David Nguyen

CS CM 121

Winter 2023

Project 1 Report

With my python script, I take in DNA-seq data from a .bam file and putative SNPs metadata from a .tsv and output two files. The first output file contains the posterior probability of the most probable genotype for each chromosome and is called output.tsv. The second file contains the posterior probability of all possible genotypes for each chromosome in a file called all_poss_genotypes_output.tsv. The first output file is more useful as it contains data about the most probably genotype and less redundant data.

In my script, we first check that the input files are correct. There are some default files for the script to fall back and use if the input files are incorrect. After getting the .bam and .tsv files, we index and parse the data from each file. Then we go through each chromosome in the metadata file and process it. From the processing, we are able to find the posterior probabilities of genotype from the chromosome. We then put all our data in a dataframe that is then output as a .tsv file.

With the processing of the data for each chromosome, first have to obtain the correct sequences to observe from the .bam file. Upon receiving the correct sequences, we need to clean the sequence to only contain the relevant alleles (the major and minor allele which are given to us in the metadata). We then find the posterior probability using math that we derive from our model. We use a biallelic model and have the derivation below.

Below, we have the model used as well as the names for variables and assumptions used.

Using biallelic model

Posterior probability derivations for biallelic model

Variables used

- A : major allele
- B : minor allele
- δ_A : amount of A (proportion of A)
- ϵ : error
- N : number of observations
- N_A : number of A observations
- maf : minor allele frequency

Assumptions made

- $\delta_A = 1/2$ or 0.5
- prior probabilities
  1) $P(AA) = (1 - maf)^2$
  2) $P(BB) = maf^2$
  3) $P(AB) = 1 - P(AA) - P(BB)$
- Error → $\epsilon = 0.05$

Posterior probability of genotype AB

$$P(AB \mid O_1, \dots, O_n) = \frac{P(O_1, \dots, O_n \mid AB)\, P(AB)}{P(O_1, \dots, O_n)}$$

↓

Note: $P(O_1, \dots, O_n) = P(O_1, \dots, O_n \mid AA) P(AA) + P(O_1, \dots, O_n \mid AB) P(AB) + P(O_1, \dots, O_n \mid BB) P(BB)$

↓

$$P(AB \mid O_1, \dots, O_n) = \frac{P(O_1, \dots, O_n \mid AB)\, P(AB)}{P(O_1, \dots, O_n \mid AA) P(AA) + P(O_1, \dots, O_n \mid AB) P(AB) + P(O_1, \dots, O_n \mid BB) P(BB)}$$

Note: Break into smaller sections to keep track of

$P(O_1, \dots, O_n \mid AA) P(AA) + P(O_1, \dots, O_n \mid AB) P(AB) + P(O_1, \dots, O_n \mid BB) P(BB)$

1) $$P(O_1, \dots, O_n \mid AA) P(AA) = \prod_{i:A} P(O_i = A \mid AA) \prod_{i:B} P(O_i = B \mid AA)\ P(AA)$$

$$= \prod_{i:A} P(\varepsilon_i = 0) \prod_{i:B} P(\varepsilon_i = 1)\ P(AA)$$

$$= (1 - \epsilon)^{N_A} (\epsilon)^{N - N_A}\ P(AA)$$

2) $$P(O_1, \dots, O_n \mid BB) P(BB) = \prod_{i:A} P(O_i = A \mid BB) \prod_{i:B} P(O_i = B \mid BB)\ P(BB)$$

$$= \prod_{i:A} P(\varepsilon_i = 1) \prod_{i:B} P(\varepsilon_i = 0)\ P(BB)$$

$$= \epsilon^{N_A} (1 - \epsilon)^{N - N_A}\ P(BB)$$

3) $$P(O_1, \dots, O_n \mid AB) P(AB) = \left(\prod_{i:A} P(\varepsilon_i = 0 \mid S = A) P(S = A) + P(\varepsilon_i = 1 \mid S = B) P(S = B) \right)$$

$$\left(\prod_{i:B} P(\varepsilon_i = 1 \mid S = A) P(S = A) + P(\varepsilon_i = 0 \mid S = B) P(S = B) \right)\ P(AB)$$

Note:

$$\left(\prod_{i:A} P(G_i=0 \mid S=A)\,P(S=A) + P(G_i=1 \mid S=B)\,P(S=B)\right)\left(\prod_{i:B} P(G_i=1 \mid S=A)\,P(S=A) + P(G_i=0 \mid S=B)\,P(S=B)\right) P(AB)$$

$P(G_i=0 \mid S=A) \to (1-\epsilon)$, $P(S=A) \to s_A$, $P(G_i=1 \mid S=B) \to \epsilon$, $P(S=B) \to 1-s_A$

$P(G_i=1 \mid S=A) \to \epsilon$, $P(S=A) \to s_A$, $P(G_i=0 \mid S=B) \to 1-\epsilon$, $P(S=B) \to 1-s_A$

raise expression to N_A power

raise expression to $N-N_A$ power

$$\Downarrow$$

$$\left((1-\epsilon)s_A + \epsilon(1-s_A)\right)^{N_A}\left(\epsilon(s_A) + (1-\epsilon)(1-s_A)\right)^{N-N_A} P(AB)$$

$$P(o_1,\ldots,o_n \mid AB)\,P(AB) = \left((1-\epsilon)s_A + \epsilon(1-s_A)\right)^{N_A}\left(\epsilon(s_A) + (1-\epsilon)(1-s_A)\right)^{N-N_A} P(AB)$$

$$P(AB \mid o_1,\ldots,o_n) = \frac{P(o_1,\ldots,o_n \mid AB)\,P(AB)}{P(o_1,\ldots,o_n \mid AA)\,P(AA) + P(o_1,\ldots,o_n \mid AB)\,P(AB) + P(o_1,\ldots,o_n \mid BB)\,P(BB)}$$

$$= \frac{\left((1-\epsilon)s_A + \epsilon(1-s_A)\right)^{N_A}\left(\epsilon(s_A) + (1-\epsilon)(1-s_A)\right)^{N-N_A} P(AB)}{(1-\epsilon)^{N_A}(\epsilon)^{N-N_A} P(AA) + \left((1-\epsilon)s_A + \epsilon(1-s_A)\right)^{N_A}\left(\epsilon(s_A) + (1-\epsilon)(1-s_A)\right)^{N-N_A} P(AB) + \epsilon^{N_A}(1-\epsilon)^{N-N_A} P(BB)}$$

$$\Downarrow$$

posterior probability for AB

↳ derivations for AA and AB will be similar

Posterior Probability AB

$$\frac{\left((1-\epsilon)s_A + \epsilon(1-s_A)\right)^{N_A}\left(\epsilon(s_A) + (1-\epsilon)(1-s_A)\right)^{N-N_A} P(AB)}{(1-\epsilon)^{N_A}(\epsilon)^{N-N_A} P(AA) + \left((1-\epsilon)s_A + \epsilon(1-s_A)\right)^{N_A}\left(\epsilon(s_A) + (1-\epsilon)(1-s_A)\right)^{N-N_A} P(AB) + \epsilon^{N_A}(1-\epsilon)^{N-N_A} P(BB)}$$

Posterior Probability AA

$$\frac{(1-\epsilon)^{N_A}(\epsilon)^{N-N_A} P(AA)}{(1-\epsilon)^{N_A}(\epsilon)^{N-N_A} P(AA) + \left((1-\epsilon)s_A + \epsilon(1-s_A)\right)^{N_A}\left(\epsilon(s_A) + (1-\epsilon)(1-s_A)\right)^{N-N_A} P(AB) + \epsilon^{N_A}(1-\epsilon)^{N-N_A} P(BB)}$$

Posterior Probability BB

$$\frac{\epsilon^{N_A}(1-\epsilon)^{N-N_A} P(BB)}{(1-\epsilon)^{N_A}(\epsilon)^{N-N_A} P(AA) + \left((1-\epsilon)s_A + \epsilon(1-s_A)\right)^{N_A}\left(\epsilon(s_A) + (1-\epsilon)(1-s_A)\right)^{N-N_A} P(AB) + \epsilon^{N_A}(1-\epsilon)^{N-N_A} P(BB)}$$