Kiley Huffman
QBIO 478
Spring 2025

QBIO 478 HW #3 Due Thursday, April 3

1. We discussed many possible explanations for the missing heritability problem in GWAS. In your own words, explain the reason scientists believe explains most of the missing heritability problem.

Scientists believe that the most likely explanation for the missing heritability problem is that there are many genetic variants with very small individual effects, which GWAS cannot detect, due to limited sample sizes and/or statistical power. This likely occurs because these variants are scattered across the genome and contribute to complex traits, but their individual effects are too small to reach genome-wide significance. Scientists also believe that rare variants, gene-gene interactions, structural variants, and environmental influences may also contribute to the missing heritability problem in GWAS. Though, the most commonly believed explanation is that GWAS methods struggle to detect the small-effects of variants that contribute to a cumulative effect.

2. What is a source of genetic variability that cannot be determined by a SNP chip? How could this source of genetic variability be determined for each individual in a GWAS study?

A source of genetic variability that cannot be determined by a SNP chip is structural variation. Structural variations include copy number variations (CNVs), insertions, deletions, inversions, and translocations. To determine structural variation for each individual in a GWAS study, we could use whole-genome sequencing (WGS) or array comparative genomic hybridization (aCGH). The WGS method sequences an individual's entire genome, allowing for the identification of all types of genetic variation, including SNPs, rare variants, and structural variations. The aCGH method, on the other hand, detects copy number variations by comparing the individual's DNA to a reference genome, identifying regions of DNA that are gained or lost. By using either or both of these methods with a SNP chip, we could get a more complete picture of genetic variability and its contribution to complex traits.

3. For problem #3, we are going to assume that there are <u>no sequencing errors.</u>

a. Suppose an individual is heterozygous for a particular SNP. If we do whole genome sequencing, and the coverage at this particular SNP is 12, calculate the probability that all 12 reads are the same allele (so that we would mistakenly think this individual is homozygous at this SNP).

Since there are no sequencing errors, the only source of variation comes from random sampling. Given that the individual is heterozygous for a particular SNP, each read has an equal probability of capturing either allele: this means each read has a 0.5 probability of being either allele.

To mistakenly call the individual homozygous, all 12 reads must capture the same allele. The probability of this happening is:

$$P(all\ 12\ reads\ are\ the\ same\ allele)\ =\ P(all\ are\ one\ allele)\ +\ P(all\ are\ the\ other\ allele)$$

Since each read is independently sampled, the probability that all 12 reads capture only one specific allele is:

$$P(all\ reads\ capture\ only\ one\ specific\ allele)\ =\ (0.5)^{12}$$

Since this can happen for either allele, we multiply by 2:

$$p\ =\ 2\ (0.5)^{12}\ =\ 0.00049$$

Thus, the probability that all 12 reads show the same allele, leading to a mistaken homozygous call, is approximately p = 0.00049.

b. Next suppose there are 3 million SNPs for which this individual is heterozygous. If we do whole genome sequencing, and the coverage at all of these 3 million SNPs is exactly 12, view the number of these SNPs such that all 12 reads are the same allele (so we would mistakenly think this individual is homozygous at these SNPs) as a random variable. What is the distribution and expected value of this random variable?

Let X represent the number of SNPs (out of 3 million heterozygous SNPs) for which all 12 sequencing reads capture the same allele, leading to a mistaken homozygous call.

Then, each SNP has two possible outcomes:

- p = 0.00049, all 12 reads show the same allele (mistaken homozygous call)
- p =1− p = 1-0.0049 = 0.995, at least one read shows the other allele (correct heterozygous call).

Since there are 3 million SNPs, and each SNP's outcome follows an independent Bernoulli trial, the total number of misclassified SNPs, X, follows a Binomial distribution:

$$X \sim Binomial(n\ =\ 3,000,000, p\ =\ 0.00049)$$

The expected value of a binomially distributed random variable is given by:

$E[X] = n\cdot p = 3,000,000\times0.00049 = 1470$

Thus, the distribution of this random variable X, follows a binomial distribution,

$X\sim Binomial(n = 3,000,000, p = 0.00049)$, and the expected value is 1470 SNPs.

c. Again suppose that there are 3 million SNPs for which an individual is heterozygous. What does the coverage at these 3 million SNPs have to be, such that the random variable described in part b has expected value less than one?

The expected number of misclassified SNPs is $E[X] = n\cdot p$ where:

- n = 3,000,000 (total heterozygous SNPs),
- P = $2\times(0.5)^C$ (probability that all CC reads are the same allele).

Need to find C so $E[X] = 3000000(2)(0.5)^C < 1$

Solve for C: $6000000(0.5)^C < 1 \rightarrow (0.5)^C < \frac{1}{6000000}$

$$(0.5)^C < \frac{1}{6000000} \rightarrow C \cdot ln(0.5) < ln(1/6000000)$$

$$C > \frac{ln(1/6000000)}{ln\,(0.5)} \rightarrow C > 22.52$$

The coverage must be an integer, so we round 22.52 up to 23. Thus, the coverage at these 3 million SNPs has to be C = 23 so that the random variable described in part b had an expected value less than 1.

4. Use the GWAS Catalog website https://www.ebi.ac.uk/gwas/ . Search for the gene PTPN11.

a. How many associations?: There are 140 associations.

b. Sort by odds ratio (OR). List the five traits with the highest odds ratios:

1. Hypothyroidism (OR = 1.5)
2. Low HDL-cholesterol levels (OR = 1.306)
3. Coffee consumption (OR = 1.293)
4. Atrial Fibrillation (OR = 1.24)
5. Bilateral cleft lip and palate (OR = 1.22)

c. Sort by beta (the slope for a quantitative trait). List the five traits with the greatest betas.

1. Hematological parameters/platelet count (Beta = 4.65 10^9/l increase)
2. Platelet Count (Beta = 3.031687 unit increase)
3. Response to alcohol consumption/flushing response (Beta = 1.7027 unit decrease)
4. LDL cholesterol levels (Beta = 1.64 unit increase)
5. HDL cholesterol levels (Beta = 1.37 unit decrease)
6. Low-carbohydrate diet (LCD) score (Beta = 1.074 decrease)

d. Scan through the associations. List five other traits not listed in parts b or c.

1. Systolic blood pressure
2. Diastolic blood pressure
3. Mean arterial pressure
4. Medication use (thyroid preparations)
5. Kynurenine levels
6. High light scatter reticulocyte count

e. What is the term when one gene affects two or more seemingly unrelated phenotypic traits?: The term is pleiotropy.

5. Suppose you want to do a GWAS to study Type 2 diabetes. For your cases, you talk to a number of medical doctors in the Los Angeles area and you enroll 10,000 people living in the Los Angeles area with Type 2 diabetes. For your cases, in order to save money, you decide to use a published reference panel of 10,000 people living in China that do not have Type 2 diabetes and have already been genotyped with SNP chips (at no cost to you). Does this seem like a good experimental design? Explain why or why not.

No, this does not seem like a good experimental design for a GWAS study. The biggest issue is that the cases and controls come from different genetic backgrounds. Since the 10,000 Type 2 diabetes cases are from Los Angeles (likely representing a diverse population), and the controls are entirely from China (predominantly Chinese or East Asian ancestry), we will not be able to tell if the genetic differences between these groups are related to Type 2 diabetes, or if they are due to the differences in the two group's ancestry. This can lead to false assumptions that genetic variants appear to be linked to the disease when they are actually just markers of population differences. For this GWAS study to have a good experimental design, the cases and controls should come from the same or similar populations. For example, a better approach would be to recruit both cases and controls from the same Los Angeles population, so that the ancestry and genetic background of the cases matches those of the controls.

6. Show the global alignment of the two sequences:

Sequence 1: GAGGACC ,   Sequence 2: AGTAC

The scoring matrix is:

	A	**C**	**G**	**T**
A	4	-2	-2	-2
C	-2	4	-2	-2
G	-2	-2	4	-2
T	-2	-2	-2	4

The space penalty is -3. Show the full score and traceback matrix:

FULL SCORE MATRIX

Create (m+1) × (n+1) matrix, where m = 7 (length of seq1) and n = 5 (length of seq2). The first row and column are initialized using the gap penalty (-3 per gap).

	-	**A**	**G**	**T**	**A**	**G**
-	0	-3	-6	-9	-12	-15
G	-3	-2	4	1	-2	-5
A	-6	4	1	-2	8	5
G	-9	1	8	5	2	5
G	-12	-2	5	12	9	6
A	-15	-5	2	9	16	13
C	-18	-8	-1	6	13	20

Row 1 (G) Calculations

- [G-A]: $\max(-3+(-3),\ 0+(-3),-3+(-2)) = -2$
- [G-G]: $\max(-6+4,\ -3+(-3),\ -2+(-3)) = 4$
- [G-T]: $\max(-9+(-2),\ -6+(-3),\ 4+(-3)\ ) = 1$
- [G-A]: $\max(-12+(-2),-9+(-3),1+(-3)) = -2$
- [G-C]: $\max(-15+(-2),-12+(-3),-2+(-3)) = -5$

Row 2 (A) Calculations

- **[A-A]:** max(−3+4, −6+(−3), −6+(−3)) = 4
- **[A-G]:** max(4+(−2), −3+(−3), 4+(−3)) = 1
- **[A-T]:** max(1+(−2), −6+(−3), 1+(−3)) = −2
- **[A-A]:** max(−2+4, −9+(−3) ,−2+(−3)) = 8
- **[A-C]:** max(−5+(−2), −12+(−3), 8+(−3)) = 5

Final Score (bottom-right corner): 20

TRACEBACK MATRIX

Start at the bottom-right corner (with score 20) and move backward using the highest scoring path.

- (6,5) → (5,4) [Match C-C]
- (5,4) → (4,3) [Match A-A]
- (4,3) → (3,3) [Match G-G]
- (3,3) → (2,2) [Match G-G]
- (2,2) → (1,1) [Match A-A]
- (1,1) → (0,0) [Match G-A] (substitution)

	-	**A**	**G**	**T**	**A**	**G**
-	↘	←	←	←	←	←
G	↑	↘	↘	←	←	←
A	↑	↘	↘	←	↘	←
G	↑	↑	↘	↘	←	←
G	↑	↑	↑	↘	↘	←
A	↑	↑	↑	↑	↘	↘
C	↑	↑	↑	↑	↑	↘

The Final alignment is: GAGGACC

-A-GTAC

And the full score is 20.

7. Find three programs used for multiple sequence alignment. When were they created?

1. CLUSTALW (1994): developed by Des Higgins and colleagues, introduce progressive alignment and sequence weighting
2. MUSCLE (2004): developed by Robert Edgar known for its speed and accuracy improvements over CLUSTALW
3. MAFFT (2002): developed by Kazutaka Katoh and colleagues, optimized for large datasets, offers high accuracy with iterative refinement strategies

8. In class, we looked at the multiple sequence alignments of vertebrate genomes. Describe a specific computational challenge with respect to one or more topics discussed in class for why such a multiple sequence alignment is challenging to create. What if only a pair of genomes is used? Give a strategy to overcome such a challenge, and relate that strategy to a topic discussed in class.

Specific Computational Challenge: One specific computational challenge in aligning vertebrate genomes is genome size and complexity. Vertebrate genomes are large (often billions of base pairs) and contain repetitive elements, insertions/deletions, and rearrangements that make it difficult to align sequences accurately across multiple species. Moreover, the presence of evolutionary divergence means that homologous regions may be highly mutated or even missing in certain species.

Only a Pair of Genomes: If only a pair of genomes is used, it becomes difficult to handle large-scale structural variations such as genome rearrangements (inversions, translocations), duplications and deletions that cause alignment gaps, and variable mutation rates between species.

Strategy to Overcome the Challenge: One strategy to handle this challenge is to use progressive alignment with guide trees. This method uses a phylogenetic tree to determine the order of alignment, aligns closely related sequences first before incorporating more divergent ones, and uses profile-based alignment, where aligned groups of sequences are treated as units to reduce complexity. This helps mitigate alignment errors caused by large evolutionary distances, by ensuring that conserved regions are properly aligned before introducing more distant sequences.

9. Give an example of four sequences for which progressive multiple sequence alignment creates an MSA that is clearly not optimal with respect to the sum of pairs scores.

To construct an example where progressive multiple sequence alignment (MSA) results in an MSA that is clearly not optimal with respect to the sum-of-pairs (SP) score, we need sequences where the guide tree leads to a misalignment. Here is a set of four example sequences:

	Example four sequences	MSA alignment	Best Alignment
Seq1:	GATT	GATT	GATT
Seq2:	GAT	GAT-	GAT-
Seq3:	GCT	GCT-	G-TTT
Seq4:	GTT	GTT-	GCTT

Calculate the sum of pairs score, assume a simple match = +1, mismatch = -1, and gap penalty = -2:

MSA Alignment

Column	**Pairs**	**Scores**
G, G, G, G	(G,G), (G,G), (G,G), (G,G), (G,G), (G,G)	1+1+1+1+1+1 = 6
A, A, C, T	(A,A), (A,C), (A,T), (A,C), (A,T), (C,T)	1 -1 -1 -1 -1 -1 = -4
T, T, T, -	(T,T), (T,T), (T,-), (T,T), (T,-), (T,-)	1+1-2+1-2-2 = -3
T, -, -, -	(T,-), (T,-), (T,-), (-,-), (-,-), (-,-)	-2 -2 -2 +0 +0 +0 = -6

Total SP score for MSA alignment = 6 - 4 - 3 - 6 = -7

Best Alignment

Column	**Pairs**	**Scores**
G, G, G, G	(G,G), (G,G), (G,G), (G,G), (G,G), (G,G)	1+1+1+1+1+1 = 6
A, A, -, C	(A,A), (A,-), (A,C), (A,-), (A,C), (-,C)	1 -2 -1 -2 -1 -2 = -7
T, T, T, T	(T,T), (T,T), (T,T), (T,T), (T,T), (T,T)	1+1+1+1+1+1 = 6
T, -, T, T	(T,-), (T,T), (T,T), (-,T), (-,T), (T,T)	-2 +1 +1 -2 -2 +1 = -3

Total SP score for better alignment = 6 - 7 + 6 - 3 = 2

Thus the alignment generated using a progressive MSA is not optimal because it introduces unnecessary gaps and does not maximize the sum-of-pairs score, as the SP = -7 for this alignment (the suboptimal alignment), whereas the best alignment had an SP = 2.