Human Immunology xxx (xxxx) xxx

Contents lists available at ScienceDirect

ASHI
American Society for Histocompatibility & Immunogenetics
www.ashi-hla.org

journal homepage: www.elsevier.com/locate/humimm

Research article

Allelic and haplotypic HLA diversity in indigenous Malaysian populations explored using Next Generation Sequencing

Timothy A. Jinam [a,b,*], Kazuyoshi Hosomichi [c], Hirofumi Nakaoka [d], Maude E. Phipps [e], Naruya Saitou [a,b], Ituro Inoue [f]

[a] *Population Genetics Laboratory, National Institute of Genetics, Mishima, Japan*
[b] *Department of Genetics, The Graduate University for Advanced Studies, SOKENDAI, Mishima, Shizuoka, Japan*
[c] *Department of Bioinformatics and Genomics, Graduate School of Advanced Preventive Medical Sciences, Kanazawa University, Kanazawa, Japan*
[d] *Department of Cancer Genome Research, Sasaki Institute, Sasaki Foundation, Chiyoda-ku, Tokyo, Japan*
[e] *Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Bandar Sunway, Selangor Darul Ehsan, Malaysia*
[f] *Human Genetics Laboratory, Department of Genomics and Evolutionary Biology, National Institute of Genetics, Mishima, Japan*

ARTICLE INFO

Article history:
Received 22 May 2021
Revised 24 August 2021
Accepted 10 September 2021
Available online xxxx

Keywords:
HLA
Balancing selection
Malaysia
NGS

ABSTRACT

The heterogenous population of Malaysia includes more than 50 indigenous groups, and characterizing their HLA diversity would not only provide insights to their ancestry, but also on the effects of natural selection on their genome. We utilized hybridization-based sequence capture and short-read sequencing on the HLA region of 172 individuals representing seven indigenous groups in Malaysia (Jehai, Kintaq, Temiar, Mah Meri, Seletar, Temuan, Bidayuh). Allele and haplotype frequencies of *HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1*, and *-DPB1* revealed several ancestry-informative markers. Using SNP-based heterozygosity and pairwise Fst, we observed signals of natural selection, particularly in *HLA-A, -C* and *-DPB1* genes. Consequently, we showed the impact of natural selection on phylogenetic inference using HLA and non-HLA SNPs. We demonstrate the utility of Next Generation Sequencing for generating unambiguous, high-throughput, high-resolution HLA data that adds to our knowledge of HLA diversity and natural selection in indigenous minority groups.

1. Introduction

Malaysia is a Southeast Asian (SEA) country made up of Peninsular Malaysia which is connected to the Mainland Asian continent, and the states of Sarawak and Sabah on Borneo Island (Fig. 1). The current populations in Malaysia range from indigenous groups with historical continuities to the region, to fairly recent migrants mostly from Southern China and India. The term ‘Orang Asli’, which translates to ‘original people’ in the Malay language, refers to a heterogenous group of indigenous peoples from Peninsular Malaysia. They are further subdivided into Negrito (also referred to as Semang), Senoi, and Proto-Malay, based on geographic, phenotypic, linguistic, and cultural factors [1]. The “Layer-cake” model [2] posits that these three groups migrated sequentially into the Malay Peninsula, with the first (i.e. oldest) migrants being the Negritos and the most recent being the Proto-Malay. The Negritos and Senoi speak Austroasiatic languages [3], whereas the Proto-Malay speak Austronesian languages. The indigenous groups in Borneo also speak Austronesian languages, implying that their ancestry may be traced back to the Austronesian expansion from Taiwan that began 5000 years before present (BP) [4–6]. However, genetic analyses using various markers paint a more complex picture, with multiple migration waves and admixture events being considered [7–9]. The genomes of these indigenous peoples bear the marks of those various demographic processes, and also potentially the hallmarks of natural selection, given their long term exposure to various endemic pathogens in the region.

The Human Leukocyte Antigen (HLA) region on chromosome 6p21 is one of the candidate loci that may be under the effects of natural selection [10]. This 4 Mb region contains genes that play

Abbreviations: BP, years before present; EW, Ewens-Watterson test; NGS, Next Generation Sequencing; NJ, Neighbor-Joining; PCS, Principal Component Scores; SBT, Sequence-based typing; SEA, Southeast Asia; WGS, Whole genome sequencing.

[*] Corresponding author at: Population Genetics Laboratory, National Institute of Genetics, Yata 1111, 411-8540 Mishima, Japan.

E-mail address: tjinam@nig.ac.jp (T.A. Jinam).

https://doi.org/10.1016/j.humimm.2021.09.005
0198-8859/© 2021 The Author(s). Published by Elsevier Inc. on behalf of American Society for Histocompatibility and Immunogenetics.

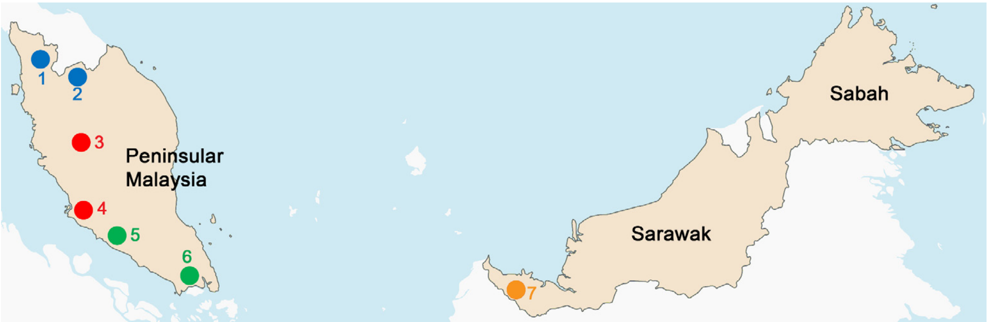

Fig. 1. Location map of populations used in this study. 1) Jehai; 2) Kintaq; 3) Temiar; 4) Mah Meri; 5) Temuan; 6) Seletar; 7) Bidayuh. The populations from Peninsular Malaysia are grouped as Negrito (blue), Senoi (red), and Proto-Malay (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

important roles in presenting processed antigens to T cells, thus eliciting immune responses. These HLA genes are very diverse, with more than 20,000 reported alleles between the nine most polymorphic genes [11], also known as classical HLA genes. Balancing selection has been proposed to explain why such high levels of polymorphism is being maintained in the HLA region [12–13]. Individuals bearing heterozygous HLA genotypes may have a higher fitness than homozygotes by being able to present a wider repertoire of antigens to immune cells, a form of balancing selection called heterozygote advantage. Selection intensity may differ between HLA genes and populations, depending on the exposure to specific pathogens endemic to a certain location. The high level of HLA diversity proved very useful for population genetics studies, by making use of allele frequency differences between populations to infer their phylogenetic relationships [14–16]. However, such phylogenetic inferences may potentially be influenced by natural selection, particularly balancing selection [17–18].

Discerning the alleles of these highly polymorphic HLA genes is not only critical for population genetics, but also for organ or tissue transplantation, and disease association studies. HLA typing methods have progressed from serology-based, to sequence-specific primer PCR and oligonucleotide probes, and to sequence-based typing. The latter was initially performed using Sanger sequencing, but now more protocols have been developed to utilize Next Generation Sequencing (NGS) technologies. NGS-based methods typically involve targeted enrichment of the HLA region using either short- or long-range PCR [19–20], or hybridization-based sequence capture [21], followed by short- or long-read sequencing using platforms such as Illumina or PacBio. These high-throughput methods may reduce typing cost per sample, but also require more bioinformatics analyses [22].

The HLA diversity in some indigenous Malaysian populations have been previously reported [23–26], with varying resolutions of allele designation, which included ambiguous allele calls. The HLA genes that were genotyped also varied by study. Here we report full HLA gene sequences in seven indigenous Malaysian populations generated using a probe-based target enrichment method followed by short-read NGS. By taking advantage of the dense SNP data generated using NGS, we aim to identify ancestry-informative HLA alleles and haplotypes, to explore the intensity of natural selection in HLA genes, and to investigate its effects on phylogenetic inferences.

2. Materials and methods

2.1. Sample information

A total of 172 individuals representing seven indigenous groups from Malaysia were recruited for this study. Six of those groups represent the Orang Asli from Peninsular Malaysia: Jehai (n = 25), Kintaq (n = 17), Mah Meri (n = 20), Seletar (n = 50), Temuan (n = 25), Temiar (n = 10); whereas the Bidayuh (n = 25) are one of the Dayak subgroups from Borneo. The Jehai and Kintaq can be further grouped as Negrito, the Mah Meri and Temiar as Semai, and the Temuan and Seletar as Proto-Malay. Their respective locations are shown in Fig. 1. Blood samples were collected after receiving approval from the respective local government agencies (JKEOA in the case of Orang Asli), and from the participants themselves. Informed consent was obtained from all participants. This study has been approved the ethics committees from Monash University Sunway, Malaysia, and the National Institute of Genetics, Japan.

2.2. Targeted HLA sequencing

DNA was extracted from blood samples using a previously described protocol [27]. Entire HLA gene sequencing was done using the sequence-capture method [21]. This method is based on hybridization between an adapter-ligated DNA library (KAPA Hyper Prep Kit, Roche) and a biotinylated DNA probe (SeqCap EZ choice kit, Roche) custom designed to target HLA genes and other loci in the MHC region. Since the terminal ends of read 2 tends to have lower quality, paired-end sequence reads were set to 350 bp for read 1 and 250 bp for read 2 in the MiSeq sequencing run to improve overall base quality.

2.3. Data analysis

2.3.1. Variant calling and HLA allele genotyping

Raw fastq files were first subjected to quality control (QC) and adapter trimming using fastp tool [28], with minimum phred score set to Q20 and other parameters set to default. Reads that passed QC were then mapped to the human reference genome hg19 using bwa v0.7.17 [29]. After marking duplicates in the mapped files using picard, variant calling was performed using HaplotypeCaller and GenotypeGVCFs programs from GATK v3.8.1 [30]. The resulting VCF file was further filtered to remove variants with less than 90% genotype call rate across all samples using vcftools v.0.0.17 [31], and multi-allelic variants using bcftools v1.9 [32]. Of the 23,690 bi-allelic variants that passed QC filtering, 67% were mapped to the HLA region, while the rest were mapped to random segments in the genome. This was possibly due to the sequencing of non-specific DNA fragments generated during the library preparation step. The average depth of coverage is 42x for genes in the HLA region and 18x for other non-HLA segments. Typing of six-digit HLA alleles was conducted in Omixon Target software version 1.9.3 (Omixon) and TypeStream Visual NGS Analysis Software

(Thermo Fisher Scientific) with IPD-IMGT/HLA Database release 3.21.0.

2.3.2. HLA allele-based analysis

Exact test for deviation from Hardy-Weinberg Equilibrium (Supplementary Table 1) and allele frequencies (Table 1) for eight HLA genes (*-A, -C, -B, -DRB1, -DQA1, -DQB1, -DPA1, -DPB1*) were calculated in each of the seven indigenous Malaysian groups. Linkage Disequilibrium (LD) was calculated between pairs of genes, and haplotype frequencies were estimated using the expectation–maximization (EM) algorithm. Allelic diversity and Ewens-Watterson (EW) test for neutrality were also calculated. All the aforementioned calculations were performed in pypop v0.7.0 [33]. Principal Component Analysis (PCA) was conducted using allele and haplotype frequency data using R version 4.0.2. To identify ancestry informative alleles and haplotypes that contributed to the observed PCA clustering, Principal Component Scores (PCS) were calculated as described in Nakaoka et al. [34].

2.3.3. SNP and sequence-based analysis

A total of 16,057 bi-allelic Single Nucleotide Polymorphisms (SNP) from the HLA region (chr6: 29,681,967–33,120,090) were extracted from the VCF file containing all Malaysian samples. Haplotype phasing was done using WhatsHap [35], a read-based phasing method. Haploid DNA sequences for each sample in FASTA format for the eight HLA genes *(-A, -C, -B, -DRB1, -DQA1, -DQB1, -DPA1, -DPB1*) were generated using bcftools, and DnaSP software [36] was used to calculate nucleotide diversity and Tajima's D. Heterozygosity and pairwise Fst were calculated using PLINK [37], and permutation tests were conducted using the coin package in R version 4.0.2.

For phylogenetic inference, all 23,690 SNPs from the Malaysian samples were merged with whole genome sequencing (WGS) data of individuals from the Simon's Genome Diversity Project (SGDP) dataset [38], 10 individuals from the Andaman Islands [39], and one Jomon-period (~4000 years BP) ancient genome from Japan [40]. The 13,518 overlapping SNPs were further divided into HLA SNPs only (dataset A; 11,520 SNPs), and non-HLA SNPs (dataset B; 1998 SNPs). To further compare phylogenetic signals between HLA and non-HLA loci, an additional dataset was prepared using genome-wide SNP data of five Orang Asli groups (Jehai, Kintaq, Mah Meri, Temuan, Seletar) genotyped using Illumina SNP array [41], a Bidayuh group genotyped using Affymetrix SNP array [42], and the same WGS samples as datasets A and B. SNPs in the HLA region were removed from this merged dataset, (dataset C; 109,821 SNPs). Linkage disequilibrium pruning was performed in PLINK [37], using 50 SNP windows, sliding every 5 SNPs, with an r^2 threshold of 0.5, for all three datasets. Nei's genetic distance was calculated from the population allele frequencies using the Phylip package [43]. Neighbor-joining (NJ) trees [44] were generated from the resulting distance matrices using MEGA X [45]. To estimate similarities between phylogenetic trees, the Robinson & Foulds (RF) metric [46] implemented in the phangorn R package was used.

3. Results

3.1. HLA allele and haplotype diversity

A total of 159 distinct alleles across 8 HLA genes were identified in the indigenous Malaysian populations (Table 1). These included *HLA-C, -DQA1, -DPA1*, and *-DPB1* genes that have not been reported in any indigenous Malaysian groups so far. All genes were in Hardy-Weinberg Equilibrium except for *HLA-A* in the Temuan, and *HLA-DPB1* in the Bidayuh (Supplementary Table 1). NGS-based HLA typing allowed for unambiguous, high resolution allele calls up to the 3rd-field level of HLA nomenclature, although no novel alleles were identified. Based on pairwise measures of LD between HLA genes (Fig. 2), haplotype frequencies were estimated using pypop [33] for *HLA-B ~ C, HLA-DRB1 ~ DQA1 ~ DQB1*, and *HLA-DPA1 ~ DPB1* (Supplementary Table 2).

3.2. Ancestry-informative HLA alleles and haplotypes

The clustering of the seven Malaysian groups based on PCA using *HLA-A* allele frequency and the three haplotype frequencies (Supplementary Table 2) is shown as colored circles in Fig. 3. The Seletar and Bidayuh showed the largest distance along the first principal component (x-axis), whereas the second principal component (y-axis) showed the largest distinction between the Negritos (Kintaq and Jehai), and the Mah Meri and Bidayuh. Ancestry informative alleles and haplotypes are shown as monochrome shapes in Fig. 3. For example, the separation of the Bidayuh from others along PC1 can be explained by the HLA-A*24:07:01 allele which has the highest frequency in that group (52%). The Bidayuh and Mah Meri are two geographically distant groups (Fig. 1), but they share high frequencies of two haplotypes: DRB1*16:02:01 ~ DQA1*01:02:02 ~ DQB1*05:02:01 (DRDQ_H39), and B*15:25:01 ~ C*04:03:01 (BC_H19). The DRB1*15:02:01 ~ DQA1*01:01:01 ~ DQB1*05:01:01 haplotype (DRDQ_H34) was in highest frequency in the Kintaq, Jehai, and Temiar, which explains their clustering in the PCA plot. When using only allele instead of haplotype frequencies, the PCA pattern is slightly different (Supplementary Fig. 1), most notably the Bidayuh and Mah Meri are clustered very closely, and the Temiar are further away from the Negritos (Jehai, Kintaq).

3.3. Natural selection in the HLA region

High heterozygosity at a genetic locus relative to neutral expectation is a simple indicator that it may be under the effects of balancing selection. We measured heterozygosity for each SNP, averaged across all seven Malaysian populations, and then grouped by gene (Fig. 4). The genes listed in Fig. 4 are the ones covered by the targeted sequencing probes (see Methods 2.2). The average heterozygosity for the 7633 non-HLA SNPs was 0.0983, indicated by a red horizontal line in Fig. 4. All genes in the HLA region had higher heterozygosity than non-HLA loci, with similar patterns observed in other global populations from the 1000 Genomes project data (Supplementary Fig. 2). Classical HLA genes have significantly higher heterozygosity compared to non-classical HLA genes ($p < 0.01$ from permutation tests), with *HLA-DPB1* showing the highest heterozygosity, followed by *HLA-A* and *HLA-C*. Some non-HLA genes in the MHC region such as *MICA* and *MICB* also have high levels of heterozygosity, possibly due to their function in immune response.

Natural selection also influences population differentiation, measured as Fst. Low Fst between divergent populations (e.g. between continents) implies that the loci may be under balancing selection [47]. Pairwise Fst was calculated between the indigenous Malaysian groups and East Asian, and European populations from the SGDP dataset [38] and averaged by gene (Fig. 5). When comparing divergent populations (Malaysian-European), *HLA-A, -C*, and *-DPB1* were the three genes that have lowest Fst compared to non-HLA loci (vertical red lines in Fig. 5), suggesting that these genes are under balancing selection. On the other hand, Fst between less differentiated populations (Malaysian-East Asian) showed significantly high values for *HLA-DQ* and *-DP* genes, implying local adaptation at those genes, possibly driven exposure to different pathogens. We also observed signals of balancing selection in HLA Class II genes, indicated by positive Tajima's D values (Sup-

Table 1
HLA allele frequencies in seven Malaysian populations.

	Bidayuh	Jehai	Kintaq	Mah-Meri	Seletar	Temuan	Temiar
n	**25**	**25**	**17**	**20**	**50**	**25**	**10**
HLA-A							
01:01:01					0.0900		
02:01:01		0.1200	0.2059	0.2500	0.0900	0.0400	0.0500
02:01:02		0.0800	0.1471				
02:03:01	0.0200	0.0200	0.0294		0.0300	0.0600	
02:05:01					0.0100		
02:06:01		0.0400					0.0500
02:07:01	0.0200	0.0200				0.0400	
03:01:01			0.0882				
11:01:01	0.1000	0.0200	0.2059	0.2250	0.2000	0.2000	0.1500
11:02:01					0.0100		
24:02:01	0.1200	0.1800	0.0294	0.0500	0.1700	0.2000	0.1000
24:07:01	0.5200	0.3600	0.2647	0.2250	0.0900	0.2600	0.4000
24:10:01					0.0600		
24:95		0.0400					
29:01:01						0.0200	
33:03:01	0.0400	0.0600	0.0294	0.0250	0.0600	0.1400	0.1000
34:01:01	0.1800	0.0600		0.2250	0.1900	0.0200	0.1500
34:05						0.0200	
HLA-C							
01:02:01	0.0200	0.0400		0.0500	0.0400	0.1400	
01:88						0.0200	
03:02:02	0.0400	0.0800			0.0200	0.0800	0.1000
03:03:01					0.0100		
03:04:01	0.0600	0.0400	0.0294	0.0500	0.1800	0.1400	0.2000
03:15			0.0294				
04:01:01	0.0400	0.0200	0.1765			0.0400	
04:03:01	0.2800		0.0294	0.3750		0.0800	
04:06	0.0200	0.0400					
04:09 N	0.1200	0.1000	0.0588			0.0200	0.1000
04:82				0.0250	0.0200	0.0200	
06:02:01				0.0500	0.0900		
06:116 N					0.0100		
07:01:01	0.0200						
07:02:01	0.0600	0.1000	0.1471	0.0250	0.2400	0.1400	0.2500
07:04:01		0.0200		0.0750	0.0600	0.1000	
07:06				0.0250		0.0200	
07:199:01		0.1800	0.2353	0.0500			
08:01:01	0.2600	0.1600	0.1471	0.2000	0.1600	0.1000	0.1500
08:22					0.0100		
12:02:02				0.0250		0.0800	
12:03:01		0.1200	0.0294				0.1500
14:02:01	0.0600	0.0800	0.1177	0.0500	0.1500		
15:02:01	0.0200	0.0200			0.0100		0.0500
15:05:02						0.0200	
HLA-B							
07:02:01			0.0882				
07:05:01						0.0200	
07:06:01	0.0200				0.0400		
13:01:01	0.0200	0.0600		0.0500	0.1800	0.1000	0.1000
15:01:01					0.0100		
15:02:01	0.1400				0.0500	0.0800	
15:13:01	0.1000	0.1400	0.1471	0.2000	0.1100	0.1000	0.1500
15:18:01	0.0200						
15:21	0.1800			0.1000			
15:25:01	0.1000		0.0294	0.2750		0.0800	
15:35		0.0600		0.0250			0.1500
18:01:01	0.0200	0.2200	0.2353	0.1250		0.1200	
18:02					0.0600		
27:06	0.0200		0.0294			0.0400	0.1000
35:01:01			0.0294				
35:05:01	0.1600	0.1000	0.2059			0.0600	0.0500
35:89		0.0200					
38:02:01		0.0200			0.0600	0.1000	0.1000
39:01:01		0.1200	0.0294		0.1500		0.1500
40:01:02	0.0600	0.0200		0.0250	0.0200	0.0600	
40:02:01		0.0200	0.0294				0.0500
40:06:01		0.0200					
44:03:02				0.0250		0.0200	
45:01:01				0.0500			
46:01:01	0.0200	0.0200			0.0100	0.0600	
48:01:01	0.0200		0.0294				

Table 1 (*continued*)

	Bidayuh	Jehai	Kintaq	Mah-Meri	Seletar	Temuan	Temiar
50:01:01					0.0100		
51:01:01	0.0400						
51:01:02	0.0200	0.0800		0.0500	0.0100		
51:02:01					0.0100		
51:02:02	0.0200		0.1177		0.1400		0.0500
52:01:01				0.0250		0.0600	
55:02:01				0.0500			
56:01:01		0.0200			0.0200	0.0200	
57:01:01					0.0900		
58:01:01	0.0400	0.0800	0.0294		0.0200	0.0800	0.1000
81:02					0.0100		
HLA-DRB1							
01:02:01				0.0750			
03:01:01	0.0200				0.0200	0.0400	
04:03:01	0.0200			0.0250	0.0200	0.0200	
04:05:01		0.1000	0.0294	0.0250			0.2500
07:01:01				0.0250	0.0100	0.0200	
08:02:01					0.0100		
08:03:02					0.0300		
09:01:02	0.0200	0.2400	0.2647	0.1750	0.0400	0.1200	
10:01:01			0.0882		0.0900	0.0200	
11:01:01	0.0200				0.0200	0.0400	
11:05					0.0100		
12:01:01					0.0100		
12:02:01	0.4400	0.1000	0.3235	0.1500	0.1200	0.1000	0.1000
13:02:01	0.0200	0.0800		0.0250		0.0600	
13:12:01					0.0400	0.0400	
14:04:01		0.0800		0.0500	0.0800	0.0200	
14:07:01					0.1400		
14:54:01	0.0200	0.0200				0.0600	
15:01:01	0.0600	0.0600	0.1177	0.0500	0.3300	0.1200	0.2000
15:02:01	0.1600	0.2800	0.1765	0.1250	0.0200	0.2000	0.3500
16:02:01	0.2200	0.0400		0.2750	0.0100	0.1400	0.1000
HLA-DQA1							
01:01:01		0.2800	0.1765	0.0250	0.0200	0.1600	0.3500
01:01:02				0.0750			
01:02:01	0.1800	0.0800	0.0882	0.1750	0.1800	0.2200	0.1000
01:02:02	0.2800	0.1000	0.0294	0.3250	0.1600	0.1400	0.2000
01:03:01		0.0200	0.0294		0.0300		0.1500
01:04:01	0.0200	0.1000		0.0500	0.2300	0.0800	
01:05:01			0.0882		0.0900	0.0200	
02:01:01				0.0250	0.0100	0.0200	
03:01:01	0.0200	0.0200	0.0294	0.0250	0.0200	0.0200	0.0500
03:02	0.0200	0.2200	0.2353	0.1750	0.0400	0.1200	
03:03:01		0.0800		0.0250			0.0500
04:01:02	0.0800		0.0588	0.0500	0.0100	0.0200	
05:01:01	0.0200				0.0200	0.0400	
05:03					0.0400	0.0400	
05:05:01	0.0200				0.0300	0.0400	
06:01:01	0.3600	0.1000	0.2647	0.0500	0.1200	0.0800	0.1000
HLA-DQB1							
02:01:01	0.0200				0.0200	0.0400	
02:02:01				0.0250	0.0100	0.0200	
03:01:01	0.4400	0.1000	0.3235	0.1000	0.1900	0.1800	0.1000
03:02:01	0.0200			0.0250	0.0200	0.0200	
03:03:02	0.0200	0.2400	0.2647	0.1750	0.0400	0.1200	
03:04:04	0.0200						
04:01:01		0.0200					
04:02:01		0.0600		0.0250	0.0100		0.1000
05:01:01		0.2600	0.2647	0.1000	0.0900	0.1600	0.3500
05:02:01	0.3000	0.1200	0.0588	0.3750	0.3600	0.3200	0.2000
05:03:01		0.1000	0.0294	0.0500	0.2300	0.0800	0.1500
05:66:01		0.0200					
06:01:01	0.1600			0.1000	0.0300		0.1000
06:02:01			0.0588				
06:04:01				0.0250		0.0600	
06:09:01	0.0200	0.0800					
HLA-DPA1							
01:03:01	0.6400	0.6000	0.7059	0.6750	0.6800	0.4200	0.5500
02:01:01	0.0800	0.2000		0.1250		0.1600	0.0500
02:02:02	0.2000	0.2000	0.1177	0.2000	0.1800	0.3400	0.3500
04:01	0.0800		0.1765		0.1400	0.0800	0.0500
HLA-DPB1							

(*continued on next page*)

Table 1 (*continued*)

	Bidayuh	Jehai	Kintaq	Mah-Meri	Seletar	Temuan	Temiar
01:01:01		0.0800	0.0294	0.0500		0.0200	0.2500
02:01:02	0.0400	0.0400	0.1177	0.1000	0.2400	0.0600	
02:02	0.0200		0.0588	0.0500	0.0200	0.0200	
03:01:01	0.1200	0.1000			0.0700	0.0600	0.1500
04:01:01	0.2600	0.3400	0.4412	0.4000	0.2300	0.1400	0.3000
04:02:01						0.0200	
05:01:01	0.1000	0.0800	0.0294		0.1100	0.1000	0.0500
09:01:01				0.0250		0.0600	
105:01:01	0.1600	0.1600	0.0588	0.1250	0.0800	0.1200	
107:01	0.0600		0.0294		0.0100	0.0400	0.0500
124:01	0.0400						
12:60:12					0.0100	0.0200	
13:01:01	0.1000	0.2000		0.1750		0.1400	0.1000
133:01	0.0200					0.0200	
135:01	0.0200				0.0500	0.0200	
21:01					0.0100		
28:01			0.0882	0.0500	0.0400	0.0800	
296:01			0.1471		0.1200	0.0400	
31:01				0.0250		0.0200	
352:01					0.0100		
463:01	0.0600					0.0200	
93:01							0.1000

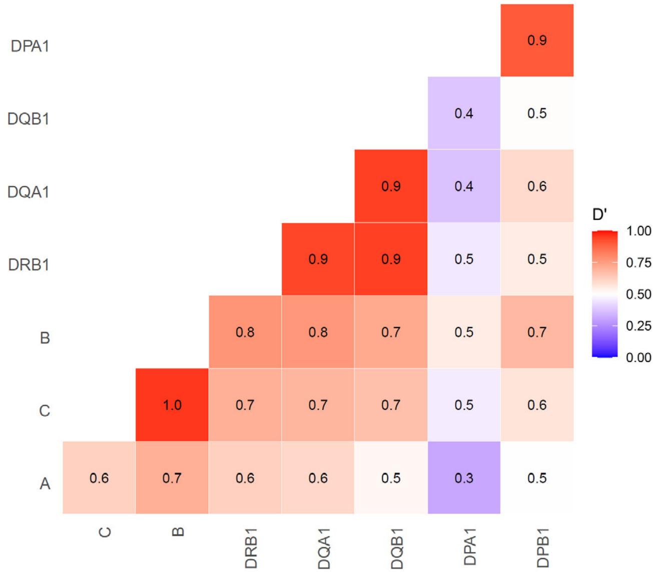

Fig. 2. Pairwise Linkage Disequilbrium (measured as D') heatmap for eight HLA genes, sorted by their genomic positions.

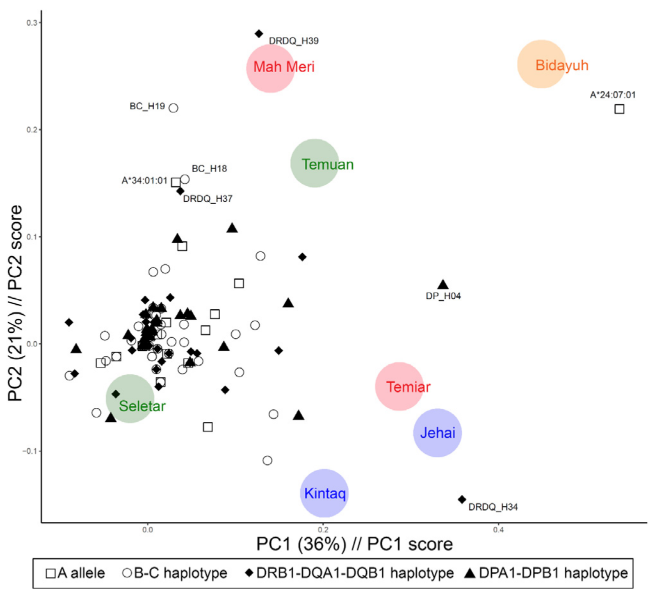

Fig. 3. Principal Component Analysis (PCA) and Principal Component Score (PCS) using HLA allele and haplotype frequencies. Population clustering using PCA, shown as colored circles, is overlaid over the PCS for the HLA alleles and haplotypes, shown as monochrome shapes.

plementary Table 3). The EW test for neutrality (Supplementary Table 4) also showed mostly negative Fnd values indicative of balancing selection, although test statistics were not significant after correction for multiple tests.

3.4. Comparison of phylogenetic signals between HLA and non-HLA loci

Based on the results in Section 3.3, we investigated how natural selection would impact phylogenetic inferences. Phylogenetic trees were generated using three sets of SNP data (Methods 2.3.3), depicted in Fig. 6. Since genome-wide SNP data for the Temiar was not available for tree c), they were also omitted from trees a) and b) in Fig. 6 for consistency. All trees were rooted using Yoruban as the outgroup, with the Papuans, French, Andamanese, and Jomon being more basal when using non-HLA SNPs (Fig. 6b and c), although the branching orders differ slightly. When using only HLA SNPs (Fig. 6a) the Han and Cambodians were more basal, and the Jomon now cluster with indigenous Taiwanese (Ami, Atayal). The clustering of Malaysian Negritos (Jehai, Kintaq) is consistent across all three trees, as is the close affinity between Mah Meri and Temuan, in agreement with HLA allele-based PCA (Fig. 3). The positions of the Seletar and Bidayuh however differ according to the dataset used. The normalized RF metric was used to measure differences between trees, with the maximum value of 1 indicating no similarities between two trees. The normalized RF between trees generated using HLA SNPs (Fig. 6a) and non-HLA SNPs (Fig. 6b and c) were both 0.9412, suggesting that those trees are quite different. The normalized RF between the trees generated using non-HLA SNPs (Fig. 6b and c) was 0.7058.

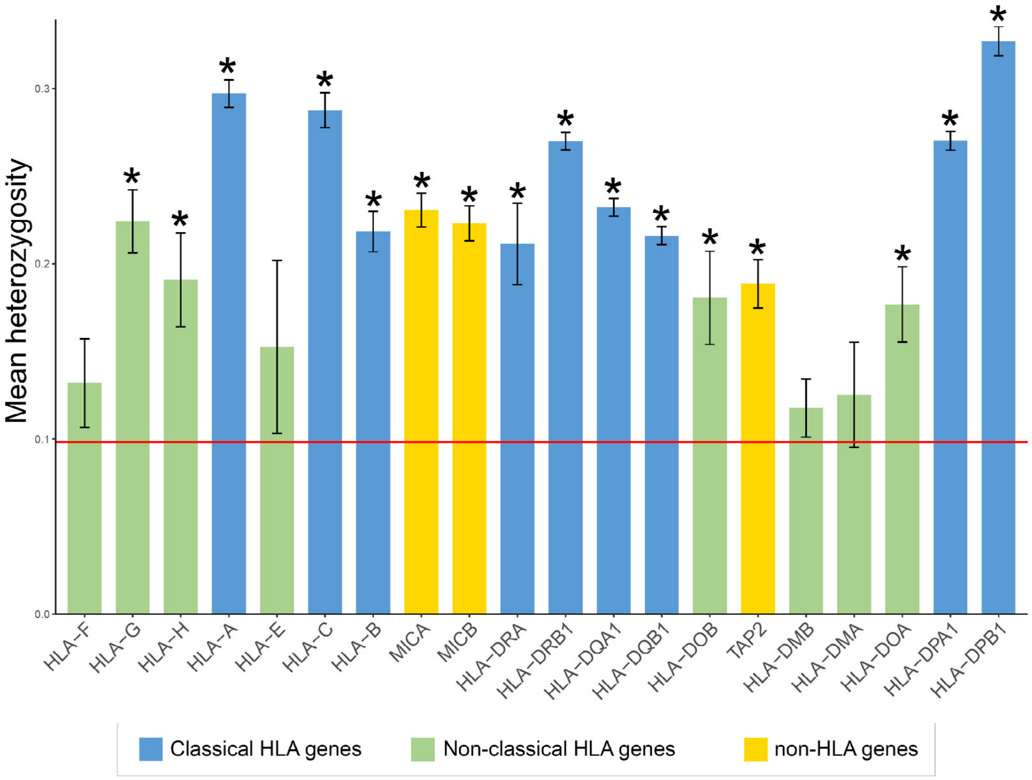

Fig. 4. SNP heterozygosity averaged across all seven Malaysian populations, grouped by genes in the MHC region. Genes are sorted according to their genomic positions and error bars are for standard error of mean. Average heterozygosity for SNPs outside the MHC region is indicated by the red horizontal line. Genes with significantly higher heterozygosity compared to non-MHC loci are indicated by asterisks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

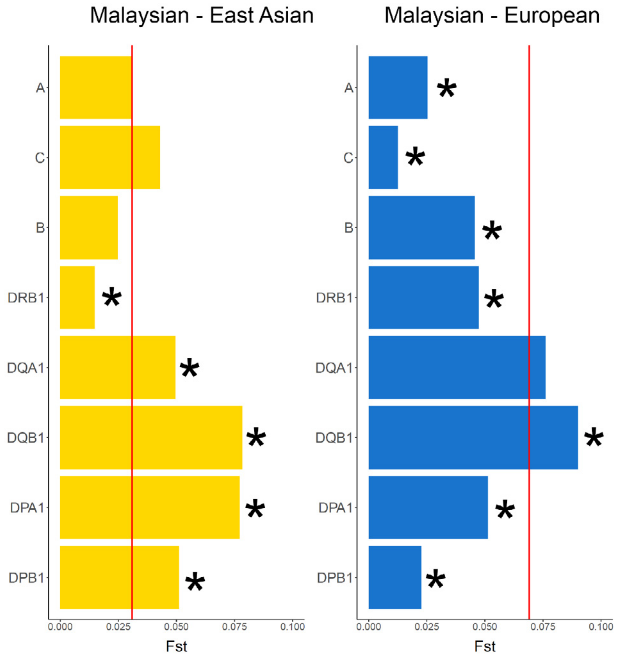

Fig. 5. Pairwise Fst between the seven indigenous Malaysian populations and East Asians, and Europeans, for eight classical HLA genes. Average Fst of SNPs outside the MHC region are indicated by vertical red lines for each pairwise comparison. HLA genes with significantly different mean Fst from the non-MHC loci are indicated by asterisks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

This study reports for the first time the HLA diversity of seven indigenous Malaysian populations using a NGS platform, including four groups for which HLA data was never reported so far: Kintaq, Seletar, Temiar, and Mah Meri. The combined use of targeted probes and NGS allows for the sequencing and genotyping of HLA genes that are less frequently reported, including *HLA-C, -DQA1, -DPA1* and *-DPB1*. This method also allows for unambiguous HLA allele designation, which was a problem for PCR and oligonucleotide probe-based tests, and even SBT methods [23].

We also report haplotype frequencies based on patterns of LD between HLA genes. Although LD in the MHC region is known to stretch over long distances [48], we report *HLA-B ~ C, HLA-DRB1 ~ DQA1 ~ DQB1*, and *HLA-DPA1 ~ DPB1* haplotypes given the very high LD (Fig. 2) and close physical distances between those gene combinations. We acknowledge that long HLA haplotypes may exist, and that these frequencies are still estimates, not based on physical haplotypes. The use of long-range NGS sequencing such as PacBio or Nanopore would be necessary to define the actual HLA haplotypes. Regardless, using both allele haplotype frequencies, we identified several ancestry informative markers for indigenous Malaysians (Fig. 3). The Kintaq, Jehai, and Temiar share high frequencies of DRB1*15:02:01 ~ DQA1*01:01:01 ~ DQB1*05:01:01 (DRDQ_H34) and DPA1*01:03:01 ~ DPB1*04:01:01 (DP_H04) haplotypes. Those two haplotypes were also frequent in Papuans and Melanesians, and in Nusa Tenggara (East of Java island, Indonesia), respectively [49].

The HLA-A*24:07:01 allele which was most frequent in the Bidayuh, is also frequent in other Orang Asli groups [23], and Indonesians from Java [50]. The DRB1*16:02:01 ~ DQA1*01:02:02 ~ DQB1*05:02:01 (DRDQ_H39) haplotype which is frequent in

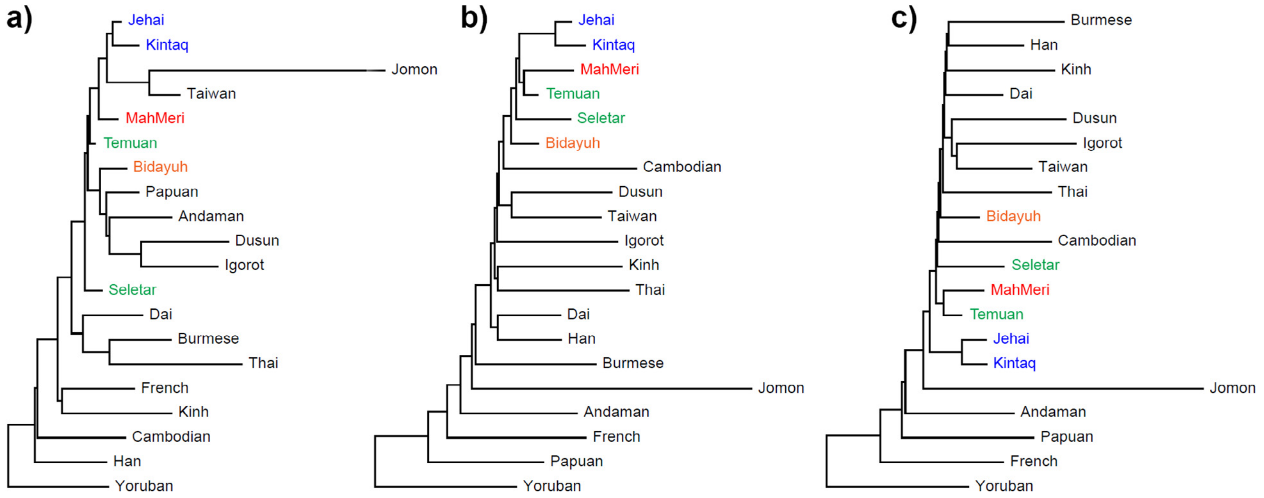

Fig. 6. Neighbor-Joining trees generated using three SNP datasets: a) HLA SNPs from targeted sequencing; b) non-HLA SNPs from targeted sequencing; c) non-HLA SNPs from genome-wide SNP arrays.

the Bidayuh and Mah Meri, was also reported in Papuans and the Muong from Vietnam [49]. On the other hand, the B*15:25:01 ~ C*04:03:01 (BC_H19) haplotype common to the Bidayuh and Mah Meri is also frequent in indigenous Taiwanese [49]. These patterns suggest a mixed ancestry from Mainland Asia and from the Austronesian expansion from Taiwan contributed to the HLA diversity in indigenous Malaysians, consistent with more recent studies [7–9]. However, population comparison using haplotype (or allele) frequencies has its difficulties, one of which is the paucity of HLA data. Even though there have been many reports detailing HLA diversity in various populations, they vary greatly in the HLA genes that were genotyped and the genotyping resolution. Furthermore, not all studies report haplotype frequencies, and even if they do, the haplotype combinations also vary by study. The allele and haplotype frequencies of eight classical HLA genes we report here would hopefully help fill the gaps of HLA data, especially for indigenous minority groups. However, we acknowledge that the sample sizes are low for some populations, for example in the Temiar (n = 10) and Kintaq (n = 17). There remains a possibility that low frequency alleles are undetected and unreported in those populations, and may potentially influence frequency-based analyses.

Balancing selection is well documented for the HLA region, and in this study we assessed the intensity of the signal in specific HLA genes using heterozygosity and pairwise Fst. We compared both parameters in HLA genes and non-HLA loci, under the assumption that these non-HLA loci are selectively neutral. Because of the small number of non-HLA loci used relative to HLA SNPs, there is a possibility that the values for these non-HLA loci may fluctuate. Both the heterozygosity and pairwise Fst, in addition to Tajima's D, points to a signal of balancing selection particularly in *HLA-A*, *-C* and *-DPB1* genes. Using the EW test for neutrality on HLA-DPA1 ~ DPB1 haplotype frequencies, Begovich et al. [52] showed excess heterozygotes in Papuans, Cameroonians, and Ugandans, although the values were not statistically significant. It may be possible that these populations from tropical climates share the same selective pressures in the form of common pathogens [53] as the indigenous groups from Malaysia, leading to observed signals of balancing selection. At the same time, pairwise Fst values were higher for *HLA-DQ* and *-DP* genes when comparing indigenous Malaysians and East Asians, suggesting the effects of directional selection. These seemingly contrasting results may be reconciled if we consider that balancing selection works over a long period of time, with signals appearing among divergent populations, whereas the signal of local positive selection may be relatively recent. Currently, it remains unclear what selective pressures lead to the observed signals of natural selection in indigenous Malaysians, but it should be the focus for future studies.

Prior to the availability of genome-wide SNP and subsequently WGS data, HLA allele frequencies were used to construct phylogenetic trees thanks to its high level of polymorphism. Here we utilized the SNPs extracted from NGS reads to compare phylogenetic trees constructed from HLA and non-HLA markers (Fig. 6). Even though trees in Fig. 6b and c were both constructed from non-HLA SNPs, there are discrepancies between them which may be attributable to the different number of SNPs used: 972 and 83,272 respectively, after LD pruning. Generally, the tree generated from the larger number of markers tends to be more reliable, and the topology of Fig. 6c is consistent with previous studies [38,51]. It contrasts greatly with the tree made from HLA SNPs (Fig. 6a), with the only consistency being the close affinities between certain indigenous Malaysian groups. The short distance between East Asians (Han and Cambodian) and Africans (Yoruban) in Fig. 6a may not be the true phylogenetic signal but instead reflects the effects of natural selection. Balancing selection results in the persistence of alleles between divergent populations (even across species), and may lead to the clustering of geographically distant populations in phylogenetic analysis, as observed in Fig. 6a. However, if we consider Fig. 6a as a gene tree, some interesting patterns can be observed. For example, the affinity between ancient Japanese (Jomon) and indigenous Taiwanese, as well as the clustering of Papuans and Andamanese with Austronesians (Bidayuh, Dusun, Igorot) may reflect shared HLA profiles between those populations, possibly driven by similar selective pressures. So while HLA markers alone may not be suitable for phylogenetic inferences of geographically separate populations, they can still be informative for inferring the substructure of closely related populations.

5. Conclusion

Here we utilized NGS technology to characterize the HLA diversity in seven indigenous Malaysian populations. This platform allowed for unambiguous and high-resolution HLA genotyping for traditional HLA allele-based descriptive analyses, as well as deeper SNP-based analysis. The identification of ancestry-informative HLA alleles and haplotypes should be useful for inferring population relationships. We also showed that HLA genes in these indigenous Malaysians were not only influenced by balancing selection, but also directional selection, particularly for HLA Class II genes. Consequently, the effects of natural selection may lead to conflicting phylogenetic signals when using only HLA SNPs for tree construction. Although the samples sizes in this study are smaller compared to other large-scale surveys, the addition of these high resolution data spanning several HLA genes should add to the growing wealth of HLA data, particularly for indigenous minority groups.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the MEXT Grant-in-Aid for Young Scientists (B) [grant number 15K18622] awarded to T.A.J, MEXT Grant-in-aid for Scientific Research on Innovative Areas “Yaponesian Genome” [grant number 18H05505] awarded to S.N, and erBiotek MOSTI grant “Genomic Variations and Their Importance In Understanding Evolution, Migration And Health in Multi-Ethnic Populations of Malaysia” [grant number 10-02-10-MB001] awarded to M.E.P. The authors would like to thank the following organizations and individuals: Jabatan Kemajuan Orang Asli (JKOA), Bau District Office, Dr. Hoh Boon Peng, Dr. Farhang Aghakhanian.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.humimm.2021.09.005.

References

[1] The Department of Orang Asli Development, Orang Asli tribes of Peninsular Malysia (in Malay). https://www.jakoa.gov.my/en/suku-kaum/, 2021 (accessed 12 April 2021).

[2] I. Carey, Orang Asli: Aboriginal Tribes of Peninsular Malaysia, Oxford University Press, Kuala Lumpur and New York, 1976.

[3] G. Benjamin, Austroasiatic Subgroupings and Prehistory in the Malay Peninsula, Ocean. Linguist. Spec. Publ. (1976) 37–128.

[4] P. Bellwood, The first farmers: the origins of agricultural societies, Wiley-Blackwell (2005).

[5] R. Blust, The prehistory of the Austronesian-speaking peoples: a view from language, J. World Prehistory. 9 (4) (1995) 453–510.

[6] I. Glover, P.S. Bellwood, Southeast Asia: From Prehistory to History, Routledge, 2004.

[7] T.A. Jinam, L.C. Hong, M.E. Phipps, M. Stoneking, M. Ameen, J. Edo, N. Saitou, Evolutionary history of continental Southeast Asians: early train hypothesis based on genetic analysis of mitochondrial and autosomal DNA data, Mol. Biol. Evol. 29 (2012) 3513–3527.

[8] M. Lipson, P.-R. Loh, N. Patterson, P. Moorjani, Y.-C. Ko, M. Stoneking, B. Berger, D. Reich, Reconstructing Austronesian population history in Island Southeast Asia, Nat. Commun. 5 (2014) 4689.

[9] P.A. Soares, J.A. Trejaut, T. Rito, B. Cavadas, C. Hill, K.K. Eng, M. Mormina, A. Brandão, R.M. Fraser, T.-Y. Wang, J.-H. Loo, C. Snell, T.-M. Ko, A. Amorim, M. Pala, V. Macaulay, D. Bulbeck, J.F. Wilson, L. Gusmão, L. Pereira, S. Oppenheimer, M. Lin, M.B. Richards, Resolving the ancestry of Austronesian-speaking populations, Hum. Genet. 135 (3) (2016) 309–326.

[10] D. Meyer, R.M. Single, S.J. Mack, H.A. Erlich, G. Thomson, Signatures of demographic history and natural Selection in the Human Major Histocompatibility Complex loci, Genetics. 173 (2006) 2121–2142.

[11] EMBL-EBI, IPD-IMGT/HLA Statistics. https://www.ebi.ac.uk/ipd/imgt/hla/stats.html, 2021 (accessed 10 May 2021).

[12] A.L. Hughes, M. Nei, Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection, Nature. 335 (6186) (1988) 167–170.

[13] N. Takahata, Y. Satta, J. Klein, Polymorphism and balancing selection at major histocompatibility complex loci., Genetics. 130 (1992) 925–938.

[14] S.J. Mack, T.L. Bugawan, P.V. Moonsamy, J.A. Erlich, E.A. Trachtenberg, Y.K. Paik, A.B. Begovich, N. Saha, H.P. Beck, M. Stoneking, H.A. Erlich, Evolution of Pacific / Asian populations inferred from HLA class II allele frequency distributions, Tissue Antigens. 55 (2000) 383–400.

[15] Y. Hatta, J. Ohashi, T. Imanishi, H. Kamiyama, M. Iha, T. Simabukuro, A. Ogawa, H. Tanaka, T. Akaza, T. Gojobori, T. Juji, K. Tokunaga, HLA genes and haplotypes in Ryukyuans suggest recent gene flow to the Okinawa Islands, Hum. Biol. 71 (1999) 353–365.

[16] L. Shi, L. Shi, Y.F. Yao, M. Matsushita, L. Yu, X.Q. Huang, W. Yi, T. Oka, K. Tokunaga, J.Y. Chu, Genetic link among Hani, Bulang and other Southeast Asian populations: evidence from HLA -A, -B, -C, -DRB1 genes and haplotypes distribution, Int. J. Immunogenet. 37 (2010) 467–475.

[17] S. V Edwards, Natural selection and phylogenetic analysis, Proc. Natl. Acad. Sci. 106 (2009) 8799 LP – 8800.

[18] J. Klein, Y. Satta, C. O'hUigin, N. Takahata, The molecular descent of the major histocompatibility complex, Annu. Rev. Immunol. 11 (1) (1993) 269–295.

[19] K. Hosomichi, T.A. Jinam, S. Mitsunaga, H. Nakaoka, I. Inoue, Phase-defined complete sequencing of the HLA genes by next-generation sequencing, BMC Genomics. 14 (2013) 355.

[20] V. Albrecht, C. Zweiniger, V. Surendranath, K. Lang, G. Schöfl, A. Dahl, S. Winkler, V. Lange, I. Böhme, A.H. Schmidt, Dual redundant sequencing strategy: full-length gene characterisation of 1056 novel and confirmatory HLA alleles, HLA. 90 (2017) 79–87.

[21] J. Hirata, K. Hosomichi, S. Sakaue, M. Kanai, H. Nakaoka, K. Ishigaki, K. Suzuki, M. Akiyama, T. Kishikawa, K. Ogawa, T. Masuda, K. Yamamoto, M. Hirata, K. Matsuda, Y. Momozawa, I. Inoue, M. Kubo, Y. Kamatani, Y. Okada, Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population, Nat. Genet. 51 (3) (2019) 470–480.

[22] S. Klasberg, V. Surendranath, V. Lange, G. Schöfl, Bioinformatics strategies, challenges, and opportunities for next generation sequencing-based HLA Genotyping, Transfus. Med. Hemotherapy Off. Organ Der Dtsch. Gesellschaft Fur Transfusionsmedizin Und Immunhamatologie. 46 (5) (2019) 312–325.

[23] T.A. Jinam, N. Saitou, J. Edo, A. Mahmood, M.E. Phipps, Molecular analysis of HLA Class I and Class II genes in four indigenous Malaysian populations, Tissue Antigens. 75 (2010) 151–158.

[24] J.S. Dhaliwal, M. Shahnaz, A. Azrena, Y.A. Irda, M. Salawati, C.L. Too, Y.Y. Lee, HLA polymorphism in three indigenous populations of Sabah and Sarawak, Tissue Antigens. 75 (2010) 166–169.

[25] A.R. Tasnim, S. Allia, H.A. Edinur, S. Panneerchelvam, Z. Zafarina, M.N. Norazmi, Distribution of HLA-A, -B and -DRB1 alleles in the Kensiu and Semai Orang Asli sub-groups in Peninsular Malaysia, Hum. Immunol. 77 (8) (2016) 618–619.

[26] K. Hirayama, A.S.M. Zaidi, S. Lokman Hakim, A. Kimura, K.J. Ong, M. Kikuchi, H. A. Nasuruddin, S. Kojima, J.W. Mak, Molecular analysis of HLA-B in the Malaysian aborigines, Tissue Antigens. 48 (1996) 692–697.

[27] S.A. Miller, D.D. Dykes, H.F. Polesky, A simple salting out procedure for extracting DNA from human nucleated cells, Nucleic Acids Res. 16 (1988) 1215.

[28] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics. 34 (2018) i884–i890.

[29] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.

[30] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303.

[31] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean, R. Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools, Bioinformatics. 27 (15) (2011) 2156–2158.

[32] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A. Whitwham, T. Keane, S.A. McCarthy, R.M. Davies, H. Li, Twelve years of SAMtools and BCFtools., Gigascience. 10 (2021).

[33] A.K. Lancaster, R.M. Single, O.D. Solberg, M.P. Nelson, G. Thomson, PyPop update–a software pipeline for large-scale multilocus population genomics, Tissue Antigens. 69 (Suppl 1) (2007) 192–197.

[34] H. Nakaoka, S. Mitsunaga, K. Hosomichi, L. Shyh-Yuh, T. Sawamoto, T. Fujiwara, N. Tsutsui, K. Suematsu, A. Shinagawa, H. Inoko, I. Inoue, M.G. Baroni, Detection of ancestry informative HLA alleles confirms the admixed origins of Japanese Population, PLoS One 8 (4) (2013) e60793.

[35] M. Martin, M. Patterson, S. Garg, S.O. Fischer, N. Pisanti, G.W. Klau, A. Schöenhuth, T. Marschall, WhatsHap: fast and accurate read-based phasing, BioRxiv. (2016) 85050.

[36] J. Rozas, A. Ferrer-Mata, J.C. Sánchez-DelBarrio, S. Guirao-Rico, P. Librado, S.E. Ramos-Onsins, A. Sánchez-Gracia, DnaSP 6: DNA sequence polymorphism analysis of large data sets, Mol. Biol. Evol. 34 (2017) 3299–3302.

[37] C.C. Chang, C.C. Chow, L.C.A.M. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, Gigascience. 4 (2015) 1.

[38] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J.P. Spence, Y.S. Song, G. Poletti, F. Balloux, G. Van Driem, P. De Knijff, I.G. Romero, A.R. Jha, D. M. Behar, C.M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O.L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M.S. Abdullah, A. Ruiz-Linares, C.M. Beall, A. Di Rienzo, C. Jeong, E.B. Starikovskaya, E. Metspalu, J. Parik, R. Villems,

B.M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J.T.S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M.F. Hammer, T. Kivisild, W. Klitz, C.A. Winkler, D. Labuda, M. Bamshad, L.B. Jorde, S.A. Tishkoff, W.S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Paäbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations, Nature. 538 (2016) 201–206.

[39] M. Mondal, F. Casals, T. Xu, G.M. Dall'Olio, M. Pybus, M.G. Netea, D. Comas, H. Laayouni, Q. Li, P.P. Majumder, J. Bertranpetit, Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation, Nat. Genet. 48 (9) (2016) 1066–1070.

[40] H. Kanzawa-Kiriyama, T.A. Jinam, Y. Kawai, T. Sato, K. Hosomichi, A. Tajima, N. Adachi, H. Matsumura, K. Kryukov, N. Saitou, K.-I. Shinoda, Late Jomon male and female genome sequences from the funadomari site in Hokkaido, Japan, Anthropol. Sci. 127 (2) (2019) 83–108.

[41] F. Aghakhanian, Y. Yunus, R. Naidu, T. Jinam, A. Manica, B.P. Hoh, M.E. Phipps, Unravelling the genetic history of negritos and indigenous populations of southeast Asia., Genome Biol. Evol. 7 (2015) 1206–1215.

[42] D. Reich, N. Patterson, M. Kircher, F. Delfin, M. Nandineni, I. Pugach, A.-S. Ko, Y.-C. Ko, T. Jinam, M. Phipps, N. Saitou, A. Wollstein, M. Kayser, S. Pääbo, M. Stoneking, Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania, Am. J. Hum. Genet. 89 (4) (2011) 516–528.

[43] J. Felsenstein, PHYLIP - Phylogeny Inference Package (Version 3.2), Cladistics 5 (1989) 164–166.

[44] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.

[45] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: molecular evolutionary genetics analysis across computing platforms, Mol. Biol. Evol. 35 (2018) 1547–1549.

[46] D.F. Robinson, L.R. Foulds, Comparison of phylogenetic trees, Math. Biosci. 53 (1-2) (1981) 131–147.

[47] D. Meyer, V.R. C. Aguiar, B.D. Bitarello, D.Y. C. Brandt, K. Nunes, A genomic perspective on HLA evolution, Immunogenetics. 70 (1) (2018) 5–27.

[48] J. Dausset, L. Legrand, V. Lepage, L. Contu, A. Marcelli-Barge, I. Wildloecher, A. Benajam, T. Meo, L. Degos, A haplotype study of HLA complex with special reference to the HLA-DR series and to Bf. C2 and glyoxalase I polymorphisms., Tissue Antigens. 12 (1978) 297–307.

[49] F.F. Gonzalez-Galarza, A. McCabe, E.J.M. dos Santos, J. Jones, L. Takeshita, N.D. Ortega-Rivera, G.M. Del Cid-Pavon, K. Ramsbottom, G. Ghattaoraya, A. Alfirevic, D. Middleton, A.R. Jones, Allele frequency net database (AFND),, update: gold-standard data classification, open access genotype data and new query tools, Nucleic Acids Res. 48 (2020) (2020) D783–D788.

[50] R. Yuliwulandari, K. Kashiwase, H. Nakajima, J. Uddin, T.P. Susmiarsih, A.S.M. Sofro, K. Tokunaga, Polymorphisms of HLA genes in Western Javanese (Indonesia): close affinities to Southeast Asian populations, Tissue Antigens. 73 (2009) 46–53.

[51] T.A. Jinam, M.E. Phipps, F. Aghakhanian, P.P. Majumder, F. Datar, M. Stoneking, H. Sawai, N. Nishida, K. Tokunaga, S. Kawamura, K. Omoto, N. Saitou, Discerning the origins of the Negritos, first Sundaland people: deep divergence and archaic admixture, Genome Biol. Evol. 9 (2017) 2013–2022.

[52] A.B. Begovich, P.V. Moonsamy, S.J. Mack, L.F. Barcellos, L.L. Steiner, S. Grams, V. Suraj-Baker, J. Hollenbach, E. Trachtenberg, L. Louie, P. Zimmerman, A.V. Hill, M. Stoneking, T. Sasazuki, V.I. Konenkov, M.L. Sartakova, V.P. Titanji, O. Rickards, W. Klitz, Genetic variability and linkage disequilibrium within the HLA-DP region: analysis of 15 different populations, Tissue Antigens. 57 (2001) 424–439.

[53] Máté Manczinger, Gábor Boross, Lajos Kemény, Viktor Müller, Tobias L. Lenz, Balázs Papp, Csaba Pál, Andrew Fraser Read, Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations, PLOS Biol. 17 (1) (2019) e3000131.