INTRODUCTION TO TRANSCRIPTOMIC AGE ANALYSIS

Welcome to our second event!

MCGILL MEDICAL STUDENTS' GENOMICS GROUP

- **Why did we start this, and why might this be helpful to you?**
  - We want to provide medical students with a working understanding of genomics and opportunities for hands-on training in genomic research, computational biology and bioinformatics.

- **Who are we?**
  - Richie Jeremian (MDCM '24)
  - Marc Henein (MDCM '24)
  - Misha Fotovati (MDCM '26)
  - Nicole Zhang (MDCM/PhD '30)

WHAT KIND OF QUESTIONS CAN WE ANSWER WITH GENOMICS?

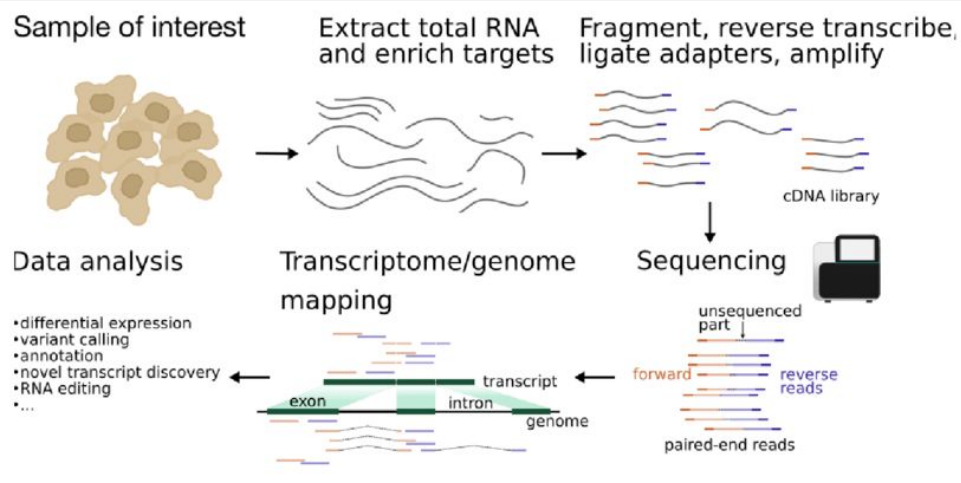
Which **genes/genetic markers** (e.g. SNPs) are associated with a disease of interest?

How are genes differentially expressed or epigenetically modified **compared to healthy control groups** or in **response to a certain treatment**?

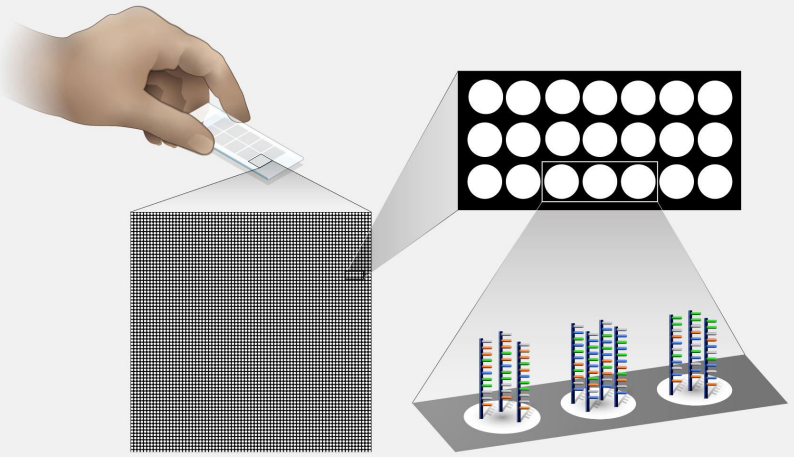
- **GWAS =** Genome Wide Association Study
- **SNP =** Single Nucleotide Polymorphism
- **mRNA =** RNA that is necessary for protein production
- **DNAm** = Epigenetic modifications to DNA that influence gene expression

DNA TECHNOLOGY

RNA Sequencing: Used to quantify the levels of gene expression in a sample, identifying the number and type of genes expressed in a particular tissue or disease.



SNP Microarray: Involves a small slide with thousands to millions of DNA fragments with known sequence. By hybridizing fluorescently labeled DNA to the microarray, one can genotype individuals at >100K SNP sites.



HYPOTHESIS TESTING & P-VALUE

- The null hypothesis (H0) is a specific hypothesis that we try to disprove.
- The alternative hypothesis (Ha) claims "H0 is false".

So what measure do we use to accept or reject a hypothesis?

The p-value! It measures the probability of obtaining the observed result or a more extreme result, assuming that the null hypothesis is true.

Type I error (α) = probability of rejecting H0 when it is true

Type II error (β) = probability of accepting H0 when it is false

A p-value of α=0.05 or lower is considered statistically significant. However, we need to adjust for the number of tests performed.

A study will have greater **power (1-β)** if we increase the sample size or add more assumptions to the null hypothesis.

Outline

1. Introduction to transcriptomic age & analysis
2. Demonstration of transcriptomic age analysis in Behçet's disease

R Basics

R is a programming language for statistical computing and graphics.

Let's go to R Studio!

- Posit Link: https://posit.cloud/
- GitHub: https://github.com/mss-genomics/first-meeting

Behçet's Disease

Behçet's disease (BD) is a rare inflammatory condition that causes vasculitis throughout the body and has numerous (oral, mucosal, GI, genital, etc) manifestations. Its etiology is poorly understood.

Using a publicly available transcriptome dataset from the Gene Expression Omnibus (GSE205867), we investigate transcriptomic age differences in neutrophils derived from **BD patients** and **healthy controls**.

Transcriptomic Age

Estimated age calculation based on the expression patterns of several ‘key’ genes.

Numerous transcriptomic age algorithms exist, trained on a variety of healthy donor tissues, generated using elastic net regression. Each clock incorporates distinct training data and ‘key’ genes.

Dysregulated transcriptomic age (ie. different from ‘true’ chronological age) may be associated with disease states (eg. inflammatory disease, cardiovascular disease, cancer).

Overview of Workflow

1. Download and import dataset https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE205867
2. Calculate transcriptomic age for 7 clock algorithms
3. Determine 'residual' from calculated age (ie. a measure of age dysregulation)
4. Perform Pearson's correlation and T-test on calculated outputs
5. Visualize data using principal component analysis, scatterplots (Pearson's correlation) and boxplots (T-test)

Additional Resources

Background on RNAAgeCalc package and transcriptomic age analysis

1. https://bioconductor.org/packages/release/bioc/vignettes/RNAAgeCalc/inst/doc/RNAAge-vignette.html
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7402472/

Learning Modules

1. R/RStudio: https://moderndive.netlify.app/1-getting-started.html
2. ggplot2: https://ggplot2-book.org/introduction.html

General Stats Learning:

1. StatQuest: https://www.youtube.com/watch?v=tlf6wYJrwKY

Thank You

Please fill out this form if you have a few minutes to help us improve and let us know what went well!

https://forms.gle/YutD2TDV1Ur8KNwG9

Some ideas for future events:

- Precision Medicine
- Finding datasets and generating research questions
- Statistical analysis basics
- Making plots/introduction to ggplot2
- Your idea here!