

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

Supported by



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 slovenščina jezik srpski (latinica) Sotho svenska
中文 漢語 (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





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](#)

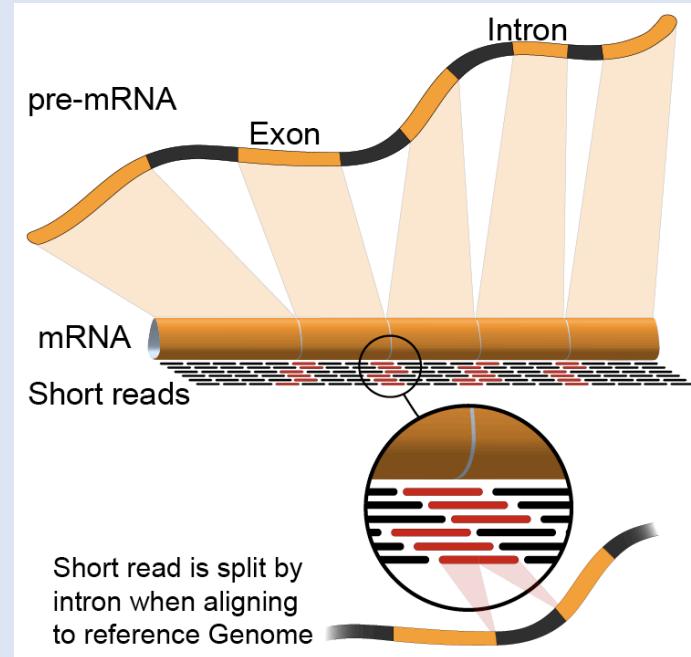
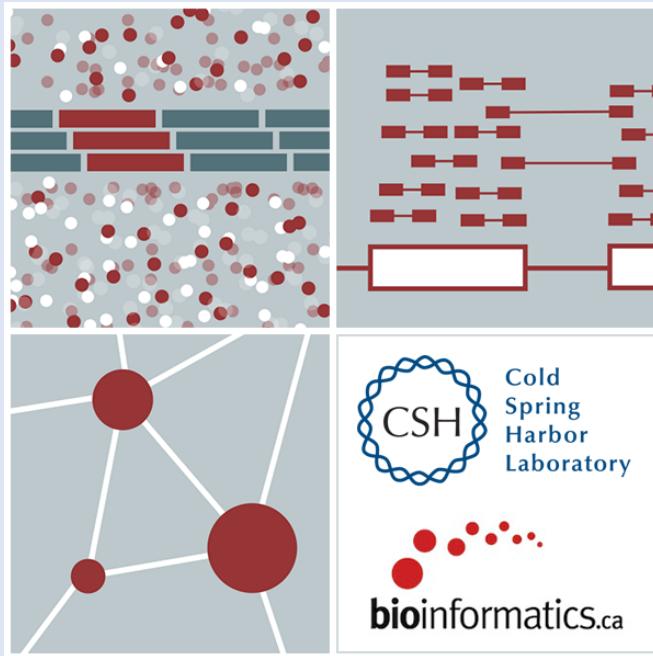
Learn how to distribute your work using this licence

Differential Expression

Kelsy Cotto, Obi Griffith, Malachi Griffith, Saad Khan, Allegra Petti, Huiming Xia

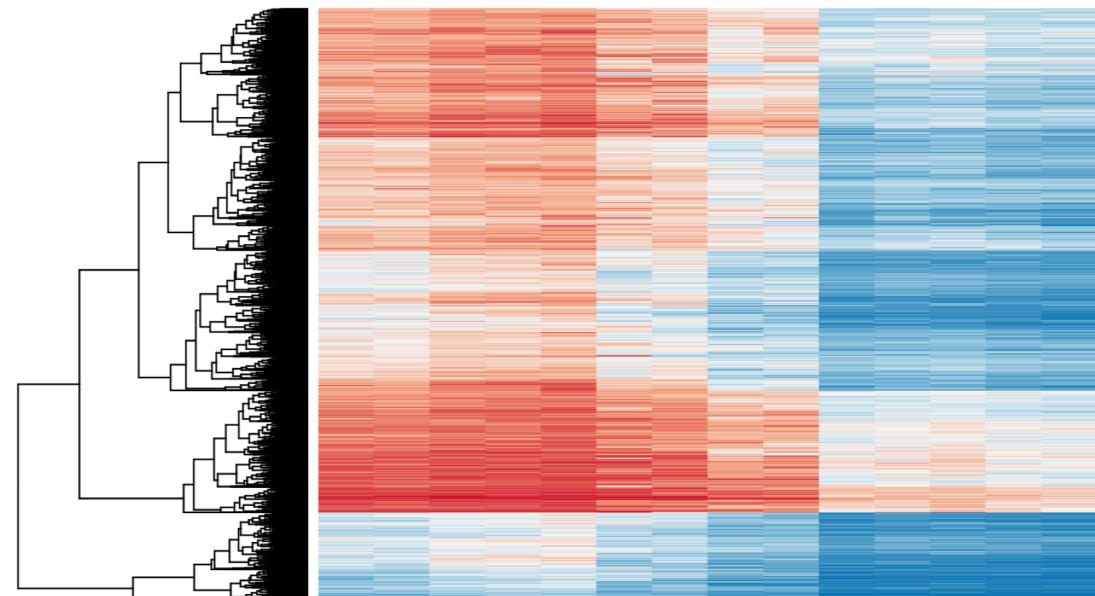
Informatics for RNA-seq Analysis

June 17-19, 2020



Differential Expression

- Tying gene expression back to genotype/phenotype
- What genes/transcripts are being expressed at higher/lower levels in different groups of samples?
 - Are these differences 'significant', accounting for variance/noise?
- Examples (used in course):
 - UHR cells vs HBR brain
 - Tumor vs Normal tissue
 - Wild-type vs gene KO cells



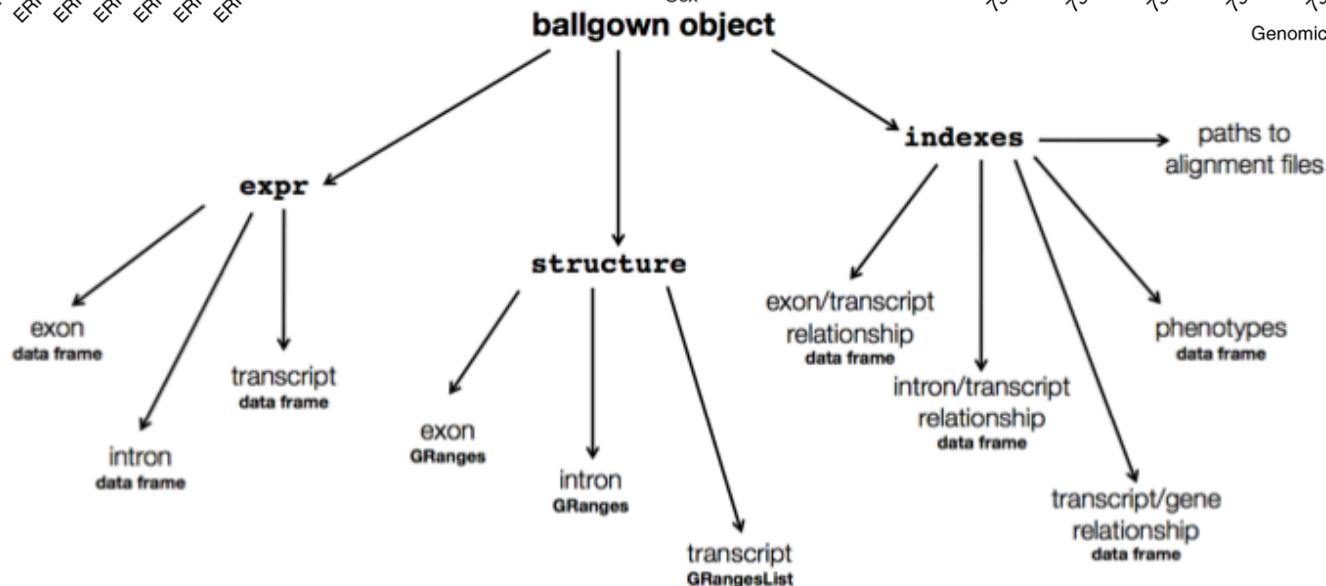
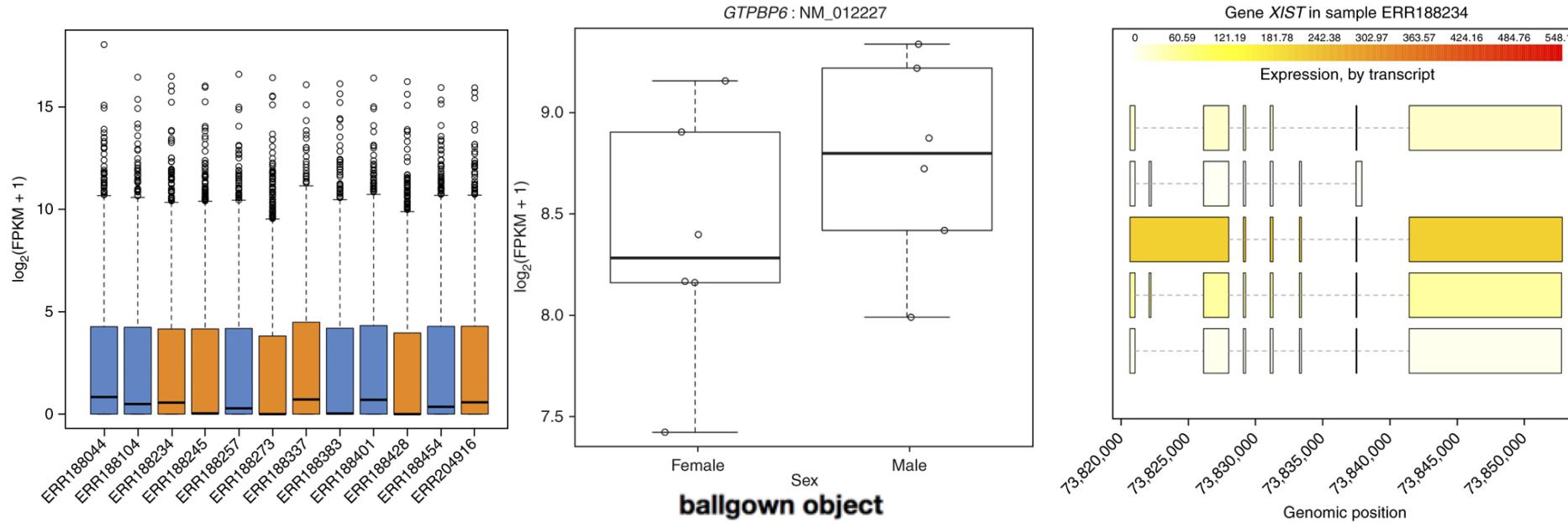
Differential Expression with Ballgown

Parametric F-test comparing nested linear models

- Two models are fit to each feature, using expression as the outcome
 - one including the covariate of interest (e.g., case/control status or time) and one not including that covariate.
- An F statistic and p-value are calculated using the fits of the two models.
 - A significant p-value means the model including the covariate of interest fits significantly better than the model without that covariate, indicating differential expression.
- We adjust for multiple testing by reporting q-values:
 - $q < 0.05$ the false discovery rate should be controlled at $\sim 5\%$.

[Frazee et al. \(2014\)](#)

Ballgown for Visualization with R



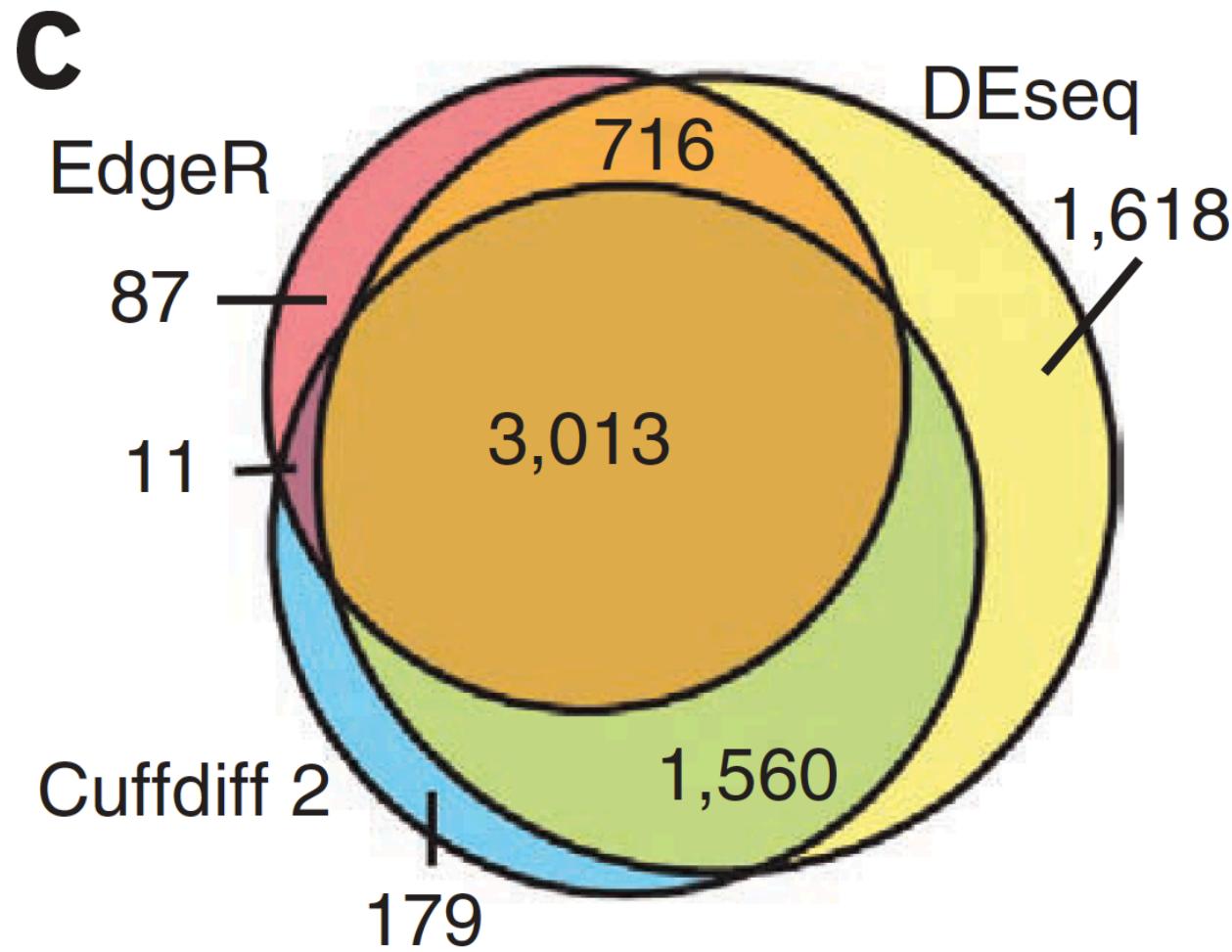
Alternative differential expression methods

- Raw count approaches
 - DESeq2 - <http://www-huber.embl.de/users/anders/DESeq/>
 - edgeR - <http://www.bioconductor.org/packages/release/bioc/html/edgeR.html>
 - Others...

'FPKM/TPM' expression estimates vs. 'raw' counts

- Which should I use?
 - Long running debate, but the general consensus:
- FPKM/TPM
 - When you want to leverage benefits of tuxedo suite
 - Isoform deconvolution
 - Good for visualization (e.g., heatmaps)
 - Calculating fold changes, etc.
- Counts
 - More robust statistical methods for differential expression
 - Accommodates more sophisticated experimental designs with appropriate statistical tests

Multiple approaches advisable



Lessons learned from microarray days

- Hansen et al. “Sequencing Technology Does Not Eliminate Biological Variability.” *Nature Biotechnology* 29, no. 7 (2011): 572–573.
- Power analysis for RNA-seq experiments
 - <http://euler.bc.edu/marthlab/scotty/scotty.php>
- RNA-seq need for biological replicates
 - <http://www.biostars.org/p/1161/>
- RNA-seq study design
 - <http://www.biostars.org/p/68885/>

Multiple testing correction

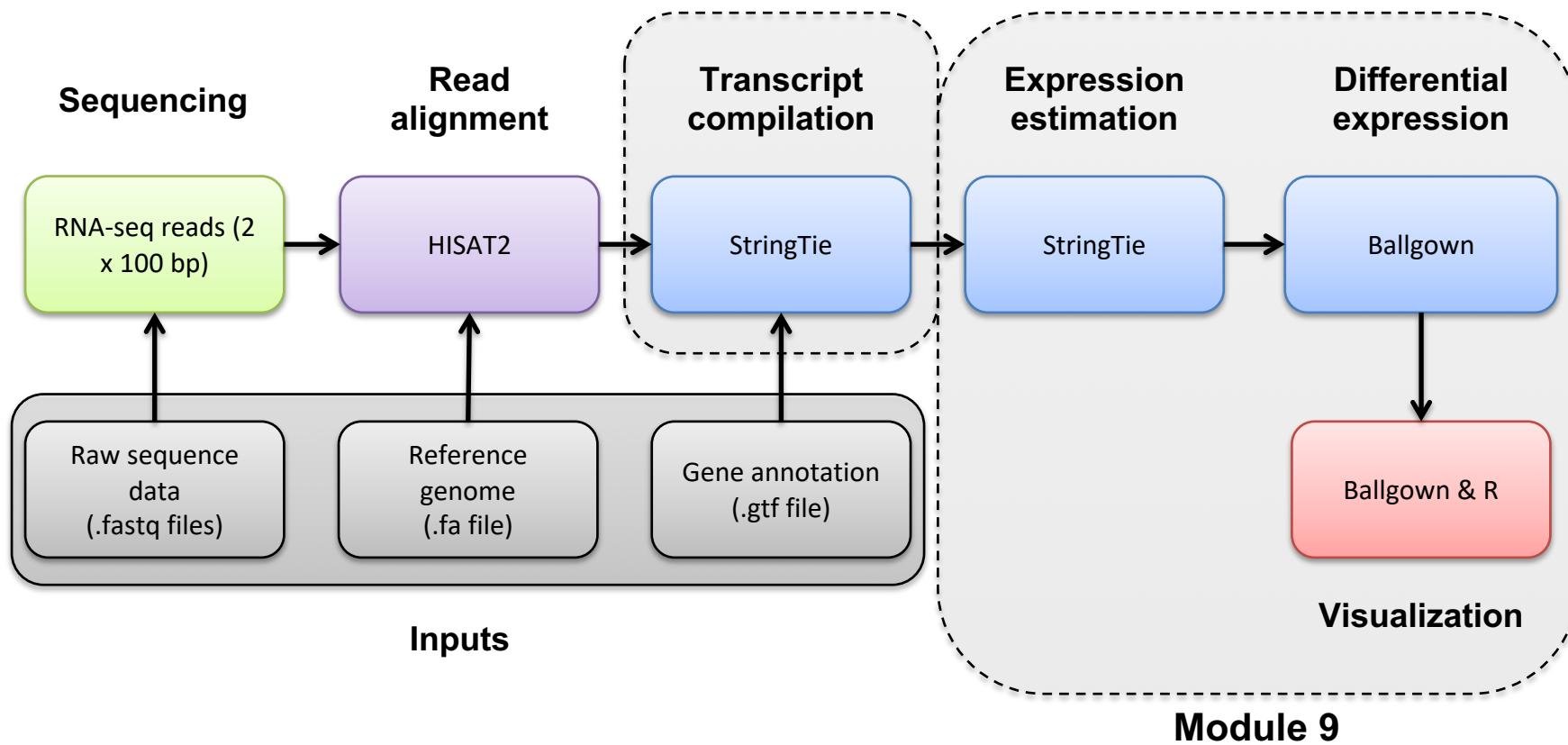
- As more attributes are compared, differences due solely to chance become more likely!
- Well known from array studies
 - 10,000s genes/transcripts
 - 100,000s exons
- With RNA-seq, more of a problem than ever
 - All the complexity of the transcriptome gives huge numbers of potential features
 - Genes, transcripts, exons, junctions, retained introns, microRNAs, lncRNAs, etc
- Bioconductor multtest
 - <http://www.bioconductor.org/packages/release/bioc/html/multtest.html>

Downstream interpretation of expression analysis

- Topic for an entire course
- Expression estimates and differential expression lists from StringTie, Ballgown or other alternatives can be fed into many analysis pipelines
- See supplemental R tutorial for how to format expression data and start manipulating in R
- Clustering/Heatmaps
 - Provided by Ballgown
 - For more customized analysis various R packages exist:
 - hclust, heatmap.2, plotrix, ggplot2, etc.
- Classification
 - For RNA-seq data we still rarely have sufficient sample size and clinical details but this is changing
 - Weka is a good learning tool
 - RandomForests R package (biostar tutorial being developed)
- Pathway analysis
 - GSEA, IPA, Cytoscape, many R/BioConductor packages:
<http://www.bioconductor.org/help/search/index.html?q=pathway>

https://genviz.org/module%204/0003/12/31/Expression_Profiling_and_Visualization/

HISAT2/StringTie/Ballgown RNA-seq Pipeline



We are on a Coffee Break & Networking Session

Workshop Sponsors:

compute | calcul
canada | canada



Canadian Centre for
Computational
Genomics

MicCM McGill initiative in
Computational Medicine