De novo transcriptome assembly

hunting for useful molecules in medicinal maggots

Ken Kraaijeveld
Ewoud Ewing
Genome versus transcriptome

- Heterozygosity
- Single nucleotide polymorphisms
- Copy number variation
- Structural variation
- Introns
- Repetitive elements

- Expression differences
- Alternative splicing
- Sense / antisense transcripts
Expression differences

Alternative splicing

(a) Alternative selection of promoters (e.g., *myosin* primary transcript)

(b) Alternative selection of cleavage/polyadenylation sites (e.g., *tropomyosin* transcript)

(c) Intron retaining mode (e.g., *transposase* primary transcript)

(d) Exon cassette mode (e.g., *troponin* primary transcript)

http://www.scfbio-iitd.res.in/tutorial/splicing.html
Even with a reference genome, transcripts need to be assembled

- Spiced alignment (Tophat, Gmap)
- Assemble transcripts (Cufflinks)
- Differential expression (CuffDiff)

Trapnell et al. 2010 Nature Biotechnology
Genome assembly versus transcriptome assembly

- **Multiple paths due to:**
  - Sequence errors
  - Gene duplications
  - Heterozygosity

- **Multiple paths due to:**
  - Alternative splice forms
  - Paralogous transcripts
  - Allelic variation
  - Sequence errors

**Which paths to keep and which to throw away?**
Transcriptome assemblers

- Trinity http://trinityrnaseq.sourceforge.net/
- Oases http://www.ebi.ac.uk/~zerbino/oases/
- Trans-ABySS
  http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss
Trinity

• Inchworm
  - k-mer based contig assembly
  - Full-length contigs for dominant transcripts
  - Unique bits of alternative transcripts

• Chrysalis
  - Groups inchworm contigs

• Butterfly
  - Reconciles chrysalis graphs with original reads
  - Reports full-length transcripts
Oases

- Build contigs
- Divide into long and short contigs
- Group long contigs into loci
- Connect the short contigs
Oases

Schulz et al. Bioinformatics 2012
Case study: medicinal maggots

http://zombieapocalypseacademy.org/2012/06/maggots/
<table>
<thead>
<tr>
<th>Platform</th>
<th>Illumina GAII</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reads post-filtering</td>
<td>8,162,882</td>
</tr>
<tr>
<td>Read length</td>
<td>35</td>
</tr>
<tr>
<td>Total bases</td>
<td>293,863,749</td>
</tr>
</tbody>
</table>
# Maggot assemblies

<table>
<thead>
<tr>
<th></th>
<th>Trinity</th>
<th>Oases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>1293</td>
<td>8578</td>
</tr>
<tr>
<td>Total bases</td>
<td>769,692</td>
<td>4,293,399</td>
</tr>
<tr>
<td>n50</td>
<td>742</td>
<td>867</td>
</tr>
<tr>
<td>Longest contig</td>
<td>8,242</td>
<td>13,942</td>
</tr>
</tbody>
</table>
Blast Trinity <> Oases assemblies

Contig length
Transcript annotation

- Blastx
  - NCBI
  - Flybase
  - Uniprot
  - Pfam

- Conserved domain database
## Transcript annotation

<table>
<thead>
<tr>
<th>Trinity</th>
<th>1293 contigs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1129 database hits</td>
</tr>
<tr>
<td></td>
<td>194 hits of interest</td>
</tr>
<tr>
<td>Oases</td>
<td>8578 transcripts from 4302 loci</td>
</tr>
<tr>
<td></td>
<td>6705 database hits</td>
</tr>
<tr>
<td></td>
<td>517 hits of interest</td>
</tr>
</tbody>
</table>
Transcripts of interest
A very poor first draft genome

- 4,857,608 contigs
- 243,607 scaffolds
- 8x coverage
Aligning transcripts to draft genome

- All transcripts had a match in genomic contigs
- Look for introns using Blast output viewer
- http://cas-bioinfo.cas.unt.edu/cgi-bin/BOV/index.cgi
Blast transcripts on genomic contigs
Blast transcripts on genomic contigs
Trinity contigs blasted against genome
n = 194
Oases contigs blasted against genome
n = 517
Validation using qPCR: designed primers for 25 serine proteases
Outlook: transcriptomics using SMRT sequencing

- PacBio reads span entire transcripts > de novo assembly obsolete
- Error correction necessary for accurate assessment of exon structure
Outlook: transcriptomics using SMRT sequencing

Koren et al. 2012 Nature Biotechnology
Outlook: transcriptomics using SMRT sequencing
Thank you

- Ewoud Ewing
- Yahya Anvar
- Jonathan den Boer
- LGTC team