CS CM124 Final Project Report

Jong Hoon Kim 405189216, Hae Sung Jeong 005146648

The genotype of a person contains the values {0, 1, 2}, indicating the number of copies of the reference allele at each SNP. Since the given file is imperfect, it will have some missing (masked) values for some position of the genome. We have to find out the correct values for the missing values at first in order to impute the haplotypes.

Since the missing values only contain either 0 or 2, we could easily find the missing values with an obvious solution by comparing with the other individuals. If all of the other individuals have all 0’s or all 2’s for the particular SNP at the same position, then we fill the missing value with the same 0 or 2. Once we complete unmasking the “obvious” unknowns, we initiate the estimation for the rest masked SNPs. Our algorithm to estimate and guess each known follow these steps:

1. For each masked SNP, calculate the similarities with other individuals by comparing the neighboring SNPs up to *N* range (*N* to the left and *N* to the right from the original position).
2. Each matching 0, 1, or 2 will add 1 to the similarity score.
3. Except 1, any matching with masked SNPs will add 0.5 to the score.
4. Once the calculation is done for every individual, we choose the one that has highest similarity score and use the SNP from that individual to unmask our current unknown SNP.

We used this algorithm because each SNP has some level of relations with neighboring SNPs. And we perform many testing and benchmark to find the hyperparameter (*N*) that will give us the highest accuracy. Hence, the range of 3 (length of 7 SNPs) has the highest accuracy with having more than 99.9% for both given example data.

In order to find the correct haplotypes, we applied similar method that we used to find the correct genotype and blend in Clark’s algorithm with following steps:

1. Divide the entire genotypes by length of *M*
2. For each index of genotype, perform Clark’s algorithm with other individuals
   1. Count the SNPs of ‘1’ appear in each genotype
   2. If the genotype contains less than two ‘1’s, create corresponding haplotypes and add them to *knowns* list
   3. Else add the genotype to *unknown* list
   4. Once finish sorting through all individuals, start iterating each genotype in *unknown*
   5. If the current genotype is possible to be assembled by one of *known* haplotype, add a matching haplotype to *known* and remove the genotype from *unknown*
   6. If there was no change being made with a full iteration of *unknown,* exit out the iteration loop and guess the remaining *unknown* genotypes with random numbers.
   7. Repeat (a) to (f) for rest of index

Due to big size of original genotype (40,000), it is ideal to cut them into smaller pieces when generating estimated haplotypes. As we did for unmasking genotype, we performed multiple testing to validate the hyperparameter and found the cutting size of 20 to achieve highest accuracy.