Canadian Bioinformatics Workshops

www.bioinformatics.ca

Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Free Cultural Works **APPROVED FOR**

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

Learn how to distribute your work using this licence

RNA-Seq Module 2
Alignment and Visualization (tutorial)

Malachi Griffith, Obi Griffith, Fouad Yousif
Informatics for RNA-seq Analysis
June 16 - 17, 2016
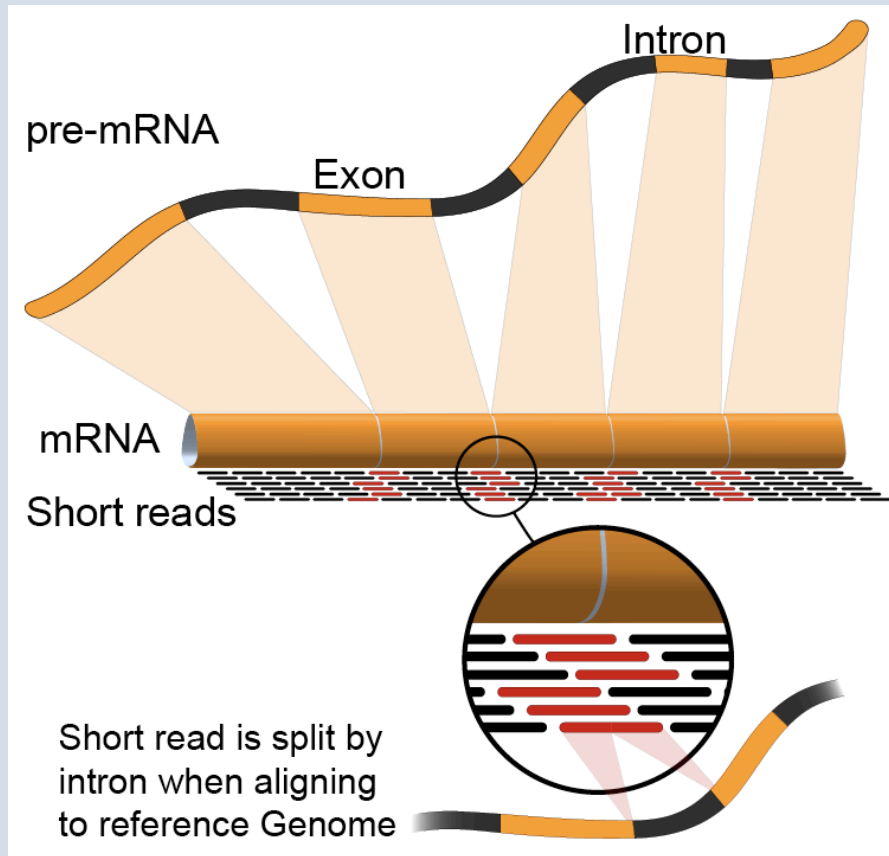
Learning Objectives of Tutorial

- 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

2-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

2-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?

2-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

bioinformatics.ca
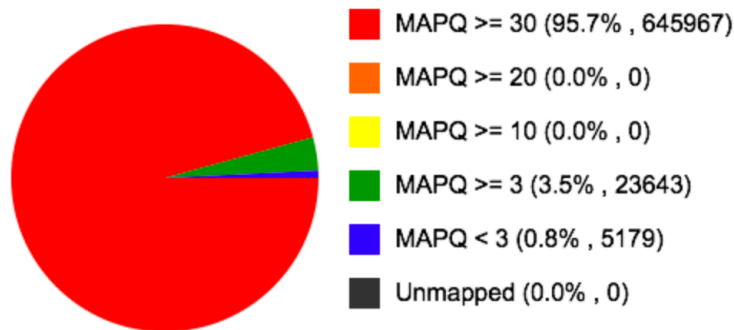
2-ii. Post-alignment visualization (IGV)

2-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
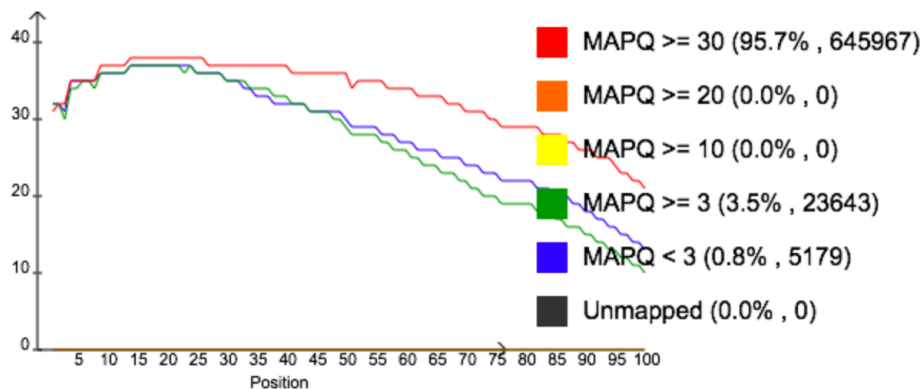- Use FastQC to perform basic QC of your alignments

bioinformatics.ca

2-iii. Post-alignment QC (samstat)

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

- MAPQ >= 30 (95.7% , 645967)
- MAPQ >= 20 (0.0% , 0)
- MAPQ >= 10 (0.0% , 0)
- MAPQ >= 3 (3.5% , 23643)
- MAPQ < 3 (0.8% , 5179)
- Unmapped (0.0% , 0)

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

- MAPQ >= 30 (95.7% , 645967)
- MAPQ >= 20 (0.0% , 0)
- MAPQ >= 10 (0.0% , 0)
- MAPQ >= 3 (3.5% , 23643)
- MAPQ < 3 (0.8% , 5179)
- Unmapped (0.0% , 0)

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