DNA sequence patterns in human major histocompatibility complex region in southern Chinese

YQ Song *, PS Sham, SP Yip, YH Fan, SY Bao

KEY MESSAGES

1. We developed a programme, quantLD, for genome-wide comparisons of linkage disequilibrium difference between two populations and applied the methods to the major histocompatibility complex region in 1000 Genomes Project phase 3 data. Results suggested that linkage disequilibrium difference exists across populations.
2. We also developed a programme PyHLA for HLA alleles association analysis and applied to a set of 246 APOE ε4 allele negative Alzheimer disease cases and 172 normal controls. HLA-DRB1*07:01 and HLA-C*03 were inversely associated with Alzheimer disease, whereas HLA-DQB1*03:01 and DRB1*09:21 were associated with Alzheimer disease.

Hong Kong Med J 2019;25(Suppl 7):S13-6
HMRF project number: 01121726

[1] YQ Song, [2] PS Sham, [3] SP Yip, [1] YH Fan, [1] SY Bao

[1] *School of Biomedical Sciences, The University of Hong Kong*
[2] *Department of Psychiatry, The University of Hong Kong*
[3] *Department of Health Technology & Informatics, The Hong Kong Polytechnic University*

* Principal applicant and corresponding author: songy@hku.hk

Introduction

The variation LD (varLD) method was developed to assess the extent of differences in linkage disequilibrium (LD) patterns between populations. The varLD method can identify candidate regions with different LD patterns between populations. However, quantification of the difference and how to use the difference to direct the replication and fine mapping of association signals from genome-wide association studies remain unclear.

The major histocompatibility complex (MHC) is a region on chromosome 6 that encodes MHC molecules. In humans, MHC is known as human leukocyte antigen (HLA). The HLA region (3.9 Mb, chromosome 6, 29587512-33516520, NCBI, Build 36.3; 28477797-33448354, GRCh37) contains 224 gene loci, 128 of which are predicted to be expressed.[1] In HLA, there are three major (HLA-A, HLA-B, HLA-C) and three minor (HLA-E, HLA-F, HLA-G) MHC class I genes, and three major (HLA-DP, HLA-DQ, HLA-DR) and two minor (HLA-DM, HLA-DO) MHC class II genes.

Alzheimer disease (AD) is a devastating neurodegenerative disease primarily affecting the elderly people. HLA-DRB1 and HLA-DRB5 alleles were reported as susceptibility loci of AD by genome-wide association study and meta-analysis.

The same disease can be associated with different MHC loci in different populations. An integrated gene map of the extended human MHC has reviewed the MHC genes in relation to paralogy, polymorphism, immune function, and disease.[2] However, the underlying mechanisms of the differences between different populations are poorly understood.

Methods

1000 Genomes Project data

The 1000 Genomes Project phase 3 data were downloaded from the European Bioinformatics Institute (ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). This dataset is based on 20130502 sequence freeze and alignments. It contains 2504 individuals from 26 populations. There are 167 082 biallelic variants within the HLA region (human chromosome 6,28,477,797 - 33,448,354 GRCh37). Only 33 722 SNPs with minor allele frequency higher than 1% in all 26 populations were kept for the subsequent analysis.

Quantification of linkage disequilibrium patterns between populations

The programme quantLD (https://github.com/felixfan/quantLD) was developed to measure the LD patterns between populations. The programme takes SNPs genotype data common in two populations as input, LD (D', or r^2, or signed r^2) between every pair of SNPs in a window is calculated. The window moves forward one SNP each time. In each window, a symmetric LD matrix for each population is calculated, and the difference of inter-population

LD difference is measured by raw score calculated using seven different methods. For genome-wide assessment, the raw score is standardised across the whole genome, and only the standardised scores that are larger or less than the predefined quantiles are defined as candidate regions of significant LD difference. The permutation test that randomly shuffles the label of populations is used to define significant LD difference regions.

Whole-exome sequencing of Alzheimer disease

A set of 246 APOE ε4 allele negative AD cases and 172 normal controls were enrolled in the study. Genomic DNA was isolated from the whole blood by using QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) in accordance with protocols. Coding regions were captured using the TruSeq Kit (Illumina, California, USA). The Illumina HiSeq 2000 (Illumina, California, USA) platform was used to generate 100 base-pair (bp) paired-end sequences, according to the manufacturer protocols.

Estimation of HLA types

First, sequence reads obtained by whole-exome sequencing were aligned to the genomic HLA sequences that were constructed from the international immunogenetics project HLA (IMGT/HLA) database using BWA-MEM. Then, the expected read counts on HLA class I (HLA-A, HLA-B, and HLA-C) and class II (HLA-DQA1, HLA-DQB1, and HLA-DRB1) alleles were estimated by variational Bayesian inference statistical framework in HLA-VBSeq.

HLA alleles association analysis

Pearson Chi-squared test or Fisher's exact test was performed on a 2×2 contingency table, which contains the counts of minor and major alleles for a single locus in cases and controls. Logistic and linear regression methods allow multiple covariates when testing for allele and amino acid association. The covariates can be either continuous or binary. A genotype is coded as 0, 1, or 2, depending on the number of effect allele it carries and which genetic model is tested.

We developed a Python package called PyHLA (https://github.com/felixfan/PyHLA) for HLA association analysis. PyHLA is a tailor-made, easy to use, and flexible tool designed specifically for the association analysis of the HLA types imputed from genome-wide genotyping and next-generation sequencing data. PyHLA provides functions for association analysis, zygosity tests, and interaction tests between HLA alleles and diseases. Monte Carlo permutation and several methods for multiple testing corrections were also implemented.

Results

A set of 246 APOE ε4 allele negative AD cases (70% female; 80.58±7.25 years old) and 172 normal controls (67% female; 78.5±6.27 years old) were analysed. HLA types were estimated from whole-exome sequencing data using HLA-VBSeq. HLA alleles association analysis was performed by using PyHLA.

Under allelic model, each allele was compared with the other alleles. HLA-DRB1*07:01 was inversely associated with AD (odds ratio=0.65, P=0.0027), whereas HLA-DQB1*03:01 was associated with AD (odds ratio=3.08, P=0.0017).

Under recessive model, individuals who carry two copy of the allele were compared to individuals who carry one or zero copy of the allele. HLA-C*03 was inversely associated with AD (odds ratio=0.29, P=0.0051), whereas HLA-DRB1*09:21 was strongly associated with AD (odds ratio=10.6, P=5.4e-7).

Under additive model, individuals who carry zero, one, and two copy of the allele were coded as 0, 1, and 2, respectively. Logistic regression was used to analyse the association. HLA-DRB1*07:01 was strongly inversely associated with AD (odds ratio=0.11, P=5.36e-7), whereas HLA-DRB1*09:21 was strongly associated with AD (odds ratio=10.6, P=1.02e-4).

The 1000 Genomes Project phase 3 data contains 2504 individuals from 26 populations. Overall, 33 722 SNPs within the HLA region with minor allele frequency >1% in all 26 populations were kept for analysis. We first examined LD difference across the HLA region between East Asian and European populations. Window size was set to 50 SNPs and r^2 was used to measure LD. We took the region with standardised score that was higher than the score at the 99th percentile or lower than the score at the 1st percentile as a candidate LD difference region. Only a small fraction of the candidate LD different regions was overlapped among different methods that measure LD difference between two populations. Most candidate LD different regions were unique to each method. The candidate LD different regions were consistent across different window sizes, although the smaller window size tended to identify higher resolution boundaries of the candidate regions.

About 50% to >90% of the candidate LD different regions were different between super population pairs, indicating strong evidence of inter super populations LD difference in MHC regions. The East Asian super population contained five populations. Even within the same super populations, the inter-population LD difference in MHC region still exists.

The tag SNP is a representative SNP in a high LD genome region. The pairwise tag SNP selection method in Tagger was used to select the tag SNPs by using the stand-alone programme Haploview,

which has implemented Tagger. LD r2 threshold of 0.8 was used for the selection of tag SNPs. Among the tag SNPs in the eleven populations, the Japanese population had the highest tagging efficiency (4.75), and the African ancestry population had the lowest tagging efficiency (Table). The results suggested that the tag SNPs are likely to differ between populations and the tagging efficiency are also different between different regions across the MHC regions (Fig 1). This may be one possible reason of failure replication of association in a different population.

Although haplotypes were mainly used as reference for imputation of unobserved genotypes in genome-wide association studies and phasing of other individuals. Haplotypes may enable susceptibility gene identification in complex diseases more effectively than individual SNPs because they can capture the LD patterns of a genomic region more completely. The results suggested that the length and frequency of the most common haplotype blocks vary greatly throughout the MHC region, as well as among different populations (Fig 2). The haplotype difference should be taken into consideration when planning to replicate haplotype-based association analysis.

Discussion

The current study investigated HLA alleles in southern Chinese patients with AD. The HLA typing association analysis suggested that HLA-DRB1*07:01 is significantly less prevalent in patients with AD, and that HLA-DQB1*03:01 is significantly more prevalent in patients with AD. The homozygous of HLA-C*03 is significantly less prevalent in patients with AD than controls, and homozygous of HLA-DRB1*09:21 is significantly more prevalent in patients with AD than controls.

TABLE. tag SNPs in the major histocompatibility complex of eleven populations

Populations	No. of samples	No. of SNPs	No. of tags	Ratio	No. of specific	No. of common
ASW	53	4480	1512	2.96	137	112
CEU	112	4405	1067	4.13	104	
CHB	137	4209	1181	3.56	92	
CHD	109	4013	1048	3.83	81	
GIH	101	4314	1248	3.46	126	
JPT	113	4069	856	4.75	80	
LWK	110	4471	1435	3.12	156	
MEX	58	4493	1326	3.39	120	
MKK	156	4506	1605	2.81	160	
TSI	102	4556	1361	3.35	111	
YRI	147	4341	1352	3.21	114	

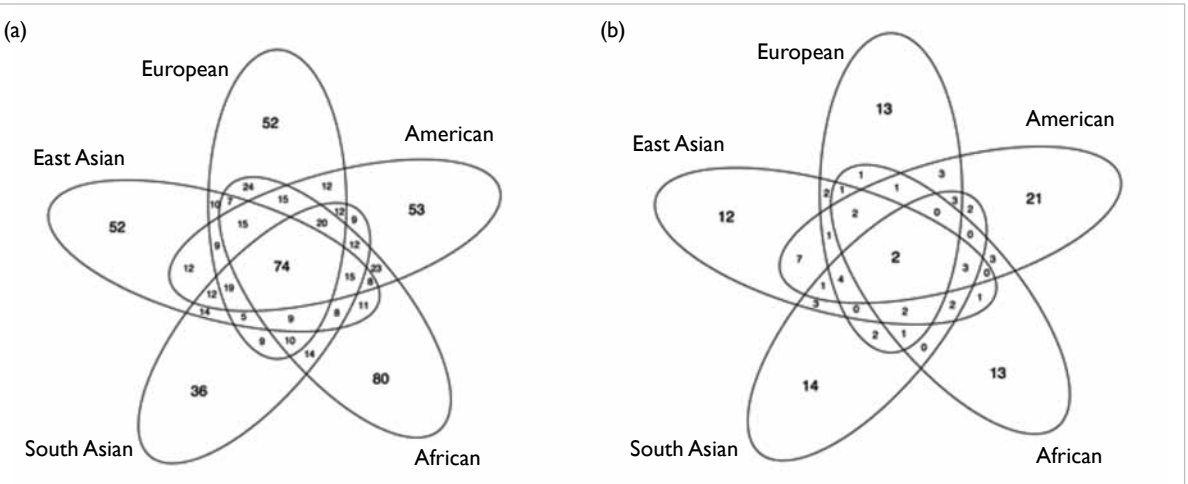

FIG 1. Number of shared tag SNPs (a) from rs1634794 to rs9265631 and (b) from rs805297 to rs805297 among different populations. In each region, 1000 SNPs with MAF of >0.05 were included. Pairwise tagging algorithm was used with r^2>0.8.

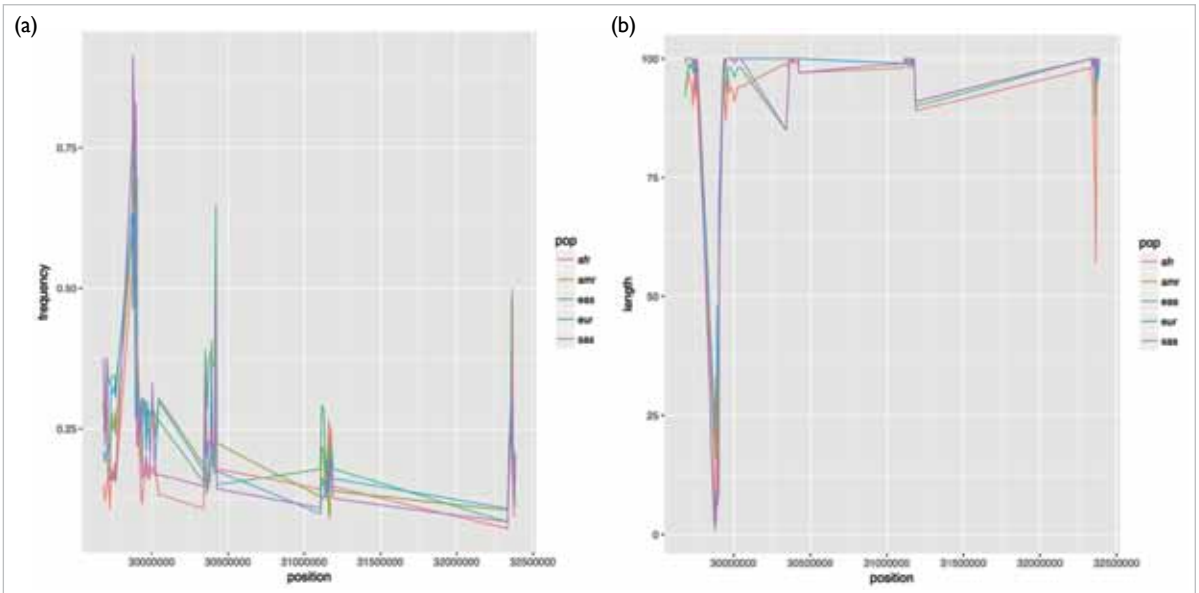

FIG 2. (a) Frequency and (b) length of the most common haplotype throughout the major histocompatibility complex region and across the populations

A meta-analysis of 74 046 individuals identified 11 new susceptibility loci for AD.[3] HLA-DRB5-DRB1 region is the most significant region. Association analysis of 48 clinically diagnosed elderly AD cases and 44 pathologically confirmed elderly controls showed an increased frequency of DRB1*03 and decreased frequency of DRB1*09 in the late-onset AD cases.[4] HLA-DRB1/DQB1 gene variants appeared to modulate the alteration of the left posterior cingulate volume, hence modulating the susceptibility of AD.[5]

In the current study, HLA-DRB1*09 was more prevalent in patients with AD, which is not consistent with previous study. This may be caused by the smaller sample size of previous study or the population structure. There is no previous study about association of HLA-C*03 and AD. The current study provides the first association study of higher resolution HLA alleles and AD in a bigger sample size.

The programme quantLD can handle larger data and is more efficient for imputation by using multithreading, compared with varLD. More methods to measure the LD difference are also available in quantLD. We compared LD difference between populations in 1000 genomes project data using quantLD. Results suggested that there are LD difference between populations. Knowledge of LD difference in a region is valuable when replicate association signals in this region across populations. For genotype imputation, the procedure for the region of LD difference may need to be different from the region with LD difference to produce confident genotypes.

Acknowledgements

This study was supported by the Health and Medical Research Fund, Food and Health Bureau, Hong Kong SAR Government (#01121726). We thank Dana Wong for technical support.

Results from this study have been published in: Li M, Li J, Li MJ, et al. Robust and rapid algorithms facilitate large-scale whole genome sequencing downstream analysis in an integrative framework. Nucleic Acids Res 2017;45:e75.

References

1. Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. Nature 1999;401:921-3.
2. Horton R, Wilming L, Rand V, et al: Gene map of the extended human MHC. Nat Rev Genet 2004;5:889-99.
3. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 2013;45:1452-8.
4. Neill D, Curran MD, Middleton D, et al. Risk for Alzheimer's disease in older late-onset cases is associated with HLA-DRB1*03. Neurosci Lett 1999;275:137-40.
5. Wang ZX, Wang HF, Tan L, et al. Effects of HLA-DRB1/DQB1 genetic variants on neuroimaging in healthy, mild cognitive impairment, and Alzheimer's disease cohorts. Mol Neurobiol 2017;54:3181-8.