Assignment 03

PUBH 8878

**Requirements:**

* Show complete mathematical work for any derivations.
* Submit well-documented R code with clear comments and set seeds.
* Interpret results in a genetic/biological context where applicable.
* Submit the rendered PDF; do not submit the source .qmd.
* You may reuse and adapt your Lab 03 code, but your write‑up must be self‑contained.

## Problem 1: EM for ABO gene frequencies; derivation, inference, and sensitivity (25 pts)

### Part A (15 pts)

Consider phenotype counts from the ABO system under Hardy–Weinberg equilibrium (HWE): with total . Let genotype frequencies be and .

Derive the EM updates shown in lecture. Specifically, show that the E‑step allocations for and phenotypes are

and similarly for , and that the M‑step is

Hint (how to derive EM generally):

1. Choose latent variables so the complete data are simple. Here, treat latent genotype counts,

as missing, with only phenotype totals observed.

1. Write the complete‑data log‑likelihood

using .

1. E‑step: replace latent counts by their conditional expectations given the observed phenotypes under current parameters, e.g., for phenotype , and form .
2. M‑step: M-step: Maximize subject to . Hint: Consider rewriting the objective in terms of allele counts rather than genotype counts.

### Part B (10 pts)

Consider two biallelic SNPs with haplotypes under HWE and unphased genotypes . Note that yields , or yields , yields , etc.

Show that if every observed genotype is , then and the likelihood depends only on the cross‑sum (a ridge; parameters not identifiable).

## Problem 2: Two‑point linkage — LOD and support intervals (25 pts)

Suppose one heterozygous transmitting parent ( and ) is crossed to an mate, yielding child haplotype counts . Let and .

### Part A: LOD from counts (10 pts).

Show that for a given recombination fraction the two‑point LOD relative to independence () is

Derive the MLE and show it equals when .

### Part B: Unknown phase and LD‑informed LOD (15 pts)

A heterozygous transmitting parent has **unknown phase**: either **coupling** or **repulsion** . Let

Let and .

**(i) Mixture likelihood (5 pts).**  
Show that with unknown phase the observed‑data likelihood is a **mixture** of the two phase‑specific binomial likelihoods:

Hence the two‑point LOD relative to independence is

**(ii) Linking LD to (5 pts).**  
Let population haplotype frequencies be (sum to 1).  
Condition on the parent being the **double heterozygote** . Use Bayes’ rule to show

Define and note that indicates whether **coupling** () or **repulsion** () phase is a priori more likely.

**(iii) Quick numerical check (5 pts).**  
Take .  
Compute and . Then, using the example counts (so , , ), evaluate and **compare** versus at .  
Briefly explain (one sentence) how LD information () can increase or decrease the peak LOD when phase is unknown.

## Problem 3: Two‑SNP haplotype EM and LD measures (25 pts)

### Part A (10 pts)

For two biallelic SNPs with haplotypes at frequencies (summing to 1), enumerate the possible haplotype pairs consistent with each unphased genotype .

Show that only is ambiguous with two possible pairs: and .

### Part B (10 pts)

Simulate individuals from true haplotype frequencies and estimate via EM from a uniform start. Report and absolute errors .

Note that the E‑step weights for are proportional to and , and the M‑step update is .

Here is a sample R code snippet to get you started:

set.seed(8878)  
  
N <- 1000  
p\_true <- c(ab = 0.40, aB = 0.10, Ab = 0.25, AB = 0.25)  
  
# helper: draw N unordered haplotype pairs, then make unphased genotypes (g1,g2)  
draw\_genotypes <- function(N, p) {  
 # code haplotypes to allele counts (B allele) at SNP1, SNP2  
 H <- rbind(ab = c(0, 0), aB = c(0, 1), Ab = c(1, 0), AB = c(1, 1))  
 hap1 <- sample(rownames(H), size = N, replace = TRUE, prob = p)  
 hap2 <- sample(rownames(H), size = N, replace = TRUE, prob = p)  
 G <- H[hap1, ] + H[hap2, ] # N x 2 matrix with entries in {0,1,2}  
 as.data.frame(G) |>  
 setNames(c("g1", "g2"))  
}  
  
# simulate data  
dat <- draw\_genotypes(N, p\_true)

### Part C (5 pts)

Compute with , , . Report and . Comment on how LD would affect single‑marker association at either SNP.

## Problem 4: Single‑marker association with QC and LD attenuation (25 pts)

Simulate unrelated individuals. Let a causal biallelic SNP have MAF and generate a quantitative trait with additive effect size (per allele) and noise . Let a tag SNP be in LD with such that and both are in HWE. Let have MAF .

### Part A (15 pts)

Generate by first simulating haplotypes for with a chosen LD structure that yields , then form genotypes and . Fit simple linear models and and report and , alongside their confidence intervals.

### Part B (10 pts)

Introduce a basic QC step: test HWE in the controls of a case–control subsample formed by thresholding at its 80th percentile to define cases (cases are the top 20% of ). Compute an exact or HWE p‑value in controls for ; state whether you would flag using a threshold of and why QC is typically done in controls only.