RNA-Seq

Lalit Ponnala
CBSU
What is RNA-Seq

• Massively parallel sequencing method for transcriptome analyses
• Complementary DNA (cDNA) generated from RNA are sequenced using next-generation “short read” technologies
• Reads are aligned to a reference genome and a transcriptome map is constructed
Transcriptome

• The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition

• Understanding the transcriptome is essential for
  – interpreting the functional elements of the genome
  – revealing the molecular constituents of cells, tissues
  – understanding development and disease

Aims of RNA-Seq

- To quantify mRNA abundance
- To determine the transcriptional structure of genes: start sites, 5’ and 3’ ends, splicing patterns
- To quantify the changing expression levels of each transcript during development and under different conditions

Technology

• Single-end, paired-end
• Typically 30-400bp reads
• Popular platforms: Illumina, 454, SOLID
• >10 million reads in a single “lane”
• Alignment tools: Bowtie, BWA, Eland etc
• Additional step: align to exon-junctions
• Automated pipeline for RNA-Seq:
  – **Tophat**: for alignment
  – **Cufflinks**: for calculating expression levels
Sequence data
Analysis Pipeline

1. Raw reads from next-gen sequencing (FASTQ)
2. Reference Genome (FASTA)
3. Genome annotation (GFF)
4. Exon-junction sequences
5. Quantify abundance of reads aligning to exons
6. Total transcript abundance
7. Calculate abundance of reads at each exon-junction
Alignment Issues

• Exon Boundaries
  – What if exon is shorter than read?

• Multiple matches
  – Simple weighting
  – Evidence-based weighting
Units of measurement

• RPKM : Reads per kilobase per million mapped reads
  
  1kb transcript with 1000 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped) will have
  
  \[ \text{RPKM} = \frac{1000}{1 \times 8} = 125 \]

• FPKM : for paired-end sequencing
  
  – A pair of reads constitute one fragment
Tophat

- Aligns sequences to the whole genome AND to exon-junctions
- Uses Bowtie, an ultrafast, memory-efficient short read aligner
- Output reported in SAM format
- Independently aligns segments of each read (default 25bp) allowing up to 2 mismatches
- Does not support indels / gapped alignments

http://tophat.cbcb.umd.edu/index.html
Tophat : junctions

• From supplied annotation file (GFF) or list of junction coordinates

• Without reference annotation
  – Sets of coverage islands : high coverage regions
  – Paired end reads: using genomic distance between mates
  – Segments of same read mapped far apart: “GT-AG” introns
Running Tophat

• **Index the genome:**
  
  ```
  bowtie-build maize_pseudo.fa maize_pseudo
  ```

• **Run tophat:**
  
  ```
  tophat -o zero -G annot.gff --no-novel-juncs maize_pseudo s_1_sequence.txt
  ```

• **Output files:**
  
  – accepted_hits.sam
  
  – annot.juncs
  
  – junctions.bed

Maize genome: [http://maizesequence.org/](http://maizesequence.org/)  
Illumina data: Tom Brutnell, BTI
Viewing the alignments (IGV)

```
samtools faidx maize_pseudo.fa
samtools view -bt maize_pseudo.fa.fai -o accepted_hits.bam accepted_hits.sam
samtools index accepted_hits.bam
```

Cufflinks

- can estimate the abundances of the isoforms present in the sample, using either:
  - a known "reference" annotation
  - an ab-initio assembly of the transcripts
- constructs a set of transcripts that "explain" the reads observed in an RNA-Seq experiment
- Input: alignments in SAM format, annotation in GTF (optional)
- Output: assembled transfrags, genes

http://cufflinks.cbcb.umd.edu/index.html
Cufflinks

• Command line:
  cufflinks -G annot.gtf accepted_hits.sam

• Output files:
  – transcripts.gtf
  – transcripts.expr
  – genes.expr
Cuffdiff

- Differential expression at the transcript “isoform” level and at the gene level

cuffdiff annot.gtf ./zero/accepted_hits.sam ./one/accepted_hits.sam

Examine output file: 0_1_gene_exp.diff
Advantages of RNA-Seq

• Does not require existing genomic sequence
  – Unlike hybridization approaches
• Very low background noise
  – Reads can be unambiguously mapped
• Resolution
  – Up to 1 bp
• High-throughput
  – Better than Sanger sequencing of cDNA or EST libraries
• Cost
  – Lower than traditional sequencing
• Can reveal sequence variations (SNPs)
Issues

Mismatch rate vs base position
Issues

• Depth of coverage depends on “sequenceability” of the genomic region
Conclusion

RNA-Seq

• Offers high-throughput quantitative measurement of transcript abundance
• Expression levels correlate well with qPCR
• Costs continue to fall due to multiplexing
• Expected to replace microarrays for transcriptomic studies
• Automated pipeline (Tophat/Cufflinks)
References
