectation Maximization (VTEM)**

- **VTEM** is based on a modification of Virtual String Expectation Maximization (VSEM) Algorithm [Mangul et al. 2011]).
  - the difference is that we consider in the panel exons instead of reads
  - Calculate observed **exon counts** based on read mapping
    - each read contribute to count of either one exon or two exons (depending if it is a unspliced spliced read or spliced read)
Input: Complete vs Partial Annotations

Complete Annotations

Partial Annotations

Observed exon frequencies

Transcript T3 is missing from annotations
Virtual Transcript Expectation Maximization (VTEM)

(Partially) Annotated Genome + Virtual Transcript with 0-weights in virtual transcript

EM

ML estimates of transcript frequencies

EM

Update weights of reads in virtual transcript

Virtual Transcript frequency change > \( \epsilon \)?

NO

Output overexpressed exons (expressed by virtual transcripts)

YES

Compute expected exons frequencies

Overexpressed exons belong to unknown transcripts
Simulation Setup

- Reads simulated from UCSC known genes
  - 19,372 genes
  - 66,803 isoforms
- Single end, error-free
  - 60M reads of length 100bp
- To simulate incomplete annotation, remove from every gene exactly one isoform
Comparison Between Methods

![Graphs showing comparison between methods for Sensitivity and PPV based on the number of transcripts per gene.](image-url)
Outline

• Background
• RNA-Seq read mapping
• Variant detection and genotyping from RNA-Seq reads
• Transcriptome quantification using RNA-Seq
• Implementing RNA-Seq analysis pipelines using Galaxy
• Novel transcript reconstruction
• Conclusions
Conclusions

• The range of NGS applications continues to expand, fueled by advances in technology
  • Improved sample prep protocols
  • 3rd generation: Pacific Biosciences, Ion Torrent

• Development of sophisticated analysis methods remains critical for fully realizing the potential of sequencing technologies
Further readings

Read mapping

Further readings

SNV discovery and genotyping

- S.Q. Le and R. Durbin: SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome research, to appear.
Further readings

Estimation of gene expression levels from RNA-Seq data

Further readings

Estimation of gene expression levels from DGE data


Further readings

**Novel transcript reconstruction**
- S. Mangul, A. Caciula, I. Mandoiu, and A. Zelikovsky. RNA-Seq based discovery and reconstruction of unannotated transcripts in partially annotated genomes, Proc. BIBM 2011, pp.118-123
Software packages

- SNV detection and genotyping from RNA-Seq reads: http://dna.engr.uconn.edu/software/NGSTools
- Inference of gene expression levels from RNA-Seq reads: http://dna.engr.uconn.edu/software/IsoEM/
- Inference of gene expression levels from DGE reads: http://www.dna.engr.uconn.edu/software/DGE-EM
Galaxy References


• UCONN Galaxy instance: http://rna1.engr.uconn.edu:7474/

• Main Galaxy server at PSU: http://galaxy.psu.edu/
Acknowledgments

- Jorge Duitama (KU Leuven)
- Marius Nicolae (Uconn)
- Pramod Srivastava (UCHC)
- Alex Zelikovsky (GSU)
- Serghei Mangul (GSU)
- Adrian Caciula (GSU)
- Dumitru Brinza (Life Tech)