CS M124 Final Project Report

Terry Ye, Kensei Kishimoto

1 Impute

1.1 Method Introduction

The first part is to unmask the masked values in the data file. We first divide the SNP into an array of SNPs with length of 5. Then for every 5-allele-long genotype we do an EM-like-approach but only uses the expectation part.
For each individual, we work out all the possible genotype for it by either 0 or 2 at every masked position. Then we count every possible genotype's probability in one person as $\frac{1}{numberofpossiblegenotypesforoneperson}$ assuming all genotypes have equal probability to occur to start with.. Then we store that probability into dictionary and repeat this process for every individual. Then we get a dictionary containing the likelihood of one genotype's occurance. So then for every individual we choose the genotype with the most likelihood out of all possible ones for it based on the first round's result.

1.2 Result

For the first example data, the impute reaches an accuracy of 94.22% for all the masked positions prediction.
For the first example data, the impute reaches an accuracy of 94.05% for all the masked positions prediction.

2 Haplotype Phase

2.1 Method Introduction

The second part is to use the imputed unmasked genotype data to work out the haplotypes for it. We also divide the long genotype into shorter ones with length of 8 so there are fewer possible combinations to consider for each part. Then we use a similar-EM-approach like first part. For every individual, we work out the possible combination of two haplotypes for its genotype. Then we write down the possibility of one's occurance for the individual assuming all combinations have equal possibility. Then repeat the process for every individual and we choose a combination suitable with maximum likelihood for one's genotype.

2.2 Result

The method achieves an accuray of 66.7% which already includes the possible errors when we predict the masked position's values.