## Lecture Title and Date

Lecture 25m7 part 2: Genome Annotation (AS, eQTL, GWAS) - 02/19/2025

## Objectives of the Lecture

- By the end of this lecture, students should be able to:
    - Understand annotation related to an individual's variants:
        - Understand how different versions of alleles present in an individual can influence TF binding, gene expression, epigenetics, etc.
        - Understand how to determine allele-specific gene expression using the binomial distribution.
        - Explain why reference bias occurs and how to mitigate it using personal diploid genomes.
    - Explain that the overarching goal of GWAS is to determine whether variants are significantly correlated with a particular trait or disease.
    - Understand how eQTLs are related to GWAS by linking specific alleles to gene expression patterns.
    - Understand in a broad sense the statistical tests and steps used to complete a GWAS analysis and the assumptions we make when carrying out such calculations.
    - Understand technical considerations in GWAS, such as accounting for covariates and multiple testing correction.
        - Outline the potential considerations and methods for multiple testing correction.

## Key Concepts and Definitions

- **Allele:** a version or variant of the nucleotide sequence at a particular location in the genome. A person inherits one allele from their father and one from their mother.
    - If both copies of the allele are the same, the individual is **homozygous** for that variant.
    - If the copies differ, that individual is **heterozygous** for the allele.
- **Diploid:** When a cell contains two copies of each chromosome. For example, in human somatic cells, there is one set of maternal chromosomes and one set of paternal chromosomes.
- **SNP:** Single nucleotide polymorphism - a variant resulting from the substitution of a single nucleotide compared to the reference.
- **Allele-specific event:** when there is a measurable difference (ex. In transcription factor binding, gene expression, etc.) in outcome between different alleles.
- **Binomial distribution:** the probability distribution describing the number of "successful" outcomes when performing a series of independent tests with a binary output, success or failure (ex. A coin toss with possible outcomes of heads or tails).

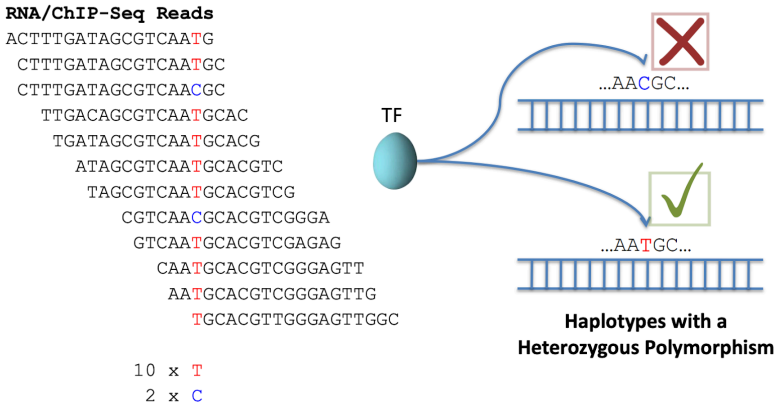
- **Reference bias:** a bias that can occur when mapping reads to the reference genome because reads containing non-ref SNPs or variants are less likely to map as well as reads containing variants matching the reference.
- **Personal diploid genome:** a personal reference genome that separates an individual's maternal and paternal variants to improve mapping and mitigate reference bias.
- **Null expectation/hypothesis:** the expected distribution of results given the effect being studied or tested for does not exist (ex. there is NO true difference between two populations being tested, such as a case and control), and all variation arises from random chance or technical errors.
- **P-value:** the probability of seeing a result as extreme or more extreme as the observation given the null hypothesis.
- **eQTL:** expression quantitative trait loci - loci in the genome with variants influencing the expression of a gene (or multiple genes).
- **GWAS:** Genome-wide association study - a research method used to link variants in the genome, typically SNPs, to traits. The trait can be disease-related or not (i.e. height). Requires genetic information for a large cohort of individuals.
- **Homoscedasticity:** a condition describing data for which the variance of the residuals or error is consistent throughout the data, i.e. for all independent variables. In contrast, with heteroscedastic data, the variance will change as the predictor variable changes. Homoscedasticity is an assumption in linear regression.
- **Normal distribution:** a distribution used to describe a random variable that takes on a "bell curve" shape. It has an equal mean and median and appears symmetrical.
- **Ordinary least squares:** a method in regression modeling used to estimate the coefficients, which represent the relationship between the independent variables and a chosen dependent variable. The goal of the method is to minimize the error between the predicted values and observed values using the sum of squared errors.
- **Covariates:** independent variables besides the variable of interest that also influence the study outcome.
- **Manhattan plot:** for GWAS, a plot depicting genomic location on the x-axis and the negative log of the adjusted p-value on the y-axis. Each point may be a SNP studied in a GWAS, with statistically significant findings standing out above a certain significance threshold on the y-axis.
- **Ancestry Principal Component Analysis (PCA):** A dimensionality reduction technique used in genetics to account for population structure by identifying major axes of genetic variation, helping to correct for ancestry-related confounding in association studies.
- **Bonferroni Correction:** A statistical method that controls the family-wise error rate by dividing the significance level by the number of independent tests to reduce false positives.
- **Family-wise error rate control:** A multiple testing correction approach that limits the probability of making at least one false positive (Type I error) across all conducted tests.
- **Polygenic Risk Scores:** A metric that aggregates the effects of multiple genetic variants, weighted by their effect sizes, to estimate an individual's genetic predisposition to a trait or disease.

## Main Content/Topics

### Allele-specific Annotation

#### 1. What is an allele-specific event?

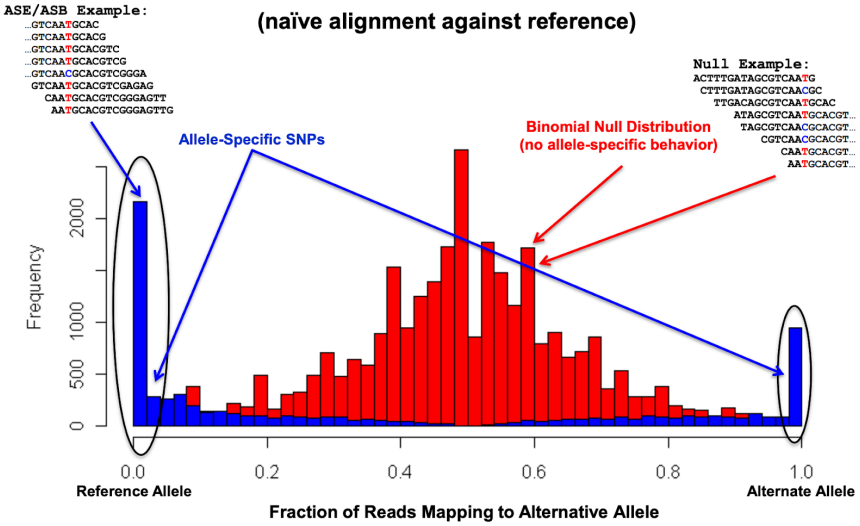
To understand **allele-specific events**, it is useful to consider the binding of a transcription factor (TF) to the genome. Our genome is **diploid**, meaning we possess two copies of each chromosome: one paternal and one maternal. This means that when a TF binds the genome, it can bind to the maternal or the paternal copy of the TF binding site. Typically, the degree of TF binding and the resulting gene expression is the same for the maternal and paternal copies of the genome. However, in some cases, one copy, the maternal copy for example, could have a variant or epigenetic modification increasing the binding affinity for the TF. How could we determine this? One method would be to carry out a ChIP-Seq experiment for that TF. Then, one can use the simple approach of counting how many reads there are for each variant, as shown in the example below:




Let's say the maternal copy contains the T variant in this case. For this ChIP-Seq experiment for the TF of interest, there are 10 reads containing the T variant and only 2 containing the C variant, indicating the TF is better able to bind the maternal, T-containing variant in this case.

While the simple counting method gives us an idea of the differences in TF binding between the maternal and paternal alleles, we can also call allele-specific events using more robust statistical methods. Say we want to annotate a SNP or variant site on the genome as being allele-specific. This can manifest as differential activity between the two alleles in a ChIP-Seq, RNA-Seq, methylation, or Hi-C experiment, for example. Given a read stack at the locus of interest, the **null hypothesis** would be that the number of maternal and paternal alleles should be about even. However, it is unlikely that *exactly* 50% of the reads will be maternal and paternal, even with the null expectation. Really, the proportion of maternal/paternal reads will be sampled from a distribution: the most simple one to consider is the **binomial distribution**. The binomial distribution can be used to model a series of tests with two possible outcomes, ex. A coin flip. The distribution can help calculate the probability of a certain number of "successful"

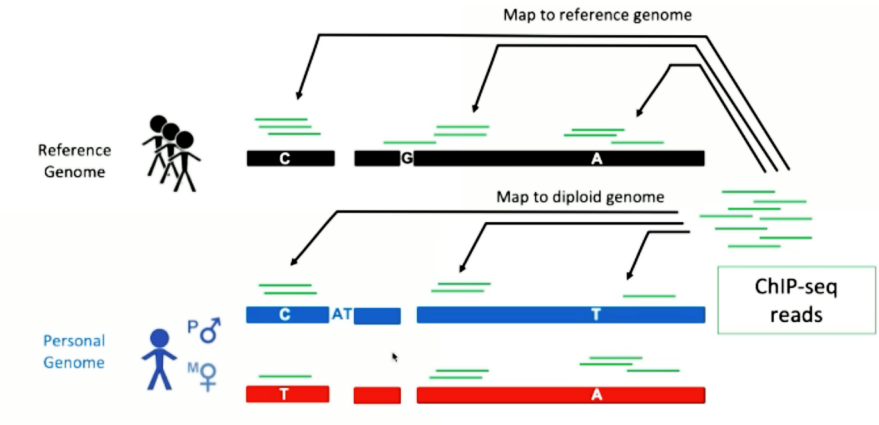
outcomes in a series of such tests, ex. The probability of seeing 5 heads after 10 independent coin flips. In the case of read mapping, if you sample 10 reads from your ChIP-Seq experiment, you would expect to see 5 maternal and 5 paternal alleles. However, you might see another result from the distribution, like 6 maternal and 4 paternal. An example of the binomial distribution in the context of calling allele-specific events is illustrated below:




Here, variants with no allele-specific behavior are shown in red, and those with allele-specific SNPs are shown in blue. The red therefore corresponds to the null expectation. In the case of the blue, there are more extreme cases observed in the read stack: ex. 10 C and 0 T, or 9 C and 1 T. In this case, we may have an allele-specific event, and we can calculate a **p-value** or probability of an allele-specific event using the null expectation. In this example, we can calculate how likely we would see that extreme data point (ex. 9 C and 1 T) or a more extreme case when sampling from the binomial distribution (i.e. by taking the area under the curve). In practice, depending on the experimental context, researchers may choose to use a slightly different distribution to model the null hypothesis, such as a beta-binomial distribution, which allows for a wider tail.

## 2. Technical considerations

Imagine you have a set of reads from an experiment like ChIP-Seq: normally, you would map the reads to the reference genome, getting a read stack, and continue on with your analysis. However, consider the example below:




In this case, the paternal allele contains a T variant relative to the maternal allele and the reference genome, which both contain an A at that locus. Since the paternal allele has a variant relative to the reference, it will not map as well. For example, imagine a read has both the T allele and a technical sequencing error downstream. Due to the combination of the error and the biological variation, reads like this will not map as well. This ultimately results in a preference for mapping the maternal alleles over the paternal ones. The concept of reads containing non-reference alleles mapping less well is known as **reference bias**. This could easily skew allele-specific event calling, whether we are using the simple counting method or the binomial distribution. To overcome this limitation, it is useful to build a **personal diploid genome** for an individual, and then map reads to the personal diploid genome instead as the reference before counting.

## GWAS

### 1. Goal of GWAS

The overarching goal of a **genome-wide association study (GWAS)** is to relate specific variants to traits or phenotypes. The trait of interest could be a complex trait such as height or a disease phenotype. It is considered a non-hypothesis-driven (or “hypothesis-free”), exploratory approach since it involves searching the entire genome to discover risk variants with prior knowledge of any specific risk regions.

**Expression quantitative trait loci (eQTL)** are specific locations in the genome that are associated with the activity of target genes. These loci therefore explain at least some part of the variation in gene expression. eQTLs can be related to GWAS because by modulating gene expression, they can also contribute to a resulting phenotype of interest. If eQTLs overlap with GWAS hits, this can help explain the mechanism behind the GWAS result. However, often only a small percentage of GWAS hits correspond to eQTLs, as GWAS and eQTL mapping results are biased for different kinds of hits (ex. GWAS hits tend to be located further from transcription start sites (TSSs), are enriched near genes with known functional roles, and are under significant selective constraint, whereas eQTLs are usually clustered more closely to TSSs of genes without a functional annotation). See here for more details:

Mostafavi, H., Spence, J. P., Naqvi, S., & Pritchard, J. K. (2023). Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nature Genetics*, *55*(11), 1866–1875. https://doi.org/10.1038/s41588-023-01529-1

## 2. Statistical Process

In a GWAS, we consider the entire genome when searching for potential variants. In a cohort, individuals are genotyped, and then the presence of different variants can be correlated with traits of interest (i.e. a disease, eye color, etc.). Using statistical tests, we can then determine which of those correlations is significant. A large cohort size is typically needed to have sufficient statistical power for a GWAS.

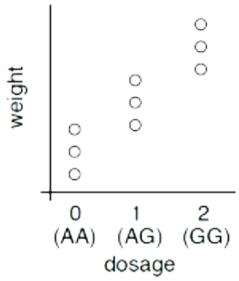
Due to the large scale of GWAS, the statistics can get challenging. Generally, we can think of GWAS analysis as a multiple linear regression problem. For example, imagine you are looking for variants influencing a quantitative trait, such as weight. For a given SNP *S* with possible alleles $\text{allele}_1$ = A and $\text{allele}_2$ = G, there are 3 possible genotypes: AA, AG, and GG. Now, when analyzing our cohort, we want to answer the following question with GWAS: is there a statistically significant difference in weight among individuals in each possible genotype group according to the $\text{allele}_2$ dosage (ex. Homozygous ref vs. homozygous alt vs. heterozygous)? To address this question, we can regress weight versus the genomic dosage, giving the following simple linear regression:

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \varepsilon_i$$

In which:

- $y_i$ represents $\text{weight}_i$, or the weight of an individual *i* (dependent variable)
- $\beta_0$ or $b_0$ represents the intercept
- $x_{1i}$ represents the $\text{dosage}_i$ or the dosage of $\text{allele}_2$ in individual *i* for SNP *S* (independent or explanatory variable)
- $\beta_1$ represents $b_1$ or the effect of $\text{allele}_2$ on the individual's weight
- $\varepsilon_i$ represents the error or residual for the weight of individual *i*

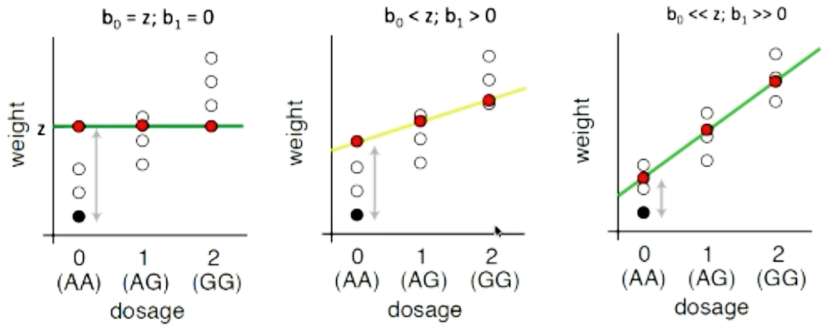
The data might look something like this, for example:



Using this model, we can do a simple linear fit or regression. It is important to keep in mind that when we do a linear regression like this, we are making some assumptions about the data, notably:

1. We assume there is a linear relationship between the dependent and independent variables and not some other kind of relationship, ex. Polynomial.
2. The residuals (i.e. error) are **homoscedastic**, meaning there is constant variance in the residuals.
3. The residuals are **normally distributed**.
4. The observations are independent and not correlated.

Using a method like **ordinary least squares**, we can get a regression line and estimate the values of the slope and intercept:




With the ordinary least squares method, this involves determining the values of $b_0$ and $b_1$ which minimize the sum of all squared residuals across the cohort. The criterion for ordinary least squares is as follows:

$$Q(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot X_i\right)^2$$

## 3. Complications: Covariates

In our large cohort of individuals, we have shown how to regress weight versus genomic dosage. However, we haven't yet considered **covariates** that could also influence weight in addition to the allele$_2$ dosage. In truth, we know there are both genetic and non-genetic factors at play in determining an individual's weight. We would like to regress those other covariates out in our model in order to study the effect of genomic dosage. For example, other factors including weight could be diet, biological sex, or geographic location. We can compose a regression model that includes those covariates so we know we are accounting for them:

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \ldots + \beta_{(p-1)} \cdot x_{(p-1)i} + \varepsilon_i$$

Similar to the previous regression model we looked at, in this case:

- *i* = 1…n observations (i.e. individuals or samples_
- $y_i$ represents weight$_i$, or the weight of an individual *i* (dependent variable)
- $\beta_0$ or $b_0$ represents the intercept
- $x_{1i}$ represents the dosage$_i$ or the dosage of allele$_2$ in individual *i* (independent or explanatory variable), could be 0, 1, or 2
- $x_{2i} + \ldots + x_{(p-1)i}$ are the covariates for that individual *i* (ex. Age, gender, diet, location)
- $\varepsilon_i$ represents the error or residual for the weight of individual *i*

Using this kind of model, we can now both model the relationship between y and all its predictors x, and we can test if our *specific* explanatory variable of interest (in this case, the dosage of allele$_2$ at SNP *S*) has a significant effect on the dependent variable (weight in our example). Accounting for covariates is critical in a biomedical context because if we ignore them, our results could be messed up by a confounding factor in the data. In this case, we have accounted for some number of covariates 2 to p.

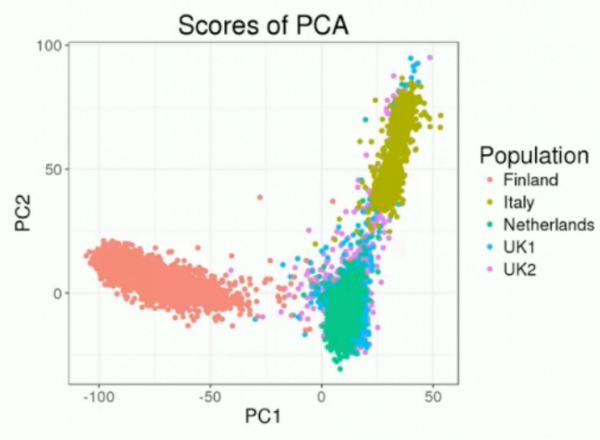
Once we have achieved such a model, we can address the fundamental question: is $\beta_1$ non-zero? The estimated value of $\beta_1$, represented as $b_1$ or $\widehat{\beta}_1$, is the estimated effect of the allele$_2$ dosage on weight based on our linear model. If $\beta_1$is zero, this indicates that genomic dosage is not significantly related to the trait of interest. $\beta_1 = 0$ is therefore the null hypothesis. We can use a simple Student's t-test to determine whether our predicted value $\widehat{\beta}_1$ is significantly different from the null hypothesis. Since the calculated *t*-statistic is part of the *t* distribution, we can use it to calculate a p-value and determine if the difference is statistically significant based on a chosen significance threshold (ex. 1%, 5%, 0.01%). If the p-value is below the given threshold, we can reject the null hypothesis and determine that the gene dosage from SNP *S* indeed has a statistically significant effect on an individual's weight. However, if the p-value is above the threshold, we would determine that the SNP does not have a significant effect on weight. P-values from a GWAS can be represented visually in a **Manhattan plot**, in which the genomic loci are plotted on the x-axis and the negative logarithm of the p-values for the tested SNPs are plotted on the y-axis. The higher a point on the Manhattan plot, therefore, the lower the p-value.

## GWAS (Additional Considerations)


### 1. Study Design

When designing a **Genome-Wide Association Study (GWAS)**, it is essential to determine whether the trait being studied is a **quantitative** or **discrete variable**, as this influences the choice of study design and analysis methods. For **discrete traits**, such as disease status, a **case-control study** is often used, where individuals are categorized into **cases (affected)** and **controls (unaffected)** to identify genetic variants associated with the condition.

One major consideration in GWAS is **population stratification**, which can introduce **confounding effects**. Some **single nucleotide polymorphisms (SNPs)** may have different **allele frequencies** across **subpopulations**, leading to **false associations** if not properly accounted for. For example, if a GWAS were conducted on the trait **"uses chopsticks"** without correcting for **ancestry**, SNPs more prevalent in **East Asian populations** might appear associated with chopstick use, even though the real association is with ancestry rather than genetics. To address this issue, researchers must include relevant **covariates** that control for **indirect effects** unrelated to the **phenotype of interest**, such as **age, sex, and genotyping batch**.




A common approach to mitigating **population stratification** is **Ancestry Principal Component Analysis (PCA)**. By analyzing **genetic variation**, PCA generates **principal components** that summarize **ancestry-related differences** in genetic data. Typically, the **first five or six principal components** are included as **covariates** in **GWAS models** to correct for **ancestry-related confounding**. This ensures that **genetic associations** identified in the study are due to the **phenotype of interest** rather than **population structure**.

### 2. Power Calculations

**Power** refers to the probability of detecting a true association between a **single nucleotide polymorphism (SNP)** and a **trait**. Statistical **significance** is assigned to **non-zero beta (β) coefficients**, which quantify the effect of a given SNP on the trait and depend on the **sample size (n)**.

Power in GWAS is influenced by several key factors, including **sample size, allele frequency, and effect size**. A **larger sample size (n)** and a **higher minor allele frequency (MAF, f)** improve the accuracy of estimating the SNP effect (**β**). Additionally, larger **absolute values of β** increase the **difference from the null model**, where no association exists (e.g., the mean value of the trait remains the same across genotype groups).

When evaluating whether a significant **association** exists between a **trait** and **genotype**, it is crucial to perform **power calculations**. These calculations help determine the necessary **sample size** given an expected **effect size** to ensure the study has enough **power** to detect true associations.

## 3. GWAS

The **type of study** in **Genome-Wide Association Studies (GWAS)** determines the appropriate **statistical model** used for analysis. If the trait being studied is **quantitative** (e.g., height, blood pressure), **linear regression** is applied, where **beta (β) values** represent the effect size of a given SNP on the trait.

For **case-control studies** (e.g., disease vs. healthy individuals), **logistic regression** is used instead, calculating **odds ratios (ORs)** to estimate the likelihood of disease presence based on genetic variation. Choosing the correct statistical model ensures accurate interpretation of **SNP-trait associations**.

## 4. Downstream Analysis

A key assumption in **linear regression for GWAS** is that all SNPs are **independent**. However, in reality, **linkage disequilibrium (LD)**—the non-random association of alleles—leads to many SNPs being highly correlated. This can result in a large number of **indirect associations**, where significant SNPs may not be causative but instead linked to the true causal variant. Out of approximately **4 million SNPs** in the human genome, only about **0.5 to 1 million** are truly independent.

To account for multiple hypothesis testing, **multiple-test correction** is required. One common method is **Bonferroni correction**, which controls the **family-wise error rate (FWER)** to maintain an acceptable false positive rate. Given that GWAS tests up to **1 million independent SNPs**, the adjusted significance threshold is calculated as:

- $\text{FWER} = \frac{\alpha}{m'}$,
  - $m$ = # of independent hypotheses
  - # of independent common variants = $10^6$
  - FWER = $0.05/10^6 = 5 \cdot 10^{-8}$

This results in a **very stringent significance threshold** to minimize false positives. However, **Bonferroni correction** is considered overly conservative, as it assumes independence among tests and does not account for the correlation structure introduced by LD. Other approaches, such as **family-wise error rate control** and alternative multiple-testing corrections, may be used to balance false discovery control with statistical power.

Beyond individual SNP associations, researchers construct **Polygenic Risk Scores (PRS)** to assess the cumulative genetic contribution to a trait. PRS is a **linear combination of effect sizes (β values) from multiple SNPs**, aiming to enhance **trait predictability** beyond single-SNP associations. To improve accuracy, PRS models include **SNPs below a certain p-value threshold** while ensuring that only **low-LD SNPs** are retained to maintain independent signals.

### 5. Replication Studies

Replication studies are essential in **validating GWAS findings** and ensuring that identified SNP-trait associations are **robust and reproducible** across different populations. Since **GWAS involves millions of statistical tests**, false positives are inevitable, making independent replication a critical step before claiming a true genetic association.

### GWAS has evolved

Since the first **Genome-Wide Association Study (GWAS)** was conducted at **Yale by Robert Kline** two decades ago to identify genetic risk factors for **macular degeneration**, GWAS has evolved into a **powerful and large-scale genomic tool**. Today, it is widely used in both **academic research and commercial applications**, driving discoveries across numerous complex traits and diseases.

Modern GWAS is now a **global enterprise**, with large datasets and **biobanks** enabling researchers to identify genetic variants associated with various conditions. **Companies like 23andMe** use GWAS to offer **direct-to-consumer genetic testing**, providing individuals with insights into their genetic predispositions. Additionally, the **NHGRI-EBI Catalog of Human GWAS Studies** serves as a comprehensive **repository of significant SNPs**, consolidating findings from thousands of GWAS studies worldwide.

## Expression Quantitative Trait Loci (eQTL) Analysis


eQTL analysis extends the principles of **Genome-Wide Association Studies (GWAS)** by identifying genetic variants that influence **gene expression levels** rather than phenotypic traits like height or weight. eQTL studies regress **genetic variants against gene expression levels** instead of regressing against a physical characteristic, typically measured as the number

of RNA sequencing reads for a particular gene. Like GWAS, eQTL studies are conducted at a **population scale** and follow similar statistical frameworks for quantitative traits.

Researchers often **simplify calculations** to manage the computational burden and multiple testing corrections given the vast number of genetic variants and genes that could be tested. The majority of eQTL studies focus on **cis-eQTLs**, where genetic variants are tested for association **only within a predefined genomic window** surrounding the gene of interest, reducing the number of statistical tests and increasing the likelihood of detecting true associations. However, some studies use **trans-eQTLs**, where variants located on **one chromosome** are tested for their effects on gene expression on **a different chromosome**. Trans-eQTL analyses are particularly used for insights into long-range regulatory interactions, such as transcription factor binding or chromatin looping. However, trans-eQTL are much more challenging due to the need for extensive multiple testing corrections. As a result, **trans-eQTL studies remain conceptually valuable but are not commonly performed**.

Another key challenge in eQTL analysis is the possibility of **unaccounted covariates** that may confound results. Researchers often adopt a **hierarchical testing approach** to address this challenge and reduce the number of statistical comparisons. Instead of testing each gene against all possible features independently (which would require correcting for a massive number of tests), hierarchical approaches **first test broader gene-level groupings (clusters) for significance**. If a cluster is found to be significant, researchers then perform additional tests within the cluster to identify the specific feature driving the association. This method significantly reduces the burden of **multiple testing correction**, making eQTL studies more computationally feasible.

## Discussion/Comments

- When we perform a linear regression, we make a few key assumptions. Which of these assumptions do you think might be violated in a GWAS? Explain.
    - Example answer: we assume that all observations are independent, but we know this isn't true because of linkage disequilibrium. Fine-mapping techniques are needed to find the true causal variants among the correlated variants.
- In previous lectures, we discussed how the genomes of African individuals generally contain more variants, and therefore more SNPs, relative to the reference. This is due to founder effects from early human populations leaving Africa at various points, leading to genetic drift. How would this affect LD for African populations (ex. Would you expect higher or lower LD compared to individuals of European descent?). How could this impact a GWAS analysis? For example, if GWAS is done using only individuals of European descent, do you think the results will apply well to African and African-American populations? Why or why not?
    - Given your answer to the question above, what are the implications for ethical execution of a GWAS, particularly in a clinical setting?
    - Example answer: Because African genomes are more diverse, they will also have lower LD. This means an analysis done only on European populations could be missing key SNPs. In a clinical setting, it is critical to use a more diverse

population, otherwise individuals from other populations, like individuals of African ancestry, will be excluded from the medical benefits of the study findings.

- Say you find a new disease-associated variant for a complex disease using GWAS. What follow-up experiments could you do to determine the mechanism by which this variant impacts disease risk? (Example answer: do a type of Hi-C analysis to see which areas of the genome that locus interacts with. For example 4C-seq allows you to see which other parts of the genome interact with a region of interest. Then, see if those interacting regions have a function related to disease, ex. Promoter for a gene involved in disease mechanism.)
    - Say you find 8 SNPs in one region of the genome that are all associated with disease. Why might this occur? How can you determine which SNP actually plays a mechanistic role in disease risk?
    - Example answer: The SNPs are in LD with each other. Determining the causal SNP in this case is known as “fine-mapping”, and there are many possible approaches to this. For example, you could build an *in vitro* system to independently test the impact of each SNP on gene expression, i.e. a massive parallel reporter assay. Using multiple ethnic groups in the analysis with different LDs can also be advantageous in narrowing down the causal variants *in silico*, though this method relies on an assumption that all populations should theoretically share the disease-causing variant. For more information on fine-mapping, see this resource:

Wang, Q. S., & Huang, H. (2022). Methods for statistical fine-mapping and their applications to auto-immune diseases. *Seminars in Immunopathology*, *44*(1), 101–113. https://doi.org/10.1007/s00281-021-00902-8

## References ISL/ESL

ISL (An Introduction to Statistical Learning, with Applications in Python):

- Relevant Chapters:
    - Chapter 3: Linear regression: OLS, assumption of linear regression,
    - Chapter 4: Classification: Logistic regression, odds ratio
    - Chapter 12: Unsupervised learning: Principal Component Analysis
    - Chapter 13: Multiple testing: Bonferroni correction, Family-wise error rate
- Chapter 3: Linear regression:
    - Simple linear regression:
    - It assumes that there is approximately a linear relationship between X and Y . a very straightforward simple linear approach for predicting a quantitative response Y on the basis of a single regression predictor variable X.
        - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$
        - Estimate betas (OLS):

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (3.4)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

- Assessing accuracy of coefficients:
  To compute the standard errors associated with beta_0 and beta_1, we use the following formulas

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (3.8)$$

- Assessing accuracy a model:
  - The residual standard error (RSE) is an estimate of the standard deviation of the error term. Roughly speaking, it is the average amount that the response will deviate from the true regression line.
    It is computed using the formula

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \qquad (3.15)$$

    Note that RSS was defined in Section 3.1.1, and is given by the formula

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \qquad (3.16)$$

  - R square:
    The R-square statistic provides an alternative measure of fit. It takes the form of a proportion—the proportion of variance explained—and so it always takes on a value between 0 and 1. and is independent of the scale of Y.
    To calculate $R^2$, we use the formula

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

    where $\mathrm{TSS} = \sum(y_i - \bar{y})^2$ is the *total sum of squares*,

- Multiple linear regression:
  Similarly with the simple regression:
  - Formula: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$
    We choose betas to minimize the sum of squared residuals

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

- Chapter 4: Classification (4.3 Logistic regression)
  - Formula: $p(X) = \beta_0 + \beta_1 X.$

    In logistic regression, we use the *logistic function*,

    $$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

    To fit the model, we use a method called maximum likelihood.
  - Log odds:

    By taking the logarithm of both sides of (4.3), we arrive at

    $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X. \qquad (4.4)$$

    The left-hand side is called the *log odds* or *logit*. We see that the logistic regression model (4.2) has a logit that is linear in $X$.
- Chapter 12: Unsupervised learning (12.2 Principal Component Analysis)
  - Principal components analysis (PCA) refers to the process by which principal components are computed, and the subsequent use of these components in understanding the data.
  - principal components (PC):

    The *first principal component* of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features

    $$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p \qquad (12.1)$$

    After the first principal component $Z_1$ of the features has been determined, we can find the second principal component $Z_2$. The second principal component is the linear combination of $X_1, \ldots, X_p$ that has maximal variance out of all linear combinations that are *uncorrelated* with $Z_1$. The second principal component scores $z_{12}, z_{22}, \ldots, z_{n2}$ take the form

    $$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots + \phi_{p2}x_{ip}, \qquad (12.4)$$

    where $\phi_2$ is the second principal component loading vector, with elements $\phi_{12}, \phi_{22}, \ldots, \phi_{p2}$.
  - And similar for the remaining PCs.
- Chapter 13: Multiple testing
  - Type I error & Type II error:

| | | Truth | |
|---|---|---|---|
| | | $H_0$ | $H_a$ |
| **Decision** | Reject $H_0$ | Type I Error | Correct |
| | Do Not Reject $H_0$ | Correct | Type II Error |

**TABLE 13.1.** *A summary of the possible scenarios associated with testing the null hypothesis $H_0$. Type I errors are also known as false positives, and Type II errors as false negatives.*

- Family-wise error rate (FWER):
  controlling the probability of making at least one Type I error. The family-wise error rate is given by:

$$\text{FWER} = \Pr(V \geq 1).$$

$$\begin{aligned} \text{FWER}(\alpha) &= 1 - \Pr(V = 0) \\ &= 1 - \Pr(\text{do not falsely reject any null hypotheses}) \\ &= 1 - \Pr\left(\bigcap_{j=1}^{m} \{\text{do not falsely reject } H_{0j}\}\right). \end{aligned} \quad (13.4)$$

- Bonferroni:

$$\text{FWER}(\alpha/m) \leq m \times \frac{\alpha}{m} = \alpha,$$

Besides Bonferroni, they also discussed other methods to control FWER in this section: e.g. Holm's step-down procedure, Tukey's method and Scheffé's method

Citation:
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An Introduction to Statistical Learning: With Applications in Python. Chapter 10 – Deep Learning. First Printing, Springer.

ESL (The Elements of Statistical Learning, Second Edition):

- Relative Chapter:
  - Chapter 3: Linear method for regression: 3.2 Linear Regression Models and Least Squares
  - Chapter 4: Linear Methods for Classification
- Chapter 3: Linear Regression
  - Except contents mentioned in ISL, ESL gave the matrix formula of multiple linear regression of OLS in linear regression:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta).$$

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2).$$

- Chapter 4: Classification
  - Similarly as mentioned in the ISL:

    Actually, all we require is that some monotone transformation of $\delta_k$ or $\Pr(G = k|X = x)$ be linear for the decision boundaries to be linear. For example, if there are two classes, a popular model for the posterior probabilities is

    $$\begin{aligned}\Pr(G = 1|X = x) &= \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)},\\ \Pr(G = 2|X = x) &= \frac{1}{1 + \exp(\beta_0 + \beta^T x)}.\end{aligned} \tag{4.1}$$

    Here the monotone transformation is the *logit* transformation: $\log[p/(1-p)]$, and in fact we see that

    $$\log \frac{\Pr(G = 1|X = x)}{\Pr(G = 2|X = x)} = \beta_0 + \beta^T x. \tag{4.2}$$

  -

Citation:
Hastie, T., Tibshirani, R., & Friedman, J. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Chapter 11 – Neural Networks. Corrected 12th Printing, Second Edition, Springer.

## Other Suggest references

A reference about power calculation:
Serdar CC, Cihan M, Yücel D, Serdar MA. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochem Med (Zagreb). 2021 Feb 15;31(1):010502. doi: 10.11613/BM.2021.010502. Epub 2020 Dec 15. PMID: 33380887; PMCID: PMC7745163.

A reference about model assumption and diagnostics:
Shatz I. Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. Behav Res Methods. 2024 Feb;56(2):826-845. doi: 10.3758/s13428-023-02072-x. Epub 2023 Mar 3. PMID: 36869217; PMCID: PMC10830673.