Assignment 2: Ancestry Based on Genotype Likelihoods

Abdullah Faqih Al Mubarok - vpx267

October 6, 2023

1 Estimate allele ancestral frequencies and ancestral proportions

1. What is the average estimated European ancestry in the African Americans?

   Answer:

   To know what population group represented by each column, we look at the table summary for the maximum ancestry proportion summary table

	pop_id				
max_K	ASW	CEU	CHB	MXL	YRI
K1	20	0	0	0	20
K2	0	20	0	20	0
K3	0	0	20	0	0

   Table 1: Table Summary of Maximum Ancestry Proportion

   From the results above, we can conclude that the K1 represents African, K2 represents European, and K3 represents (Asian + Native American). Then, we can calculate the average estimated European ancestry (K2) in the African Americans (ASW).

   Based on the calculation in R (attached), the average (mean) of the estimated European ancestry in the African Americans is 0.1965986.

2. For each of the five populations calculate the average estimated ancestry proportion belonging to the 3 ancestral populations. Summaries the results in a table.

   Answer:

pop_id	afr_anc_prop	eu_anc_prop	asia_america_anc_prop
ASW	0.792	0.197	0.012
CEU	0.000	0.999	0.001
CHB	0.000	0.000	1.000
MXL	0.017	0.752	0.231
YRI	1.000	0.000	0.000

Table 2: Average of Estimated Ancestry for Each Population

2 Genotype calling based on different priors

From the genotype likelihoods you should write code to call genotypes for all 50000 sites for individual NA19700 using 4 different approaches. For each approach you should first calculate the posterior probability of the 3 possible genotypes and then call the genotype with the highest posterior probability.

1. Assume a uniform prior (assume all 3 genotypes are equally likely). Plot a histogram of the 50000 posterior probabilities for the called genotypes.

   Answer:

   The posterior probability for each genotype $g \in \{0, 1, 2\}$ on site j would be:

$$\begin{aligned} P(G_j = g|X_j) &= \frac{P(X_j|G_j = g)P(G_j = g)}{P(X_j)} \\ &\text{by law of total probability,} \\ &= \frac{P(X_j|G_j = g)P(G_j = g)}{\sum_{g' \in \{0,1,2\}} p(X_j|G_j = g')p(G_j = g')} \end{aligned} \tag{1}$$

   where

$$\begin{aligned} P(X_j|G_j = g) &= \text{calculated genotype likelihood data for site } j \text{ given genotype } g \\ P(G_j = g) &= \tfrac{1}{3} \text{ for all genotypes of all sites } j \text{ (uniform prior)} \end{aligned}$$

   Next, we call the genotype for each site j by:

$$\underset{g \in \{0,1,2\}}{\arg\max}\, P(G_j = g|X_j) \tag{2}$$

   If there is any same values, we chose the first occurrence. The figure 1 below is the histogram of 50000 posterior probabilities for the called genotypes:

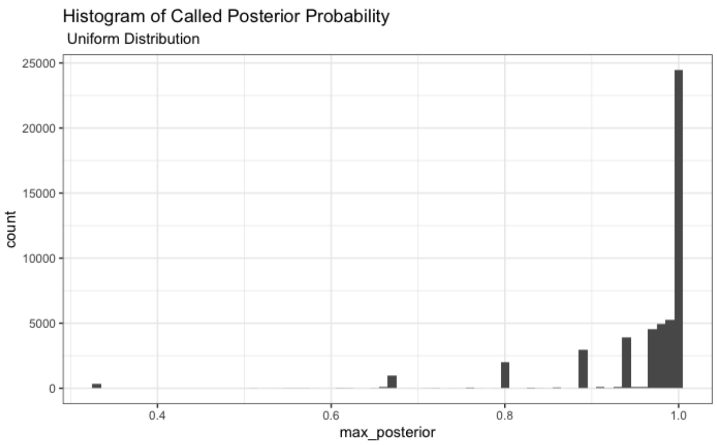

Figure 1: Posterior Probabilities for the Called Genotypes Under a Uniform Prior

2. Use estimated ancestral allele frequency of the African ancestry as a prior assuming Hardy- Weinberg equilibrium. Write the formula that you are using to obtain the posterior probability 3 and plot a histogram of the posterior probabilities for the called genotypes.

   Answer:

   The posterior probability for each genotype $g \in \{0, 1, 2\}$ on site j would be:

$$\begin{aligned} P(G_j = g|X_j, f_j) &= \frac{P(X_j|G_j = g, f_j)P(G_j = g|f_j)}{P(X_j|f_j)} \\ &= \frac{P(X_j|G_j = g)P(G_j = g|f_j)}{P(X_j|f_j)} \\ &\text{by law of total probability,} \\ &= \frac{P(X_j|G_j = g)P(G_j = g|f_j)}{\sum_{g' \in \{0,1,2\}} p(X_j|G_j = g', f_j)p(G_j = g'|f_j)} \\ &= \frac{P(X_j|G_j = g)P(G_j = g|f)}{\sum_{g' \in \{0,1,2\}} p(X_j|G_j = g')p(G_j = g'|f_j)} \end{aligned} \tag{3}$$

   where

$$\begin{aligned} f_j &= \text{estimated African allele frequency of NA19700 at site } j \\ p(G_j = g|f_j) &= \begin{cases} f_j^2 & g = 2 \\ 2 * f_j * (1 - f_j) & g = 1 \\ (1 - f_j)^2 & g = 0 \end{cases} \\ P(X_j|G_j = g) &= \text{calculated genotype likelihood data for site } j \text{ given genotype } g \end{aligned}$$

   Next, we call the genotype for each site j by:

$$\arg\max_{g \in \{0,1,2\}} P(G_j = g|X_j, f_j) \tag{4}$$

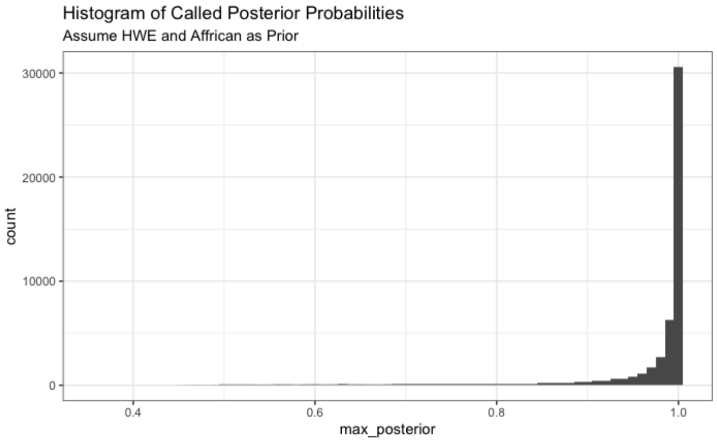

Figure 2: Posterior Probabilities for the Called Genotypes Under HWE Using African Ancestry

3. Try to make a prior that is better for genotypes calling by using a combinations of the 3 ancestral frequencies. Write the formulation for how you make this prior and plot a histogram of the posterior probabilities for the called genotypes.

   Answer: To get the better calling, we have to combine the information of the frequencies of all $k \in \{1,2,3\}$ ancestrals. Therefore, we introduce h_j which is the allele frequency of site j of individual NA19700:

$$\begin{aligned} h_j &= P(allele|F_k^j, Q_k^{NA19700}) \\ &= \sum_{k\in\{1,2,3\}} q_k^{NA19700} f_k^j \end{aligned} \tag{5}$$

   where

$$\begin{aligned} Q_k^{NA19700} &= \{q_1^{NA19700}, q_2^{NA19700}, q_3^{NA19700}\} \text{ NA19700 admixture proportion} \\ F_k^j &= \{f_1^j, f_2^j, f_3^j\} \text{ allele frequency of each ancestral } k \text{ and site } j \end{aligned}$$

   Therefore, the posterior probability for each genotype $g \in \{0,1,2\}$ on site j would be:

$$\begin{aligned} P(G_j = g|X_j, h_j) &= \frac{P(X_j|G_j = g, h_j)P(G_j = g|h_j)}{P(X_j|h_j)} \\ &= \frac{P(X_j|G_j = g)P(G_j = g|h_j)}{P(X_j|h_j)} \\ &\text{by law of total probability,} \\ &= \frac{P(X_j|G_j = g)P(G_j = g|h_j)}{\sum_{g'\in\{0,1,2\}} p(X_j|G_j = g', h_j)p(G_j = g'|h_j)} \\ &= \frac{P(X_j|G_j = g)P(G_j = g|h_j)}{\sum_{g'\in\{0,1,2\}} p(X_j|G_j = g')p(G_j = g'|h_j)} \end{aligned} \tag{6}$$

where

$$\begin{aligned} h_j &= \text{allele frequency of site j of NA19700 as in equation (5)} \\ p(G_j = g|h_j) &= \begin{cases} h_j^2 & g = 2 \\ 2 * h_j * (1 - h_j) & g = 1 \\ (1 - h_j)^2 & g = 0 \end{cases} \\ P(X_j|G_j = g) &= \text{calculated genotype likelihood data for site } j \text{ given genotype } g \end{aligned}$$

Next, we call the genotype for each site j by:

$$\underset{g \in \{0,1,2\}}{\arg\max}\, P(G_j = g|X_j, h_j) \tag{7}$$

The figure 1 below is the histogram of the called posterior probabilities:

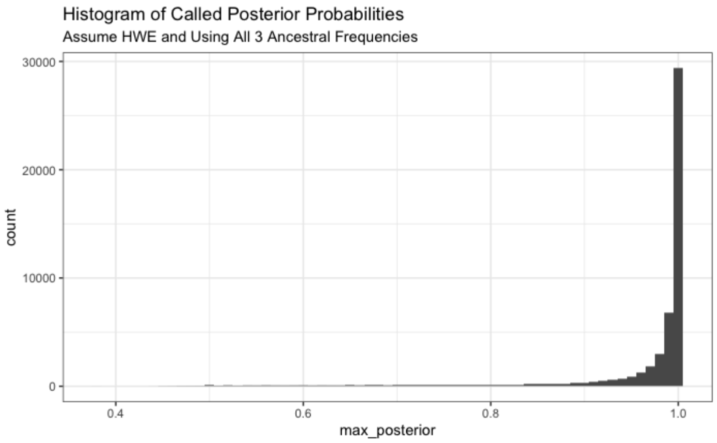

Figure 3: Posterior Probabilities for the Called Genotypes Under HWE Using All Ancestry

4. Use haplotype imputation to call the genotypes. You should run Beagle4 on all individuals to obtain posterior probabilities for of each possible genotype. Make a histogram of the posterior probabilities for the called genotypes for the NA19700 individual

   Answer:

   First, we need to to obtain the posterior probabilities using Beagle4

   ```
   #beagle in the same folder
   BEAGLE=./beagle.jar

   java -jar $BEAGLE like=input.gz out=imputation
   ```

   After that, we find the maximum posterior probabilities for each site j. Figure 4 shows the called posterior probabilities.

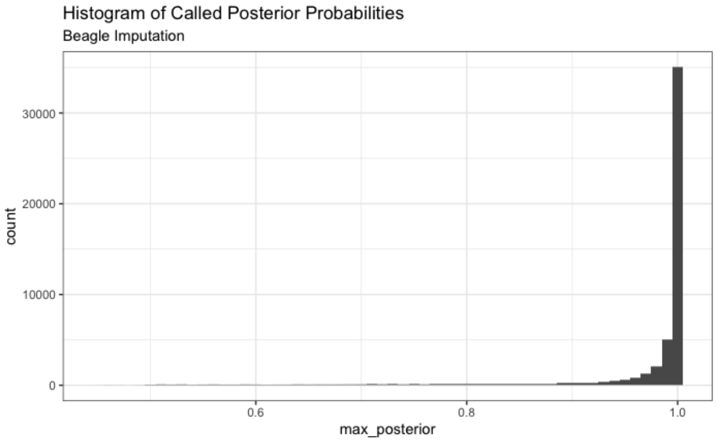

Figure 4: Posterior Probabilities for the Called Genotypes with Beagle Imputation

3 Evaluation

1. Make a plot with the accuracy of the four genotyping approaches.

   Answer:

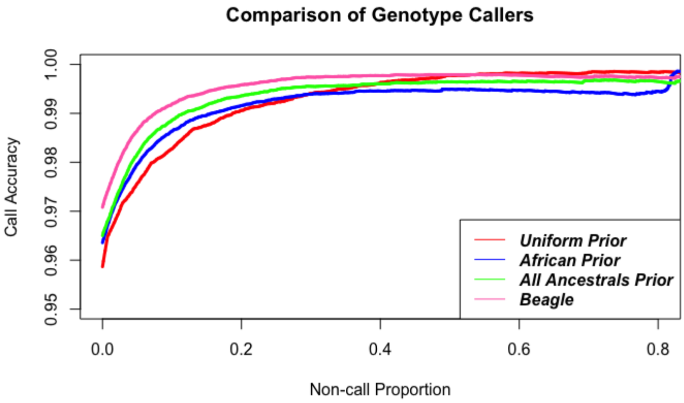

Figure 5: Comparison of Different Callers