**Haplotype Admixture**

**Background**

Non-negative Matrix Factorization (NMF) is a group of algorithms where a matrix V is factorized into two matrix W and H, all tree matrices have no negative elements. NMF is an unsupervised learning technique that has finds applications in several fields, such as computer vision, document clustering, astronomy and bioinformatics. Most of the methods that used to investigate the haplotype admixture models

Population substructure is often estimated using admixture models, where each population is broken down into a mixture of original populations. Multiple methods have been developed to detect population admixture (e.g. ADMIXTURE and STRUCTURE). These models are not adapted to a single highly polymorphic region in contrast to large regions of limited polymorphism. We here present a method to estimate population admixture, in the presence of background LD, based on Non-Negative-Matrix Factorization (NMF).

**Conclusions/Results**

We present an algorithm that dissects population genetic admixture, based on HLA haplotype frequencies, into OP components in the presence of background LD and high polymorphism. We demonstrate that HLA is polymorphic enough to allow for a clear delineation of population composition using a single genomic region. The admixture problem, shown equivalent to a Non-Negative Matrix Factorization analysis, is more computationally efficient than traditional admixture models and allows for the admixture of a large number of populations with an extensive number of haplotypes. To our knowledge, this is the first algorithm to dissect population admixture with specific focus on HLA and a validation framework using a dataset with the presented magnitude. We developed and applied our method to the haplotype frequencies of 56 populations from different adult volunteer stem-cell registries representing over 3.5 million donors. We showed that the resulting admixture is consistent with the known ethnic composition, recent history and SNP based admixture

**The File In**

In this repo contain all you need to performing the NMF analysis using HLA haplotype frequencies from 56 populations.  
The code in R, you need NMF and ggplot2 packages .  
Clone this repository or download zip version, you need open the NMF\_to\_PurePop.r

1. Freq\_norm.csv- matrix of with all haplotype frequencies in each population, the frequency are normalized
2. Haplo\_S.csv- a scrambled haplotypes name
3. Ppopulation\_name.csv- All population names-, not sorted
4. Population\_name\_and\_index\_Sort.csv- Population named that was sorted, and the index of the original indexes
5. Population\_size.csv- this file contain the size of each population.