**Machine Learning in Genetics Final Project Report**

Xuesen Cui UID:

Tong Xia UID:304946753

Methods:

* First, we make the masked genotypes to unmasked genotypes by choosing the highest frequencies, either 0 or 2.
* Second, after getting unmasked genotypes, we cut SNPS by 20 as a group, then in a group, we have 50\*20 matrix.
* Third, in each matrix, we generate haplotypes according to their genotypes.
* Forth, after getting all possible haplotypes of all genotypes, we use Expectation Maximalization (EM) to get the most likely haplotypes of each genotypes. For the detailed of EM: We assume all the haplotypes have same probability at first, which we use to generate E step. After E step, we get maximized haplotype probability by M step. Then we repeat this EM step for 10 times, to get the most likely haplotypes of the genotypes.

Results: Our solution has an accuracy of 86.2%