Statistical Genetics, Spring 2022

Class notes 5: LD, multiplicity, test statistics

Reminder: 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:

  with test statistic:

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

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	$S(n_0 + n_1)/n$	$S(n_2)/n$

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

Dealing with multiplicity in GWAS

Recall the GWAS multiple testing problem:

$$H_{0k} : \theta_k = 0 \ , \ H_{1k} : \theta_k \neq 0 \ , \ k = 1, \ldots, K,$$

where θ_k is a measure of association of the kth SNP with the phenoytpe. Typical number is $K = 10^6$.

The outcome of each test can be denoted by $D_k \in \{0, 1\}$ to denote non-reject or reject decision. We want to perform the testing in such a way to avoid false discoveries. Denote by $R = \sum_{k=1^K} D_k$ the total number of null rejections, and by $V = \sum_{k=1}^{K} D_k \mathbb{I}\{H_{0k}\}$. Then the most common measures of false discovery are:

$$FWER = \mathbb{P}\left(V > 0\right) \ , \ \ FDR = \mathbb{E}\left(\frac{V}{\max(R, 1)}\right).$$

A method for controlling FWER is *Bonferroni's method*, which amounts to performing each test at level α/K, and guarantees $FWER \leq \alpha$, while a well known approach for controlling FDR assuming the test statistics are independent (or dependent in specific ways) is Benjamini-Hochberg's suit of methods.

A common practice in GWAS is to perform all tests at level 5×10^{-8}, which corresponds to a Bonferroni correction to guarantee $FWER \leq 0.05$ with 10^6. Since tests may have complex dependence due to LD, this can be very conservative, and a common approach is to estimate the null distribution of the smallest p-values by permutations, where we repeat the following M times:

1. Permute the class labels $y_1, \ldots, y_n$ between the observations
2. Calculate all p values and record the smallest one (or l smallest ones)

Then we can guarantee $FWER \leq \alpha$ by using the αth quantile of the distribution of smallest p values as the cutoff for our study.

A common design is a two-stage design, where we perform all K tests on a subset $m < n$ of our data, apply a much less stringent p-value threshold $p < \alpha_1$ (for example $p < 0.001$), choosing some $L << K$ "candidates", then on the rest $n - m$ of the data (or in a follow-up study) test only these variants at level $\frac{\alpha/K}{\alpha_1}$. For example, if $K = 10^6, \alpha = 0.05$, and $\alpha_1 = 0.001$, the second test will be at level 5×10^{-5}. This guarantees that $FWER \leq \alpha$ (proof: HW2, problem 3). The two-stage approach has several advantages:

1. Cost: On the set of $n - m$ samples, we only need to genotype (measure) the roughly $\alpha_1 K$ SNPs that pass the first threshold, not the entire K
2. Replication: If the second set of data is genotyped independently of the first one, errors in genotyping or lab problems may not repeat, so results are less sensitive to those

One aspect that is sometimes mistakenly considered as an advantage: power (probability of discovering a true association) is generally *decreased* by the two-stage policy compared to the single-stage one (Proof - HW extra credit).

Expectation-Maximization (EM) to estimate stratification by ancestry

(*This section is primarily based on the paper Estimation of Individual Admixture by Tang et al., Genetic Epidemiology, (2005)*).

For this section we will assume that we have:

- I individuals from K different ethnic origins. For simplicity we assume $K = 2$,, mixture of European (Eu) and African (Af) ancestry, as in African-Americans. We assume in the I we have:
  - I_0 of mixed ancestry (unknown mixture proportions)
  - $I_1 = I - I_0$ of known ancestry (typically 100% from one ancestry), $I_1 = 0$ is possible
- On each individual we observe M genetic markers ($\times 2$ for two chromosomes), which may have a different distribution in Eu and Af, and therefore carry information on ancestry
- Each marker m has L_m possible values. For SNPs usually $L_m = 2$, but the markers can also be other elements like STRs with $L_m > 2$.

Notations:

- $G = \{G_{ima}\}$ – Value of the m marker in the i individual, copy $a \in \{1, 2\}$. This is a random variable.
- $P = \{P_{mlk}\}$ – Proportion of value l for marker m in population k. For example, if SNP j is always A in African and has 50% A in Europe, then $P_{j,A,Af} = 1,\ P_{j,A,Eu} = 0.5$. These are unknown parameters.
- $Q = \{Q_{ik}\}$ – Proportion of ancestry k in individual i. for $i > I_0$ this is a known binary vector $Q_{ik} \in \{0, 1\}$, while for $i \leq I_0$ this is an unknown parameter vector on the simplex.

Assumptions:

- $G_{im_1a_1}, G_{im_2a_2}$ are independent $\forall m_1, m_2, a_1, a_2$. This entails two assumptions:
  1. No LD between the markers m_1, m_2. This may not be very problematic if the M markers were samples for the sole purpose of estimating ancestry, so there are not too many of them and they are far apart.
  2. The two chromosomes of the same individual are independent, so for $m_1 = m_2$ the two copies are still independent. This is known as the Hardy-Weinberg Equilibrium (HWE) assumption, and is violated for example by marriages between relatives.

The resulting log-likelihood function:

$$\ell(P, Q; G) = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{a=1}^{2} \sum_{l=1}^{L_m} \mathbb{I}\{G_{ima} = l\} \log \left(\sum_{k=1}^{K} P_{mlk} Q_{ik} \right),$$

where the last term is the probability of the value l in the m for person i, summarized over her ancestry distribution.

The paper describes several interesting solutions for this estimation problem, we will focus on one that uses a well known approach we can refresh and use: Expectation-Maximization (EM).

Reminder: EM algorithm

Assume we have a parameter vector Θ, some observed data X and some unobserved data Y. We want to calculate the MLE of Θ given the observed data X, however the calculations are much easier if we had known Y as well, that is calculating $\ell(\Theta; X, Y)$ is easier than directly $\ell(\Theta; X)$.

Then the EM algorithm is an iterative algorithm. At stage r, we have a "current guess" $\Theta^{(r)}$, and we use it to calculate:

$$\text{E-Step: } \ell_r^E(\Theta) = \mathbb{E}_{\Theta^{(r)}}\left(\ell(\Theta; X, Y)|X\right),$$

that is, the expected value of the log-likelihood, integrated over the unknown Y, and using the current vector $\Theta^{(r)}$ in the distribution of $Y|X$. Note that Θ plays two roles here – one, where $\Theta^{(r)}$ is used to calculate conditional expectation, and two, where Θ is used symbolically in the likelihood. For example, if Y appears only linearly in the log-likelihood, then we simply plug $\mathbb{E}_{\Theta^{(r)}}Y|X$ into this to obtain ℓ_r^E.

The next step is the M-step, which finds the best value of Θ given the current integrated likelihood ℓ_r^E :

$$\text{M-Step: } \Theta_{r+1} = \arg\max_{\Theta} \ell_r^E(\Theta).$$

The theoretical guarantee we get is that $\ell(\Theta^{(r)}; X)$ is an increasing function of r, which converges to a local maximum (not necessarily the MLE, which is the global maximum). For convex problems, it will eventually converge to the MLE.

EM for our problem

Define as unobserved data: $Z = \{Z_{ima}\} \in \{1, \ldots, K\}$ the ethnic origin (e.g., Eu or Af) of the ath copy of the mth marker in the ith individual.

For $i > I_0$, $Z_{ima} = \{k : Q_{ik} = 1\}$ is in fact known, since $Q_{ik} \in \{0, 1\}$. For $i \leq I_0$, under our assumptions $Z_{ima} \sim multinom(Q_i)$.

The log-likelihood of the complete data:

$$\ell\left(P, Q; G, Z\right) = \sum_i \sum_m \sum_a \sum_l \sum_k \mathbb{I}\left\{G_{ima} = l, Z_{ima} = k\right\} \log\left(P_{mlk} Q_{ik}\right).$$

From this it is easy to see the form of the E-step:

$$\begin{aligned} \ell_r^E(P, Q) &= \mathbb{E}_{Q^{(r)}, P^{(r)}}\left(\ell(P, Q;\ X, Y)|X\right) = \\ &= \sum_i \sum_m \sum_a \sum_l \sum_k \mathbb{I}\left\{G_{ima} = l\right\} \mathbb{P}_{Q^{(r)}, P^{(r)}}\left(Z_{ima} = k|G\right) \log\left(P_{mlk} Q_{ik}\right). \end{aligned}$$

We have to calculate the probability / expectation, denote

$$E_{imak}^{(r)} = \mathbb{P}_{Q^{(r)}, P^{(r)}}\left(Z_{ima} = k|G\right),$$

and assume $G_{ima} = l$ is given, then:

$$\begin{aligned} E^{(r)}_{imak} &= \mathbb{P}\left(Z_{ima} = k | G_{ima} = l;\ P^{(r)}_{mlk}, Q^{(r)}_{ik}\right) \overset{(*)}{=} \frac{\mathbb{P}\left(Z_{ima} = k, G_{ima} = l\right)}{\mathbb{P}\left(G_{ima} = l\right)} = \\ &= \frac{P^{(r)}_{mlk} Q^{(r)}_{ik}}{\sum_{u=1}^{K} P^{(r)}_{mlu} Q^{(r)}_{iu}}, \end{aligned}$$

where the equality (*) is due to the independence assumptions we made above (each Z_{ima} is drawn independently than all other Z's, and G_{ima} depends only on Z_{ima}.)

Now we can write the explicit integrated likelihood to move to the M-step:

$$\begin{aligned} \ell^{E}_{r}(P, Q) &= \sum_{i,m,a,l,k} \mathbb{I}\left\{G_{ima} = l\right\} E^{(r)}_{imak} \log\left(P_{mlk} Q_{ik}\right) = \\ &= \sum_{m,k,l} \left[\log\left(P_{mlk}\right) \sum_{i,a} \mathbb{I}\left\{G_{ima} = l\right\} E^{(r)}_{imak} \right] + \sum_{i,k} \left[\log\left(Q_{ik}\right) \sum_{m,l,a} \mathbb{I}\left\{G_{ima} = l\right\} E^{(r)}_{imak} \right], \end{aligned}$$

and maximizing this to find $P^{(r+1)}, Q^{(r+1)}$ is easy:

$$\begin{aligned} P^{(r+1)}_{mlk} &= \frac{\sum_{i=1}^{I} \sum_{a=1}^{2} \mathbb{I}\{G_{ima} = l\} E^{(r)}_{imak}}{\sum_{i=1}^{I} \sum_{a=1}^{2} E^{(r)}_{imak}} \\ Q^{(r+1)}_{ik} &= \frac{\sum_{m=1}^{M} \sum_{a=1}^{2} E^{(r)}_{imak}}{2M} \ \text{ for } i \leq I_0 \\ Q^{(r+1)}_{ik} &= Q_{ik} \ \text{ known for } i > I_0. \end{aligned}$$