Lecture 6
Introduction to GWAS (with examples using R)

CREIGS 2020

Lewis E Tomalin, PhD
Assistant Professor of Biostatistics

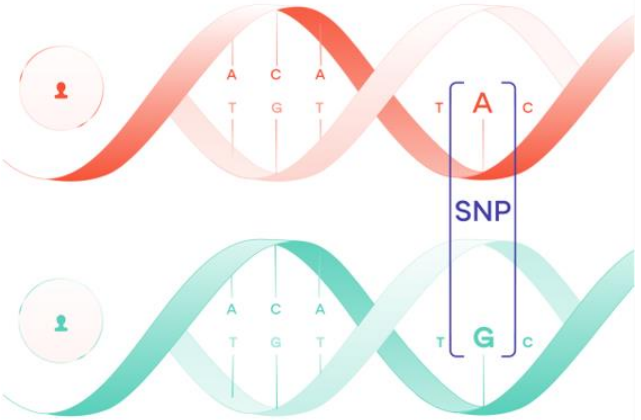
Icahn School of Medicine at Mount Sinai
Department of Population Health Science & Policy

Mount Sinai

Lecture Overview

1. **Part 1: Introduction to GWAS**
    1. What is GWAS?
    2. Basic GWAS methodology
    3. Interpreting GWAS results
2. **Part 2: Basic GWAS in R, key points**
    1. Installing packages
    2. Looking at the data
    3. Basic QC
    4. Fitting Regression models

What is GWAS?



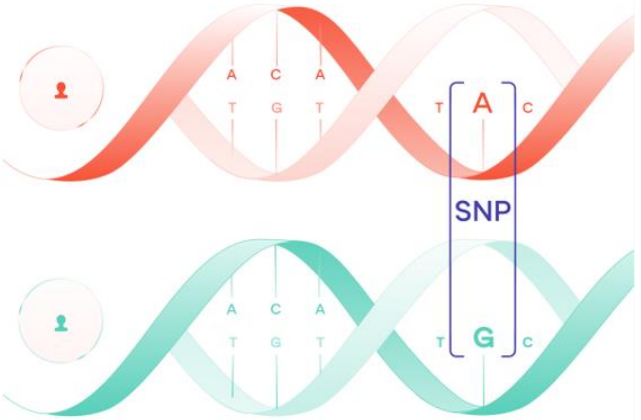
Genome-wide association studies (GWAS) use the genetic profile of many individuals to find significant associations between phenotypes and SNPs across the entire genome.

Key Terminology:

- **Phenotype:** A biological characteristic/trait (eg height, eye color, disease)
- **SNP:** Single Nucleotide Polymorphism, polymorphisms involving variation of a single base pair.

GWAS often uses genetic profiles from a disease group and a control group, with the aim of finding disease markers and novel drug targets.

Basic GWAS methodology



The Data

Genomic profile data: The raw sequencing data is processed to create a binary indication for the presence of each SNP in each sample.

(eg: a matrix where each row is a sample, and each column is SNP)

This data is often filtered to remove very rare SNPs and remove patients that might be related to each other, other filters might also be applied.

Phenotype data: Usually a matrix or vector containing the phenotype of each sample.

In our example we will use the cholesterol levels of each patient.

Basic GWAS methodology: Quality Control (QC)

Principle Component Analysis (PCA):

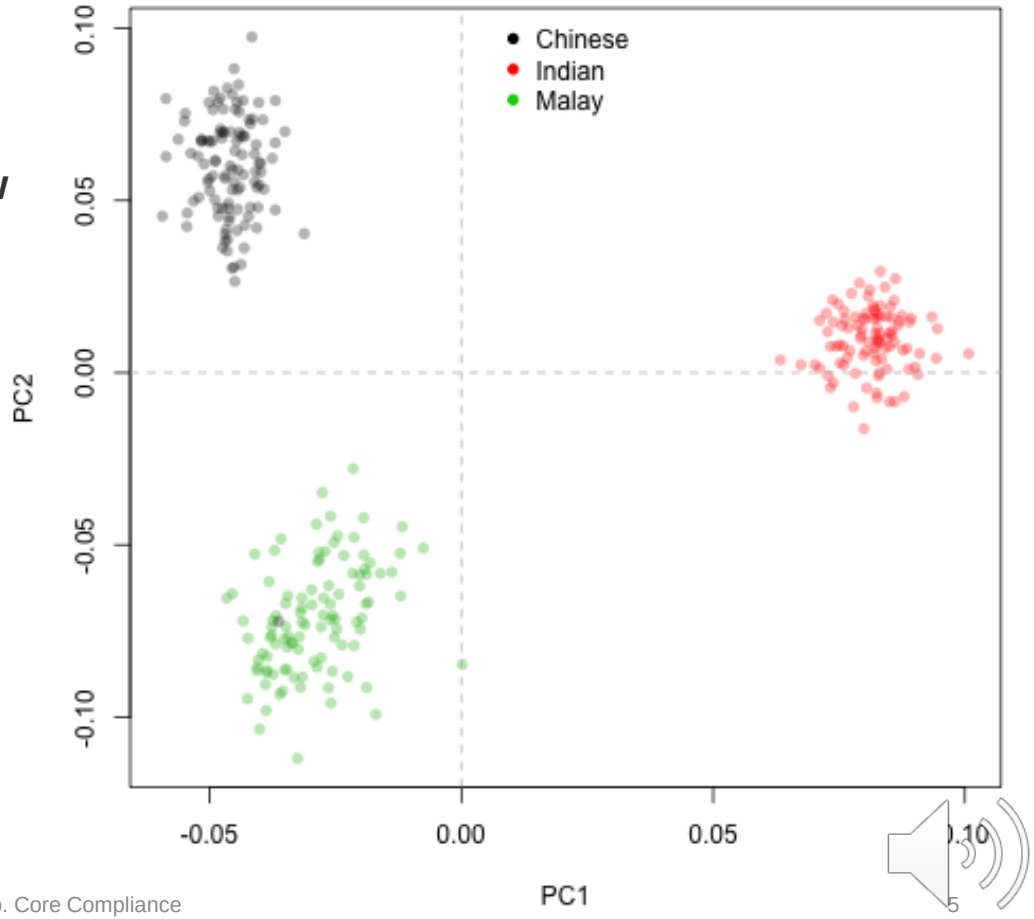
The first principal-component (PC1) is a plane/line in the data that explains most of the variation on the data. PC2, PC3 etc … explain the 2nd and 3rd most variation…

(more next week)

PCA allows us to visualize the data 2-dimensions and thus get an overview of the data.

Is the data behaving as expected?

Are there any outliers?



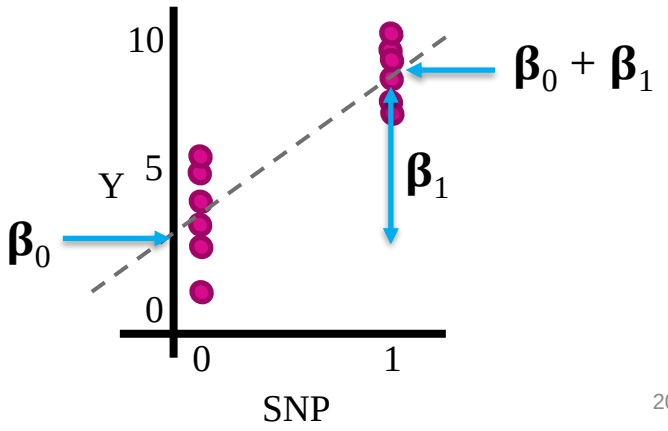
Basic GWAS methodology: Regression

Key Question: Is there a significant association between a SNP (X) and the phenotype.

GWAS answers this question by **fitting a separate regression model for each SNP (X) against (Y)**. This can a linear or a logistic regression depending the the phenotype Y is continuous or binary.

The formula for this linear regression is $\boldsymbol{Y = \beta_0 + \beta_1 X}$ where Y is a vector containing the phenotype of each sample (n) and X is a binary indicator for the presence of the SNP in each sample.

When this model is fitted, β_0 will represent the mean level of Y in the subject without the SNP and $\beta_0 + \beta_1$ will represent the mean level of Y in subjects with the SNP.



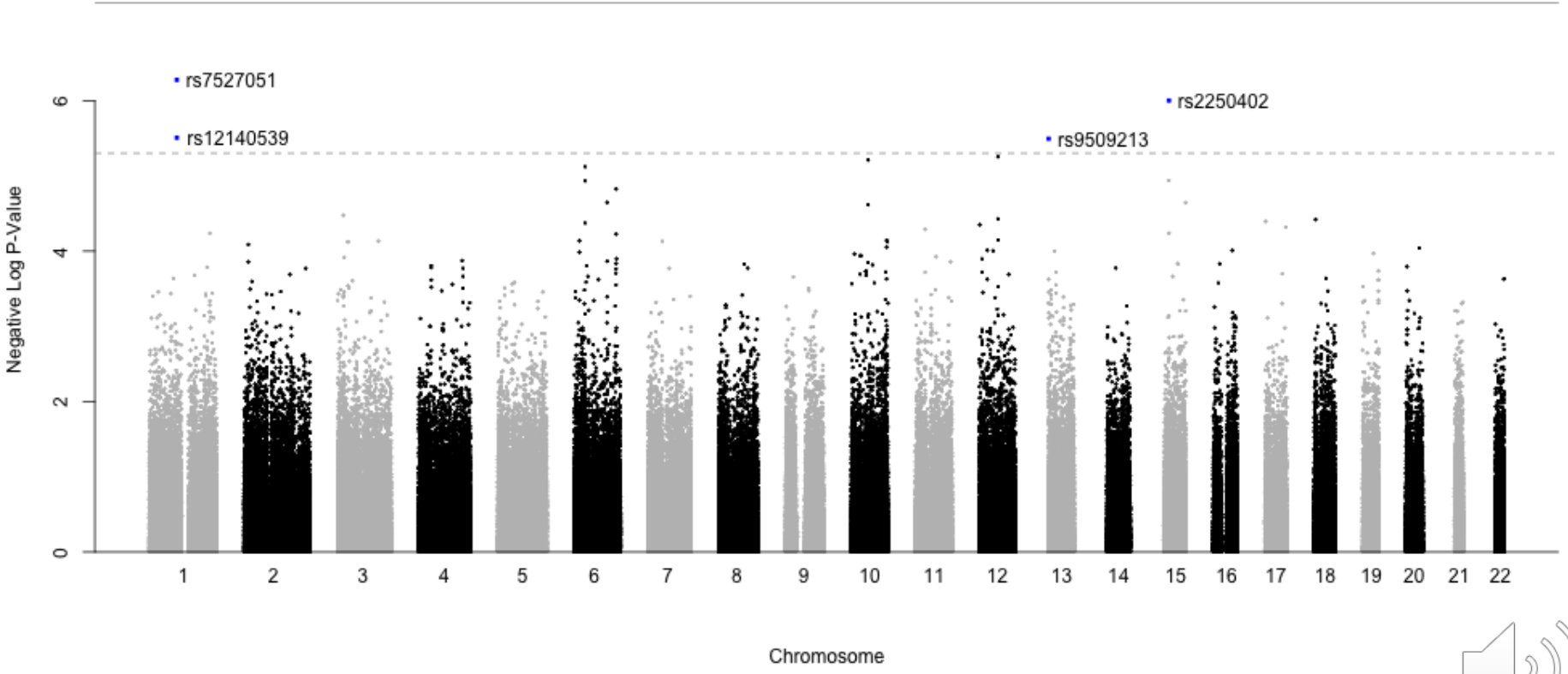
Thus if β_1 is significantly different from 0 ($p < 0.05$) we can say that the SNP is associated with the phenotype.

In practice this p-value is adjusted to account for multiple comparisons.

Common adjustments include Bonferroni or Benjamini-Hochberg.

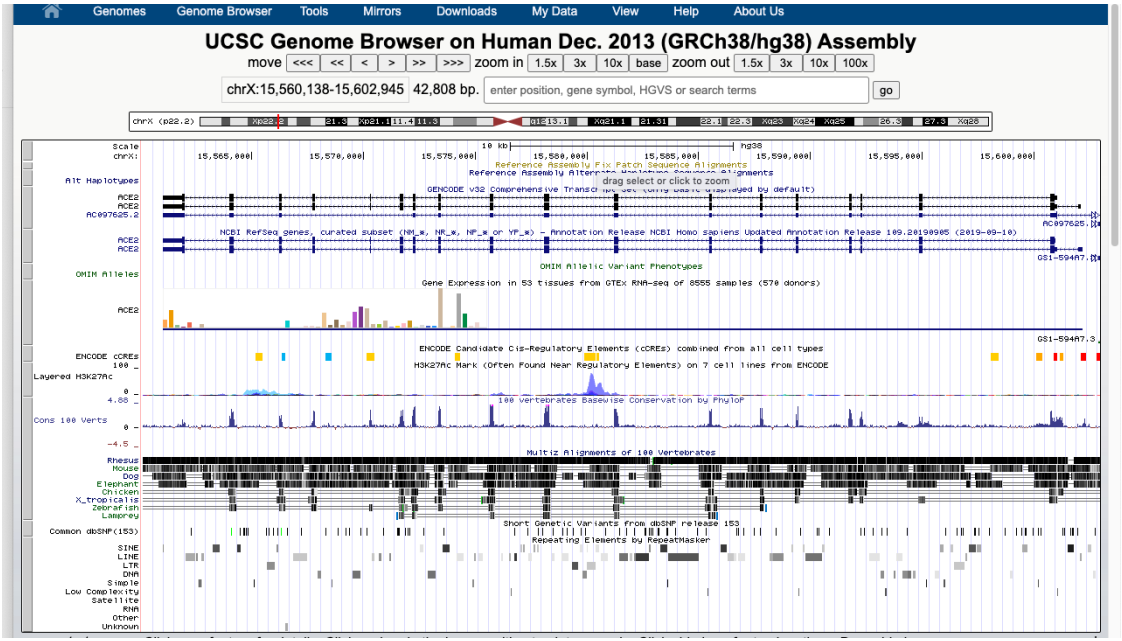
Results Visualization: Manhattan plot

To visualize the results of GWAS analysis, researchers commonly use Manhattan plots. Each point on the plot indicates a different SNP, with the x-axis showing the chromosomal location and the y-axis indicting the *p*-value for that SNP. Various significance thresholds are indicated with by each line. In this example 4 SNPs reached the soft significance threshold.



Next Video and Further Analysis

The USCS Genome Browser (https://genome.ucsc.edu/) can be used to find more details about the identified SNPs. Just type them into the browser to find more information.



In the next lecture I will show some key points when performing GWAS analysis in R.

The full analysis is at (https://www.r-bloggers.com/2017/10/genome-wide-association-studies-in-r/).

I recommend you attempt the full analysis if you can, then consult my video for help some specific sections.

Thank You