Canadian Bioinformatics Workshops

www.bioinformatics.ca

Creative Commons

This page is available in the following languages:
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

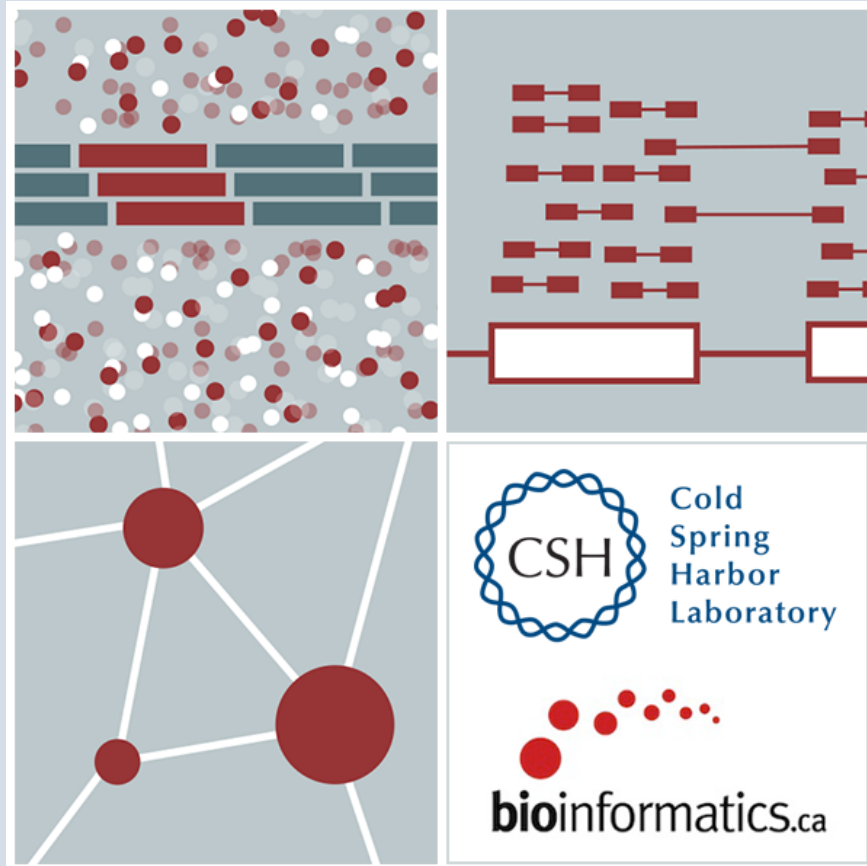Learn how to distribute your work using this licence

RNA-Seq Module 2
Alignment and Visualization (tutorial)

Malachi Griffith, Obi Griffith, Fouad Yousif
High-Throughput Biology: From Sequence to Networks
March 20-26, 2017

Learning Objectives of Tutorial

- Run HISAT2 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

bioinformatics.ca

2-i. Adaptor trim

- Use Flexbar to trim sequence adapter from the read FASTQ files
  - The output of this step will be trimmed FASTQ files for each data set.
- Compare the FastQC reports for fastq files before and after trimming
- http://sourceforge.net/projects/flexbar/

2-ii. Align reads with HISAT2

- 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 HISAT2 for the alignment
  - Supply the bowtie indexed genome obtained in section 1-iv
  - The '-dta' option tells HISAT2 to report alignments tailored for transcript assemblers
- Since there are 6 libraries in the test data set, 6 alignment commands are run
- On a test system, each of these alignments took ~4 seconds using 8 CPUs
- Each alignment job outputs a SAM/BAM file
  - http://samtools.sourceforge.net/SAM1.pdf

2-iii. 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
- Try to find variant positions

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

bioinformatics.ca

2-iii. Post-alignment visualization (IGV)

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

- Optional: explore RSeQC for alignment QC

2-iv. Post-alignment QC (RSeQC)
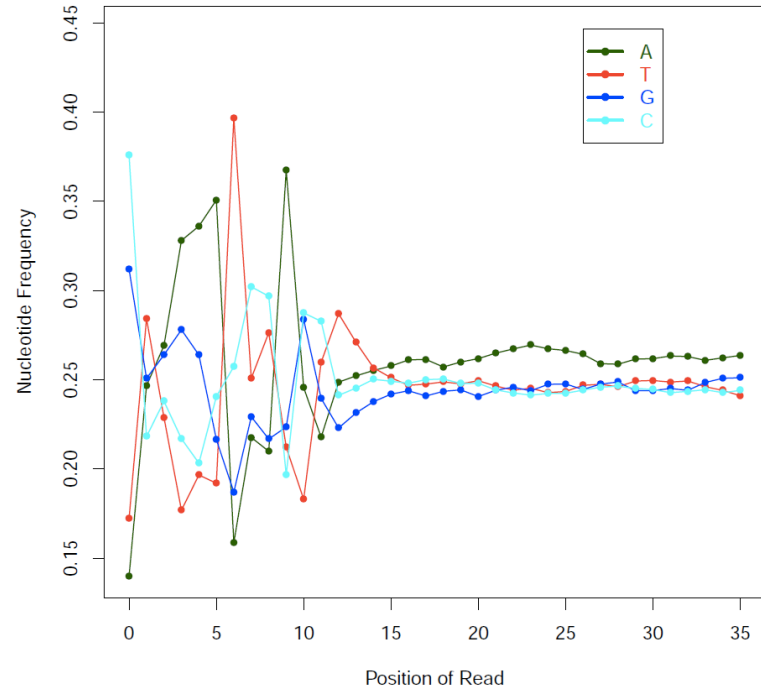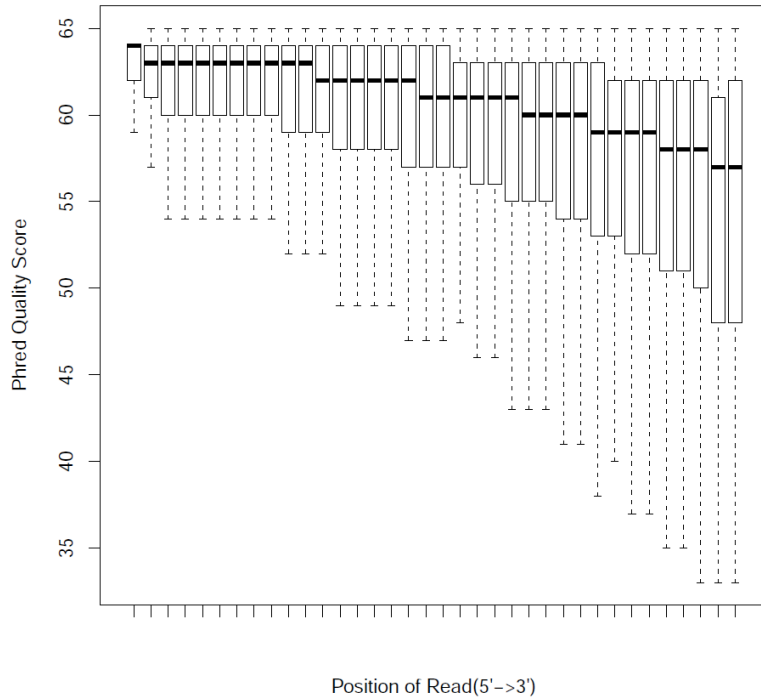
We are on a Coffee Break & Networking Session