Statistical Genetics, Fall 2025

Class notes 5: Introduction to GWAS

We are interested in general in understanding the mechanisms underlying phenotypes (properties of organisms). The factors affecting phenotypes can be broadly categorized into:

1. Genetic factors G
2. Environmental factors E : Diet, radiation, pollution, etc.
3. "Random" noise ϵ – can be truly random or due to unmeasured effects, typically assumed to be independent of G, E.

$E' = (E, \epsilon)$ are sometimes put together in genetic studies, assuming most of E is due to unmeasured covariates anyway.

For a continuous pheontype, say for Y =height we can write $Y = F(G, E) + \epsilon = G(G, E')$.

For $Y \in \{0, 1\}$ (say 0=healthy, 1=sick in disease) we can write $\mathbb{P}(Y = 1|G, E) = F(G, E)$, where the randomness due to unmeasured error comes in through $F(G, E)$ being far from 0 or 1.
GLM Example: Assume G, E are scalar, and a logistic or probit model

$$\mathbb{P}(Y = 1|G, E) = F(G, E) = \frac{\exp(f(G, E))}{1 + \exp(f(G, E))}, \text{ or } \mathbb{P}(Y = 1|G, E) = F(G, E) = \Phi(f(G, E)).$$

A different representation of the same idea: The *Liability threshold model*:

$$l = F(G, E) + \epsilon,\ Y = \mathbb{I}\{l > t\} \text{ (indicator function)}.$$

This assumes there is an *unobserved* continuous phenotype, and the binary phenotype (disease) is a threshold on this unobserved phenoytpe.

Equivalence between liability threshold and GLM view: Assume $\epsilon \sim N(0, \sigma^2)$, then

$$\mathbb{P}(Y = 1|G, E) = \mathbb{P}(F(G, E) + \epsilon > t|G, E) = \mathbb{P}(\epsilon > t - F(G, E)) = \Phi(F(G, E) - t), \text{a probit model.}$$

Other ϵ distributions can give logit or other GLMs.

Genetic factors affecting phenotypes

Mutations and other changes in the genome are expected to affect phenotypes:

1. Mutations in genes which change the protein they generate:

   - Non-synonymous mutations changing an amino-acid in the gene
   - Mutations that change the *splicing* of the gene by confusing the bounds between exons (coding) and introns (non-coding) – for example, by adding or removing a stop codon. Also insertions or deletions that change the *reading frame* (division into codons).

   Example: Most well known are the mutations in BRCA1, BRCA2 genes (about 100 known rare mutations) which confer significantly increased risk for breast cancer and other female reproductive cancers.
   Since such mutations often have a strong effect, and hence strong selection against, they are usually rare or very rare.

2. Mutations which affect the genes activation and modulation mechanisms (promoters, micro-RNA, etc.)
   Many of the GWAS findings of "common variants" which affect diseases, probably belong to this class

In the naive view, mutations which do not belong to these classes are unlikely to affect phenotypes. In practice, we find many statistical connections between genetics and phenotypes that do not clearly correspond to biology. This may be due to:

- Confounding mechanisms like linkage disequilibrium and stratification, which we will discuss in detail
- Other biological mechanisms in the genome which we don't fully understand – there is a lot of emerging knowledge on issues like epigenetics, long-range effects, diversity of functional classes in the genome, etc., which we won't really discuss in detail

Genome-wide Association Studies (GWAS)

General idea: A *statistical* search for connections between genetics (the genome) and phenotypes of interest:

- Collect a large number n of samples, measure their phenotype:
  - Quantitative phenotype (like height): collect random people, measure their phenotype
  - Binary phenotype like disease, especially when one class (sick) is rare: typically employ a case-control strategy, to make sure we have a good number of sick individuals

  Typical numbers: $2005:\ n=10^3,\ 2010:\ n=10^4,\ now:\ n=10^5-10^6$.
- Measure the genome of these individuals at a large number M of points with *common variation*: $2005:\ M\ \approx 10^5,\ 2010:\ M\approx 10^6,\ now:$ whole genome sequencing

- Search for statistical connections and associations in the resulting $n \times M$ matrix with the n vector of phenotypes

Basic questions which have to be addressed in designing GWAS and analyzing GWAS data:

1. How to select the M loci to sample? (Now somewhat obsolete)
2. How to test for association and determine statistical significance?
3. How to differentiate correlation from causation?
4. Which type of effects are we expecting to find:
   (a) Effect of one mutation at a time, independent of others?
   (b) Mode: dominant, recessive, additive?
   (c) Combination of mutations acting together (interaction/epistasis)?
   (d) More generally: which "statistical language" is appropriate to describe the relevant associations?

There are three critical elements to consider when analyzing GWAS and trying to answer the above questions:

1. **Linkage disequilibrium (LD):** Mutations that are close to each other tend to be inherited together due to non-perfect recombination. Hence if a mutation is associated causally with the phenotype, its neighbors in the genome will be associated statistically with it as well
2. **Stratification:** If in studying a disease, all our cases are African, and all our controls are European, then any genetic difference between Europeans and Africans will be statistically associated with the disease! So we have to be able to neutralize this, either by careful sampling, or more likely, by modeling and taking into account stratification in the sampling.
3. **Multiplicity:** If $M = 10^6$ and we test each locus (column) for association with the phenotype, we perform 10^6 hypothesis tests — severe problem of false discovery. The standard solution in the GWAS community is to perform all tests at level 5×10^{-8}, implicitly doing Bonferroni correction for 10^6 tests. We will discuss this and other strategies in more detail.

Measures of LD

Assume we have two binary loci, one denoted X with genotypes a, A and Y with b, B. Assume we are either considering haploid organisms, or more likely, looking at each copy of the genome (so one diploid organism is two samples). We can describe the joint distribution of the two loci via 2×2 table:

X \Y	b	B	Total
a	p_{ab}	p_{aB}	p_a
A	p_{Ab}	p_{AB}	p_A
Total	p_b	p_B	

(We can add hats and write $\hat{p}_{ab}, \hat{p}_{aB}, \ldots$ to differentiate observed data distributions from theoretical distributions).

We are interested in understanding whether the sites X, Y are "associated" by LD and how much. Intuitively this means that by knowing X we have information on Y.

A simple measure: **Lewontin's D:** $D = p_{ab} - p_a p_b = -(p_{aB} - p_a p_B) = \ldots = Cov(X, Y)$.

Example: MRCA is AB, mutation $A \to a$, followed by $B \to b$ giving:

X \Y	b	B	Total
a	0.3	0.2	0.5
A	0	0.5	0.5
Total	0.3	0.7	

For this table $D = 0.3 - 0.15 = 0.15$. However this tree has gone through no recombination!

An alternative measure which respects the phylogenetic order is D' which is D, normalized to the range $-1 \leq D \leq 1$ given then marginal distributions of X, Y:

$$D' = \frac{D}{m(p_a, p_b, sign(D))}, \quad m = \begin{cases} \min(p_a, p_b) - p_a p_b & \text{if } D > 0 \\ p_a p_b - \max(p_a + p_b - 1, 0) & \text{if } D < 0 \end{cases}.$$

For the example above we would get $m = 0.3 - 0.15 = 0.15$, so not surprisingly $D' = 1$.
Claim: For a pair of loci with no recombinations, $D' = 1$.

The problem with D' (to some extent also D): Not really clear how the values relate directly to the "amount of information X carries on Y.".

Squared correlation / variance explained r^2:

$$r^2 = cor^2(X, Y) = \frac{D^2}{p_a p_A p_b p_B}.$$

Recall the interpretation from regression as the "variance explained" by regressing Y on X or X on Y.

For the example above: $r^2 = \frac{0.15^2}{0.21 \times 0.25} = 0.42$.

r^2 and D combine information on:

1. Whether there is recombinations breaking the correlation
2. The "phylogenetic context", i.e., whether the mutations happened in a similar place in the tree

$r^2 \approx 1$ means that both conditions hold – few or no violations of the tree, and similar phylogenetic context.

Important Note: r^2 and D are not monotone decreasing as X, Y move further away along the genome — recombinations are increasing for sure, but far away mutations can still have similar phylogenetic context!
Conclusion: If X is causative for some disease, and $r^2(X, Y)$ is big, then Y is likely to also be associated with the disease only due to this correlation. This should be taken into account:

- What happens if we did not measure X at all, only Y?
- What should we conclude if we see many associated loci close together: are there independent associations, or is it all due to one association and LD? How can we use r^2 values to distinguish?

Statistical testing in case-control GWAS

Given we have collected M loci (say 10^6 in traditional GWAS), the simplest approach is to look at the data in case-control GWAS as a collection of M 2×3 tables:

Genotype	AA	AG	GG	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	n

The first and most important task is **identifying statistical association**. Most obvious solution: **Chi-squared test**.

- A chi-squared test on the 3×2 table with $2 - df$.
- Reduce to a 2×2 table by choosing inheritance mode (recessive / dominant). For example, if we assume A is the risk allele, and mode is dominant, so AA and AG both confer risk, we get:

Observed:

Genotype	AA+AG	GG	Total
Case	$r_0 + r_1$	r_2	R
Control	$s_0 + s_1$	s_2	S
Total	$n_0 + n_1$	n_2	n

Expected:

Genotype	AA+AG	GG
Case	$R(n_0 + n_1)/n$	$R(n_2)/n$
Control	$S(n_0 + n_1)/n$	$S(n_2)/n$

with test statistic:

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{H_0,\cdot}{\sim} \chi^2_1.$$

In the 2×2 case we can alternatively perform a Fisher's exact (hypergeometric) test.

Concerns and limitations with the Chi squared approach:

1. How can we efficiently test under the assumption that the effect is monontone/additive: $AA < AG < GG$ in terms of risk?

2. If we only get a p-value, what do we know about the magnitude of the effect? Can use odds ratios like:
$$\frac{(r_0 + r_1)/(s_0 + s_1)}{r_2/s_2},$$
but these are separate from the testing

3. Most important: how do we deal with having additional knowledge or assumptions, like:
   - That multiple SNPs might have simultaneous effect
   - That there are important measured environmental and other effects (smoking for lung cancer, age) that can increase power or correct stratification
   - Specific stratification due to ethnic origin

The obvious solution:

Testing using a regression approach. Can use logistic (or other relevant) regression, for example fit model of the form:

$$\log\left(\frac{\widehat{\mathbb{P}(Y=1)}}{\mathbb{P}(Y=0)}\right) = \hat{\beta}_0 + \hat{\beta}_1 SNP + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \ldots,$$

where SNP can be encoded as recessive, dominant, additive etc. and X_2, X_3 can be ethnic origin, smoking, or even another SNP, etc.
Then we can both estimate the effect of the SNP and test it for significance using standard methodology (e.g., Wald tests).

Advantages:

1. Account for possible confounders and stratification variables (testing is for each effect *given* all others)

2. Test and estimate at the same time

3. Extensive flexibility in types of variables and types of association that can be covered

4. Interpretation of coefficients as log-odds change