Gene Construction (Alignment) vs. Assembly

Genome assemblers expect even coverage. Assembled regions with high coverage are assumed to be repeats.

But individual genes within a transcriptome will have very different amounts of coverage…
## Transcriptome Assemblers

<table>
<thead>
<tr>
<th>Short-read (Illumina, SOLiD)</th>
<th>Longer-read (454, Sanger, etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Trinity (Broad)</td>
<td>• Mira3</td>
</tr>
<tr>
<td>• Velvet/Oases</td>
<td>• Est2assembly</td>
</tr>
<tr>
<td>• TransAbyss</td>
<td>• GS/Newbler (Roche)</td>
</tr>
<tr>
<td>• UnoSeq</td>
<td>• SMRT Pipe (PacBio)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Software Packages</strong></td>
<td><strong>Others of Note</strong></td>
</tr>
<tr>
<td>• CLCBio</td>
<td>• STM (Scaffolding using Translation Mapping)</td>
</tr>
<tr>
<td>• NextGEne (softgenetics)</td>
<td>• tgcil/CAP3 (can be useful for consolidation)</td>
</tr>
<tr>
<td>• SeqMan NGen (DNASTAR)</td>
<td></td>
</tr>
</tbody>
</table>
Transcriptome Assembly with

- De novo transcriptome assembler
  - Also has genome-guided mode
- Input Read Normalization
  - Algorithm specifically tailored to Trinity
  - Reduces run times and memory usage
  - Improves assemblies by removing kmers/reads that contain likely errors
- Written for Illumina reads (longer reads may need pre-processing)
- Three modules:
  - **Inchworm**: Assembles reads into contigs
  - **Chrysalis**: Clusters contigs, constructs de Bruijn graph for each cluster, and partitions the full read set among the graphs
  - **Butterfly**: Processes the individual de Bruijn graphs in parallel, incorporates paired-end information to generate transcripts for alternatively spliced isoforms and paralogous genes
- Don’t “re-assemble” assemblies (run everything at once)

http://bioinformatics.ucdavis.edu
Transcriptome Assembly Post-Processing

“I have too many contigs --- what do I do now??”

- Many assemblers (particularly Trinity) generate high numbers of putative transcript contigs (>100k)
- The REAL bioinformatics analysis begins at this step
- To reduce the contigs down to a more concise transcript set...

**FILTER, DON’T CLUSTER**

- Clustering will collapse isoforms, paralogs, and gene families
- Clustering will introduce chimeras.
- Filter on isoform %, coverage, ORFs, blast hits, and anything else you might find useful. Then align reads back to the subset of contigs.
Transcriptome Assembly Challenges

• Contamination in original RNA sample (other genomes represented)
• Paralogs vs. splice variants
• Coverage (highly expressed vs. low expressed genes)
• Repetitive sequences (not as much a problem as with genome assembly)
• Chimeric contigs
Transcriptome Assessment

• Map reads back to the assembly
  – >60% reads mapped is good
  – >80% reads mapped is great
  – If low, may need more sequence, better reference

• Blast to proteins/transcriptome of related organism
  – Can infer fraction of transcriptome assembled

• Annotation tools (BLAST2GO, etc.)

• Trinity has downstream pipeline
  – Transdecoder for ORF-finding
  – Supports RSEM (RNA-Seq by Expectation-Maximization) for abundance estimation calculation
  – Provides edgeR wrapper script for differential expression
  – Trinotate for functional annotation
Pop Quiz

There is a VERY SIMPLE Trinity exercise in Galaxy. (It is the small sample dataset distributed with the software.)

Find the reads (no need to run any trimming, though you can if you want).

Assemble the reads with Trinity (keep default parameters)

Look at the output!
Discussion