RNAseq

BCB 504: Applied Bioinformatics

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

April 18, 2012

Outline

1. Introduction

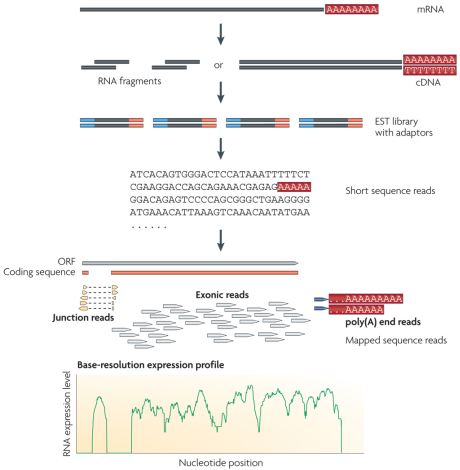
2. de novo RNA assembly
   - Trinity
   - Newbler
   - Fusion transcripts

3. RNA mapping

4. Digital Transcriptomics

Inroduction

RNA-Seq: a revolutionary tool for transcriptomics



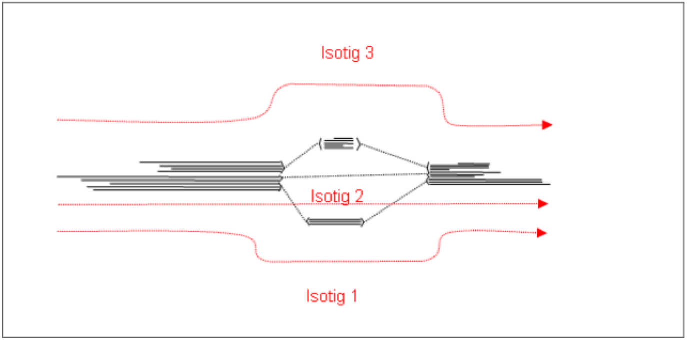
Library Types

1-5% of total RNA is messanger RNA (mRNA)

- poly(A)$^+$ RNA - Oligo (dT)$_{25}$ is used to bind to the polyA tail of mature messenger RNA and other material is washed away. cDNA is generated and adapters ligated.
- rRNA depleted - Ribosomal RNAs are depleted by hybridization of rRNA specific oligos and separated using magnetic beads. cDNA is generated and adapters ligated.
- strand-specific RNAseq - cDNA is generated from mRNA using dUTP, adapters are ligated. The second strand containing dUTP is eliminated through UNG digestion rendering the library directional.

Splice Variation

Ambiguities in transcriptome assembly contigs usually correspond to spliced isoforms, or minor variation among members of a gene family.



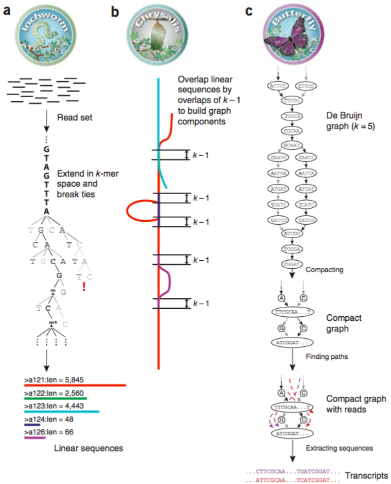
Trinity

Inchworm assembles the RNA-Seq data into transcript sequences, often generating full-length transcripts for a dominant isoform, but then reports just the unique portions of alternatively spliced transcripts.

Chrysalis clusters the Inchworm contigs and constructs complete de Bruijn graphs for each cluster. Each cluster represents the full transcriptional complexity for a given gene (or a family or set of genes that share a conserved sequence). Chrysalis then partitions the full read set among these separate graphs.

Butterfly then processes the individual graphs in parallel, tracing the paths of reads within the graph, ultimately reporting full-length transcripts for alternatively spliced isoforms, and teasing apart transcripts that corresponds to paralogous genes.

Trinity


a
Inchworm
Read set
Extend in k-mer space and break ties
>a121:len = 5,845
>a122:len = 2,560
>a123:len = 4,443
>a124:len = 48
>a126:len = 66
Linear sequences
b
Chrysalis
Overlap linear sequences by overlaps of k – 1 to build graph components
k – 1
k – 1
k – 1
k – 1
k – 1
c
Butterfly
De Bruijn graph (k = 5)
Compacting
Compact graph
Finding paths
Compact graph with reads
Extracting sequences
...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...
Transcripts

Newbler

Isotig (isoform) initiation:

- Contigs lacking reads connecting them to other contigs on one or both ends are used to initiate the traversal of an individual isotig. When a contig has no reads connecting it to any other contigs, it may become an isotig composed of a single contig.
- Spike detection
  1. The alignment depth is at least 10 reads.
  2. A minimum of 20% of the aligned reads must be in the opposite orientation relative to the more abundant orientation of the aligned reads.
  3. A spike may not occur within 10 bases of an already detected spike.
  4. A change in alignment depth between one alignment column and the next of at least 50% signals the location of a "spike".

Newbler I

Isotig (isoform) extension is then done by following reads that join together contigs. Finally, isotigs are terminated by one of the following conditions:

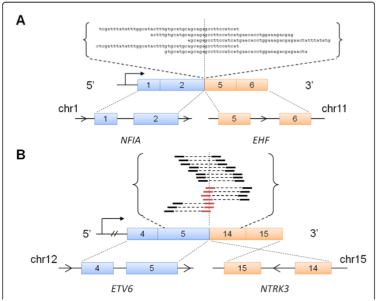
1. No reads are found that extend from the contig currently at the end of the isotig path.
2. The number of reads connecting two contigs is less than 5% of the alignment depth of either.
3. The Isotig Contig Count Threshold is reached. In this case, the further traversal of a particular isotig in an isogroup will be stopped.

Newbler II

4. A contig is reached whose length is below the Isotig Contig Length Threshold. If a contig is reached with a length shorter than this threshold, the further traversal of a particular isotig in an isogroup will be stopped. The contig shorter than the icl threshold will be marked as such and reported in the output files.
5. A cyclic path is encountered. Recursive path traversal will stop if cyclic structures are detected, i.e. revisiting one contig which has already been included in an earlier part of the isotig being traversed. Such cyclic structures will be marked in the output files by assigning cyclic status for the first contig detected.

Fusion transcripts

Fusion transcript is a chimeric RNA encoded by a fusion gene or by two different genes by subsequent trans-splicing. Certain fusion transcripts are commonly produced by cancer cells, and detection of fusion transcripts is part of routine diagnostics of certain cancer types.



Gapped Alignment

Problem: Burrows-Wheeler transform (BWT) facilitate very efficient ungapped alignment of short reads.
Gaps greatly increase the size of the search space and reduce the effectiveness of pruning, thereby substantially slowing aligners built solely on index-assisted alignment.

For RNA, aligners must be able to produce gapped alignments, reads are expected to span introns.

- BarraCUDA
- BWA
- gMAP (gSNAP)
- Mosaik
- Novoalign
- tophat/cufflinks extension to bowtie

Digital Transcriptomics

So the more you can count - and next-generation systems can count a lot - the better the measure of copy number for even those rare transcripts in a population.

- Most techniques deal with Count data.
- Reads are mapped to the reference genome (transcriptome), and the number of reads that map to a gene,
- Read counts for a gene are roughly proportional to the gene's length and transcript abundance.
- technical artifacts then must be considered and the data normalized
  - the sample depth of sequencing
  - GC count (PCR type bias)
  - mappability (uniqueness)

Digital Transcriptomics

- Differential Expression between phenotypes is then determined from Count data.
- Count data is modeled by a discribution (ie. Negative Bionomial Distribution, Poisson, etc.)
- Generally speaking differential expression analysis is performed in a very similar manner to microarrays, once bias and normalization has been performed.