Canadian Bioinformatics Workshops

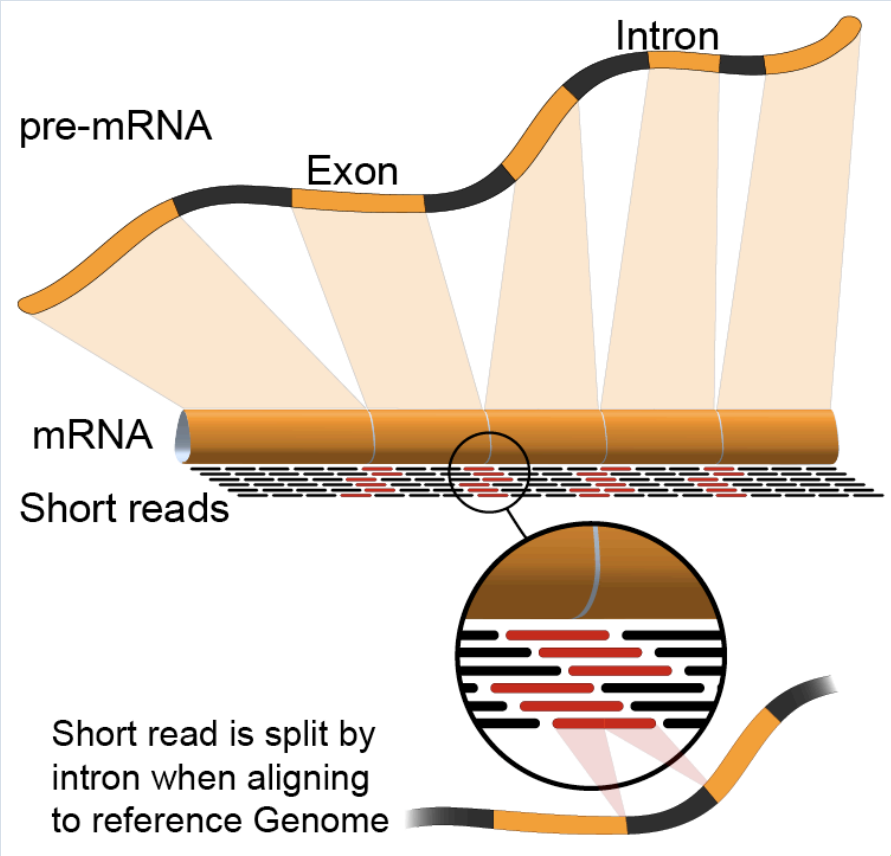
www.bioinformatics.ca

Module 2
RNA-seq alignment and visualization (tutorial)

Malachi Griffith & Obi Griffith

Informatics for RNA-seq Analysis

June 8-9, 2015



bioinformatics.ca

Learning Objectives of Tutorial

- Run Bowtie2/TopHat2 (or STAR) with parameters suitable for gene expression analysis
- Use samtools to demonstrate the features of the SAM/BAM format and basic manipulation of these alignment files (view, sort, index, filter)
- Use IGV to visualize RNA-seq alignments, view a variant position, etc.
- Determine BAM-read counts at a variant position
- Use samtools flagstat, samstat, FastQC to assess quality of alignments

3-i. Align reads with tophat

- Align all reads in the 6 libraries of the test data
  - 6 libraries with two files each (one for each read1 and read2 of the paired-end reads)
- Use tophat for the alignment
  - Supply the gene GTF file obtained in step 3
  - Supply the bowtie indexed genome obtained in step 4
  - The '-G' option tells tophat to look for the exon-exon junctions of known transcripts. It will still look for novel exon-exon junctions as well
- Since there are 6 libraries in the test data set, 6 alignment commands are run
- On a test system, each of these alignments took ~1.5 minutes using 8 CPUs
- Each alignment job outputs a SAM/BAM file
  - http://samtools.sourceforge.net/SAM1.pdf

3-i. Align reads with STAR

- Again, align all reads in the 6 libraries of the test data, now with STAR
  - Supply the same gene GTF file obtained in step 3
  - Supply the STAR indexed genome obtained in step 4
  - The '-outSAMstrandField intronMotif' is needed so that STAR produces an alignment compatible with cufflinks
- How long did the alignment take compared to tophat?
- What additional steps are needed?

3-ii. Post-alignment visualization

- Create indexed versions of bam files
  - These are needed by IGV for efficient loading of alignments
- Visualize spliced alignments
  - Identify exon-exon junction supporting reads
  - Identify differentially expressed genes
  - Compare tophat and STAR alignments
- Try to find variant positions

- Create a pileup from bam file
- Determine read counts at a specific position

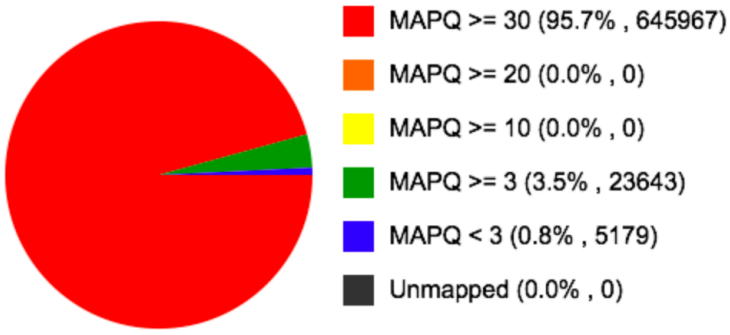
3-ii. Post-alignment visualization (IGV)

3-iii. Post-alignment QC

- Use 'samtools view' to see the format of a SAM/BAM alignment file
  - Use ‘FLAGs’ to filter out certain kinds of alignments
- Use 'samtools flagstat' to get a basic summary of an alignment
- Run samstat on Tumor/Normal BAMs and review the resulting report in your browser
- Use FastQC to perform basic QC of your alignments

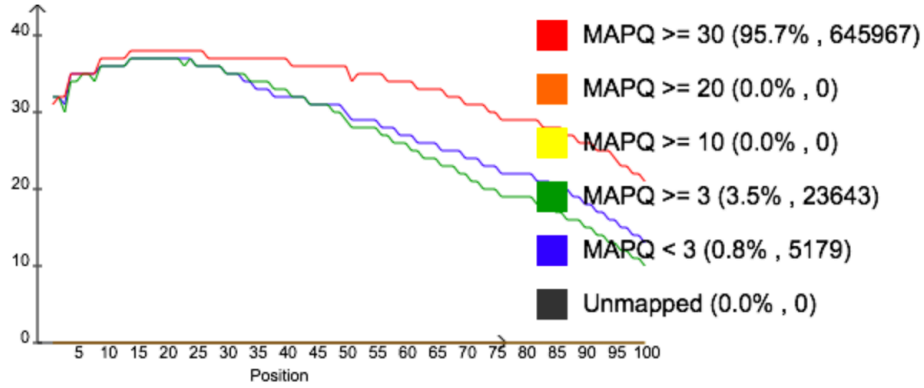
3-iii. Post-alignment QC (samstat)

Mapping stats: 100% aligned (0.7M aligned out of 0.7M total)



Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences. The percentage and number of alignments in each category is given in brackets.

Mean Base Quality



Mean base quality of reads with low and high mapping quality.

We are on a Coffee Break & Networking Session