Genomic map of blood group alleles in Malaysian indigenous Orang Asli population from whole genome sequences

Mercy Rophina[1,2*], *Lay Kek Teh* [3,4*], *Sridhar Sivasubbu*[1,2], *Vinod Scaria*[1,2$], *Mohd Zaki Salleh* [3,4$]

[1] *CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mathura Road, Delhi 110025, INDIA*

[2] *Academy of Scientific and Innovative Research (AcSIR), CSIR-HRDC Campus, Sector 19, Kamla Nehru Nagar, Ghaziabad, Uttar Pradesh 201002, INDIA*

[3] *Integrative Pharmacogenomics Institute, Universiti Teknologi MARA Selangor Branch, Puncak Alam Campus, Malaysia*

[4] *Faculty of Pharmacy, Universiti Teknologi MARA Selangor Branch, Puncak Alam Campus, Malaysia*

[*] *Equally contributed*

[$] *Address for correspondence*

Corresponding author Email:

Dr. Vinod Scaria - vinods@igib.in (Scaria V) ; Tel: 91-11-29879 109

Dr. Mohd Zaki Salleh - zakisalleh@uitm.edu.my (Salleh MZ) ; Tel: +603-3258 4652

Email of authors

Mercy Rophina - mercywilliams1608@gmail.com

Dr. Teh Lay Kek - tehlaykek@uitm.edu.my

Dr. Sridhar SIvasubbu - sridhar@igib.in

Abstract

Purpose

Differences in the distribution of RBC antigens defining the blood group types among different populations have been well established. However, very few studies exist that have explored the blood group profiles of indigenous populations worldwide. With the rapid advent of next generation sequencing techniques and availability of population scale genomic datasets, we have successfully explored the blood group profiles of the Orang Aslis, who are the indigenous population of Malaysia and provide a systematic comparison of the same with major global population datasets.

Methods

Variant call files from whole genome sequence data (hg19) of 114 Orang Asli were retrieved from The Orang Asli Genome Project (OAGP). Systematic variant annotations were performed using ANNOVAR and only those variants spanning genes of 43 blood group systems and transcription factors KLF1 and GATA1 were filtered. Blood group associated allele and phenotype frequencies were determined and were duly compared with other datasets including Singapore Sequencing Malay Project (SSMP), aboriginal western desert Australians and global population datasets including The 1000 Genomes Project and gnomAD.

Results

This study reports 4 alleles *(rs12075, rs7683365, rs586178 and rs2298720) of* DUFFY, MNS, RH and KIDD blood group systems which were significantly distinct between indigenous Orang Asli and cosmopolitan Malaysians. Eighteen (18) alleles which belong to 14

blood group systems were found distinct in comparison to global population datasets. Although not much significant differences were observed in phenotypes of most blood group systems, major insights were observed on comparing Orang Asli with aboriginal Australians and cosmopolitan Malaysians.

Conclusion

This study serves as the first of its kind to utilize genomic data to interpret blood group antigen profiles of the Orang Asli population. In addition, systematic comparison of blood group profiles with related populations were also analysed and documented.

Introduction

There are over 43 blood group systems in the world, and differences in the distribution of blood antigens and blood groups between populations have been well established. The antigenic determinants expressed on the surface of the Red Blood Cells (RBCs) are often regulated either by a single gene or by closely linked homologous genes. 345 unique human blood group antigens defined by about 1700 alleles across 50 genes have been recognized and duly approved by the International Society for Blood Transfusion (ISBT) till date.[1] This diversity has enormous implications especially in countries which have diverse populations since. Mismatches in blood group antigens can lead to clinically significant alloimmunization and adversities during transfusion or pregnancy[2,3]. Accurate and extensive characterization of blood group antigen profiles to ensure safe and effective blood transfusions therefore becomes important. Owing to the rapid discovery of novel RBC antigens and their underlying genetic diversities, conventional serological techniques and medium throughput DNA assays become inadequate. Utilization of Next generation sequencing (NGS) based whole genome or exome data to extensively evaluate human RBC antigens encoding genes has been explored in recent years. [4,5,6]

Malaysia is an abode of diverse populations in Southeast Asia, comprising 3 major ethnic groups of the Malays, Chinese and Indians alongside minority groups of the aboriginal Orang Asli and natives in the East of Malaysia. These native populations have remained underrepresented in major global population genome sequencing projects. Recent years have witnessed the efforts of understanding the genetic architecture of these native populations using high throughput DNA sequencing techniques.[7,8,9,10] The Orang Asli population representing ~0.7% of peninsular Malaysia comprises three major tribes namely

Negrito, Senoi and Proto-Malays. Each of the tribes are further divided into subtribes based mainly on linguistic, physical, economical and cultural differences. The subtribes of Negrito were found historically associated with the initial wave of modern humans who migrated out of Africa ~25,000 to ~60,000 years ago forming the earliest descendants of Peninsular Malaysia.[11,12,13] Genome sequencing data of these indigenous groups were generated by The Orang Asli Genome Project (OAGP) which was initiated to unravel the genomic architecture and environmental impacts in selection pressures. Although there have been a handful of efforts in deciphering various genetic signatures of this semi-isolated population, little is known regarding the prevailing blood group profiles.

A recent study by Schoeman and colleagues portraying distinct blood group profiles of indigenous western desert Australians[14] has provided a systematic way of assessing genomic data to elucidate the distribution of blood group antigens in various indigenous populations. In this study, we aim to curate and annotate the comprehensive collection of blood alleles prevailing in the aboriginal Orang Asli population along with systematic prediction of complete blood group phenotypes from whole genome data. In addition, we also intend to filter population specific novel and rare variants with potential impacts in blood group profiles.

Materials and Methods

Reference datasets of human blood group genes and alleles

Genomic coordinates (GRCh37/hg19) of 50 genes associated with 43 human blood groups and 2 erythroid specific transcription factors were duly fetched from Locus Genomic Reference.[15] Detailed summary of genomic coordinates is tabulated in **Table 1.** A systematically compiled reference data comprising ISBT approved blood group related alleles were fetched and documented in a pre-formatted template.[16] The reference dataset extensively includes Single Nucleotide Variations (SNVs), Insertions, Deletions, Copy Number Variations (CNVs) and combination mutations.

Genomic variation datasets

Genome sequencing data generated by The Orang Asli Genome Project (OAGP), was used in the study. The dataset comprised of sequence variations of 114 Malaysian Orang Asli individuals including the major subtribes namely Negrito *(Bateq - 23, Lanoh - 16, Kensiu - 20)*, Senoi *(Che Wong - 19, Semai - 17)* and Proto Malay *(Kanaq - 19)*. OAGP comprises a total of 21089667 unique genetic variations. In addition, genome sequencing data of 96 healthy Malays contributing to 13539289 unique variants were obtained from Singapore Sequencing Malay Project (SSMP) [17] and were used for comparison in the study. All the datasets corresponded to Human genome 19 (GRCh37/hg19) assembly. Comprehensive list of genetic variations was retrieved from the variant call format (VCF) files and were used for further analyses.

Data processing and annotation

The initial level of data processing involved fetching all variants which were found to span the LRG coordinates of human blood group related genes and erythroid specific transcription factors. All filtered variants were annotated for their functional consequences using Annotate Variation (ANNOVAR).[18] An extensive range of computational tools including SIFT,[19] Polyphen,[20] LRT, MutationTaster, Mutation Assessor,[21] FATHMM,[22] PROVEAN,[23] CADD,[24] GERP,[25] PhyloP [26] and PhastCons were used to assess the functional impact of the variations.

Identification of known and novel blood group alleles

The second level of data analysis was aimed at classifying the filtered variants into those with known blood group phenotype associations and novel/rare variants. Blood group phenotype associated variants were primarily retrieved from ISBT.[27] A preformatted compilation of human blood group associated alleles was also obtained [16] and used for comparison. Phenotypes of blood group systems with no reported blood group related variants were conferred the same nomenclature as that of the reference genome as described previously. [28]

In addition, a list of potentially novel variants was fetched based on the SNP identification numbers (dbSNP ID). A variant was termed potentially novel if it lacked the dbSNP ID. Exonic novel variants were further filtered based on their functional impact predicted using a range of computational tools and minor allele frequencies. Variants with MAF < 5% with absence of reported blood phenotypes were deemed as rare variants. A schematic representation of

the methodology followed is shown in **Figure 1.** Complete blood group profiles were also predicted for each sample used in the study.

Estimation and comparison of allele frequencies

Filtered blood group associated variants from all the datasets were systematically compiled in Variant Call Format with corresponding genotype information. Allele frequencies were estimated using PLINK.[29] Summary on the number of samples with homozygous and heterozygous genotypes were also generated using bespoke scripts. In addition, allele frequencies of the variants were fetched from major global population datasets including 1000 Genomes project[30], Exome Aggregation Consortium (ExAC v.0.3)[31] and Genome Aggregation Database (gnomAD)[32] and were used for comparison. In addition to the global comparison, frequencies of blood group variants were systematically compared and analysed between the Singaporean Malays and among the subtribes of Orang Asli.

Statistical analysis

With the aim of identifying significantly distinct blood group alleles specific for Orang Asli population, minor allele frequencies were compared with other global populations and statistical significance was observed using Fisher's exact test with a P-value < 0.05. Distinct differences in blood alleles among Orang Asli subtribes were also checked. Alleles filtered as significantly distinct were further checked for their clinical relevance in transfusion procedures and in pregnancy settings.

Results

Overview of blood group associated variants and annotations in study dataset

In the primary level analysis, all variants spanning the LRG coordinates of blood group related genes were fetched from the datasets. A total of 12542 and 9616 variants were filtered as potential blood group associated variants from the Orang Asli and SSMP datasets, respectively. **Supplementary Table 1** provides the complete summary of variant counts for each blood group system for the above mentioned datasets. In the Orang Asli dataset, about 226 of the total variants were detected across the exonic and splicing sites. Of these, 119 were found to be nonsynonymous SNVs, 80 synonymous SNVs and 1 stopgain mutation. A schematic representation of the variant summary and functional classifications across the datasets is shown in **Figure 2.**

Identification of blood group alleles and prediction of blood group phenotypes

Systematic comparison of filtered variants with reference resources revealed that a total of 33 variants belonging to 15 blood groups systems possessed blood group associated phenotypes. Twenty-one (21) of the total 33 variants were SNVs and the rest were combination mutations. For the remaining 28 blood groups including 2 erythroid specific transcription factors, the predicted phenotypes were reported the same as that of the human reference genome (hg19).[28] Similar methodologies were followed in predicting blood group phenotypes of cosmopolitan Malaysian population dataset used in the study. **Table 2** details the phenotypes and genotypes of blood group systems reported to be the same as that of the reference genome in the Malaysian Orang Asli population. Comprehensive allele and phenotype frequencies of variants which were found to match reported blood group phenotypes are compiled in **Table 3**. **Supplementary Tables 2-4** provide the complete blood

type profiles of each sample used in the study, a comprehensive compilation of each blood group system phenotype observed in the Orang Asli population and the distribution of blood phenotypes in Orang Asli subtribes, respectively.

Impact of novel and rare variants in blood group profiles

Variant classification based on dbSNP identifiers revealed that a total of 3060 variants were found potentially novel. This systematically includes 21 exonic variants primarily belonging to LAN, PEL, JUNIOR, GLOB, LUTHERAN, CROMER, KNOPS, H, LEWIS, KELL, KLF1 and JMH blood groups with no global population frequencies reported and 6 variants of CROMER, KNOPS, JUNIOR, LEWIS and H blood groups which were computationally predicted to be deleterious by at least three or more tools. **Supplementary Figure 1** provides the distribution of novel variants filtered from various blood group systems. Description of the six novel variants predicted to have potential impacts on blood group profiles is provided in **Table 4.** There was a total of 74 rare blood group variants belonging to 25 unique blood group systems. Seventeen (17) variants out of the total were potentially novel which included 13 nonsynonymous SNVs and 4 synonymous SNVs. List of rare and potentially novel blood group variants are listed in **Supplementary Table 5.**

Comparison of blood group profiles among sub-tribes and across datasets

Minor allele frequencies of blood group associated alleles were fetched from all datasets used in the study along with their corresponding frequencies in major global population datasets and significant differences across datasets were observed. **Supplementary Table 6** summarizes the frequencies of blood group alleles across various datasets. A total of 18 variants belonging to 14 blood groups were found significantly distinct in the Orang Asli

population in comparison to global population datasets (1000 Genomes Project and gnomAD). List of these variants along with their P-value is tabulated in **Table 5.** In addition, 4 variants (rs12075, rs7683365, rs586178 and rs2298720) belonging to 4 unique blood group systems namely DUFFY, MNS, RH and KIDD were found distinctly different between the aboriginal Orang Aslis and cosmopolitan Malaysians. The observed P values corresponding to the variants are 0.0270, 0.0030, 0.0000 and 0.0162 respectively. A complete overview of the distinctly different blood group alleles and corresponding phenotypes is depicted in **Figures 3A and 3B.** Blood group allele frequencies distributed among the sub tribal populations of Orang Asli is shown in **Figure 3C.**

Weak and partial antigens in Orang Aslis

There is a potential risk of hemolytic transfusion reactions in case of an antigen-negative recipient receiving an antigen-positive donor or the vice versa. Mistyping of weakly or partially expressed RBC antigens in donor RBCs as antigen-negative state can immensely alter the clinical phenotype thereby inducing adverse immune reactions. Interestingly in our study, we were able to observe weak alleles in Kidd and Junior blood group systems.

In Kidd blood group system, JK*01W.01 allele, which is predicted to weaken the antigen expression even in heterozygous genotype [33] and responsible for the JK^{a+w} phenotype, is observed in about 31% of the Orang Asli population. The observed allele frequency is found comparable to African *(1000 Genomes - 21% ; gnomAD genomes - 20%)*, East Asian *(1000 Genomes - 40% ; gnomAD genomes - 40%)* and South Asian *(1000 Genomes - 29%)* populations whereas distinctly varies from European *(1000 Genomes - 8% ; gnomAD*

genomes - 7%) and American *(1000 Genomes - 19% ; gnomAD genomes - 18%)* populations.

Similarly, in Junior blood group system, the weak allele ABCG2*01W.01, manifesting the Jr^{a+w} phenotype,[34,35] is observed in 36% of the indigenous population. Frequency of this allele was found to vary distinctly from rest of the global populations (*African : 1000 Genomes - 1.3% , East Asian : 1000 Genomes - 3.0% , South Asian : 1000 Genomes - 9.7 % , European : 1000 Genomes - 9.4% , American : 1000 Genomes - 14%)*

Discussion

This study serves first of its kind to provide the most comprehensive genetic blood group profiles of indigenous Malaysian Orang Asli population including all 43 human blood group systems and 2 erythroid specific transcription factors. Blood group alleles which are significantly different between the indigenous Orang Asli and Singaporean Malays as well as global populations were filtered. It was interesting to note that distribution of blood group allele frequencies was comparatively similar between the cosmopolitan Malaysians and Orang Asli than other global populations. In addition, although not many differences were observed in most blood group phenotypes, blood group systems with distinct changes in the manifested phenotypes among populations were also fetched. Evidence of similarities in blood group phenotypes between the Malaysian Orang Asli and the aboriginal western desert Australians were also observed. A precise compilation of alleles encoding weak/partial antigens, novel and rare alleles specific to Orang Asli was performed.

Our analysis is mainly limited by the fact that RBC antigenic expressions regulated by large deletions and insertions (especially in RH and MNS blood groups) have not been profiled owing to the limitations in the datasets. In addition, the level of concordance with serology based phenotype predictions to investigate the novel and rare variants remains to be defined.

Conclusion

Level and trend of healthcare settings often differs between indigenous and non-indigenous populations. Proportion of deaths avoidable through primary, secondary and tertiary services has always been observed higher in indigenous populations worldwide.[36] There arises increased blood transfusion requirements in cases of chronic disorders. One of the early studies which was aimed at exploring the allelic diversity of human platelet antigens of this isolated population stated that obvious similarities and differences were observed between the Orang Asli and other major Malaysian subpopulations owing to their ancestral founders.[37] This study emphasizes the importance of population scale sequencing efforts towards elucidating the comprehensive blood group antigen profiles of Malaysian Orang Asli population.

Funding

This work was supported by The Council of Scientific and Industrial Research, India (Grant : MLP2001/GenomeApp) and the Ministry of Higher Education, Malaysia (Grant : 600-RMI/LRGS 5/3 (1/2011)-1).

Author Contributions

VS conceived and designed the project with MZS and LKT. MR and VS contributed in writing the manuscript. All authors approved the final manuscript. Authors acknowledge funding from CSIR India and Ministry of Higher Education, Malaysia. The funders had no role in the preparation of the manuscript or decision to publish.

Acknowlegments

The authors acknowledge Arvinden VR and Srashti Jyoti Agrawal for their constructive comments and suggestions.

Conflict of interest

None declared.

Figures and Tables

Blood group ID	Gene name	LRG ID	LRG status	LRG curated genomic coordinates (hg19)
ABO	ABO	LRG_792	Public (23 Dec, 2013)	9:136113606-136155787
MNS	GYPA	LRG_793	Pending - Under curation	4:145028456-145066904
MNS	GYPB	LRG_794	Public (11 Feb, 2021)	4:144914325-144948014
MNS	GYPE	No LRG status		
P1PK	A4GALT	LRG_795	Public (11 Feb, 2021)	22:43086118-43122307
RH	RHCE	LRG_797	Public (24 Feb, 2021)	1:25686740-25761683
RH	RHD	LRG_796	Public (24 Feb, 2021)	1:25593981-25658936
LUTHERAN	BCAM	LRG_798	Public (01 Dec, 2016)	19:45307338-45326678
KELL	KEL	LRG_799	Public (11 Feb, 2021)	7:142636201-142664503
LEWIS	FUT3	LRG_800	Pending - Under curation	19:5840899-5862133
LEWIS	FUT6	No LRG status		
LEWIS	FUT7	No LRG status		
DUFFY	ACKR1	LRG_801	Public (11 Feb, 2021)	1:159168803-159178290
KIDD	SLC14A1	LRG_802	Public (11 Feb, 2021)	18:43261990-43334485
DIEGO	SLC4A1	LRG_803	Public (08 Feb, 2021)	17:42323757-42350502
YT	ACHE (YT)	LRG_804	Public (11 Feb, 2021)	7:100485615-100498753
XG	XG	LRG_805	Public (11 Feb, 2021)	X:2665093-2736541
SCIANNA	ERMAP	LRG_806	Public (11 Feb, 2021)	1:43277776-43312660
DOMBROCK	ART4	LRG_807	Public (23 Dec, 2013)	12:14976503-15001413
COLTON	AQP1	LRG_808	Public (20 Dec, 2016)	7:30888010-30967131
LW	ICAM4	LRG_809	Public (08 Feb, 2021)	19:10392650-10401198
CHIDO/RODGERS	C4A	LRG_137	Public (12 Apr, 2011)	6:31944834-31972457

CHIDO/RODGERS	C4B	LRG_138	Public (12 Apr, 2011)	6:31977571-32005194
H	FUT1	LRG_810	Pending - Under curation	19:49249268-49263647
H	FUT2	LRG_811	Public (11 Feb, 2021)	19:49194228-49211191
KX	XK	LRG_812	Public (11 Feb, 2021)	X:37540133-37593383
GERBICH	GYPC	LRG_813	Public (08 Feb, 2021)	2:127408684-127456246
CROMER	CD55	LRG_127	Public (24 Sep, 2010)	1:207489817-207536311
KNOPS	CR1	LRG_814	Public (08 Feb, 2021)	1:207664473-207817110
IN	CD44	LRG_815	Public(23 Dec, 2013)	11:35155417-35255949
OK	BSG	LRG_816	Public(23 Dec, 2013)	19:566325-585493
RAPH	CD151	LRG_817	Public (08 Feb, 2021)	11:827952-840835
JMH	SEMA7A	LRG_818	Public (01 Dec, 2016)	15:74699630-74731299
I	GCNT2	LRG_819	Public (08 Feb, 2021)	6:10487456-10631601
GLOB	B3GALNT1	LRG_820	Public (08 Feb, 2021)	3:160799671-160828160
GIL	AQP3	LRG_821	Public (08 Feb, 2021)	9:33439158-33452590
RHAG	RHAG	LRG_822	Public (01 Dec, 2016)	6:49570888-49609587
FORS	GBGT1	LRG_826	Public (08 Feb, 2021)	9:136026335-136044332
JUNIOR	ABCG2	LRG_823	Public (08 Feb, 2021)	4:89009416-89157474
LANGEREIS	ABCB6	LRG_824	Public (08 Feb, 2021)	2:220072488-220088712
VEL	SMIM1	LRG_827	Public (08 Feb, 2021)	1:3684325-3694546
CD59	CD59	LRG_41	Public (15 July, 2010)	11:33722556-33763024
AUGUSTINE	SLC29A1	LRG_1027	Public (07 Oct, 2015)	6:44182242-44203888
KANNO	PRNP	No LRG status		
SID	B4GALNT2	No LRG status		
CTL2	SLC44A2	No LRG status		
PEL	ABCC4	LRG_1183	Public (26 Jan, 2021)	13:95670083-95958700

MAM	EMP3	No LRG status		
GATA1	GATA1	LRG_559	Public (08 Oct, 2015)	X:48639982-48654718
KLF1	KLF1	LRG_825	Public (01 Dec, 2016)	19:12993236-13003017

Table 1. Summary of LRG genomic coordinates (hg19) for all the human blood group associated genes

Blood Group System	ISBT Blood Group ID	ISBT Blood Group System Number	Predicted Genotype call	Predicted Phenotype
DIEGO	DI	010	DI*02/DI*02	Di(a–b+)
YT	YT	011	YT* 01/YT* 01	Yt(a+b-)
DOMBROCK	DO	014	DO*01/DO*01	Do(a+b-)
COLTON	CO	015	CO*01.01/CO*01.01	Co(a+b-)
LW	LW	016	LW*05/LW*05	LW(a+b-)
CHIDO/RODGERS	CH/RG	017	C4B*03/C4B*03	Ch+Rg– or CH:1,2,3,4,5,6 RG:–1,–2
GERBICH	GE	020	GE*01/GE*01	GE:2,3,4
CROMER	CROM	021	CROM*01/CROM*01	Cra+
IN	IN	023	IN*02/IN*02	In(a–b+)
OK	OK	024	OK*01.01/OK*01.01	Ok(a+)
RAPH	RAPH	025	RAPH*01/RAPH*01	MER2+
I	I	027	GCNT2*01/GCNT2*01	I
GIL	GIL	029	GIL*01/GIL*01	GIL+
VEL	VEL	034	VEL*01/VEL*01	Vel+
CD59	CD59	035	CD59*01/CD59*01	CD59.1+

AUGUSTINE	AUG	036	AUG*01/AUG*01	At(a+)
KANNO	KANNO	037	KANNO*01/KANNO*01	KANNO1+
SID	SID	038	SID*01/SID*01	Sd(a+)
CTL2	CTL2	039	CTL2*01/CTL2*01	VER+
PEL	PEL	040	ABCC4*01/ABCC4*01	PEL+
MAM	MAM	041	MAM*01/MAM*01	MAM+
EMM	PIGG	042	PIGG*01/PIGG*01	Emm+
ABCC1	ABCC1	NA	NA	NA
JMH	SEMA7A	026	JMH*01	JMH:1 or JMH+
RHAG	RHAG	030	RHAG*01	RHAG:1 or Duclos+
KLF1	KLF1	TF	KLF1*01	Common
XG	XG	012	NA	NA
KX	XK	019	NA	NA

Table 2. Tabulation of blood group phenotypes predicted to match the reference genome in 114 Malay Orang Asli samples.

Matched Variants data summary								
Blood Group ID	**Chromosome**	**Position**	**ID**	**Ref**	**Alt**	**Gene**	**OA Allele Frequency**	**ISBT Phenotype**
P1PK	chr22	43089849	rs11541159	T	C	A4GALT	0.169643	P1+/–, P k +; A4GALT*02†
LANGERIS	chr2	220080845	rs60322991	C	T	ABCB6	0.0178571	Lan weak; ABCB6*01W.02
JUNIOR	chr4	89052323	rs2231142	G	T	ABCG2	0.606061	Jr(a+w); ABCG2*01W.01
ABO	chr9	1361329	rs817671	T	TC	ABO	0.357143	O phenotype

		08	9					
DUFFY	chr1	159175354	rs12075	G	A	ACKR1	0.0714286	FY:2 or Fy(b+) ; FY*02 or FY*B
GLOB	chr3	160804167	rs2231257	C	T	B3GALNT1	0.102679	GLOB:1 (P+); GLOB*02
LUTHERAN	chr19	45322744	rs1135062	A	G	BCAM	0.3125	LU:-18,19 or Au(a−b+) ; LU*02.19
KNOPS	chr1	207760773	rs3737002	C	T	CR1	0.147321	KN:5 or Yk(a−); KN*01.−05
KNOPS	chr1	207782931	rs6691117	A	G	CR1	0.544643	KN:−9 or KCAM− ; KN*01.−09
H	chr19	492545044	rs2071699	G	A	FUT1	0.169643	H+ ; FUT1*02
FORS	chr9	136037742	rs2073924	G	A	GBGT1	0.46875	FORS:−1 (FORS−); GBGT1*01N.02
MNS	chr4	144920566	rs374811215	G	C	GYPB	0.0357143	MNS:23 or sD+
MNS	chr4	144920596	rs7683365	G	A	GYPB	0.178571	MNS:3 or S+ ; GYPB*03 or GYPB*S
KELL	chr7	142640916	rs8176034	G	T	KEL	0.0491071	KEL:2 or k+ ; KEL*02.00.02
RH	chr1	257173605	rs609320	C	G	RHCE	0.169643	c-e-
RH	chr1	257472300	rs586178	G	C	RHCE	0.169643	RH:5 (e+ weak) ; RHCE*01.01,RHCE*ce.01
RH	chr1	256110355	rs2301153	G	C	RHD;RSRP1	0.834821	Del ; RHD*01EL.32 RHD*DEL32
KIDD	chr18	43310415	rs2298720	G	A	SLC14A1	0.308036	Jk(a+ W) ; JK*01W.01 / Jk(b+ W) ; JK*02W.04

KIDD	chr18	43316538	rs2298718	A	G	SLC14A1	0.946429	Jk(a+ W); JK*01W.06 / Jk(b+ W) ; JK*02W.03
KIDD	chr18	43319519	rs1058396	G	A	SLC14A1	0.638393	JK:2 or Jk(b+)
SCIANNA	chr1	43296456	rs146429994	G	A	ERMAP	0.0223214	SC:–7 or SCAN– ; SC*01.–07

Table 3. Tabulation of blood group alleles predicted to match ISBT approved phenotypes in 114 Malay Orang Asli samples.

Chr	Start	Ref	Alt	OA Allele Frequency	Functional classification	Gene	Blood Group System	Variant type	HGVS nomenclature
1	207500122	A	T	0.004464 29	exonic	CD55	CROMER	nonsynonymous SNV	CD55:NM_000574:exon5:c.A604T:p.S202C,CD55:NM_001114752:exon5:c.A604T:p.S202C,CD55:NM_001300902:exon5:c.A604T:p.S202C,CD55:NM_001300903:exon5:c.A604T:p.S202C,CD55:NM_001300904:exon5:c.A604T:p.S202C
1	207739215	G	T	0.008928 57	exonic	CR1	KNOPS	nonsynonymous SNV	CR1:NM_000573:exon16:c.G2549T:p.C850F,CR1:NM_000651:exon24:c.G3899T:p.C1300F
1	207789932	C	G	0.058035 7	exonic	CR1	KNOPS	nonsynonymous SNV	CR1:NM_000573:exon33:c.C5324G:p.P1775R,CR1:NM_000651:exon41:c.C6674G:p.P2225R
4	890224	A	T	0.004464	exonic	ABCG2	JUNIOR	nonsynonym	ABCG2:NM_001257386:exon11:c.T1

	09			29				ous SNV	340A:p.L447H,ABCG2:NM_001348986:exon11:c.T1340A:p.L447H,ABCG2:NM_001348987:exon11:c.T1334A:p.L445H,ABCG2:NM_001348989:exon11:c.T1340A:p.L447H,ABCG2:NM_004827:exon11:c.T1340A:p.L447H,ABCG2:NM_001348985:exon12:c.T1340A:p.L447H,ABCG2:NM_001348988:exon12:c.T1340A:p.L447H
19	5844372	C	A	0.00446429	exonic	FUT3	LEWIS	nonsynonymous SNV	FUT3:NM_001097641:exon2:c.G479T:p.R160L,FUT3:NM_000149:exon3:c.G479T:p.R160L,FUT3:NM_001097639:exon3:c.G479T:p.R160L,FUT3:NM_001097640:exon3:c.G479T:p.R160L
19	49206547	C	T	0.00446429	exonic	FUT2	H	nonsynonymous SNV	FUT2:NM_000511:exon2:c.C334T:p.P112S,FUT2:NM_001097638:exon2:c.C334T:p.P112S

Table 4. Description of 6 novel SNVs filtered in the dataset with predictions of potential impact in blood group profiles.

						1000 Genomes Project	gnomAD dataset
dbSNP ID	**Blood group**	**Chr**	**Pos**	**Ref**	**Alt**	**kg_pvals**	**gnomad_pvals**
rs11541159	P1PK	chr22	43089849	T	C	0.00000000000	0.00000000000
rs60322991	LANGEREIS	chr2	220080845	C	T	0.02499034407	0.03912390193
rs2231142	JUNIOR	chr4	89052323	G	T	0.00000000004	0.00000000000
rs8176719	ABO	chr9	136132908	T	TC	0.00403758315	0.00092759555

rs12075	DUFFY	chr1	159175354	G	A	0.00000000000	0.00000000000
rs2231257	GLOB	chr3	160804167	C	T	0.00874323306	0.00000081996
rs3737002	KNOPS	chr1	207760773	C	T	0.00000452580	0.00012439751
rs6691117	KNOPS	chr1	207782931	A	G	0.00000036869	0.04252848864
rs2071699	H	chr19	49254504	G	A	0.02717534426	0.00000000000
rs2073924	FORS	chr9	136037742	G	A	0.95123501277	0.00008768730
rs374811215	MNS	chr4	144920566	G	C	0.00000003127	0.00000000000
rs7683365	MNS	chr4	144920596	G	A	0.00124393838	0.00000312738
rs8176034	KELL	chr7	142640916	G	T	0.00021110873	0.00004909480
rs586178	RH	chr1	25747230	G	C	0.00000000000	0.00000000000
rs2301153	RH	chr1	25611035	G	C	0.00000000005	0.00000006605
rs2298720	KIDD	chr18	43310415	G	A	1.00000000000	0.00000266663
rs2298718	KIDD	chr18	43316538	A	G	0.00000000000	0.00000029145
rs146429994	SCIANNA	chr1	43296456	G	A	0.00032058550	0.00393641299

Table 5. Summary of blood group alleles found significantly distinct between the aboriginal Malays and the global population datasets

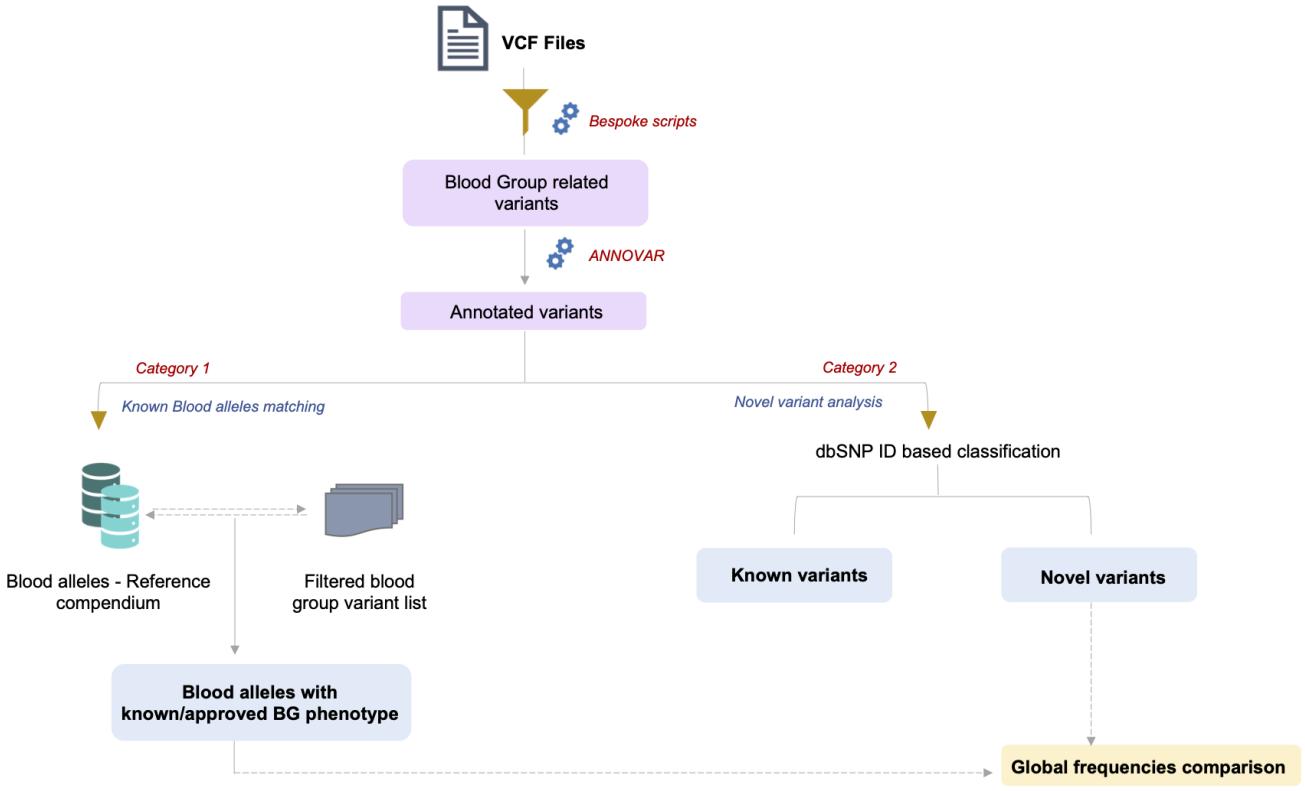


Figure 1. Schematic representation of the methodology followed in the study

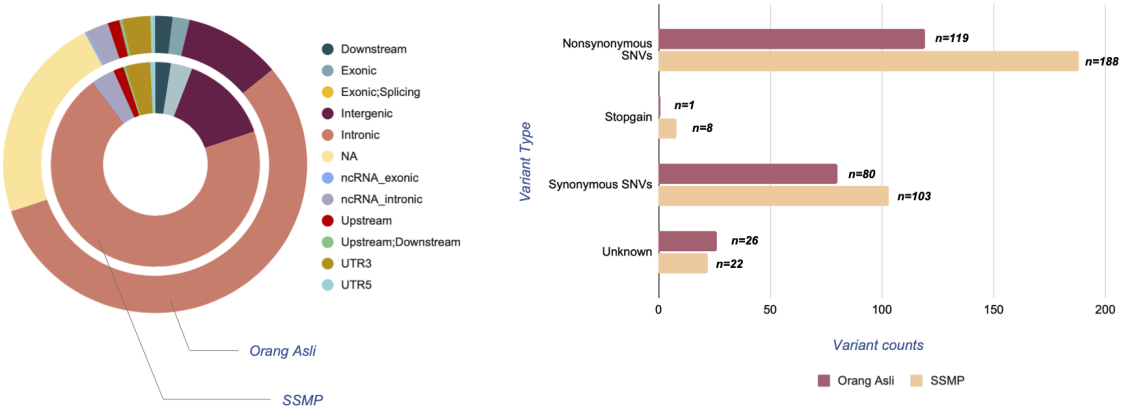


Figure 2. Complete overview of the blood group genes spanning variants and their corresponding functional classifications observed across the datasets used in the study

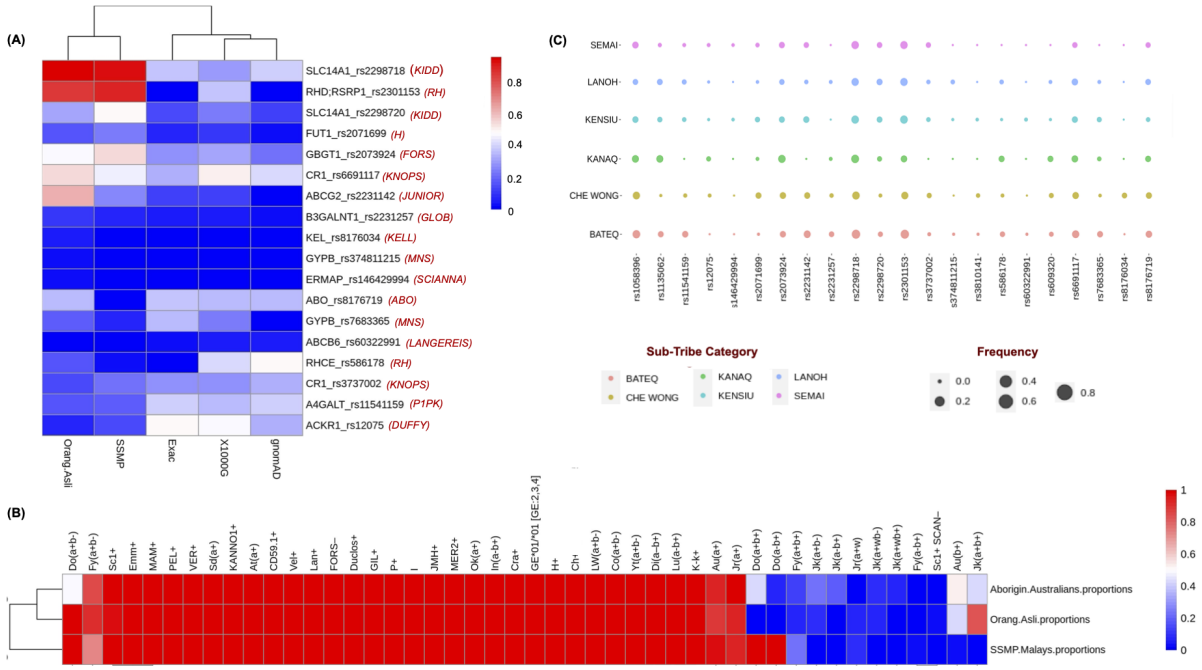
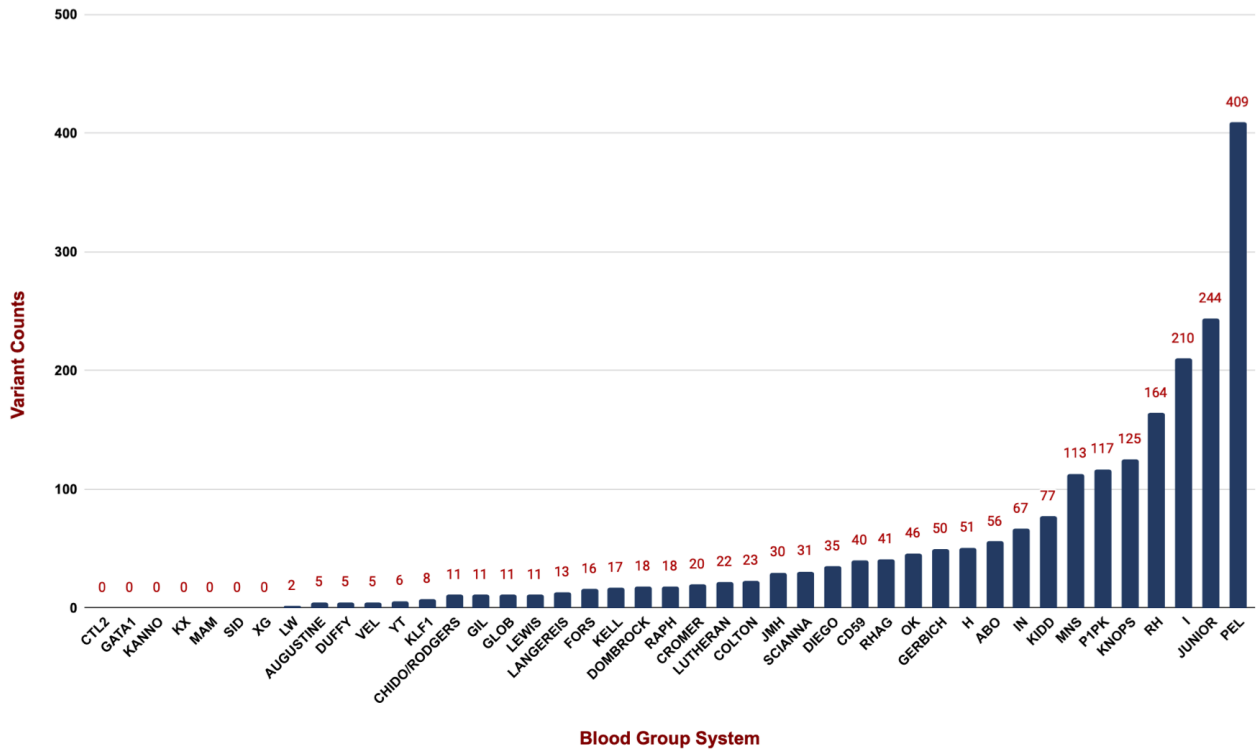


Figure 3. Distribution of blood group alleles and phenotypes among various global populations and among Orang Asli sub tribes. (A) Significantly distinct blood group alleles between Orang Aslis, cosmopolitan Malaysians and other global populations. (B) Pattern of distribution of blood group phenotypes among Orang Aslis, cosmopolitan Malaysians and aboroginal western desert Australians. © Distribution of blood group alleles in Orang Asli sub-tribal populations.

Supplementary Figures and Tables



Supplementary Figure 1. Distribution of potentially novel alleles across the human blood group systems

Supplementary Table 1. Brief tabulation of number of variants found associated with human blood group genes in the datasets used in the study.

Summary of blood group variants in study datasets

Supplementary Table 2. Complete blood type profiles of 114 Malaysian Orang Asli samples used in the study

Sample wise complete blood type profiles

Supplementary Table 3. Summary of predicted phenotypes of each blood group system in Malay Orang Asli population

Summary of overall predicted phenotypes of blood group systems

Supplementary Table 4. Distribution of predicted blood phenotypes in Orang Asli subpopulations

Summary of observed blood group phenotypes in various subpopulations of Orang Asli

Supplementary Table 5. Summary of rare and potentially novel blood group variants in Orang Asli population

List of rare and potentially novel blood group variants

Supplementary Table 6. Frequencies of blood group alleles in various population scale datasets used in this study

Population scale frequencies of filtered blood group variants

References

1. Daniels G. *Human Blood Groups*. John Wiley & Sons; 2013.

2. Pandey H, Das SS, Chaudhary R. Red cell alloimmunization in transfused patients: A silent epidemic revisited. *Asian J Transfus Sci*. 2014;8(2):75-77.

3. Webb J, Delaney M. Red Blood Cell Alloimmunization in the Pregnant Patient. *Transfus Med Rev*. 2018;32(4):213-219.

4. Stabentheiner S, Danzer M, Niklas N, et al. Overcoming methodical limits of standard RHD genotyping by next-generation sequencing. *Vox Sang*. 2011;100(4):381-388.

5. Rieneck K, Bak M, Jønson L, et al. Next-generation sequencing: proof of concept for antenatal prediction of the fetal Kell blood group phenotype from cell-free fetal DNA in maternal plasma. *Transfusion* . 2013;53(11 Suppl 2):2892-2898.

6. Giollo M, Minervini G, Scalzotto M, Leonardi E, Ferrari C, Tosatto SCE. BOOGIE: Predicting Blood Groups from High Throughput Sequencing Data. *PLoS One*. 2015;10(4):e0124579.

7. Deng L, Hoh BP, Lu D, et al. The population genomic landscape of human genetic structure, admixture history and local adaptation in Peninsular Malaysia. *Hum Genet*. 2014;133(9):1169-1185.

8. Deng L, Hoh B-P, Lu D, et al. Dissecting the genetic structure and admixture of four geographical Malay populations. *Sci Rep*. 2015;5:14375.

9. Yew CW, Hoque MZ, Pugh-Kitingan J, et al. Genetic relatedness of indigenous ethnic groups in northern Borneo to neighboring populations from Southeast Asia, as inferred from genome-wide SNP data. *Ann Hum Genet*. 2018;82(4):216-226.

10. NurWaliyuddin HZA, Norazmi MN, Edinur HA, Chambers GK, Panneerchelvam S, Zafarina Z. Ancient Genetic Signatures of Orang Asli Revealed by Killer Immunoglobulin-Like Receptor Gene Polymorphisms. *PLoS One*. 2015;10(11):e0141536.

11. HUGO Pan-Asian SNP Consortium, Abdulla MA, Ahmed I, et al. Mapping human genetic diversity in Asia. *Science*. 2009;326(5959):1541-1545.

12. Barker G, Barton H, Beavitt P, et al. Prehistoric Foragers and Farmers in South-east Asia: Renewed Investigations at Niah Cave, Sarawak. *Proceedings of the Prehistoric Society*. 2002;68:147-164. doi:10.1017/s0079497x00001481

13. Hill C, Soares P, Mormina M, et al. Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol Biol Evol*. 2006;23(12):2480-2491.

14. Schoeman EM, Roulis EV, Perry MA, Flower RL, Hyland CA. Comprehensive blood group antigen profile predictions for Western Desert Indigenous Australians from whole exome sequence data. *Transfusion* . 2019;59(2):768-778.

15. MacArthur JAL, Morales J, Tully RE, et al. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res*. 2014;42(Database issue):D873-D878.

16. Rophina M, Pandhare K, Jadhao S, Nagaraj SH, Scaria V. BGvar - a comprehensive resource for blood group immunogenetics. *bioRxiv*. Published online February 5,

2021:2021.02.04.429861. doi:10.1101/2021.02.04.429861

17. Wong L-P, Ong RT-H, Poh W-T, et al. Deep whole-genome sequencing of 100 southeast Asian Malays. *Am J Hum Genet*. 2013;92(1):52-66.

18. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.

19. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31(13):3812-3814.

20. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*. 2013;Chapter 7:Unit7.20.

21. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics*. 2013;14 Suppl 3:S7.

22. Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat*. 2013;34(1):57-65.

23. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 2015;31(16):2745-2747.

24. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019;47(D1):D886-D894.

25. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high

fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*. 2010;6(12):e1001025.

26. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010;20(1):110-121.

27. Website. Blood group systems. ISBT Science Series, 15(S1), 123–150. https://doi.org/10.1111/voxs.12593

28. Lane WJ, Westhoff CM, Uy JM, et al. Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion* . 2016;56(3):743-754.

29. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575.

30. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.

31. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291.

32. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443.

33. Wester ES, Storry JR, Olsson ML. Characterization of Jk(a+(weak)): a new blood group phenotype associated with an altered JK*01 allele. *Transfusion* . 2011;51(2):380-392.

34. Hue-Roye K, Zelinski T, Cobaugh A, et al. The JR blood group system: identification of alleles that alter expression. *Transfusion* . 2013;53(11):2710-2714.

35. Ohto H, Wada I. Variations of three single nucleotide polymorphisms in ABCG2 modify Jra expression. *International Journal of Blood Transfusion and Immunohematology*. 2019;9:1. doi:10.5348/100047z02yo2019ra

36. Ring I. The health status of indigenous peoples and others. *BMJ*. 2003;327(7412):404-405. doi:10.1136/bmj.327.7412.404

37. Syafawati WUW, Zefarina Z, Zafarina Z, et al. Human platelet antigen allelic diversity in Peninsular Malaysia. *Immunohematology*. 2016;32(4):143-160.