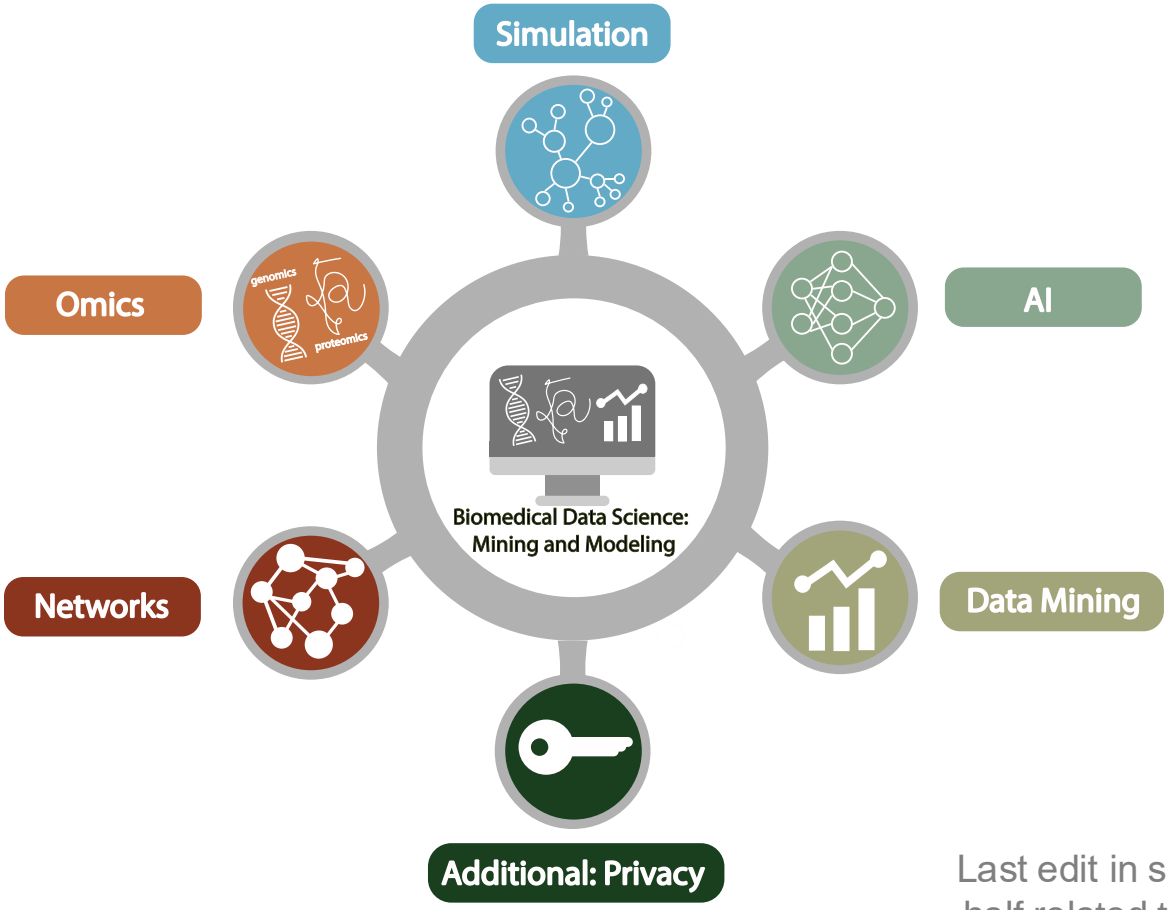
Biomedical Data Science (GersteinLab.org/courses/452)

Genome Annotation (AS, eQTL, GWAS) (25m7-part2)



Mark Gerstein
Yale U.

Last edit in spring '25. Just second half related to AS, eQTL & GWAS. Added in many GWAS slides relative to 2023. Now loosely related to 2nd half of 2021's M7 [which has a video].

Outline

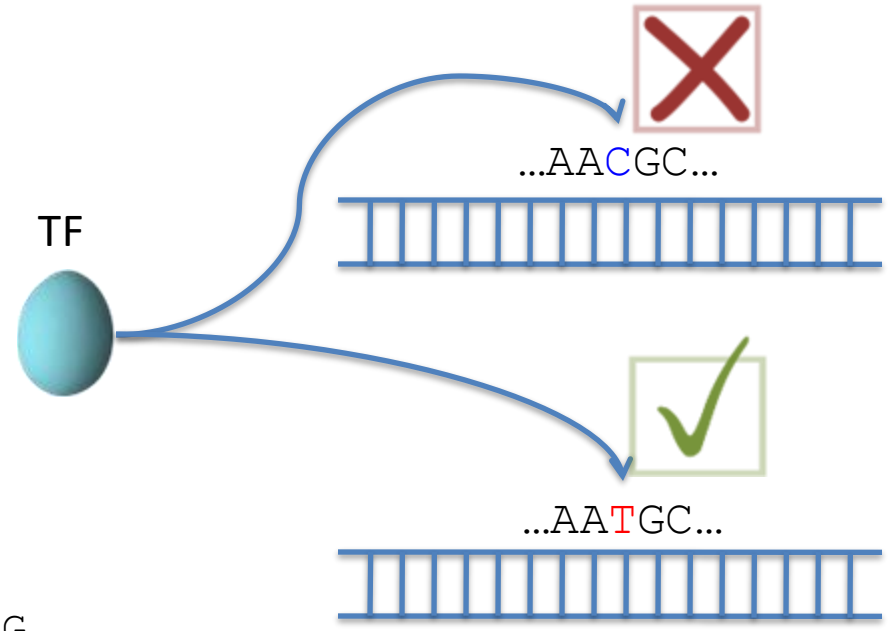
- Part 1 : Generic Annotation
  (not related to an individual's variants)
  - RNA-seq, Chip-seq
  - Integration
  - , Hi-C
- **Part 2 : Annotation related to an individual's variants**
  - ASE/ASB
  - GWAS & eQTL

Allele-specific Events

Inferring Allele Specific Binding/Expression using Sequence Reads

RNA/ChIP-Seq Reads

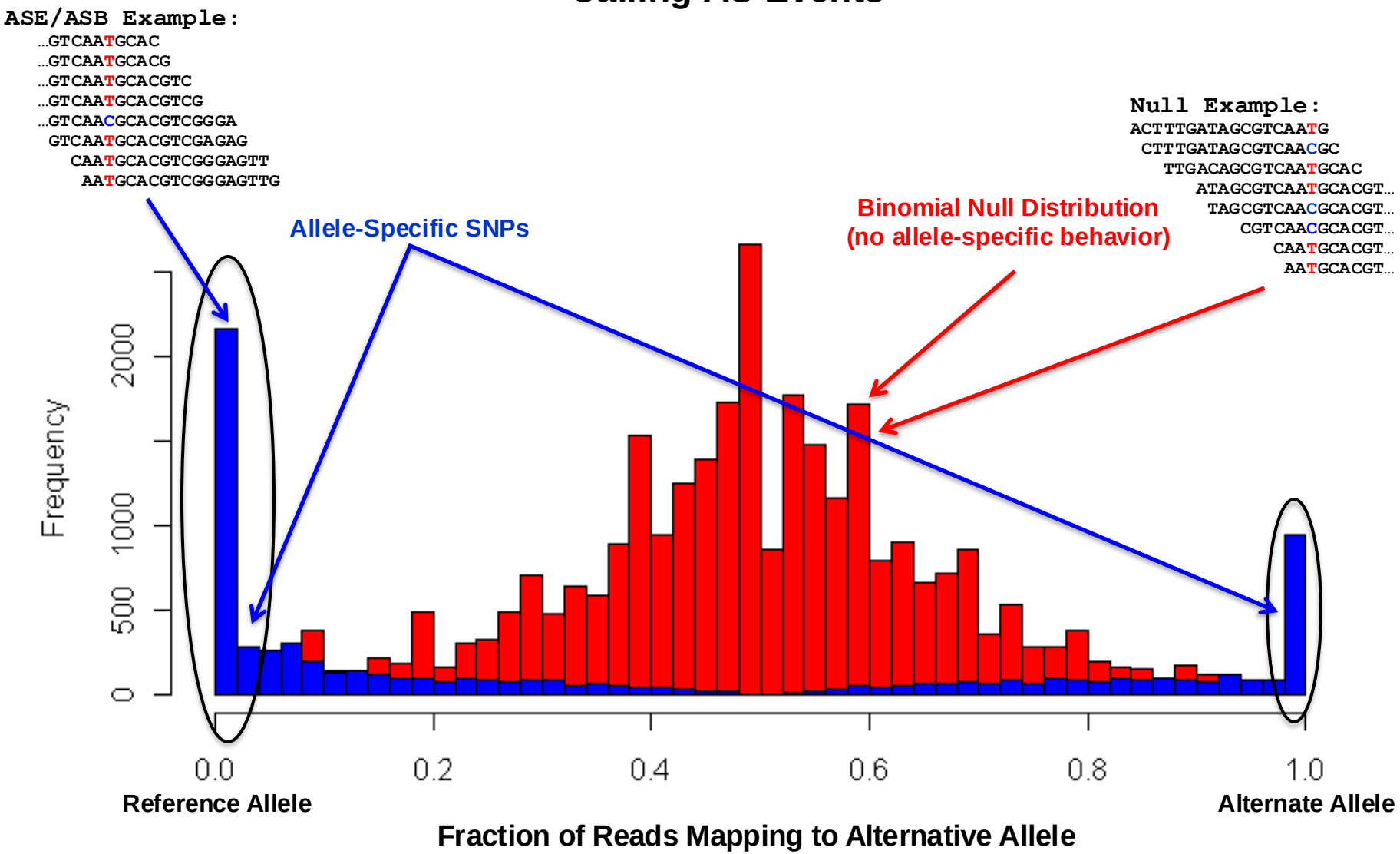
```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAATGC
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
    TGATAGCGTCAATGCACG
      ATAGCGTCAATGCACGTC
       TAGCGTCAATGCACGTCG
          CGTCAACGCACGTCGGGA
           GTCAATGCACGTCGAGAG
             CAATGCACGTCGGGAGTT
              AATGCACGTCGGGAGTTG
                TGCACGTTGGGAGTTGGC

          10 x T
           2 x C
```



Haplotypes with a Heterozygous Polymorphism

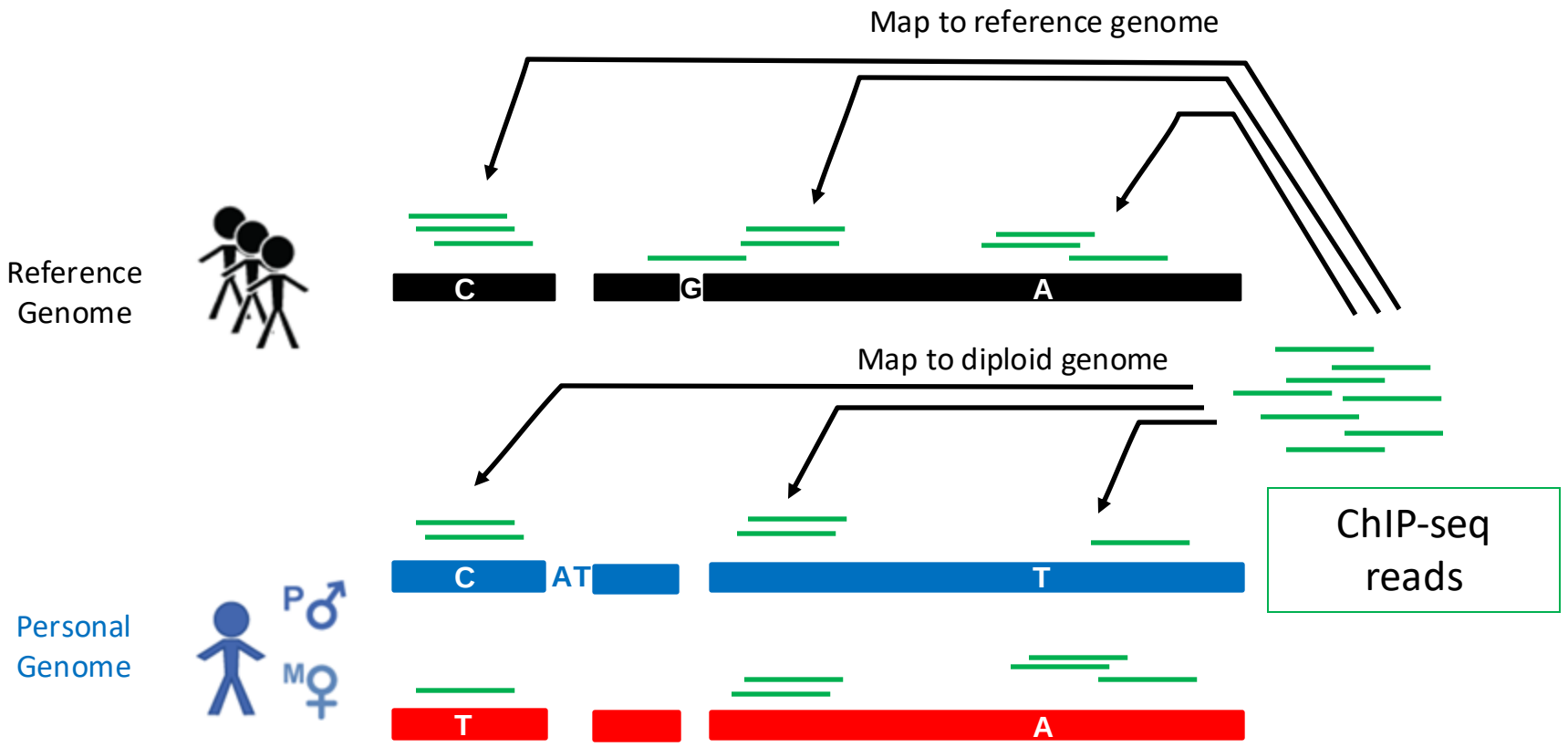
Interplay of the annotation and individual sequence variants

Calling AS Events



[Rozowsky et al., MSB ('11)]

Many Technical Issues in Determining ASE/ASB: Reference Bias (naïve alignment against reference v using a personal genome)



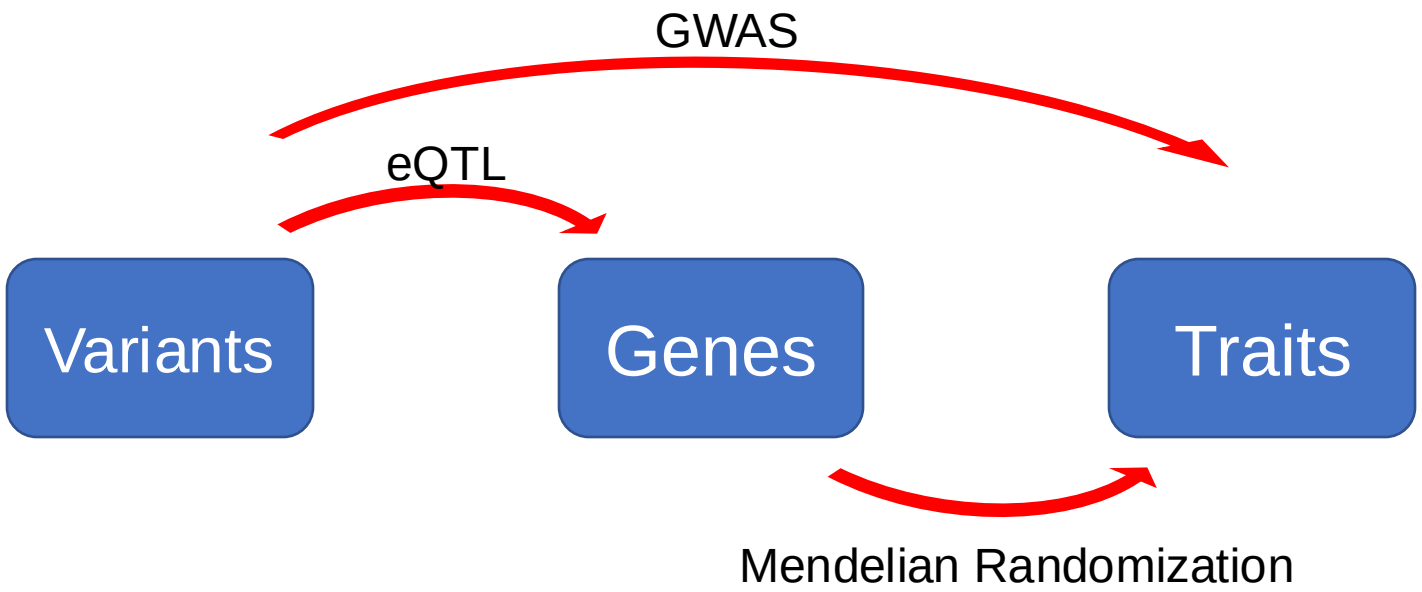
Outline

- Part 1 : Generic Annotation
  (not related to an individual's variants)
  - RNA-seq, Chip-seq
  - Integration
  - , Hi-C
- **Part 2 : Annotation related to an individual's variants**
  - ASE/ASB
  - GWAS & eQTL

GWAS (Basic Workflow)

Relating to Variants on a Population Level in a Cohort

Relating to Variants on a Population Level in a Cohort

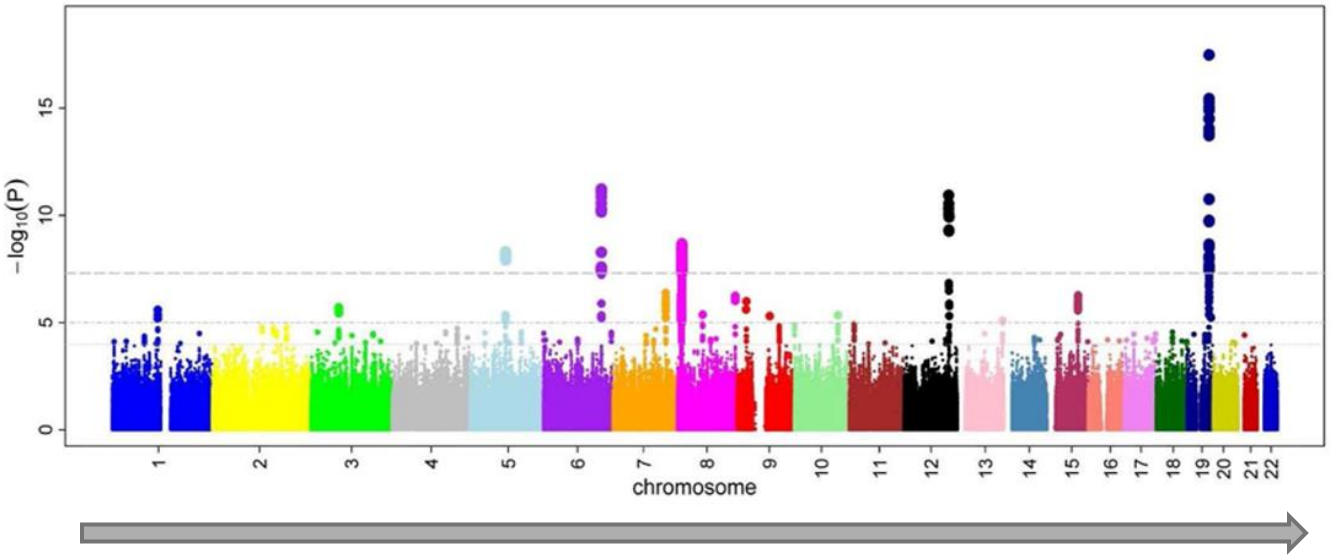


Genome-Wide Association Studies (GWAS)

The basic idea behind a GWAS is to find significant associations between genetic markers and phenotypes (disease / traits) → exploratory "genome-wide" research, non-hypothesis based

2. Testing each SNP for significant association with the trait

Manhattan plot



1. Scanning SNPs across the genome

Slides from B Borsari & S Liu

GWAS: a (multiple) linear regression problem

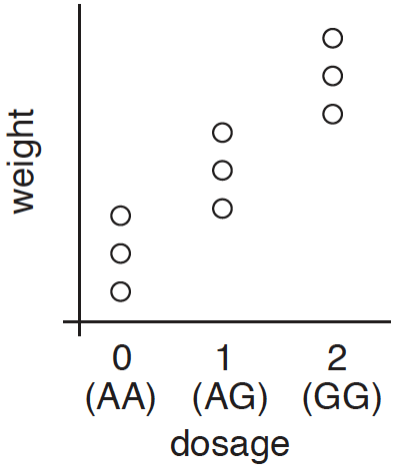
Consider a quantitative trait (eg: weight)

- Consider a SNP *S* with $allele_1$ = A, $allele_2$ = G
- Define three groups of individuals with genotype AA, AG, GG
- The question we try to answer when conducting a GWAS: do we see a significant difference in the weight between these three groups of individuals that correlates with the dosage of $allele_2$?

We can treat this as a linear regression problem:

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

$$weight_i = b_0 + b_1 \cdot (dosage_i \text{ of } allele_2) + error_i$$



- $weight_i$ = weight of individual *i* = dependent variable
- b_0 = intercept
- $dosage_i$ of $allele_2$ = dosage of $allele_2$ in individual *i* = explanatory or independent variable
- b_1 = effect of $allele_2$ on the weight of the individual

(More Later)

Theoretical model: assumptions

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

$$\text{weight}_i = b_0 + b_1 \cdot (\text{dosage}_i \text{ of allele}_2) + \text{error}_i$$

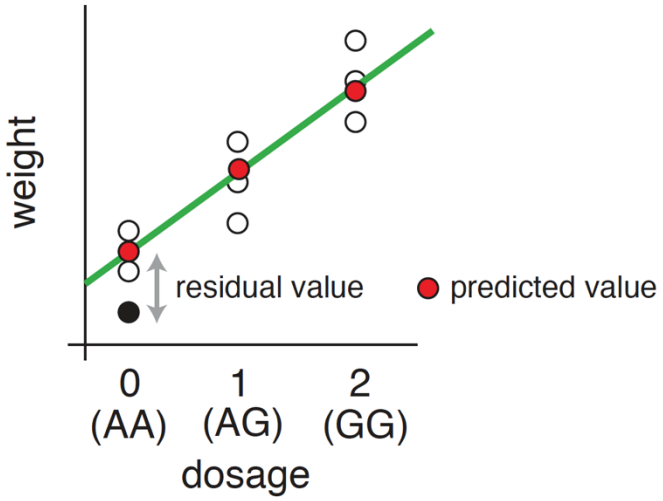
error_i is also more commonly called **residual**

Assumptions

- Linear relationship between y and x
- Homoscedastic residuals (= constant variance)
- Normally-distributed residuals
  - ε_i = ~ Normal(0, σ^2)
- Independent observations

more stringent

less stringent

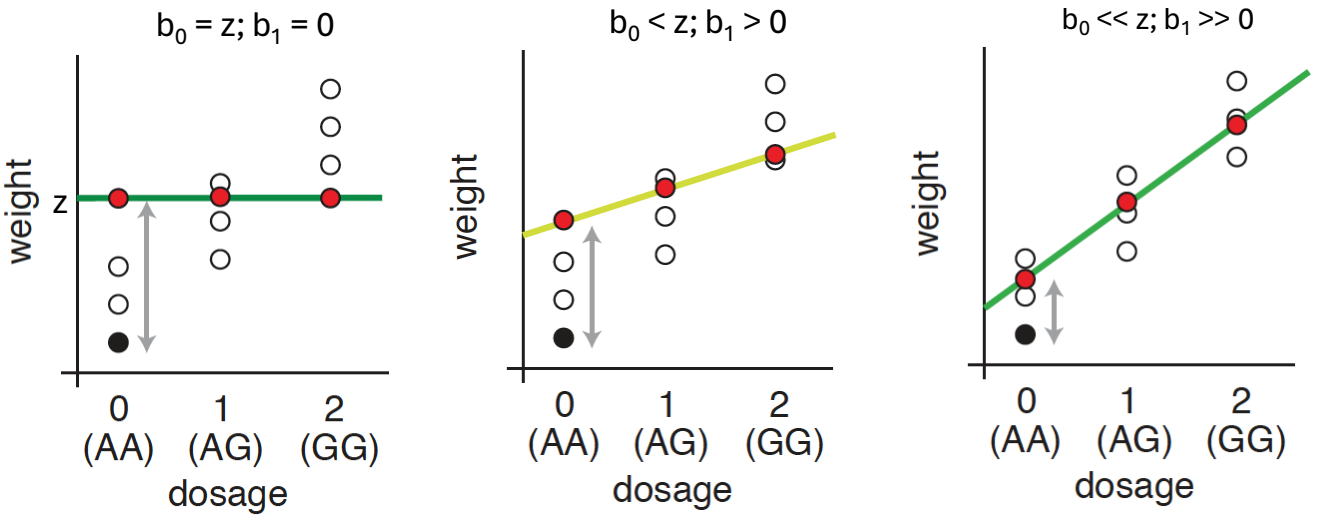


Estimation by Least Squares

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

$\text{weight}_i = b_0 + b_1 \cdot (\text{dosage}_i \text{ of allele}_2) + \text{error}_i$

To solve this equation, we apply the **Ordinary Least Squares** criterion: $Q(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 \cdot X_i)^2$



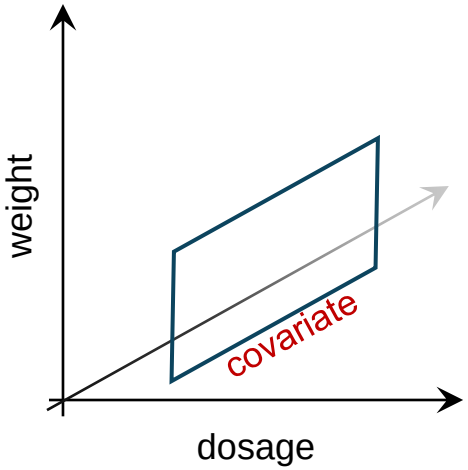
In other words, we need to find the combination of b_0 and b_1 that minimizes the sum of squared residuals across all individuals

GWAS: a (multiple) linear regression problem

A multiple regression problem:

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

- $i = 1 \ldots n$ observations (individuals / samples)
- y_i = weight of individual i
- x_{1i} = dosage of allele$_2$ of SNP S in individual i (0/1/2)
- $x_{2i} + \ldots + x_{(p-1)i}$ = covariates (age, gender, diet) in individual i
- ε_i = error or residual of the estimated weight for individual i



Caveat: a phenotype is given by the contribution of both genetic and non-genetic effects

- it might be that, by coincidence, there are more males than females in the GG group, thus we can't know a priori if the difference in weight is purely given by the effect of the SNP
- it might be that, by coincidence, the diet fatty-acid content varies between the three groups

Goals when performing multiple linear regression:

- Obtain the equation that models the relationship between y and the predictors x
- Test if a specific explanatory variable x has a significant effect in predicting y
  - We are interested in evaluating the effect of SNP *S* on weight

Determining the effect of a SNP on the trait

Question: Does the genotype of SNP *S* (x_1) have a significant effect on the weight of an individual?

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

The estimated effect of SNP *S* on weight is b_1 (or $\widehat{\beta_1}$)

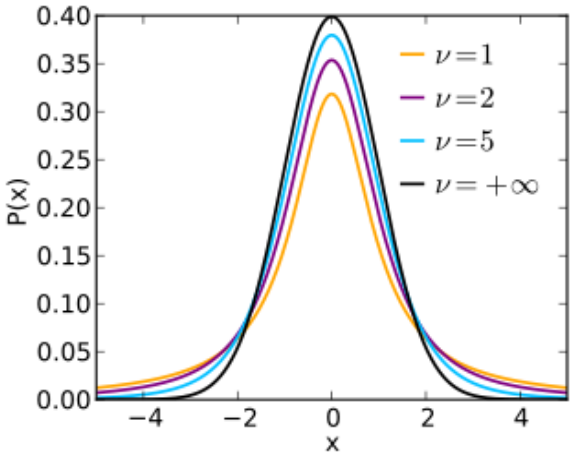
- Under the null hypothesis (no effect of SNP *S* on weight), $\beta_1 = 0$
- We can use the *t*-statistic to compute whether b_1 is significantly different from β_1 (0)

$$t = \frac{b_1 - \beta_1}{SE_{b1}}$$

$t \sim t_{STUDENT}$

$v = n - 2$

$n =$ n of indivs.



- p-value < α: reject the null hypothesis, the SNP has a significant effect on weight
- p-value ≥ α : accept the null hypothesis, the SNP does not have a significant effect on weight
- α can be 0.05, 0.01, 0.001

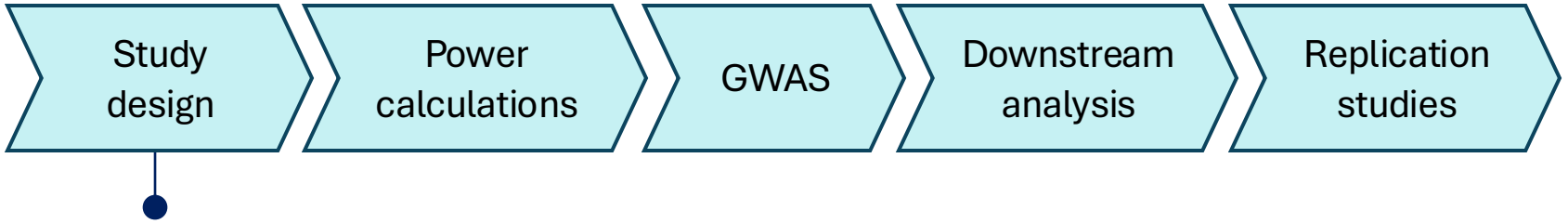
Slides from B Borsari & S Liu

Outline

- Part 1 : Generic Annotation
  (not related to an individual's variants)
  - RNA-seq, Chip-seq
  - Integration
  - , Hi-C
- **Part 2 : Annotation related to an individual's variants**
  - ASE/ASB
  - GWAS & eQTL

GWAS (Additional Considerations)

GWAS workflow



Type of study

- Quantitative trait
- Case-Control study (example: disease vs. healthy)

Balding, Nat Rev Genet, 2006

GWAS workflow

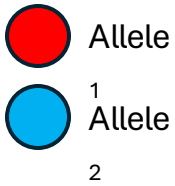
Study design → Power calculations → GWAS → Downstream analysis → Replication studies

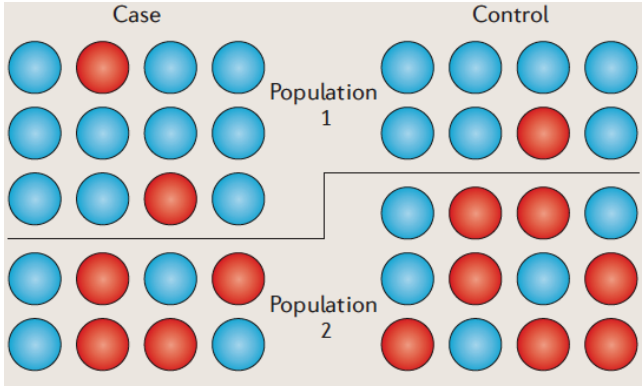
Type of study

- Quantitative trait
- Case-Control study (example: disease vs. healthy)

Population stratification

- Some SNPs might have different allele frequencies in different subpopulations
- EX: In comparing Asian vs. European, what if one did GWAS for “uses chopstick” without correction





- $allele_2$ is enriched in cases
- BUT cases are enriched in population 1, where $allele_2$ is more frequent

Balding, Nat Rev Genet, 2006

Uffelmann et al. (2021). Nature Reviews Methods Primers

GWAS workflow

Study design → Power calculations → GWAS → Downstream analysis → Replication studies

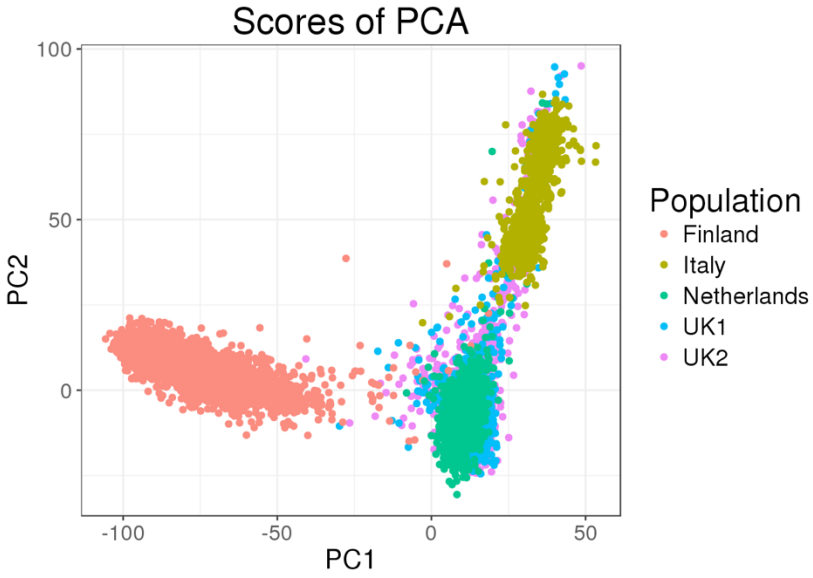
Type of study

Population stratification

- Some SNPs might have different allele frequencies in different subpopulations (eg. Asian vs. European)

Choice of relevant covariates

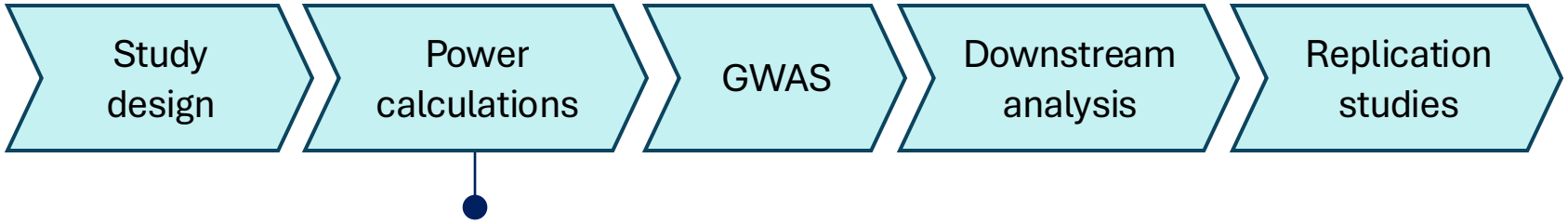
- Purpose: control for indirect effects unrelated to the phenotype of interest and eliminate the influence of confounders
- e.g., age, sex, genotyping batch
- First 5 or 6 Principal Components based on ancestry are usually included as model covariates in order to control for ancestry-related genetic variation that could confound results



https://privefl.github.io/bigsnpr/articles/how-to-PCA.html

Slides from B Borsari & S Liu

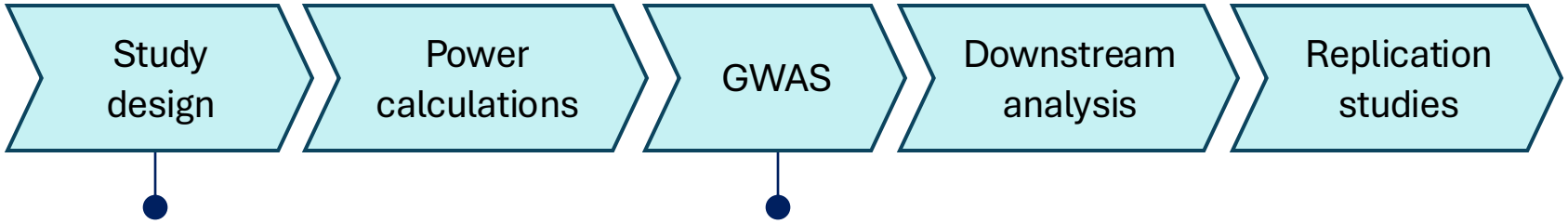
GWAS workflow



- Power is the probability that a SNP is truly associated with a trait
- It depends on sample size, allele frequency and effect size

  - Larger sample size *n* and MAF *f* result in a more accurate estimate of the SNP effect β
  - Larger absolute values of β increase the difference from the null model (e.g. same mean value of the trait across genotype groups)

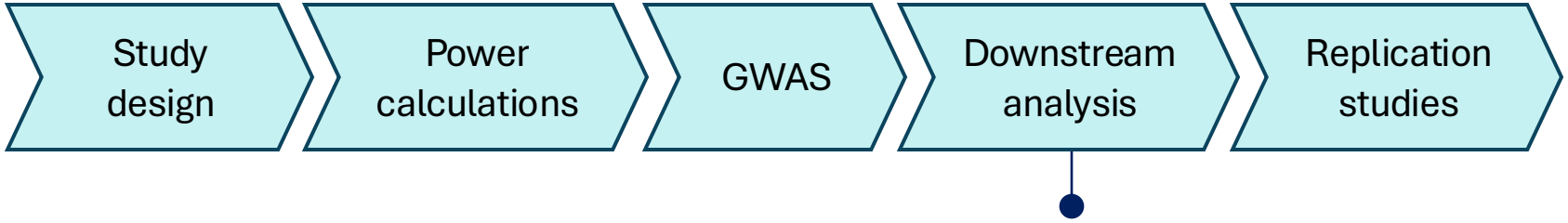
GWAS workflow



Type of study

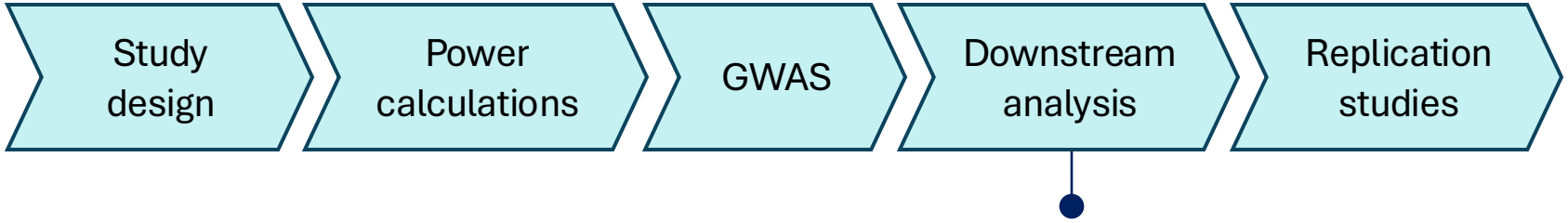
- Quantitative trait → **Statistical model**: Linear regression (beta values)
- Case-Control study → **Statistical model**: Logistic regression (OR)

GWAS workflow



- Because of LD, many significant SNPs are indeed the result of indirect associations

GWAS workflow



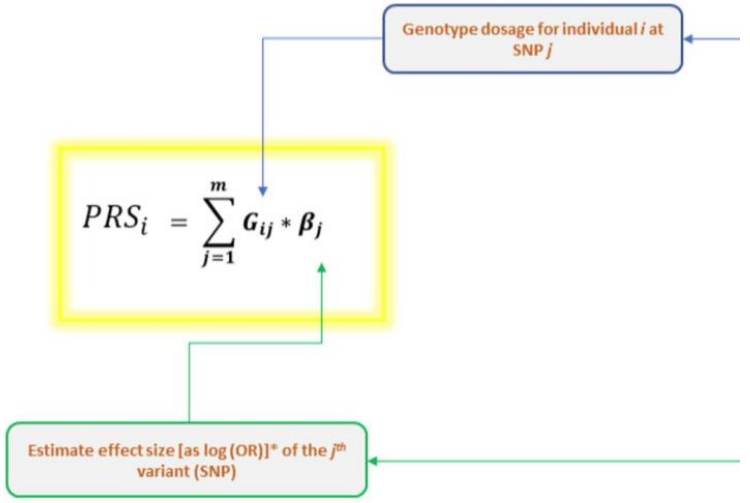
- Multiple testing Bonferroni correction:
  - GWAS test millions of SNPs for association with traits (multiple hypotheses)
  - Without correction, the chance of obtaining false positives increases dramatically
  - Controlling Family- Wise Error Rate(FWER) ensures the overall rate of false positives remains at a desired significance level (eg. 5%)
    - FWER = $\frac{\alpha}{m}$,
      - m = # of independent hypotheses
      - # of independent common variants = 10^6
      - FWER = $0.05/10^6 = 5 \cdot 10^{-8}$

Slides from B Borsari & S Liu

GWAS workflow

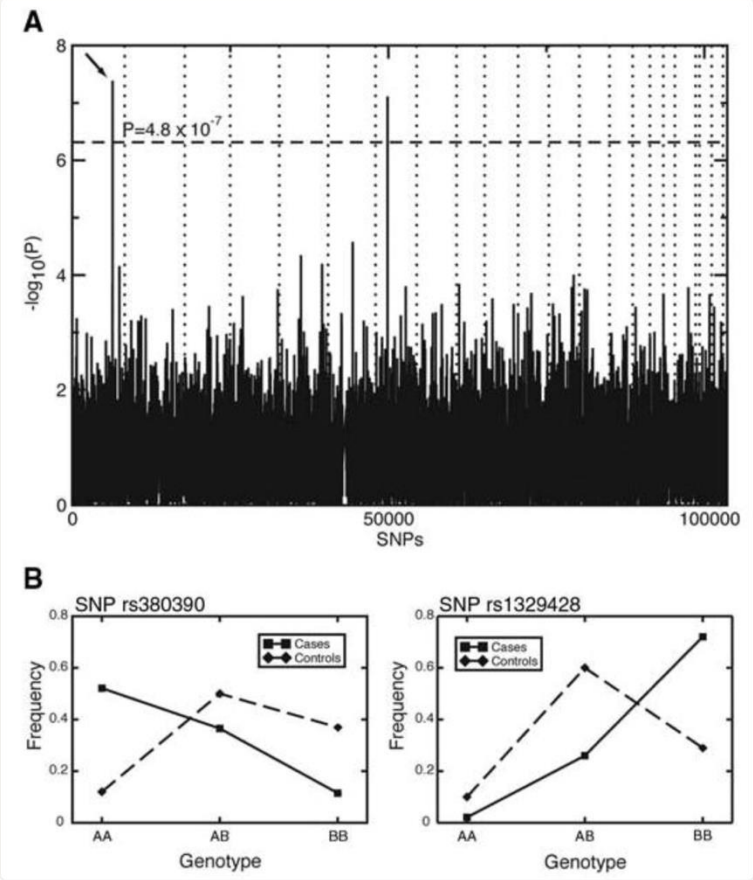
Study design → Power calculations → GWAS → Downstream analysis → Replication studies

- Constructing Polygenic Risk Score (PRS)
  - Include SNPs below a p-value threshold + in low LD to retain independent signals



Sima et al, Human Genetics, 2024

First GWAS: at Yale

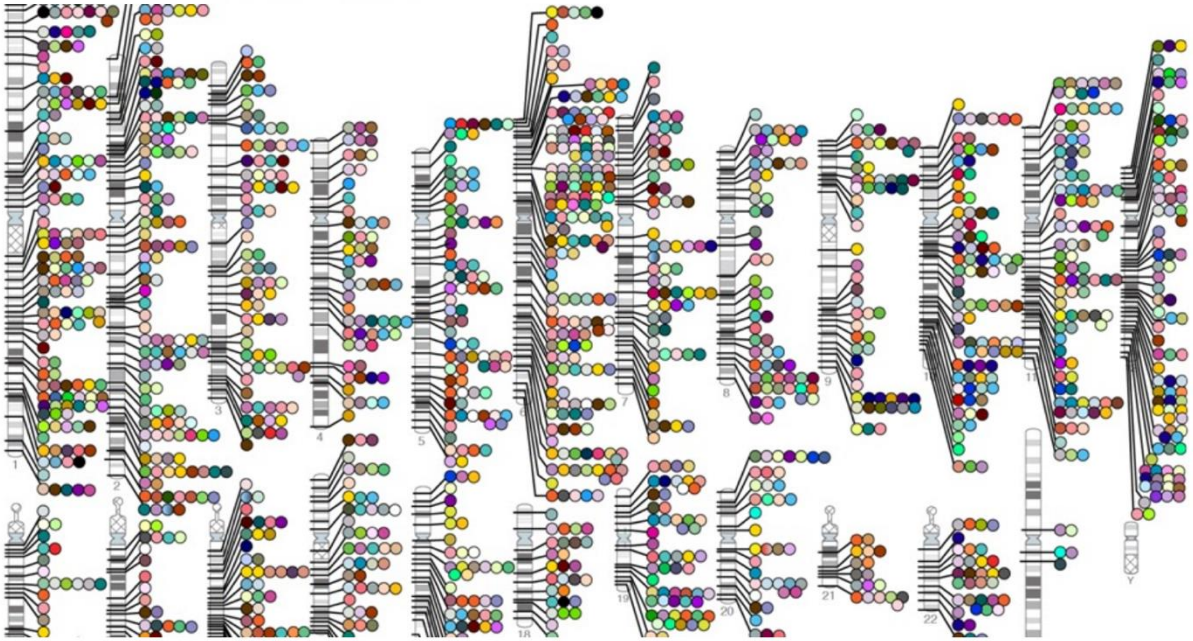


Scientists used genome-wide association to identify genes that affect the risk of developing Age – related macular degeneration

Klein et al, Science 2006

The NHGRI-EBI GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies: https://www.ebi.ac.uk/gwas/



As of 2022-10-08, the GWAS Catalog contains 6041 publications and 427870 associations.
GWAS Catalog data is currently mapped to Genome Assembly GRCh38.p13 and dbSNP Build 154.

Outline

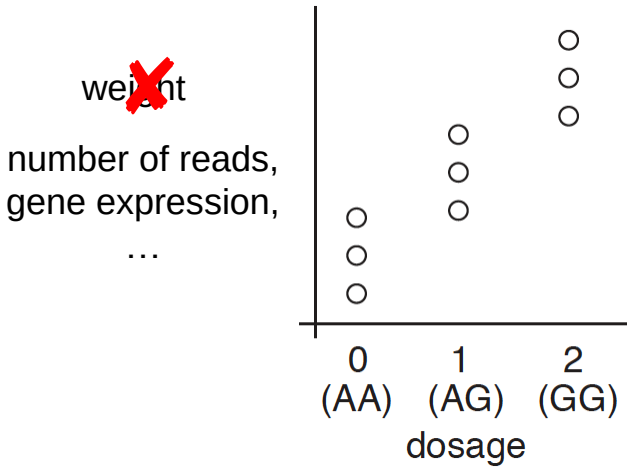
- Part 1 : Generic Annotation
  (not related to an individual's variants)
  - RNA-seq, Chip-seq
  - Integration
  - , Hi-C
- **Part 2 : Annotation related to an individual's variants**
  - ASE/ASB
  - GWAS & eQTL

eQTL

Molecular quantitative trait loci

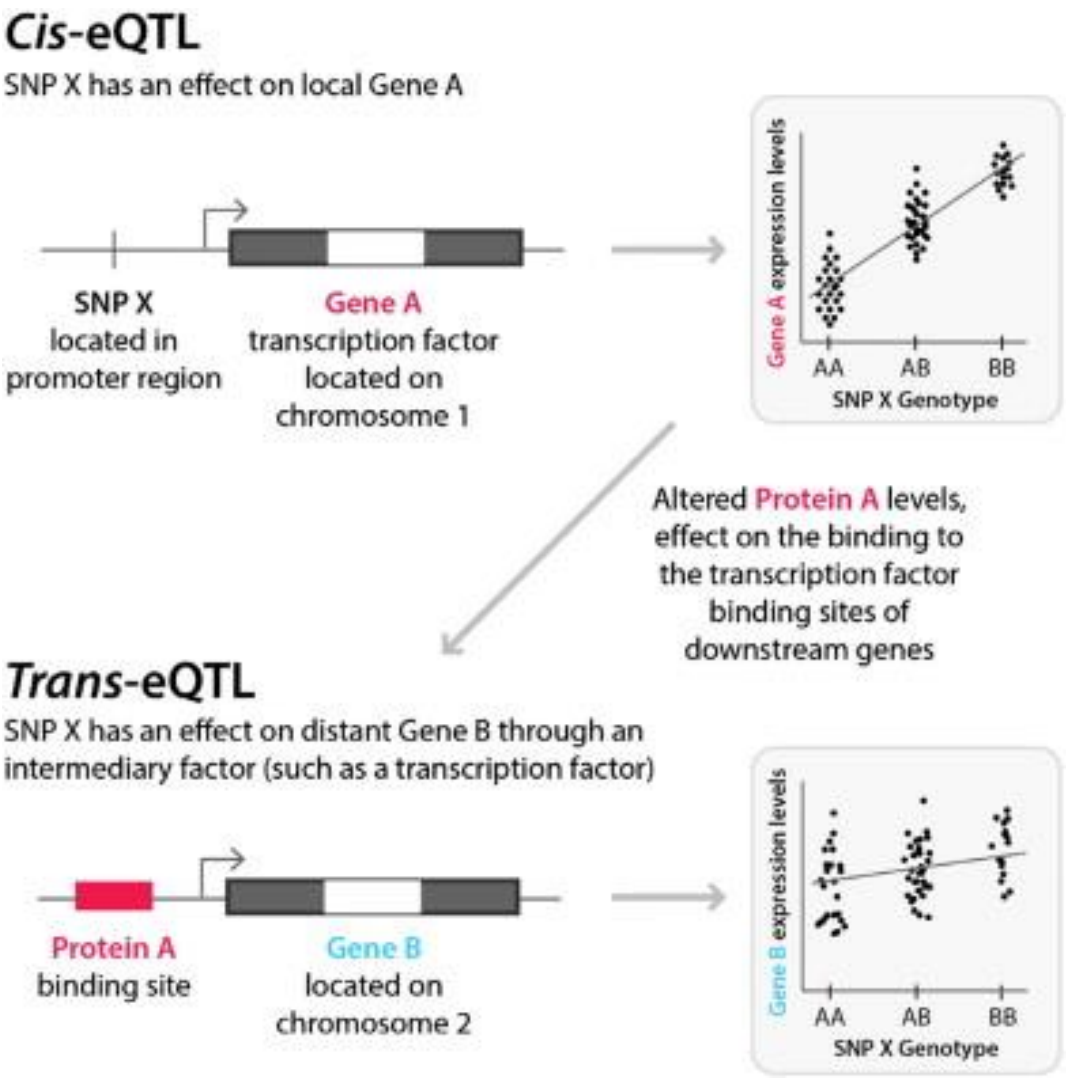
Genome sequence → **Molecular phenotypes** → Organismal traits / diseases

- **molecular quantitative trait loci (QTLs)**
- allele-specific (AS) events



- Population-scale analysis
- Same concept as GWAS for quantitative traits (linear models, effects modeled as beta coefficients)

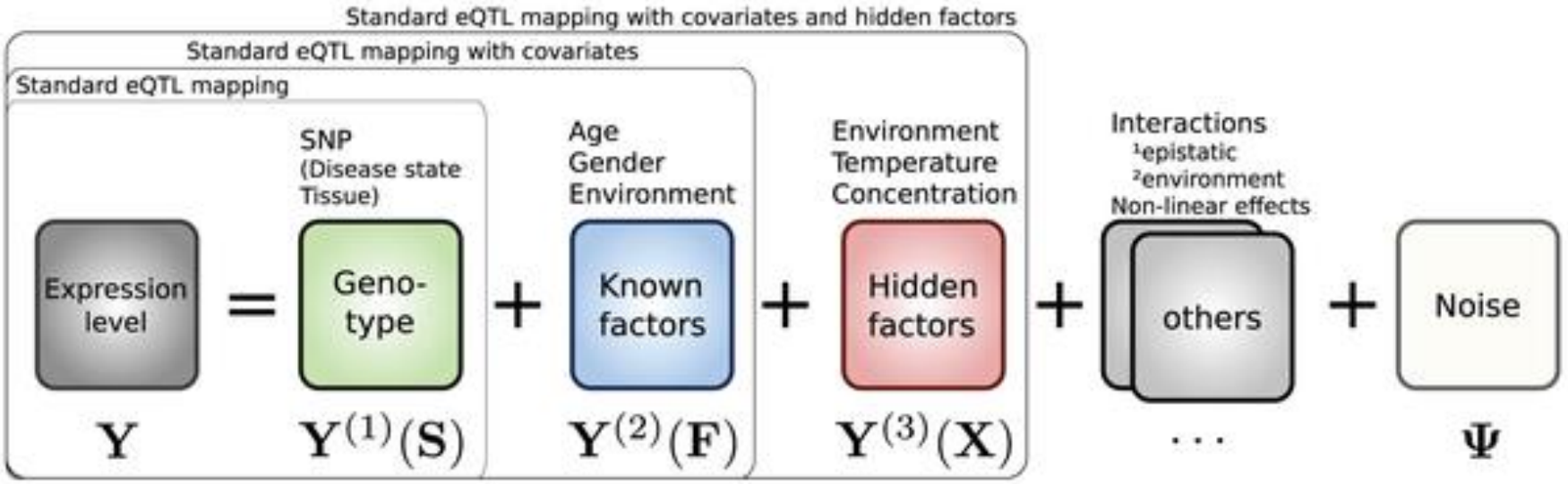
Expression quantitative trait locus (eQTL)



Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 2014, 10:1896-1902

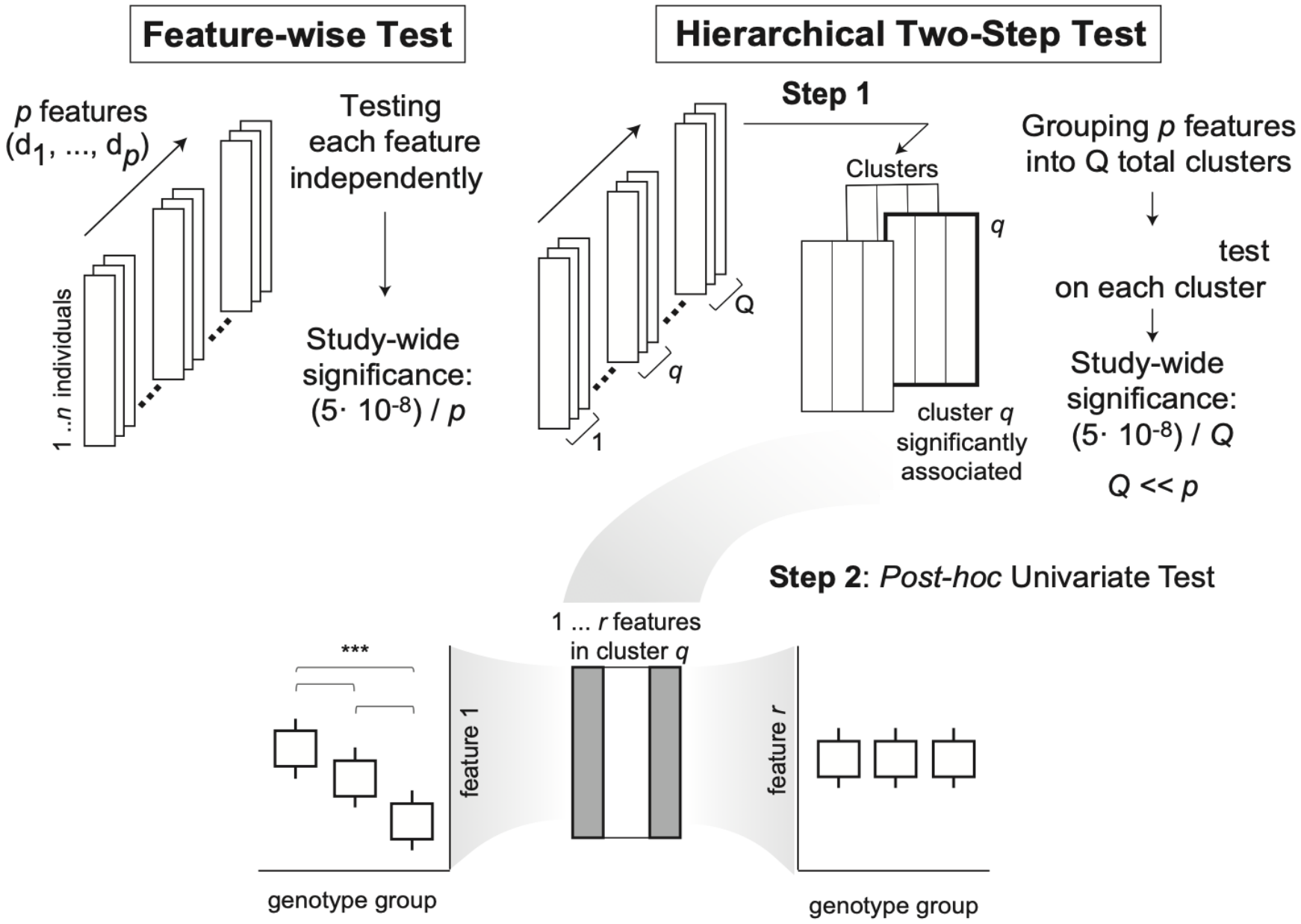
Aspects of Scaling eQTL calculation to Many SNPs & Many Samples

Taking into account covariates
General additive model for sources of gene expression variability.



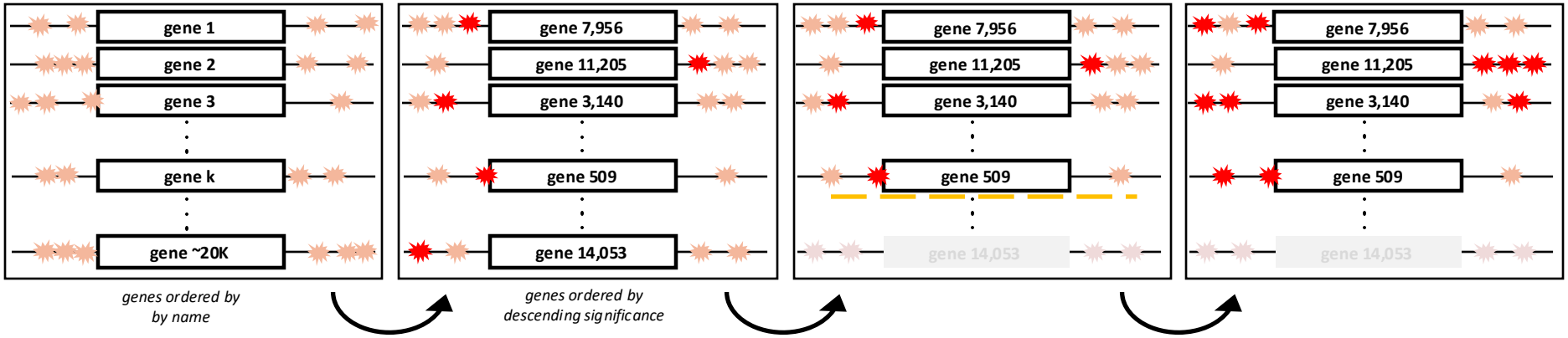
Stegle O, Parts L, Durbin R, Winn J (2010) A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. PLOS Computational Biology 6(5): e1000770. https://doi.org/10.1371/journal.pcbi.1000770
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000770

Benefits of Hierarchical Testing

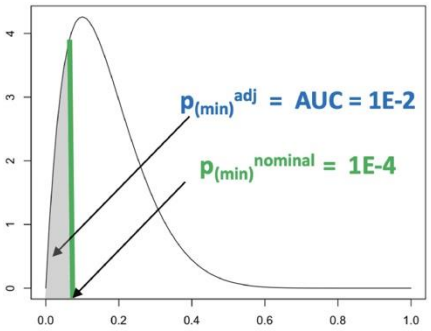


[Liu et al. ('25). Cell]

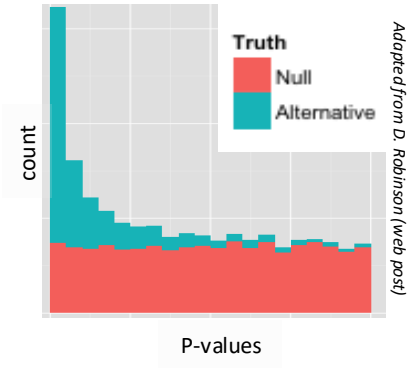
Huge burden of multiple testing in genome wide eQTLs.
Thus, use of a "cis" window around gene for cis-eQTLs +
multi-step (hierarchical) scheme to identify significant eGenes & their associated eSNPs



Step 1: Identify the **most significant eSNP** per gene, and then correct p-values for multiple testing within each gene to derive adjusted *gene-level p-values*



Step 2: Multiple testing correction (BH to estimate FDR) is applied to the set of all adjusted gene-level p-values to yield the **threshold** for defining *significant eGenes* (FDR 0.05)



Step 3: Pull in all *significant eSNPs* associated with each significant eGene by using the scheme adopted by GTEx: for each gene, a nominal p-value threshold (derived using the beta distribution in **Step 1**) is used to pull in **the full set of significant eSNPs** for each significant eGene

References for 25m7 part-2 (Annotation Related to Variants)

- Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Nature Reviews Methods Primers, 1(1).
  Genome-wide association studies.
  https://doi.org/10.1038/s43586-021-00056-9
  (Focus on the beginning up to Fig 2. Stop at the results section.)

- Aguet, F., Alasoo, K., Li, Y. I., Battle, A., Im, H. K., Montgomery, S. B., & Lappalainen, T. (2023).
  Nature Reviews Methods Primers, 3(1).
  Molecular quantitative trait loci.
  https://doi.org/10.1038/s43586-022-00188-6
  (Focus on the beginning. Stop at the results section.)

- James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert
  An Introduction to Statistical Learning: with Applications in R [ISLR (2nd edition)]
  https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1071614177/ +
  https://www.statlearning.com
  (Chapter 3.1 & 3.2 gives basic background on linear regression.
  Likewise, chap 13 (13.1 to 13.3) gives background on multiple testing.)

- Chen, J., Rozowsky, J., Galeev, T. R., Harmanci, A., Kitchen, R., Bedford, J., Abyzov, A., Kong, Y., Regan, L., & Gerstein, M. (2016). Nature Communications, 7(1).
  A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals.
  https://doi.org/10.1038/ncomms11101
  (Methods section up to “AlleleDB” part)