INTRODUCTION TO R & BIOINFORMATICS

Welcome to our first 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)

WHAT KIND OF QUESTIONS CAN WE ANSWER WITH GENOMICS?

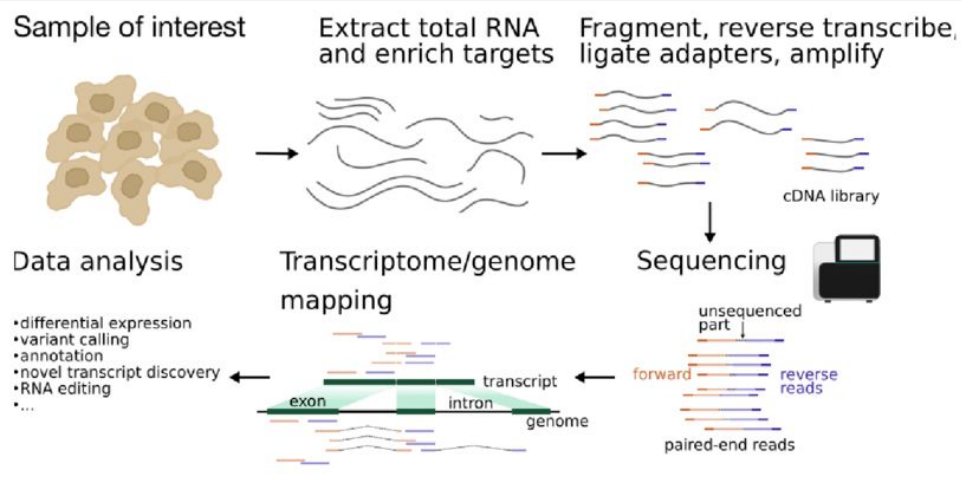
Which **genes/genetic markers** (e.g. SNPs) are associated with a disease of interest?

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

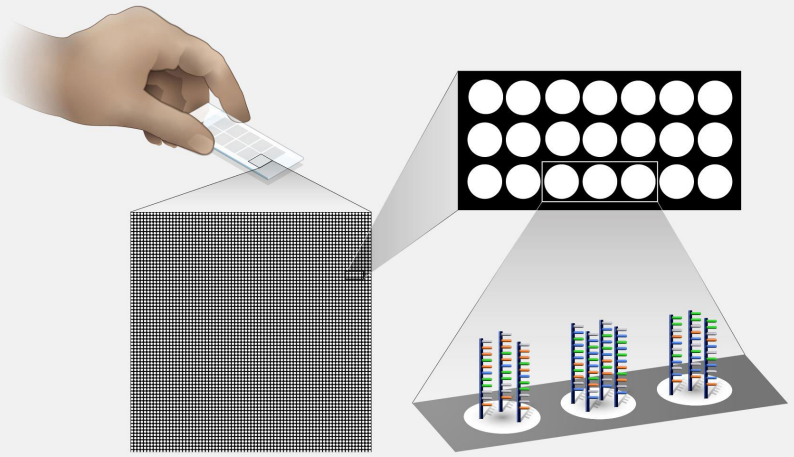
- **GWAS =** Genome Wide Association Study
- **SNP =** Single Nucleotide Polymorphism
- **mRNA =** RNA that is necessary for protein production

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. R basics
2. Genetic association study for late-onset Alzheimer disease
3. Gene expression profiling of atopic dermatitis

Part 1: 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

Part 2: Genetics of Alzheimer Disease

Early-onset Alzheimer disease (age <65) has a substantial component with autosomal **dominant** inheritance, due to mutations in the genes *PSEN1*, *PSEN2* and *APP*.

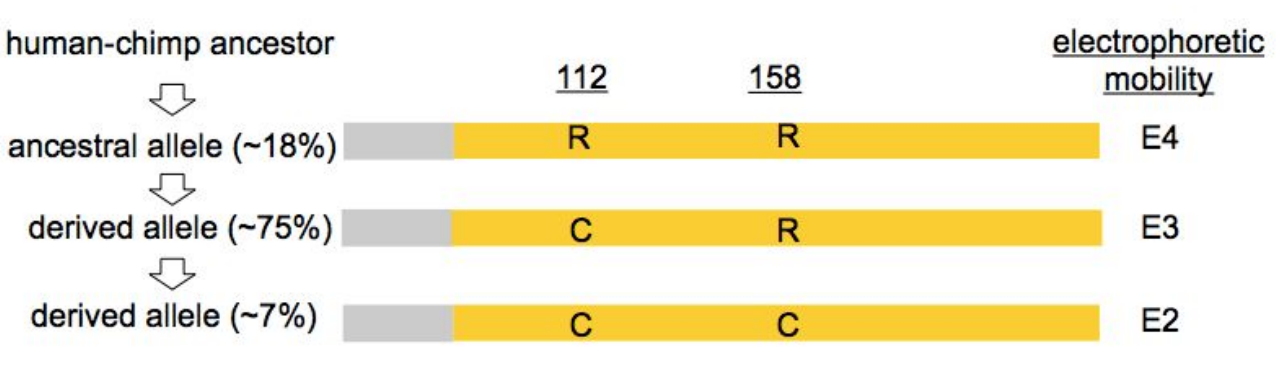
Late-onset Alzheimer disease is a **complex** disease with environmental and genetic risk factors. The broad-sense **heritability** of LOAD, i.e., the correlation between monozygotic twins raised independently*, is **between 0.4 and 0.8**.

*correcting for shared maternal environment

APOE

ApoE is a component of lipoprotein particles and can be found in Aβ plaques.

Among common variants, the E4 allele (C112R) of *APOE* is the strongest genetic risk factor for LOAD.



Genotype	OR
E4/E4	14.9
E3/E4	3.2
E3/E3	1.0
E2/E3	0.6

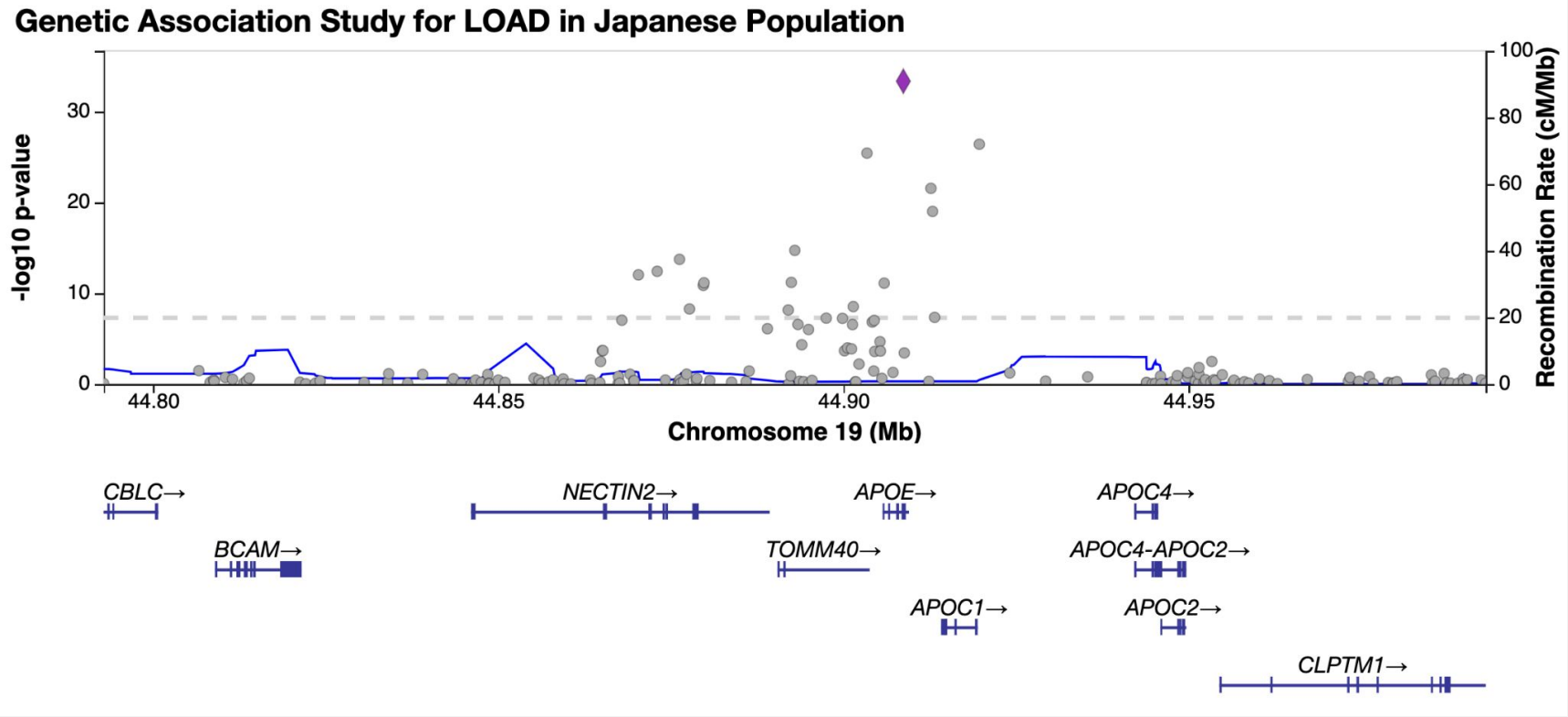
Genetic Association Study for LOAD around the *APOE* Locus in a Japanese Population

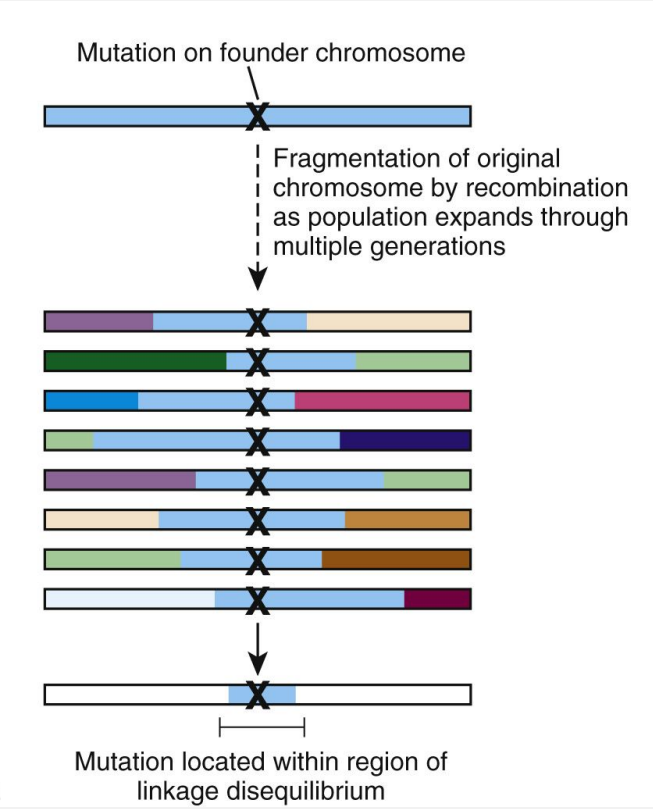
Using a dataset of 547 patients with LOAD and 715 controls genotyped in a 200 kb region including *APOE*, we illustrate the principles of genetic association studies.

Let's go to R Studio!

https://rpubs.com/mss_genomics/1031330

Linkage Peak Around *APOE* E4





Part 3: Gene Expression Profiling of Atopic Dermatitis

Atopic dermatitis (eczema) is a chronic, inflammatory skin disease characterized by immune dysregulation driven by a type 2 inflammatory phenotype.

Using a publicly available dataset from the Gene Expression Omnibus (GSE224783), we investigate gene expression differences in **chronic skin lesions** vs. **non-lesional skin**.

Overview of Workflow

1. Download and import dataset https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE224783
2. Normalize (log2 transform) dataset
3. Visualize data using principal component analysis and hierarchical clustering plots
4. Use DESeq2 package to perform differential expression analyses

Additional Resources

Full video tutorial for RNAseq normalization and differential expression

1. https://www.youtube.com/watch?v=5z_1ziS0-5w
2. https://www.youtube.com/watch?v=ZjMfiPLuwN4

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:

- Making plots/Introduction to ggplot2
- Finding datasets and generating research questions
- Statistical analysis basics
- Epigenetics
- Single cell genomics
- Whole exome sequencing

References

https://www.genome.gov/genetics-glossary/Microarray-Technology

Takei, Norihiro, et al. "Genetic association study on in and around the APOE in late-onset Alzheimer disease in Japanese." *Genomics* 93.5 (2009): 441-448.

https://www.cureffi.org/2016/03/02/what-do-we-know-about-apoe/

Thompson, Robert L., and Margaret W. Thompson. Genetics in Medicine. 8th ed., Elsevier, 2016.