Trinity On TACC For Everyone

Benni Goetz
CCBB Bioinformatics Consultant
Oct 9, 2014

(Joint work with John Hawkins)
Genomes vs Transcriptomes

• Genomes: mostly uniform coverage across genome; target is usually tens of chromosomes.

• Transcriptomes: widely varying coverage depending on gene expression; splice variants for thousands of transcripts.

• Genome and transcriptome assemblers are based similar ideas, but must use different algorithms.
Figure 3 | Overview of the FG\textsubscript{PQXQ} transcriptome assembly strategy.
a | All substrings of length k (k-mers) are generated from each read.
b | Each unique k-mer is used to represent a node (or vertex) in the De Bruijn graph, and pairs of nodes are connected if shifting a k-mer by one character creates an exact k–1 overlap between the two k-mers. Note that for non-strand-specific RNA sequencing data sets, the reverse complement of each k-mer will also be represented in the graph. Here, a simple example using 5-mers is shown. The example illustrates a SNP or sequencing error (for example, A/T) and an example of an intron or a deletion. Single-nucleotide differences cause ‘bubbles’ of length k in the De Bruijn graph, whereas introns or deletions introduce a shorter path in the graph.
c, d | Chains of adjacent nodes in the graph are collapsed into a single node when the first node has an out degree of one and the second node has an in degree of one. Last, as in the reference-based approach, four alternative paths (blue, red, yellow and green) through the graph are chosen.
e | The isoforms are then assembled.
Trinity Is A Pipeline

- Inchworm and Chrysalis are responsible for cutting up short reads into k-mers, and generating the de Bruijn graphs (one graph per gene).

- Butterfly determines putative transcripts (and isoforms) from the de Bruijn graphs that Chrysalis generates.
Trinity Parallelized On TACC

- Inchworm/Jellyfish
- Chrysalis
- Quantify Graph
- Butterfly
- Findcat

{ largemem nodes (1 TB RAM) }
An Actual Command

```
login2:~$ assemble_trinity -a DNAdenovo -l R1_reads -r R2_reads -o Fish
```

- allocation
- R1 reads
- R2 reads
- output directory
Future Development

• Set up `assemble_trinity` to work on Stampede.

• Find a way around the time limits in Lonestar and Stampede. (Some ideas, nothing working yet.)