Statistical methods and considerations for genome-wide association testing

Seth D. Temple
Sharon R. Browning
Timothy A. Thornton

BIOST 550
Seattle, WA, USA
June 3, 2022

UNIVERSITY of WASHINGTON

Agenda

1. **Review**
   - 1.1 Tests on contingency tables
   - 1.2 Asymptotically χ^2 statistics
   - 1.3 Linear modeling
   - 1.4 Multiple testing
2. Genome-wide association studies (GWAS)
   - 2.1 Binary response
   - 2.2 Quantitative response
   - 2.3 Fixed effects
   - 2.4 Random effects
     - 2.4.1 Heritability
   - 2.5 Fine mapping
3. Genetic epidemiology
   - 3.1 Linkage mapping
   - 3.2 IBD mapping
   - 3.3 Admixture mapping
   - 3.4 TWAS

Allelic, genotypic tests for case/control data

Principle: bin counts in a contingency table follow some multinomial distr.

- Allelic tests
  - No close relatives
  - HWE assumed
  - One degree of freedom $\rightarrow$ more power
  - Pearson χ^2 statistic; LRT; exact test; normal approx.
  - PLINK option `--assoc`
- Genotypic tests
  - No close relatives
  - HWE **not** assumed
  - Two degrees of freedom $\rightarrow$ less power
  - Pearson χ^2 statistic; LRT; exact test; trend test
  - PLINK option `--model`

Asymptotically χ^2 statistics

- Pearson χ^2 statistic
  - Contigency table data
  - $\sum_{\text{types}}(O-E)^2/E$ for O observed, E expected counts
  - $(r-1)(c-1)$ df where r, c are row, column size
- Likelihood ratio test statistic
  - $-2 * (\ell(\hat{\theta}_0) - \ell(\hat{\theta}))$ where ℓ is log-likelihood, $\hat{\theta}_0$ is MLE under null, $\hat{\theta}$ is unconstrained MLE
  - $d - d_0$ df where $d_0(d)$ are size of (un)constrained space
- Score test statistic
  - $S(\hat{\theta}_0)^T \mathcal{I}(\hat{\theta}_0)^{-1} S(\hat{\theta}_0)$ where S is score (derivative of ℓ), $\mathcal{I}$ is information matrix
  - $\hat{\theta}_0$ available; $\hat{\theta}$ not available
- Wald test statistic
  - $\mathcal{I}(\theta_0)^{1/2}(\hat{\theta} - \theta_0)$ normally distributed
  - $(\hat{\theta} - \theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)$ χ^2 distributed

Geometry of χ^2 test statistics

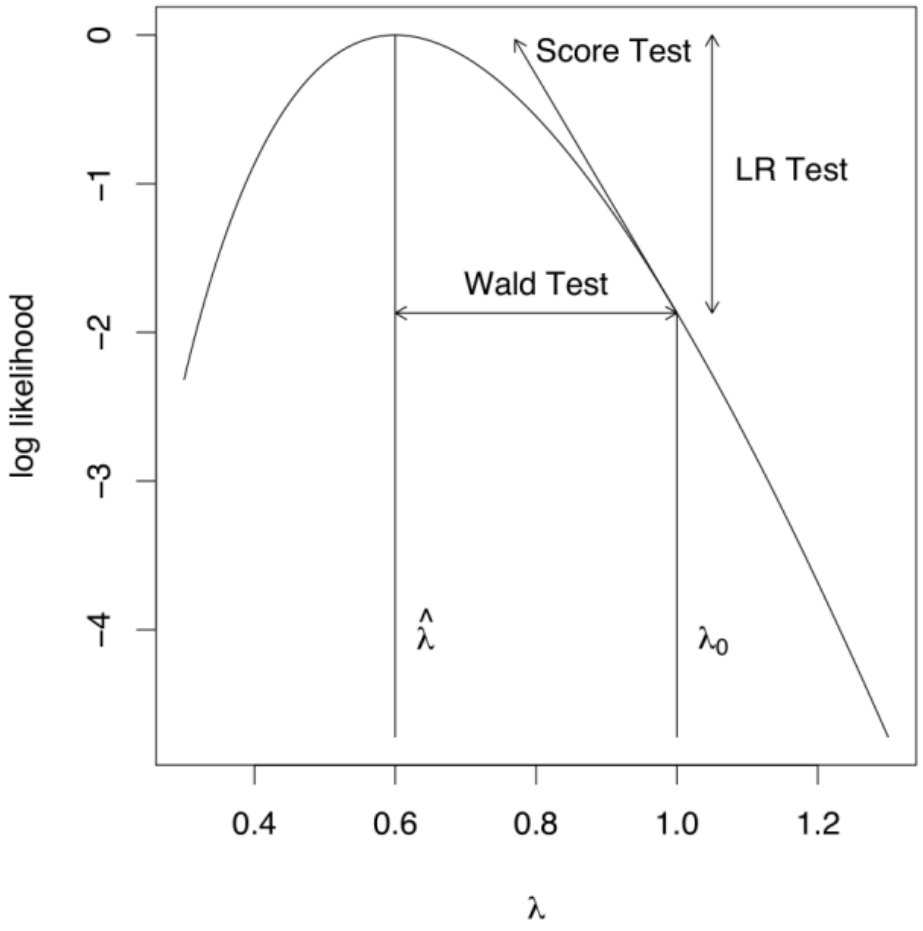


Figure: Geometric interpretation from (Wakefield, 2013)

Linear modeling

Notation

- Y is trait
- X_1 is covariate of interest (0/1 or 0/1/2 valued)
- $X_2, \ldots$ are other covariates (age, sex, etc.)
- ε is error term
- $g(\cdot)$ is link function (identity, log odds)

Model

$$g(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$
$$g(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

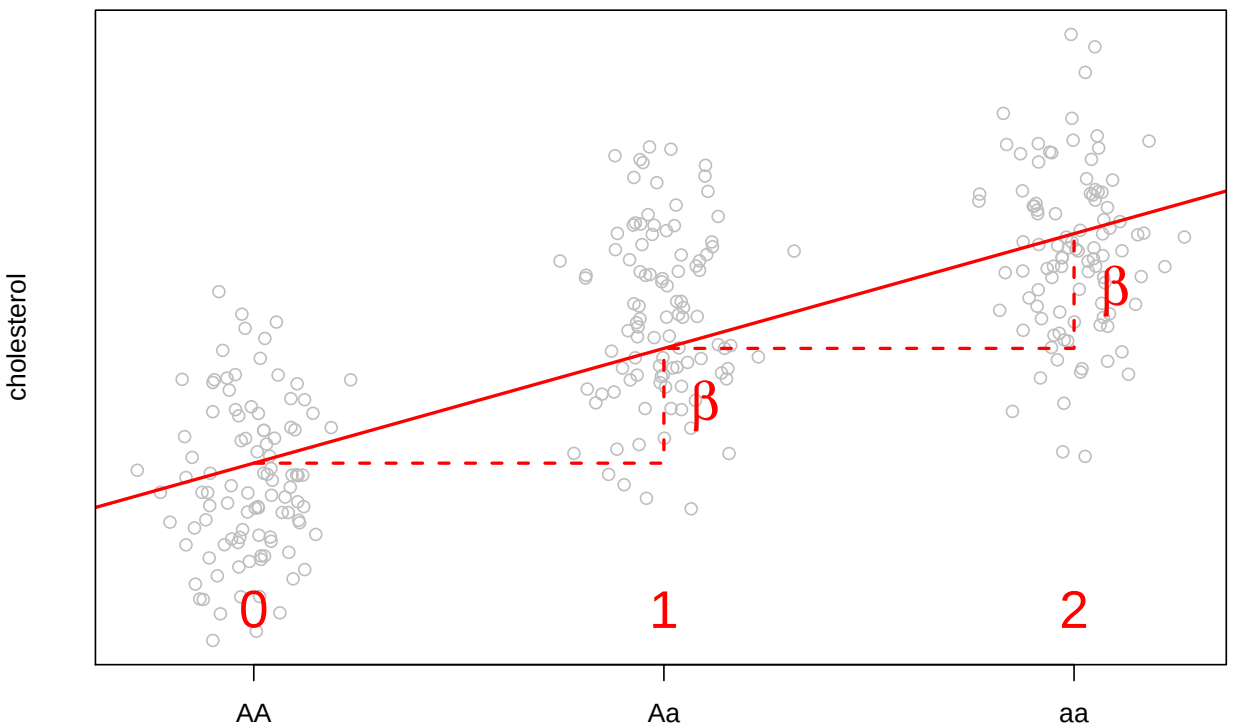
Hypothesis testing

- $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$
- Wald, score, or LR test

Linear regression with SNPs

Many analyses fit the 'additive model'

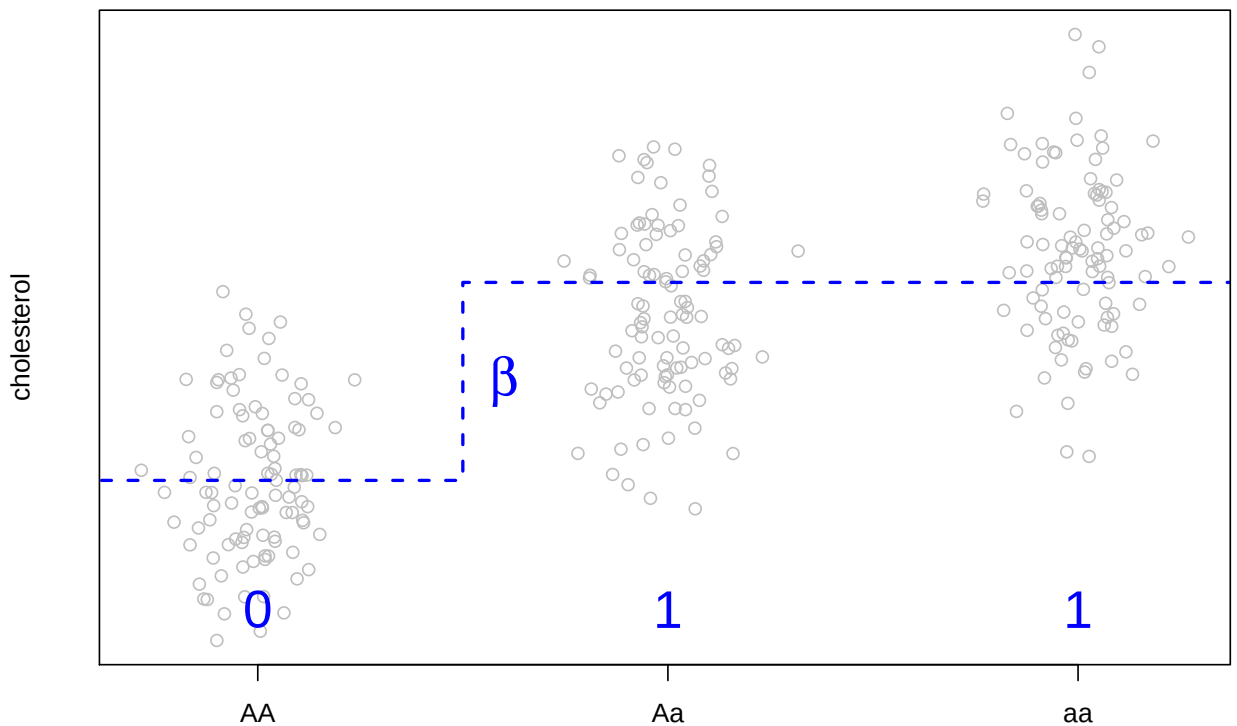
$$y = \beta_0 + \beta \times \#\text{minor alleles}$$



Linear regression, with SNPs

An alternative is the 'dominant model';

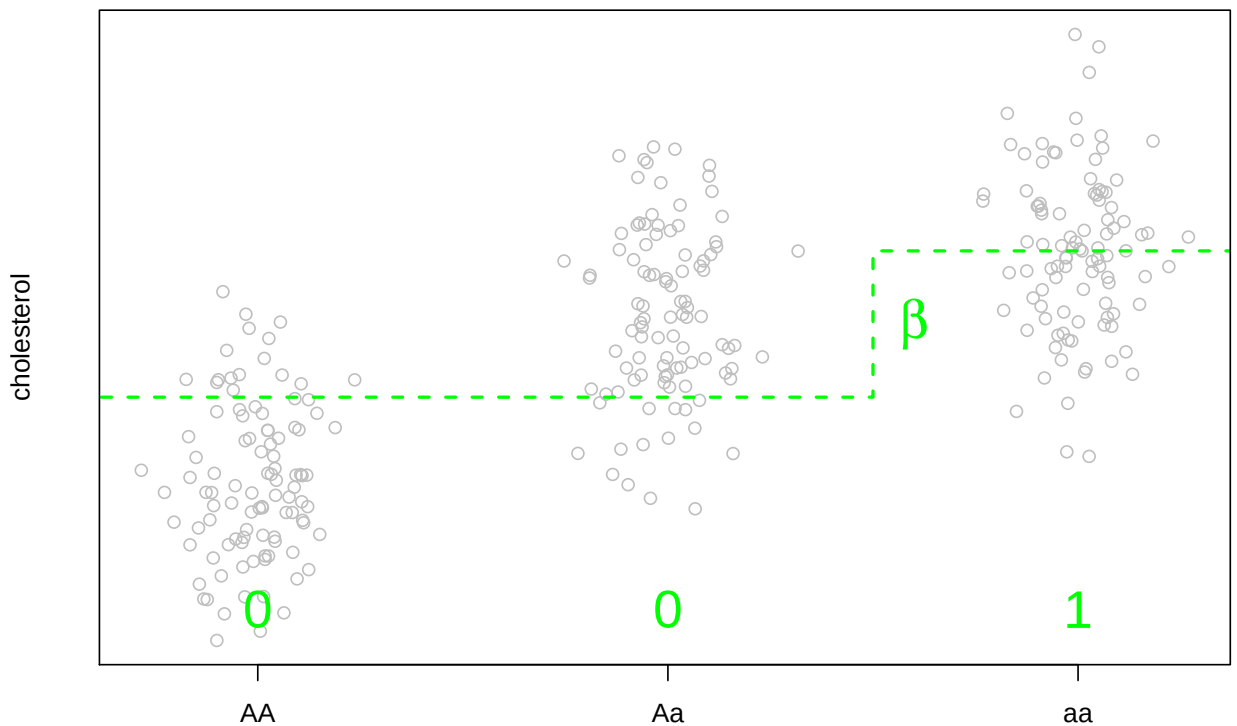
$$y = \beta_0 + \beta \times (G \neq AA)$$



Linear regression, with SNPs

or the 'recessive model';

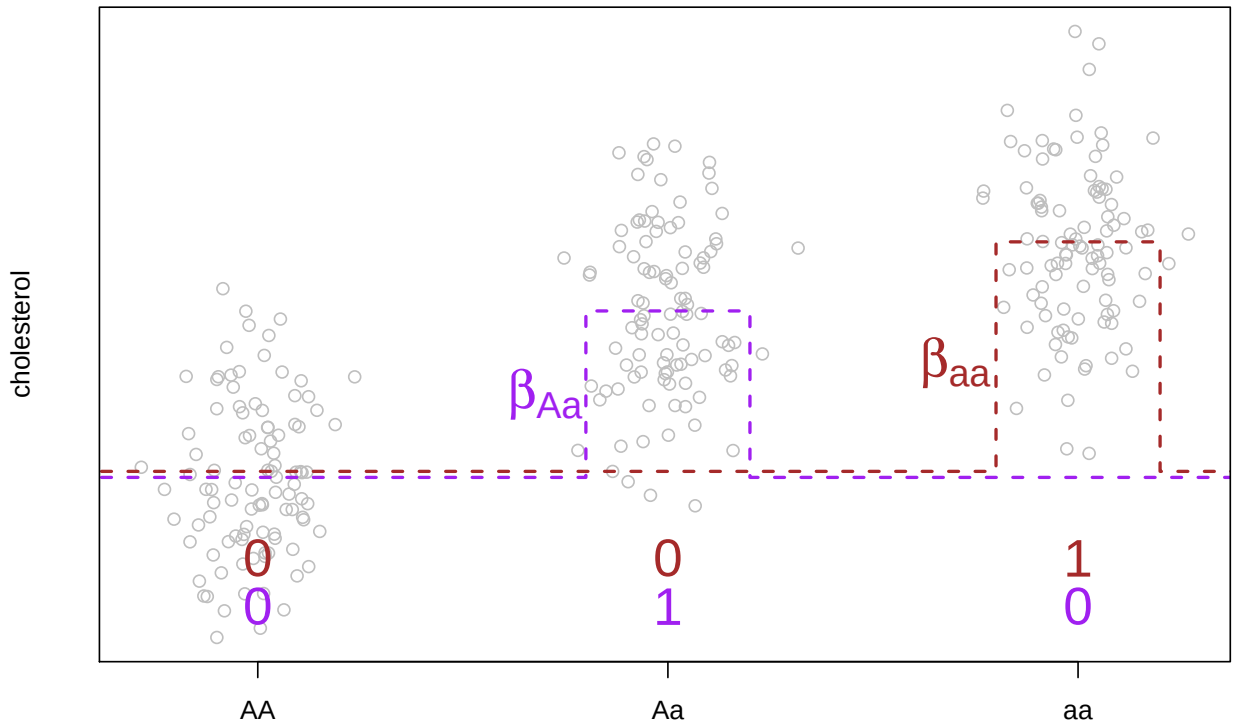
$$y = \beta_0 + \beta \times (G == aa)$$



Linear regression, with SNPs

Finally, the 'two degrees of freedom model';

$$y = \beta_0 + \beta_{Aa} \times (G == Aa) + \beta_{aa} \times (G == aa)$$



Multiple testing

Family-wise error rate (FWER):
prob. making 1 or more false positives

Problem: **control FWER** at level α
Solution:
- Bonferroni method: $\alpha_0 = \alpha/$ (# independent tests)
- α_0 is level for each marginal test

Challenge:
- Hypothesis tests in genetics are correlated.
- In GWAS, how many **independent tests** are conducted?
- In admixture (IBD) mapping, how many independent tests?

Agenda

1. Review
   - 1.1 Tests on contingency tables
   - 1.2 Asymptotically χ^2 statistics
   - 1.3 Linear modeling
   - 1.4 Multiple testing
2. **Genome-wide association studies (GWAS)**
   - 2.1 Binary response
   - 2.2 Quantitative response
   - 2.3 Fixed effects
   - 2.4 Random effects
     - 2.4.1 Heritability
   - 2.5 Fine mapping
3. Genetic epidemiology
   - 3.1 Linkage mapping
   - 3.2 IBD mapping
   - 3.3 Admixture mapping
   - 3.4 TWAS

Genome-wide association study

1. Data collection
2. Genotyping, quality control
3. Imputation (phasing)
4. **Association testing**
   - Some χ^2 test for $\beta_1 \neq 0$ in linear model
5. Meta analysis, replication studies
6. Follow-up analyses
   - Laboratory experiments
   - Mendelian randomization
   - and so much more ...

More details in Tam et al. (2019) and Uffelman et al. (2021).

GWAS: binary response

Model

$$
\begin{aligned}
g(Y) &= \beta_0 + \beta_1 X_1 + \varepsilon \\
g(\mathbb{E}[Y]) &= \beta_0 + \beta_1 X_1 \\
\text{Var}(\varepsilon) &= \sigma_e^2 I_n \\
g(p) &= \log(p/(1-p)) \quad \text{(logit link)}
\end{aligned}
$$

Hypothesis testing

$H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$

Interpretation

$$\mathbb{E}[Y] = (1 - \exp(-(\beta_0 + \beta_1 X_1)))^{-1}$$

β_1 models odds ratio $\mathbb{E}[Y|X_1 = 1]$ versus $\mathbb{E}[Y|X_1 = 0]$

GWAS: quantitative response

Model

$$
\begin{aligned}
g(Y) &= \beta_0 + \beta_1 X_1 + \varepsilon \\
g(\mathbb{E}[Y]) &= \beta_0 + \beta_1 X_1 \\
\text{Var}(\varepsilon) &= \sigma_e^2 I_n \\
g(y) &= y \quad \text{(identity link)}
\end{aligned}
$$

Hypothesis testing

$H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$

Interpretation

β_1 models difference $\mathbb{E}[Y|X_1 = 1] - \mathbb{E}[Y|X_1 = 0]$

GWAS: fixed effects

Fixed effects

- Matrix $\mathrm{X}_{2:p}$ contains covariates
- E.g., sex, age, batch, self-identified race?!?
- PCs for global ancestry
- Known causal genotype (e.g., APOE for AD)

Model

$$
\begin{aligned}
g(Y) &= \beta_0 + \beta_1 X_1 + \boldsymbol{\beta}_{2:p} \mathrm{X}_{2:p} + \varepsilon \\
g(\mathbb{E}[Y]) &= \beta_0 + \beta_1 X_1 + \boldsymbol{\beta}_{2:p} \mathrm{X}_{2:p} \\
\mathrm{Var}(\varepsilon) &= \sigma_e^2 I_n
\end{aligned}
$$

Interpretation

Mean model conditional on covariates, namely $\mathbb{E}[Y|\mathrm{X}_{0:p}]$

GWAS: random effects

Random effects

- $\boldsymbol{\alpha} \sim N(0, \sigma_g^2 \boldsymbol{\Psi})$, where σ_g^2 is phenotypic variance attributable to additive genetic effects
- $\boldsymbol{\Psi}$ = standardized kinship matrix
  or genetic relatedness matrix (GRM)
- Phenotypic variance = $\sigma_g^2 + \sigma_e^2$
- Heritability $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$
- Indicator matrix Z

Model

$$g(Y) = \beta_0 + \beta_1 X_1 + \boldsymbol{\beta}_{2:p} \mathrm{X}_{2:p} + \boldsymbol{\alpha} \mathrm{Z} + \varepsilon$$
$$g(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \boldsymbol{\beta}_{2:p} \mathrm{X}_{2:p}$$
$$\mathrm{Var}(\varepsilon) = \sigma_e^2 I_n$$

Heritability estimation

(Narrow-sense) heritability $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$

- $_g$ for (additive) genetic
- $_e$ for environment (or error)
- Impacts **power** to detect causal effects in GWAS !

- See Min, Thompson, and Basu (2021)
  - Definition of GRM and LD matrices
  - Details on estimators

- See Gogarten et al. (GENESIS; 2019)
  - Sparse GRM for efficient matrix inversion
  - Matrix inversion can be $O(n^3)$

Fine mapping

Goal: find small set of variants that explain association signal

Challenge: variants are in LD

- Frequentist
  - Perform cond. assoc. analyses on lead variant(s)
  - Forward stepwise selection (link)
- Bayesian
  - Credible set of variants that explain $100(1-\alpha)\%$ of signal
  - Based on posteriors or Bayes factors (link)
  - Priors can consider additional info:
    imputation accuracy, MAF, etc.

Trans-ethnic and admixed populations can refine location if causal variants are shared.

Agenda

1. Review
   - 1.1 Tests on contingency tables
   - 1.2 Asymptotically χ^2 statistics
   - 1.3 Linear modeling
   - 1.4 Multiple testing
2. Genome-wide association studies (GWAS)
   - 2.1 Binary response
   - 2.2 Quantitative response
   - 2.3 Fixed effects
   - 2.4 Random effects
     - 2.4.1 Heritability
   - 2.5 Fine mapping
3. **Genetic epidemiology**
   - 3.1 Linkage mapping
   - 3.2 IBD mapping
   - 3.3 Admixture mapping
   - 3.4 TWAS

GWAS variants

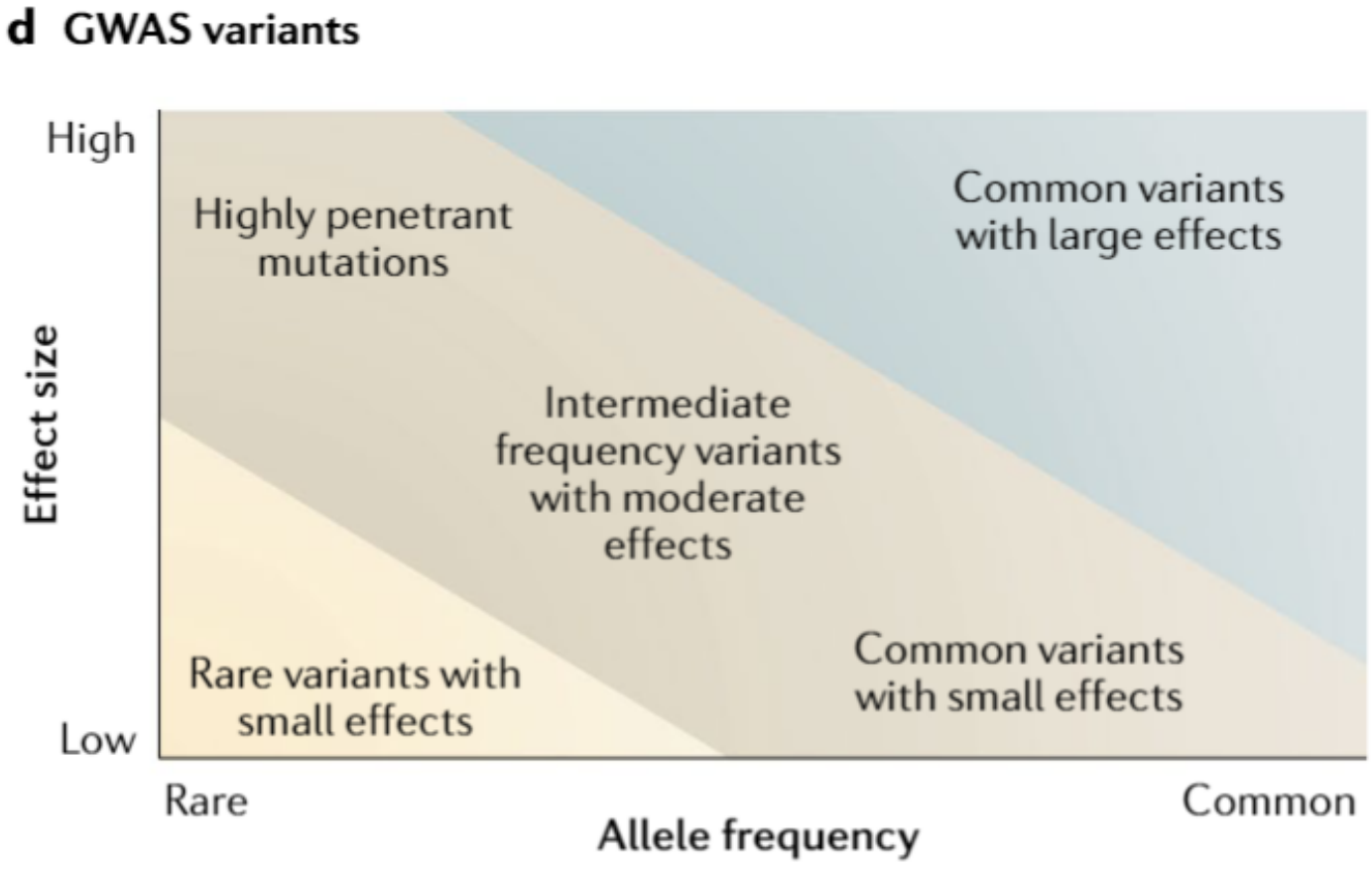


Figure: Variant types in between the two diagonals are found in GWAS (Tam et al., 2019)

Linkage mapping

Focus of lectures in weeks 6 and 7

- Test if marker is linked to trait
  - Mendelian trait acts like causal locus
- Powerful for **rare familial** diseases
- Requires **pedigrees**
  - Hard to ascertain large samples
  - Computationally intensive
- Find general location of causal variant
  - About 10 cM regions

IBD mapping

Test if cases share more IBD segments around causal variant.

- Compare to nonparametric linkage analysis
- Powerful for **multiple rare** vars. of **moderate effect** size
  - Middle ground between GWAS and linkage mapping
- Does not require pedigrees $\rightarrow$ bigger sample size
- Find general location of causal variant(s)
- Amenable to mixed effects model
- See Browning and Thompson (2012)

Admixture mapping

Test if local ancestry associates with trait.

- Compare to IBD mapping
- Follow-up study to GWAS for complex traits
  - Find signals for variants not genotyped, imputed
  - Characterize disease etiology + demography
- Must ascertain admixed sample $\rightarrow$ smaller sample size
- Find general location of causal variant(s)
- Amenable to mixed effects model
- See “Overview of Admixture Mapping” (Shriner, 2017)

Transcriptome-wide association study (TWAS)

Test if predicted gene expression associates with trait.

- Complementary/supplementary to GWAS
  - Gene-based $\rightarrow$ lower multiple testing burden
  - Interpretable transcription hypotheses
  - Relates complex traits to regulation
- Depends on prediction model from GTEx project (link)
  - Which may be a black box ...
- Results may be tissue-specific
- Be cautious making causal claims !
- See references (link, link)

Polygenic risk score (PRS)

It's just a linear model (IJALM) !

Twitter thread

Based on summary statistics from GWAS

- Tested SNPs, locations
- Effect sizes and standard errors
- Test statistics and *p*-values
- Minor allele frequencies
- Sample size

Report the above! Make it publicly available if possible!

GWAS catalog

GWAS: pros and cons

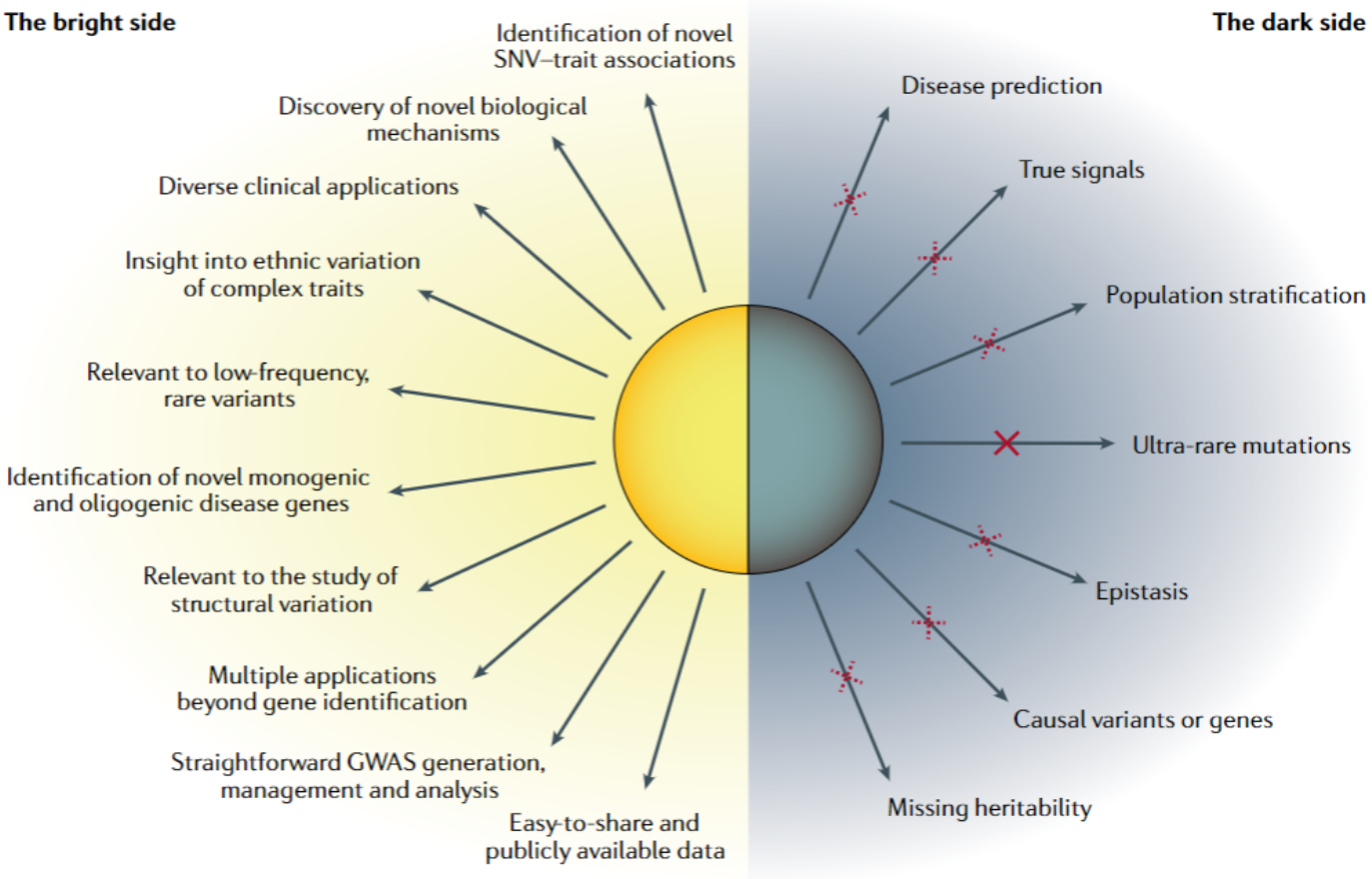


Figure: Pros and cons of GWAS (Tam et al., 2019)

GWAS: the iceberg

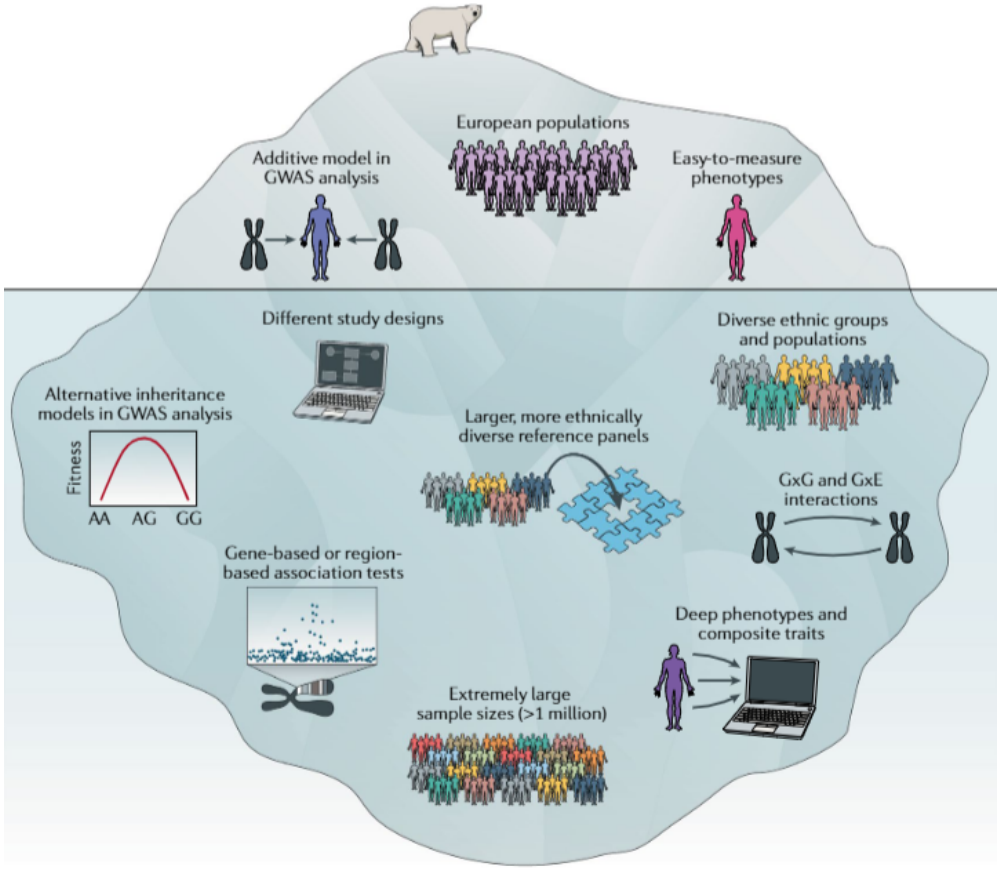


Figure: GWAS performed to date represent the tip of the iceberg (Tam et al., 2019)