DIVERSITY OF MHC GENES IN THE 1000 GENOMES DATASET

Tünde Vágó, Péter Tóth, Tim Hague, Szilveszter Juhos
Omixon LTD, Budapest, Hungary
Corresponding author: szilveszter.juhos@omixon.com

Introduction

The 1000 Genomes Project (1KG) was the first large-scale sequencing project to obtain comparable whole genome sequences from diverse human populations. Previously we validated our HLA typing method using these datasets for Class-I and Class-II HLA genes for 3 fields accuracy. Furthermore, using family trio data, it was concluded that our genotyping method can be used for genes other than HLA-A,B,C and HLA-DQB1,-DRB1. In this study we are presenting genotyping results obtained from 1KG dataset with different EUR (European), EAS (East Asian) and AFR (African) ethnic background. Besides the previous five loci, the less diverse MICA and MICB genes were also considered in the calculations.

Methods

Whole-exome FASTQ files from the 1KG data repository were pre-filtered for HLA typing; the filter discarded reads that were shorter than 75 base pairs, or if the read was not mappable to the IMGT/HLA reference sequences. Generally, a few dozen thousand reads were retained for HLA typing. The algorithm produces 3 fields precision results for all the genes in the IMGT/HLA database.

Three super populations were investigated: EUR (European), EAS (East Asian) and AFR (African). Since the Simpson's D diversity index is relatively insensitive to data size, it was calculated for each population like:

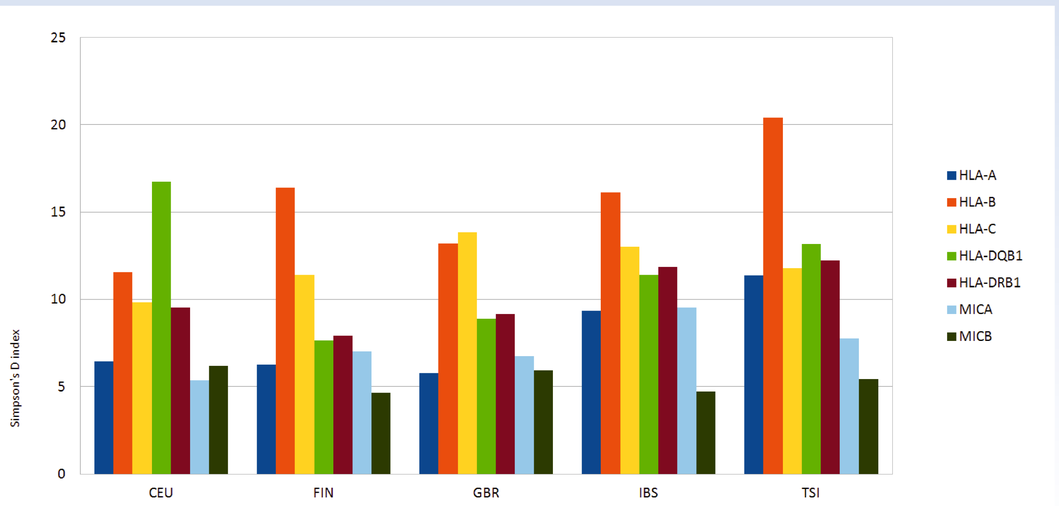
$$D = \frac{1}{\sum_{i=1}^{s} p_i^2}$$

where p is the allele frequency and s is the number of alleles.

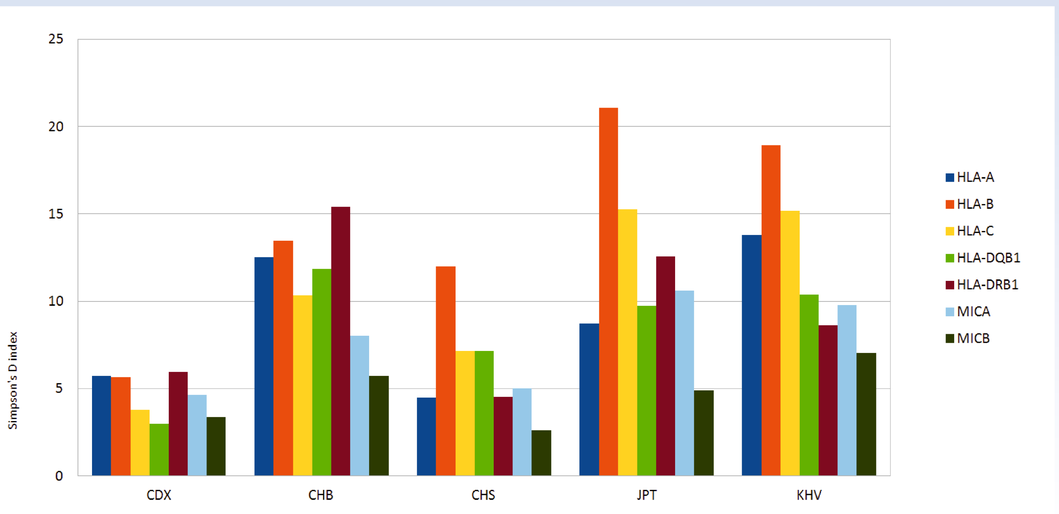
Simpson's D Diversity for the Three Super Population

Although the number of the possible alleles for the evaluated loci differs greatly, the diversity of the loci is comparable. From the 1000 Genomes data the most diverse locus is HLA-B for most of the groups in the EUR super population. Similarly, HLA-B is the most diverse for three out of the five EAS populations. On the other hand, for the most of the AFR groups HLA-A is the most diverse.

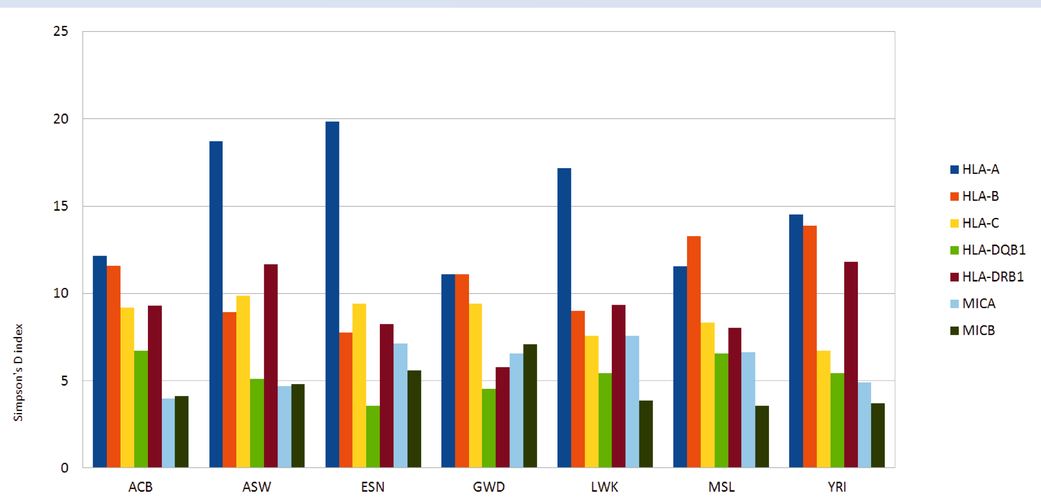
Diversity of the EUR Super Population



Diversity of the EAS Super Population



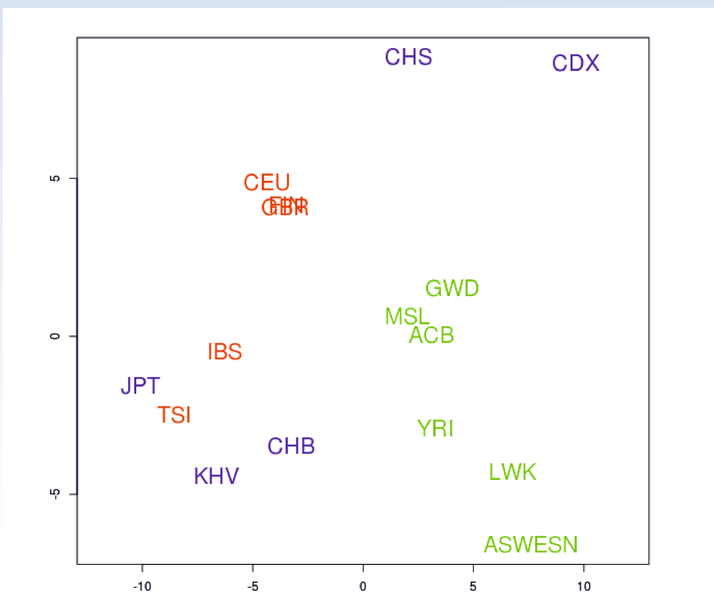
Diversity of the AFR Super Population



Metric Multidimension Scaling (MDS) of Diversity Indices

The three super population is clearly separated on the MDS graph. The calculation is based on the diversity values, the Simpson's D index of HLA-A, -B, -C, -DQB1, DRB1 and MICA, MICB. The diversity values were was considered as the elements of the vectors pointing to the samples in the parameter space, and the distances of these points were fed into the R cmdscale() routine. Red letters are data from the EUR, green letters are the AFR, blue are the EAS super population data points respectively.

Metric MDS



Population	Population code	Number of samples
Han Chinese in Beijing, China	CHB	30
Japanese in Tokyo, Japan	JPT	38
Southern Han Chinese	CHS	6
Chinese Dai in Xishuangbanna, China	CDX	21
Kinh in Ho Chi Minh City, Vietnam	KHV	26
Utah Residents (CEPH) with Northern and Western European Ancestry	CEU	58
Toscani in Italy	TSI	58
Finnish in Finland	FIN	30
British in England and Scotland	GBR	34
Iberian Population in Spain	IBS	58
Yoruba in Ibadan, Nigeria	YRI	39
Luhya in Webuye, Kenya	LWK	35
Gambian in Western Divisions in the Gambia	GWD	19
Mende in Sierra Leone	MSL	33
Esan in Nigeria	ESN	49
Americans of African Ancestry in SW USA	ASW	26
African Caribbean in Barbados	ACB	27

Conclusion

In spite of the limited sample number, the multidimensional scaling results are indicating that the diversity indices derived are estimating the real allele diversity correctly.