bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io


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
APPROVED FOR
Works
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

Basic Differential Expression Analysis

Shraddha Pai

Analysis Using R

June 14, 2024

bioinformatics.ca

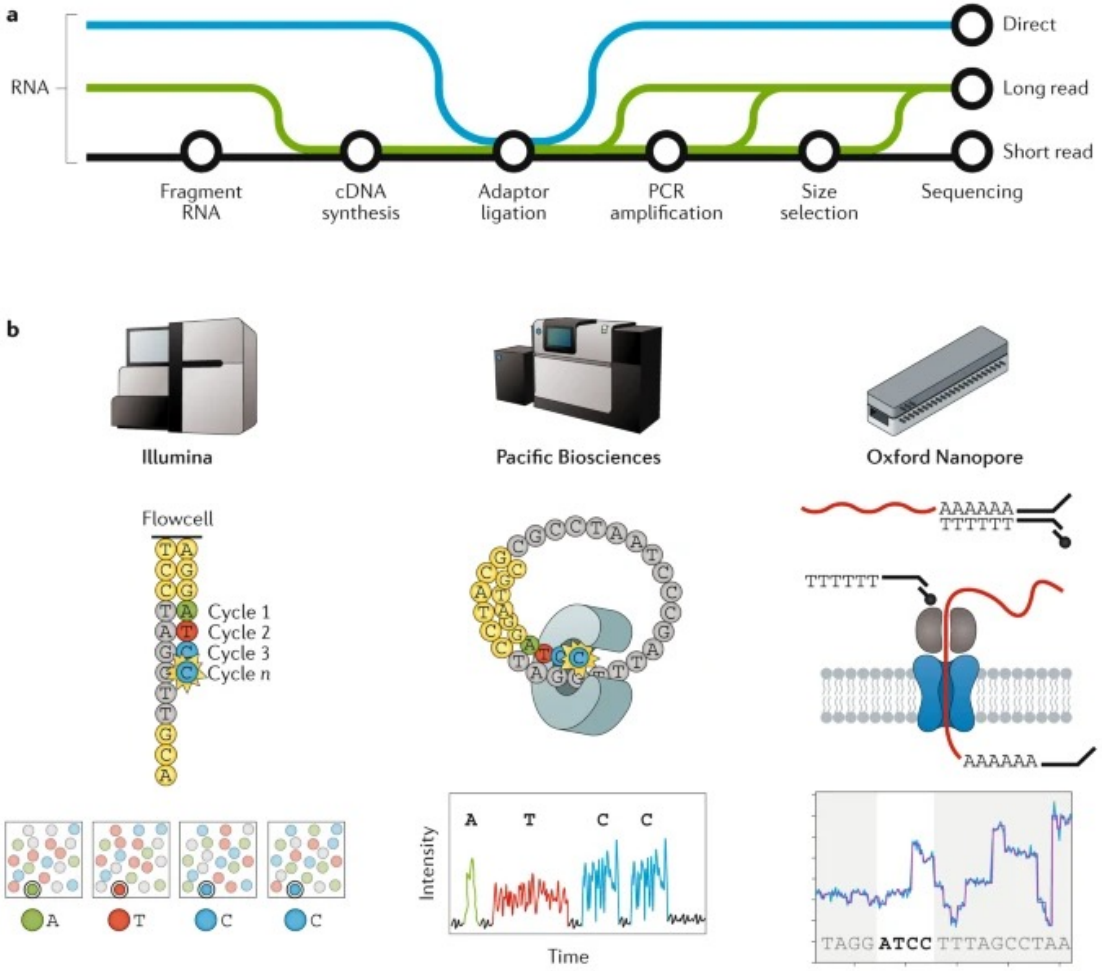
OICR
Ontario Institute
for Cancer Research

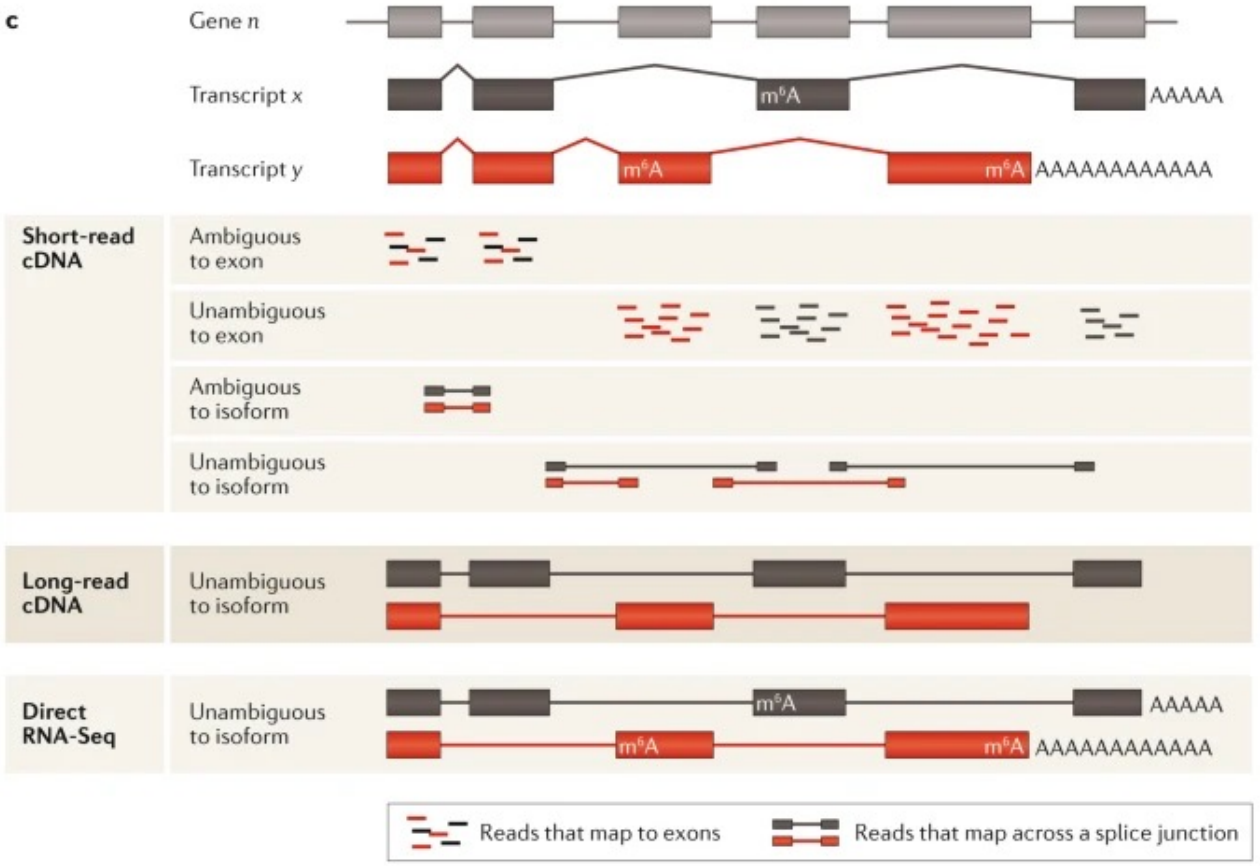
UNIVERSITY OF
TORONTO

Learning Objectives

- By the end of this lecture, you will:
  - Understand the key steps in identifying differentially expressed genes in RNAseq
  - Learn how to use p-value histograms and QQ-plots to gauge how much signal you have after multiple hypothesis testing
  - Learn to create volcano plots to visualize results of differential expression analysis

RNAseq data generation





Stark, Grzelak, Hadfield. (2019). *Nat Rev Gen.*

RNAseq data processing: High-level overview

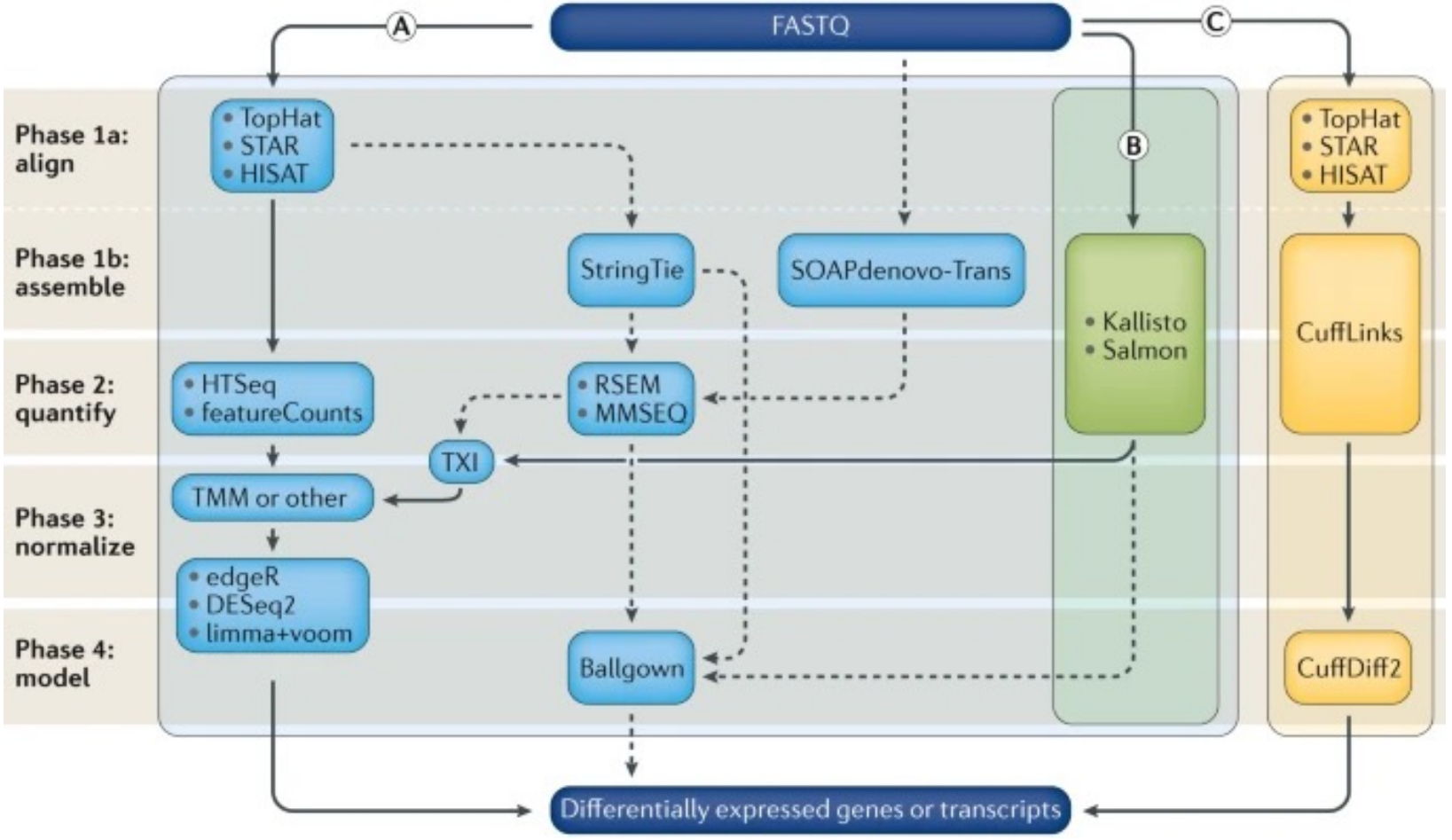
Raw reads

Align reads to genome

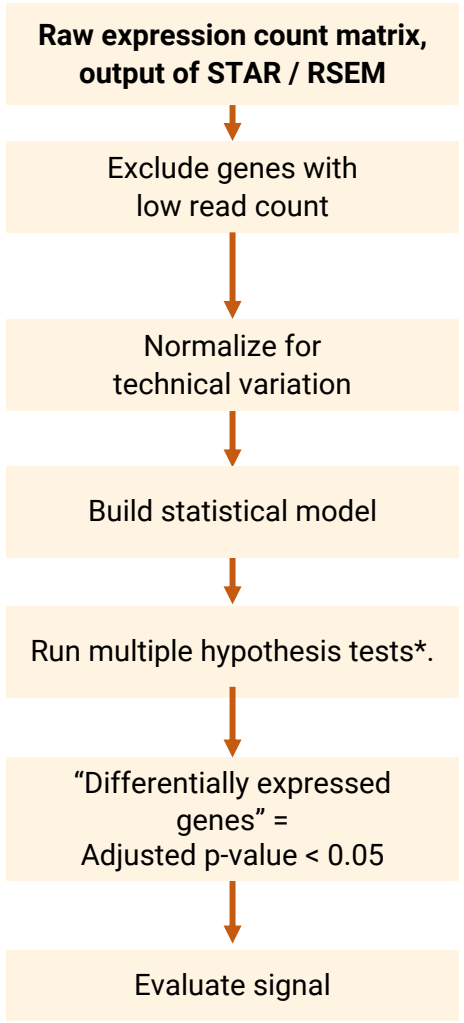
Obtain transcript counts

Normalize

Statistical model for differential gene expression analysis



RNA-seq analysis



~ 20,000 measures (protein-coding genes) → 20,000 individual statistical tests

Multiple testing burden

20,000 p-values

Reduce multiple testing burden

Variation in sequencing depth

$log(xpr) = \beta_0 + \epsilon$* *null model*

$log(xpr) = \beta_0 + \beta_1(disease) + \epsilon$ *full model*

} Likelihood ratio test (LRT)

Negative binomial distribution *(class of generalized linear models)*

Signal in count data

Correct for multiple tests

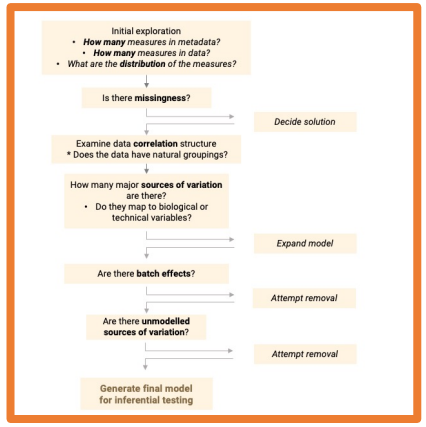
* $\epsilon \sim log(Normal)$

Let’s look at a worked example for RNAseq analysis.

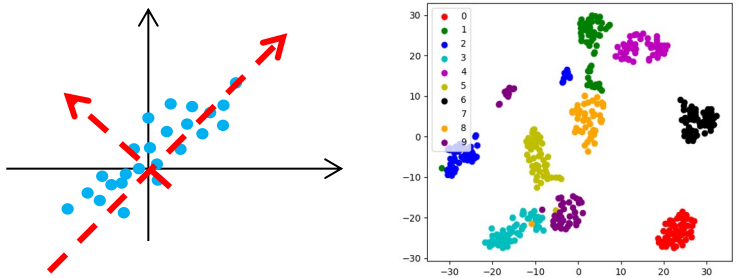
Exercise time.

Recap course

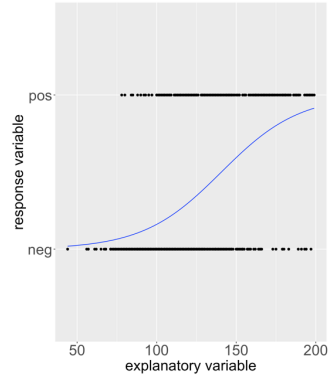
Module 1: Systematic **exploratory data analysis**



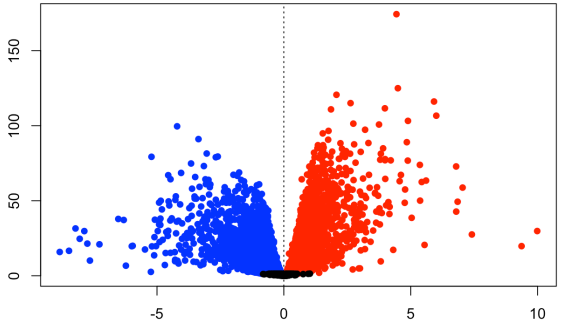
Module 2: **Dimensionality reduction** to identify major sources of variation in your data



Module 3: **Generalized linear models** to fit binary response variables (and RNAseq data!)



Module 4: **Differential expression analysis**, multiple hypothesis testing



Enjoy exploring your data!

Workshop Sponsors: