Asian Journal of Medicine and Biomedicine
eISNN: 2600-8173
https://journal.unisza.edu.my/ajmb

Asian Journal of Medicine and Biomedicine, Vol 5:2.

Original Article Open Access

Optimizing multiplex PCR for a set of Malay ancestry informative marker-single nucleotide polymorphisms (AIM-SNPs) and preliminary analysis of genotypes between Malay and non-Malay population

*Yi-Ting Cheng[1], Sharifah-Nany Rahayu-Karmilla Syed-Hassan[1], Padillah Yahya[1,3], Azian Harun[2], Nazihah Mohd Yunus[1], Sarina Sulong[1], & [1]Bin Alwi Zilfalil [1]**

[1] Human Genome Centre, School of Medical Sciences, Universiti Sains Malaysia,
16150 Kota Bharu, Kelantan, Malaysia
[2] Department of Medical Microbiology & Parasitology, School of Medical Sciences,
Universiti Sains Malaysia, 16150 Kota Bharu, Kelantan, Malaysia
[3]Halal Analysis Division, Center for Industrial Analysis and Customs Tariffs, Department of Chemistry Malaysia, Ministry of Science, Technology and Innovation, 46661 Petaling Jaya Selangor.

*zilfalil@usm.my

Abstract

Inference of genetic ancestry is of great interest in many fields and one of the markers in these analyses is ancestry informative marker single nucleotide polymorphisms (AIMSNPs). The Malay population is an ethnic group located mainly in South East Asia and comprises the largest ethnicity in Malaysia. SNPs that were informative for ancestry were previously selected from the genotyping data collected by the Malaysian Node of the Human Variome Project and Singapore Genome Variation Project. The SNPs were compiled into AIM-SNP panels, and from these a few SNPs were selected for optimization with multiplex PCR. Selected SNPs determined to be informative for Malay ancestry were chosen and optimized in a multiplex PCR. The chosen AIMSNPs were optimized and validated on Malay and non-Malay populations. Genotyping w as carried out on participants of self-reported Malay and non-Malay ancestry respectively with five participants in each group, and the data were compared for Malay and non-Malay population to investigate for significant differences in the genotype between Malay and non-Malay participants. The results showed great similarities between the Malay and non-Malay population, which may arise from many factors, and further optimization of more SNPs and genotyping is required to definitively conclude the validity of the AIM-SNP panels for Malay population. Knowledge of ancestry is important to minimise spurious association. This pilot study gives a brief account of the optimization process and offers an insight into how this may be done in South East Asian populations.

Keywords: *Ancestry-informative markers, single nucleotide polymorphisms, Malay, population*

*Author for Correspondence:

- Cite as: Cheng, Y. T., Syed-Hassan, S. R., Yahya, P., Harun, A., & Nazihah, M. Y. (in press). Optimizing multiplex PCR for a set of Malay ancestry informative marker-single nucleotide polymorphisms (AIM-SNPs) and preliminary analysis of genotypes between Malay and non-Malay population. Asian Journal of Medicine and Biomedicine, 5(2), 52-61. https://doi.org/10.37231/ajmb.2021.5.2.436

DOI: https://doi.org/10.37231/ajmb.2021.5.2.436

Introduction

Ancestry of an individual refers to the genetic information inherited from the individual's ancestors, in the immediate or remote past (Phillips, 2015). Inferring genetic ancestry and population genetic structure can be useful in forensics, genealogy, disease association and susceptibility, pharmacogenomics and personalized medicine (Hatin et al., 2014; V. Pereira, Mogensen, Borsting, & Morling, 2017; Salleh et al., 2013). Knowledge of ancestry and identification of population substructure is important to minimise spurious association (Yahya et al., 2017) which occurs when the study design does not take into consideration different ethnicities in the population of interest, and the subsequent proportion of different ethnicities in case and control groups (Pritchard & Rosenberg, 1999). If a disease is more prevalent in one subpopulation as compared to another, the affected subpopulation may be overrepresented in the group, thus leading to a spurious association between disease phenotype and marker loci (Pritchard & Rosenberg, 1999). This is especially of importance when the population of interest is composed of subpopulations of different genetic backgrounds (Ding et al., 2011; F. Pereira et al., 2019). When a locus is associated with disease in only one ethnic subpopulation but not another, this may indicate ethnic differences in the frequency of the risk allele (Tam et al., 2019).

Biogeographical analysis of ancestry focuses on an individual's genetic variation from which the person's ancestry or origin from a particular geographic region can be inferred (Phillips, 2015). Different populations share a great amount of genetic variation and only a small amount of variations are specific to a population (Zhao et al., 2019). Markers used to infer ancestry or biogeographic origins are known as ancestry informative markers (AIM) (Zhao et al., 2019).

Different markers has been used to infer ancestry, among them Y chromosome and mitochondrial DNA haplotypes, as well as single nucleotide polymorphisms (Phillips, 2015; Xavier & Parson, 2017; Zhao et al., 2019). Though Y-chromosome and mitochondrial DNA is strongly associated with continental regions, it is possible to misrepresent an individual's ancestry if analysis is based solely on these single markers, especially if the person has an atypical ancestry (Phillips, 2015). Thus, autosomal markers are the main markers when investigating individual ancestry (V. Pereira et al., 2017).

As genome-wide databases of human genetic variation in different populations like the International HapMap Project (Altshuler, Donnelly, & Consortium, 2005; Frazer et al., 2007; Gibbs et al., 2003; Pemberton, Wang, Li, & Rosenberg, 2010) and the Pan-Asian SNP Consortium (Ngamphiw et al.,2011) are widely and publicly available, AIM-SNPs is an attractive and convenient opportunity for the study of genetic ancestry. Ideally, for a SNP to be considered as an AIM it should be fixed in one ancestral population and be completely absent in the other population, but this is rarely the scenario in real life (Ding et al., 2011). Instead, SNPs with large allele frequency differences between different populations are chosen instead (V. Pereira et al., 2017; Yahya et al., 2017; Zhao et al., 2019).

One of the advantages of choosing SNPs as the marker of choice in ancestry informative panels is their abundance, where a polymorphism occurs approximately every 1000 bases, with negligible rate of recurrent mutation (Fondevila et al., 2017; Kidd et al., 2006; Zhao et al., 2019). They are also easy to genotype and their bi-allelic nature lends to accurate automated typing and allele calling (Fondevila et al., 2017; Kidd et al., 2006; Zhao et al., 2019).

There has been many panels developed for ancestry informative marker SNPs (AIM-SNP) and validated to various degrees (Jung et al., 2019; Li et al., 2016; Nakanishi et al., 2018; Pardo-Seco, Martinón-Torres, & Salas, 2014; V. Pereira et al., 2017; Phillips et al., 2014; Santos et al., 2016; Wei et al., 2014;Zhao et al., 2019). Commercial kits such as the Precision ID Ancestry Panel by Thermo Fisher Scientific (V. Pereira et al., 2017) and the ForenSeq DNA Signature Prep Kit by Verogen (Xavier & Parson, 2017) are also available for inferring biogeographical ancestry. An optimal AIM-SNP panel should have high discriminatory power while minimizing the number of SNPs in the assay (Zhao et al.,2019).

Malaysia is a multi-ethnic country on two landmasses, with the West Malaysia (Peninsular Malaysia) and East Malaysia separated by the South China Sea. According to the census of the Department of Statistics Malaysia in August 2011, the population of Malaysia comprises Bumiputera 67.4% (including Malays 63.1%), Chinese 24.6% and Indians 7.3%.

In Malaysia, though Malays have a formal political definition (Article 160(2), Constitution ofMalaysia), this definition does not take into full account their Austronesian origins, the different ethnic sub-groups and their genetic ancestry (Norhalifah, Syaza, Chambers, & Edinur, 2016).

The Malay population in peninsular Malaysia consists of several ethnic sub-groups that share a common Austronesian origin (Norhalifah et al., 2016). The sub-ethnicities are unique in their respective geographical origins, migration patterns and genetic affinities, which has developed over many centuries through mixing with indigenous populations (Orang Asli) and populations from further abroad (Deng et al., 2015; Hatin et al., 2014; Hoh et al., 2015; Norhalifah et al., 2016; Yahya et al., 2017). It has been theorized that modern Malays are descended from the Proto-Malays of the Orang Asli (Deng et al., 2014; Hatin et al., 2014; Hoh et al., 2015; Norhalifah et al., 2016), and this admixture has been demonstrated in a number of studies(Deng et al., 2015; Hatin et al., 2014; Hoh et al., 2015).

As Peninsular Malaysia stood at the crossroad of trade between the east and the west since ancient times, populations from other regions of Asia involved in trade and spread of religion have also left their cultural and genetic mark on the Malays. These populations include the Chinese, Indians, Arabians, and more recently Europeans after the fall of the Malacca Sultanate. (Deng et al., 2015).

As a result, populations that have contributed to the admixed ancestry of the Malay population include populations of East Asia, South Asia, Austronesian and South East Asia aboriginal people (Deng et al.,2015; Hatin et al., 2011; Norhalifah et al., 2016). Together these four major ancestral components allows differentiation of the Malay genome from most of the other South East Asian populations (Deng et al., 2015).

Yahya et al. (2017) demonstrated that by compiling 50 to 250 SNPs, the population structure of Peninsular Malay could be differentiated from the other studied population of Yoruba, Indian, Aboriginal, Chinese and Indonesian populations. Differentiating Malay ethnicity from other ethnicity is of interest in the medical field, as recent studies have shown that patients of Malay ethnicity presented with different risk variants and susceptibility to diseases. Of note, Maran et al. (2013) demonstrated that ethnic Malays in the northern state of Kelantan have an exceptionally low prevalence of Helicobacter pylori infection, which is a precursor to precancerous lesions of the stomach. These precancerous lesions are associated with H. pylori infection and is the first step of a cascade leading from precancerous lesions to full-blown malignancy. The authors further demonstrated that different gene variants manifest at different stages of the disease progression, and theorised that the disparity follows a similar pattern of disease progression.

Ethnic differences were also demonstrated in a study assessing susceptibility to mental disorders in association with gene polymorphisms (Lim et al., 2014). With ethnic stratification, significant differences were demonstrated between controls and patients suffering from bipolar disorder and schizophrenia in the Malay population .

Multiplex PCR allows for simultaneous amplification of different SNPs in one sample. A multitude of SNPs can be amplified at the same time for many samples, thus providing higher throughput and saving on both time and cost.

This pilot study aims to demonstrate and validate some of the chosen AIM-SNPs from Yahya et al.(2017), by genotyping participants of self-reported Malay ancestry and non-Malay ancestry, to detect any significant differences between the genotyping results of the two populations, and to further develop the validated SNPs into a multiplex assay kit.

Methodology

Study design

Selection of SNPs for Malay ancestry AIM panel

The selection of SNPs for the development of the AIM-SNP array kit for Malaysian Malays was based on data from Yahya et al. (2017). The SNPs selected in the above paper for an AIM panel for Malaysian Malays were extracted from genotyping data collected by the Malaysian Node of the Human Variome Project and Singapore Genome Variation Project with a total of 165 Malay individuals analyzed on the Affymetric SNP-6 SNP array platform and OMNI 2.5 Illumina SNP array platform. These data were then referenced against data from the International HapMap Project Phase 3 database. After quality control filtering, the data was merged and an SNP dataset consisting of 1766 individuals and 37,487 common SNPs was obtained. Panels of AIM-SNPs were selected using different methods: informativeness of assignment In, and PCAIM, and pairwise FST. Using WEKA and ADMIXTURE analysis, the accuracy of each panel of AIM-SNPs selected using the different methods were assessed.

From this study, it is postulated that based on the different approaches, AIM-SNP panels of approximately 200 SNPs were able to correctly classify Malay individuals into their appropriate group with an accuracy of more than 80%, though the number of SNPs required to achieve that accuracy differ with method. While a set of 157 SNPs with the highest FST has an accuracy of more than 80%, the number of SNPs to achieve the same accuracy using In and PCAIMs methods respectively were 200. Among these SNPs, one SNP was shared.

Table 1: AIM-SNPs chosen and associated ancestry coefficient, chromosome position, and associated genes

Reference SNP ID	Ancestry Coefficient method	Chromosome Position (GRCh38.12)	Gene
Rs4599414	In	Chr4:7461577	Intron variant in SORCS2
Rs752625	PCAIM	Chr4:169050766	None
Rs1978241	FST	Chr17:61510491	None
Rs12550668	In and PCAIM	Chr8:11722445	Intron variant in GATA4
Rs4134376	PCAIM and FST	Chr15:85885026	Intron variant in LOC105370953

To validate the AIM-SNP panels selected, 5 SNPs were chosen (Table 1), where one SNP was shared between In and PCAIM results (rs12550668), and another shared between PCAIM and FST results (rs4134376). There were no shared SNP between In and FST in the analysis. Among these, three SNPs (rs4599414, rs12550668 and rs4134376) are variants located in the intronic region of genes while another two SNPs (rs752625 and rs1978241) are not located within any genes.   Each of these SNPs were not detected to be of clinical significance in the National Centre for Biotechnology Information (NCBI), thus eliminating the possibility of selective pressures on the frequency of the allele.

Study subjects

This is a cross-sectional study where the participants were of Malay and Non-Malay descent. There are five participants each of Malay and Non-Malay ancestry. All the participants were from Peninsular Malaysia and can trace their ancestry for at least 3 generations, with no history of admixture occurring with other ethnicities. Detailed information was collected from the participants including age, gender, and ethnicity. Family pedigrees of all participants were obtained by interview to ensure that there was no history of inter-ethnicity marriages within the family in the last 3 generations. Informed consent was obtained from all participants prior to blood taking. Ethical clearance was obtained from the Human Research Ethics Committee of USM (USM/JEPeM/18080381).

Results

Forward and reverse primers were designed for each selected SNP (Table 2). As the overarching aim of this study is the eventual development of a multiplex array kit for Malaysian Malay ancestry, the primer sets were designed to allow a difference in the length of its amplicon for differentiation of the PCR products on gel electrophoresis.

Table 2: Design of primers used in PCR

SNP	Forward primer (5’→3’)	bp	Reverse primer (3’→5’)	bp	Amplicon size
Rs4599414	CACACTGCGTGTGATCAGTG	20	CGGATCATGCAAACATGCTA	20	178
Rs752625	AAATGCCTGACGTTGTTTCC	20	CAAGCCGGGATATTGTTCTC	20	245
Rs1978241	TCTGACGTGGCAAGAAGCTA	20	CCAGTCTCTCGGCCTATTTG	20	336
Rs12550668	ACTCACTTTCCCCCACACAG	20	TACATGAGCAACAGGGGACA	20	449
Rs4134376	CACTGGGCCGTAGATGAGAT	20	GATCCTCCCCCATGACTTCT	20	557

DNA is was extracted from whole blood samples using the GeneAll Exgene ™ Blood SV Mini (GeneAll, Germany) according to manufacturer’s proposal. The purity and concentration of the obtained DNA was measured using the NanoQuant Infinate M200 (TECAN, Switzerland). The purity of the DNA samples was determined by the ratio A260/A280 and all samples were all determined to be within 1.7 and 2.0. Gel electrophoresis was carried out on 1% agarose gel to check for the quality of the genomic DNA.

DNA amplification was carried out using Thermocycler 9600 ABI, with every primer optimized and amplified in singleplex PCR. Optimization of the singleplex PCR involves gradual trial-and-error, with the final concentration of reagents and conditions of PCR provided below. These apply to four SNPs (rs752625, rs1978241, rs12550668 and rs4134376) as multiple attempts at DNA amplification for the last SNP (rs4599414) was not successful. In addition, one SNP (rs12550668) required a lower concentration of primers to reduce non-specific amplification. Concentration of DNA sample was maintained at 50 ng/uL while the amount of distilled water was adjusted to bring to a final volume of 25 uL.

Table 3: Final concentration and volume of reagents used for singleplex PCR for rs752625, rs1978241, rs4134376 and rs12550668

	Stock concentration	Final concentration	Final volume
Rs752625, rs1978241, rs4134376			
Distilled water			Variable
Buffer	5x	1x	5.00
MgCl2	25 mM	2.00 mM	2.00
dNTP	10 mM	0.2 mM	0.50
Forward primer	10 mM	1.0 mM	1.00
Reverse primer	10 mM	1.0 mM	1.00
DNA	Variable	50 ng/uL	Variable
Taq polymerase	5 units/ul	1.25 units	0.25
Rs12550668			
Distilled water			Variable
Buffer	5x	1x	5.00
MgCl2	25 mM	2.00 mM	2.00
dNTP	10 mM	0.2 mM	0.50
Forward primer	10 mM	0.4 mM	0.40
Reverse primer	10 mM	0.4 mM	0.40
DNA		50 ng/uL	Variable
Taq polymerase	5 units/ul	1.25 units	0.25

PCR was performed by preparing the PCR master mix (Table 3) with distilled water, 5 uL of 5x colourless GoTaq ® Flexi Buffer (Promega, USA), 2.00 uL of 25 mM magnesium chloride (Promega, USA), 0.50 uL of deoxyribonucleotide triphosphate (dNTP) (Promega, USA), 0.25 uL of GoTaq ® Flexi DNA polymerase (Promega, USA), DNA sample volume equivalent to 50 ng/uL and concentration of primers as corresponding to the desired SNP. The PCR parameters for rs752625, rs1978241, rs12550668 and rs4134376 were listed in Table 4.

Table 4: PCR conditions

	Temperature (°C)	Duration	Cycles
Initial denaturation	95	2 minutes	-
Denaturation	95	30 seconds	35
Annealing	56	30 seconds	35
Extension	72	30 seconds	35
Final extension	72	5 minutes	-

The singleplex PCR products were electrophoresed on 2% agarose gel after staining with SybrGreen (Cambrex Bioscience Rockland Inc, USA) and photographs taken by ultraviolet transillumination. The size of the PCR amplicons was compared against the 100-bp DNA ladder to gauge the approximate size of the PCR product, which could indicate that PCR was successful in obtaining amplicons of the desired size, which are then confirmed with sequencing. An example of the gel electrophoresis for rs1250668 is provided in Figure 1. The design of the primers for rs1250668 is expected to produce amplicons 449 bp long, which are confirmed by gel electrophoresis.

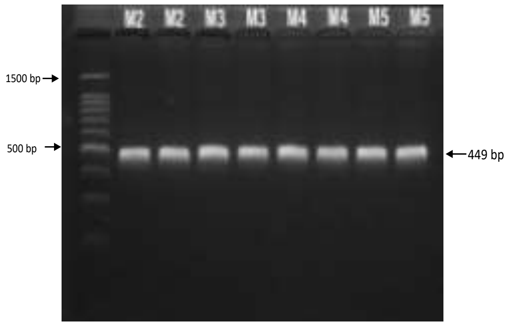

Figure 1: Gel electrophoresis of PCR product rs1250668 (449bp) for Malay samples in duplicate with primer concentration of 0.4mM , with the leftmost lane as the 100 bp DNA ladder.

After singleplex PCR had been optimized for the four SNPs, five Malay and five non-Malay samples were amplified at each SNP and the PCR products sequenced to determine the polymorphism present at each SNP for the Malay and non-Malay samples.

The sequencing results were analysed using Bioedit 7.2 software to ensure the correct amplicons have been amplified. Figure 2, Figure 3 and Figure 4 show the sequencing results with the SNP loci highlighted in red boxes , respectively, for rs752625 showing homozygosity T/T in two samples, rs12550668 showing homozygosity A/A and heterozygosity A/G in each of two samples, and rs4134376 showing heterozygosity A/G in two samples.

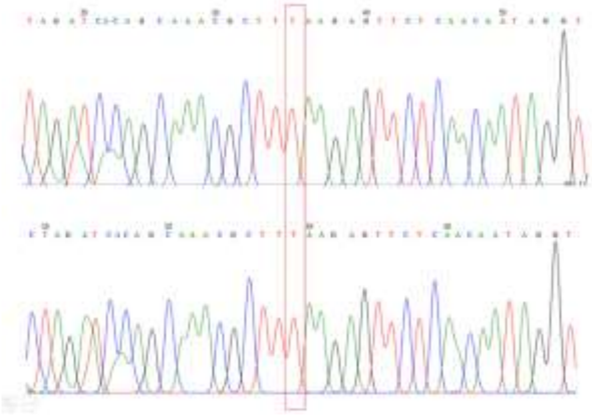

Figure 2: Sequencing results for two non-Malay subjects for rs752625 with SNP loci highlighted in the red box, both showing homozygosity.

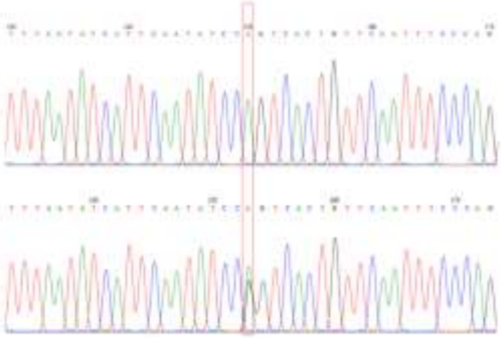

Figure 3: Sequencing results for two Malay subjects for rs12550668 with SNP loci highlighted in the red box, showing homozygosity AA and heterozygosity A/G.

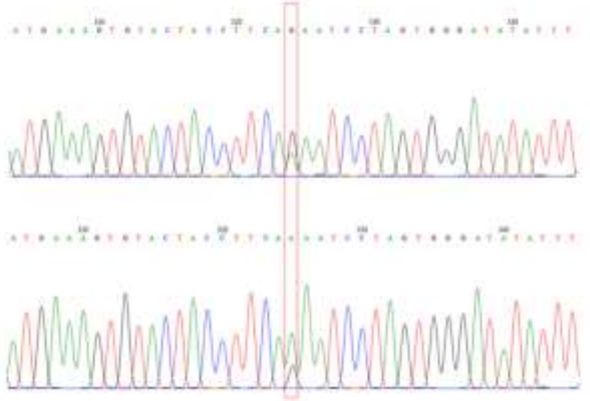

Figure 4: Sequencing results for two Malay subjects for rs4134376 with SNP loci highlighted in the red box, both showing heterozygosity A/G.

At the same time, the four SNPs that were successfully optimized in singleplex were subsequently optimized in multiplex with the final concentration and volume of reagents as detailed in Table 5.

Table 5: Final concentration and volume of reagents used for multiplex PCR for rs752625, rs1978241, rs4134376 and rs12550668

	Final concentration	Final volume (1 sample, uL)
Distilled water		Variable
Buffer	1x	5.25
MgCl2	2.00 mM	5.00
dNTP	0.2 mM	0.50
Forward primers		
Rs752625	1.0 mM	1.00
Rs1978241	1.0 mM	1.00
Rs4134376	1.0 mM	1.00
Rs12550668	0.4 mM	0.40
Reverse primers		
Rs752625	1.0 Mm	1.00
Rs1978241	1.0 mM	1.00
Rs4134376	1.0 mM	1.00
Rs12550668	0.4 mM	0.40

DNA	50 ng/uL	Variable
Taq polymerase	1.25 units	0.25
TOTAL		25.00

A temperature gradient for the annealing temeprature for the multiplex PCR was done to determine the optimal annealing temperature. Gel electrophoresis was performed to verify the sizes of the four amplicons (245 bp, 336 bp, 449 bp, 557 bp) in the multiplex PCR (Figure 5).

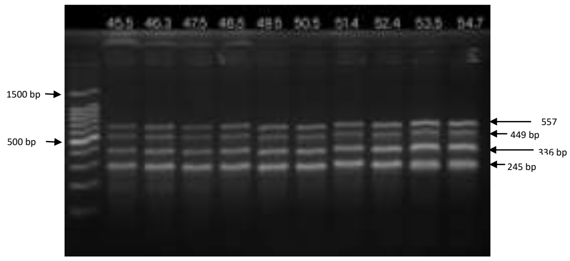

Figure 5: Gel electrophoresis with annealing temperature gradient (45.5 – 54.7°C) of multiplex of five SNPs, showing bands for rs752625 (size 245 bp) , rs1978241 (size 336 bp), rs4134376 (size 449 bp) and rs12550668 (size 557 bp). SNP rs4599414 was not amplified successfully and therefore not seen in this gel electrophoresis . The leftmost lane indicates the 100 bp DNA ladder.

The genotypes of the Malay and non-Malay samples for the four SNPs were determined, and statistical analysis using IBM SPSS Statistics for Windows, Version 26.0. The results of the genotypes for the four SNPs and the statistical analysis using chi-square test between Malay and non-Malay populations are presented in Table 6.

Table 6: Analysis of genotype and allele frequency of 4 SNPs rs7525625, rs1978241, rs12550668 and rs4134376 in Malay and non-Malay population.

RefSeq	Genotype /Allele	Population		p value
		Malay,n (%)	Non-Malay, n (%)	
Rs752625	TT	2 (40%)	5 (100%)	0.167
	CT	1 (20%)	-	
	CC	2 (40%)	-	
	C-allele	5 (50%)	0	0.033
	T-allele	5 (50%)	10 (100%)	
Rs1978241	GG	2 (40%)	1 (20%)	0.524
	AG	3 (60%)	4 (80%)	
	A-allele	3 (30%)	4 (40%)	1.000
	G-allele	7 (70%)	6 (60%)	
Rs12550668	AA	4 (80%)	5 (100%)	0.500
	AG	1 (20%)	-	
	A-allele	9 (90%)	10 (100%)	1.000
	G-allele	1 (10%)	0	
Rs4134376	AA	2 (40%)	4 (80%)	0.683
	AG	1 (20%)	-	
	AC	1 (20%)	1 (20%)	
	GG	1 (20%)	-	
	A-allele	6 (60%)	9 (90%)	0.211
	G-allele	3 (30%)	0	
	C-allele	1 (10%)	1 (10%)	

Fisher exact test p-value <0.05 is considered significant. Significant p-values are bolded.

Polymorphic variants for 4 SNPs rs752625, rs1978241, rs12550668 and rs4134376 were outsourced and sequenced by Sanger sequencing. The samples are stratified into the Malay and Non-Malay population. No statistical significance was seen between the different genotypic frequencies between these two groups for the four SNPs of interest as the p-values were above 0.05.

For rs752625, all the non-Malay samples presented with TT genotype while only 40% of the Malay samples had the same genotype. The remaining 60% were either heterozygous (CT genotype at 20%) or homozygous for the C allele (40%). The C allele with an allele frequency of 0.5 is significantly associated with the Malay population samples (p-value 0.033)

Many of the samples in both Malay (60%) and non-Malay population (80%) was heterozygous AG at rs1978241 while the remaining were homozygous for G allele. Most samples (both Malay and non-Malay) are homozygous for A allele at SNP loci rs12550668 whereby all the non-Malay samples and 80% of the Malay samples showed homozygosity AA, with allele frequency for A allele being more than 0.8 for both populations.

Rs4134376 had 3 variant alleles at the locus – A/G/C. Most of the non-Malay samples (80%) were homozygous for the A allele with only 20% being heterozygous AC. In contrast, the Malay population shows more polymorphism at this locus than the non-Malay population, with heterozygous genotypes of AG and AC each being present in 20% the samples. The allele frequency in the non-Malay population is 0.9 for the A allele but decreases to 0.6 in the Malay population.

The following genotypes were not detected in the samples: homozygous A genotype in rs1978241, homozygous G genotypes in SNP rs12550668, and heterozygous genotype GC in SNP rs4134376.In comparing the sequencing results in SNPs rs752625 and rs4134376, the homozygosity in the non-Malay population is much higher than the Malay population, with all being homozygous at rs752625 and 80% being homozygous at rs4134376 as compared to 40% in the Malay population.

The multiplex PCR for the 5 SNPs (rs4599414, rs752625, rs1978241, rs4134376 and rs12550668) performed showed the best resolution of bands at annealing temperatures above 51.4°C, with clear bands on gel electrophoresis (Figure 5). Unfortunately, the SNP rs4599414 was not amplified successfully in either the multiplex or singleplex PCR, hence it was excluded from subsequent analysis.

Discussion

This study had revealed a tri-allelic SNP rs4134376, that may prove to be significant with increase in the number of samples genotyped. Tri-allelic SNPs can perform as ancestry markers as they follow patterns of population divergence and their high heterozygosity maximizes the information that can be obtained at the locus, and there has been attempts to incorporate them into a panel for forensic identification (Kidd et al., 2006; Phillips et al., 2020).

Among the four AIM-SNPs that were genotyped, one SNP rs752625 showed a significant difference in the allele frequency (p-value < 0.05), with the C-allele being much more common in Malays. Otherwise, there is no significant difference in the genotypes of the Malay and non-Malay population. The similarities in the SNP variation between these two populations may stem from the history of admixture in the Malay population in south east Asia, where multiple different populations in Asia have contributed to the genetic ancestry to the Malay population, among which are populations from east Asia and south Asia (Deng et al., 2015; Hatin et al., 2014; Norhalifah et al., 2016).

Maritime trade between the east and the west in ancient times focused on the passage through south east Asia, whether by sea through the Straits of Malacca, or the overland route across the neck of the Malay Peninsula (Gungwu, 1958), and trading routes stretched from China to Africa, involving Indian, Arabian and Chinese merchants (Andaya & Andaya, 2016). There are archaeological evidences of Indian traders coming into contact with populations in the Malay Peninsula by 5th century BCE, and later trade with Chinese merchants followed by 10th century BCE (Andaya & Andaya, 2016). The spread of religion and the expansion of influences from kingdoms based in India and China also served to contribute to the gene flow from these regions into the Malay population. As mentioned earlier, two of the four major contributors to the genetic ancestry of the Malay population were populations of South Asia and East Asia populations of east Asia, south Asia, Austronesian and south east Asia aboriginal people (Deng et al., 2015; Hatin et al., 2011; Norhalifah et al., 2016), and these may be the reason for the similarities. Furthermore, Li et al. (2016) has also demonstrated that some southern Chinese populations are admixed with ancestry from south east Asia. Population admixture, whether due to modern globalization and ease of travel or international trade in ancient times and lack of geographic barriers, has always been a challenging factor in population analyses (Phillips et al., 2014).

The proximity of the populations and history of admixture between the populations of interest may be a factor in the similarity between the genotypes. While many AIM-SNP panels have developed in recent years, many of these focused on intercontinental populations of significant geographic distance (V. Pereira et al., 2017; Phillips et al., 2014; Santos et al., 2016; Zhao et al., 2019). For example, the panel developed by the European Forensic Genetics Network of Excellence (EUROFORGEN-NOE) (http://www.euroforgen.eu/) concentrated on the five different populations that is Asian, Africans, Europeans, Native Americans and Pacific Islanders, all populations that are divided by great distances. Analyses of populations in close geographical proximity to each other has always been difficult, as seen in Phillips et al. (2013) where differentiation of Middle Eastern populations were problematic, as these populations occupy a region historically prominent in between the east and the west.

Due to the limited number of samples and SNPs genotyped in this pilot study, the applicability of these results in discerning ancestry may be diminished. Zhao et al. (2019) has determined that by increasing the number of AIM-SNPs in a panel, the discriminatory power for populations in close geographical proximity can be improved. They developed a 36-AIM-SNP panel that was able to distinguish between five major intercontinental populations with an average accuracy of 99%, but when differentiating between four different European populations, the SNP panel required 175 SNPs to achieve the same accuracy. This higher uncertainty and difficulty in developing a highly discriminatory SNP panel with minimum number of SNPs for intracontinental populations has been noted in other studies as well (Liet al., 2016; V. Pereira et al., 2017; Santos et al., 2016).

The AIM SNPs chosen and analysed for this AIM-SNP panel were obtained from the data in a previously published paper by Yahya et al. (2017) utilizing a merged database compiled from the International HapMap Project Phase 3, Malaysian Node of the Human Variome Project and the Singapore Genome Variation Project. For the Malaysian Node of the Human Variome Project, the Malay subjects were genotyped on two platforms, the Affymetrix SNP-6 SNP array platform and OMNI 2.5 Illumina SNP array platform whereas the Singapore Genome Variation Project was genotyped on the Affymetrix SNP-6 array platform only. The International HapMap Phase 3 used both Illumina Human 1M BeadChip and the Affymetrix SNP-6 array platform. When data from different databases are merged, the number of SNPs will reduce as SNPs available in some populations may not be available in other populations (Yamaguchi-Kabata et al., 2008). It is possible that these excluded SNPs may be informative for Malay ancestry.

The International HapMap Project categorized their Asian population as Japanese in Tokyo, Han Chinese in Beijing, and Chinese in Colorado (Altshuler et al., 2005; Frazer et al., 2007; Pemberton et al., 2010). However, it has been demonstrated that the Chinese population shows a distinct genetic cline along a north-south axis, with significant variations detected between northern populations and southern populations corresponding roughly to geographic locale (Chen et al., 2009; Xu et al., 2009). The ethnic Chinese population of Malaysia is descended mainly from the Chinese population in the south of China who settled in Malaysia mostly during the time of the European colonization (Andaya & Andaya, 2016; Hoh et al., 2015; Teo et al., 2009). Furthermore, a South Asia population form southern India, from whom Malaysians of Indian ethnicity most claim descent from are not represented in the HapMap project. AIMs are limited by their inability to discriminate between unknown populations (Kidd et al., 2014) and bias may occur when inference of ancestry is performed with AIMSNP panels that do not include the population of interest in the initial training data (Zhao et al., 2019).

The underrepresentation of non-European ancestry in genome-wide association studies is well documented, and efforts are ongoing to reduce this bias (Santos et al., 2016). Among these are the 1000 Genomes Project Consortium, the Pan-Asian SNP Consortium and the GenomeAsia 100K Project, all of which aim to improve and expand the knowledge of variation across Asia. The 1000 Genomes Project provide genotype data for Indian Telugu population and Sri Lankan Tamil population from the UK, together with a Southern Han Chinese population, while the GenomeAsia 100K project is more ambitious in genotyping more South Asian populations. These projects genotyped populations that may be of more relevance in the inference of ancestry in the Malay population, as they cover more Asian populations near and far from the Malay Peninsula, and thus may offer more choices for AIM-SNPs of relevance.

To improve on this study and further develop a multiplex capable of determining Malay ancestry, more samples from Malay and non-Malay populations should be genotyped with a higher number of SNPs in a multiplex to determine the discriminatory power of the SNP panel. It has been shown that these 4 target SNPs were able to be amplified in a PCR-multiplex successfully, and these products should then be sequenced to validate the amplicons produced. Upon successful PCR-multiplex PCR amplification, further subjects can then be genotyped. There are many methods for genotyping SNPs, broadly categorized based on the method of genotyping and the assay format (Sobrino, Brión, & Carracedo, 2005). One of the most common method is single base extension, followed by an assay either in solution or on solid support (Fondevila et al., 2017; Geppert & Roewer, 2012; Lundsberg, Johansen, Børsting, Morling, & Consortium, 2013; Wei et al., 2014). A popular assay for this is the SNaPshot® Multiplex Kit (Thermo Fisher Scientific, USA), which encompasses single base extension reaction and minisequencing step for SNP genotyping.

SNP genotyping with single base extension (SBE) reaction allows for high multiplexing capability, robustness, sensitivity, and reproducibility. The protocol is relatively simple, and uses equipment that are widely available in most laboratories (Fondevila et al., 2017; Geppert & Roewer, 2012)

As seen in this study, the genomic region with the SNP of interest is amplified, preferably in a multiplex PCR. After confirming the successful amplification, the PCR products are cleansed with shrimp alkaline phosphatase to remove the primers and unincorporated dNTPs, which can interfere with subsequent reactions. The cleansed PCR product then enters the SBE reaction as a template for SBE primers. SBE primers are designed to anneal to a position immediately adjacent to the SNP of interest, so that fluorescently labelled didesoxynucleotides (ddNTPs) will bind during the PCR and inhibit further elongation. A second clean-up step is needed to remove the unincorporated ddNTPs with shrimp alkaline phosphatase to avoid interference during capillary electrophoresis. The results can then be analysed with Genemapper®.

This method is not without its disadvantages however, and one of its limitation is the number of SNPs that can be multiplexed in one reaction (Li et al., 2016). This can be overcome by designing SNP panels with different SNPs, where a panel is first used to differentiate between continental populations, then a second-tier ancestry panel is utilized to distinguish intracontinental populations (Li et al., 2016). Alternatively, modern massively parallel sequencing and DNA array technology can perform SNP genotyping and provide more information without the limitation in the number of SNPs that can be genotyped simultaneously (Daniel et al., 2015; Li et al., 2016; Xavier & Parson, 2017). Massively parallel sequencing may be better suited to investigate admixture samples with greater power of discrimination, and may discover new SNPs of interest during the sequencing process (Xavier & Parson,2017

Conclusion

AIMs are limited by their inability to discriminate between unknown and bias may occur when inference of ancestry is performed with AIM-SNP panels that do not include the population of interest in the initial training data. When data from different databases are merged, the number of SNPs will reduce as SNPs available in some populations may not be available in other populations. It is possible that these excluded SNPs may be informative for Malay ancestry. Therefore, by choosing only five AIMSNP panel for Malay population is not enough and further studies need to be done to determine the reliability of the AIM-SNP.

Funding

This study was funded by Universiti Sains Malaysia Research University Individual (RUI) Grant: 1001/PPSP/8012257 and 1001/PPSP/8012301

Authors' Contributions

Study design: YTC, PY, AH, NMY, SS, BAZ; Data collection: YTC, PY, SS; Contribution of new reagents or analytical tools: YTC, SS, AH, NMY, BAZ; Data analysis: YTC, SRS; Manuscript preparation: YTC, SRS, SS, BAZ

Competing interests

The authors would like to declare that there was no conflict of interest in this study.

Acknowledgements

The authors thank all the subjects in this study for their participation. the staff of and colleagues in the Human Genome Centre, Universiti Sains Malaysia.

References

1. Altshuler, D., Donnelly, P., & Consortium, I. H. (2005). A haplotype map of the human genome.Nature, 437(7063), nature04226.
2. Andaya, B. W., & Andaya, L. Y. (2016). A history of Malaysia: Macmillan International Higher Education.
3. Chen, J., Zheng, H., Bei, J.-X., Sun, L., Jia, W.-h., Li, T., . . . Zhang, X. (2009). Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. The American Journal of Human Genetics, 85(6), 775-785.
4. Consortium, G. P. (2015). A global reference for human genetic variation. Nature, 526(7571), 68-74.
5. Daniel, R., Santos, C., Phillips, C., Fondevila, M., Van Oorschot, R., Carracedo, A., . . . McNevin, D. (2015). A SNaPshot of next generation sequencing for forensic SNP analysis. Forensic Science International: Genetics, 14, 50-60.
6. Deng, L., Hoh, B.-P., Lu, D., Saw, W.-Y., Ong, R. T.-H., Kasturiratne, A., . . . Wickremasinghe, A. R. (2015). Dissecting the genetic structure and admixture of four geographical Malay populations. Scientific reports, 5, 14375.
7. Deng, L., Hoh, B. P., Lu, D., Fu, R., Phipps, M. E., Li, S., . . . Jin, L. (2014). The population genomic landscape of human genetic structure, admixture history and local adaptation in Peninsular Malaysia. Human genetics, 133(9), 1169-1185.
8. Ding, L., Wiener, H., Abebe, T., Altaye, M., Go, R. C., Kercsmar, C., . . . Baye, T. M. (2011).
9. Comparison of measures of marker informativeness for ancestry and admixture mapping.BMC Genomics, 12, 622. doi:10.1186/1471-2164-12-622
10. Fondevila, M., Børsting, C., Phillips, C., De La Puente, M., Carracedo, A., Morling, N., . . .
11. Consortium, E. (2017). Forensic SNP genotyping with SNaPshot: technical considerations for the development and optimization of multiplexed SNP assays. Forensic Sci Rev, 29(1), 57-76.
12. Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., . . . Leal, S. M.(2007). A second generation human haplotype map of over 3.1 million SNPs. Nature, 449(7164), 851-861.
13. Geppert, M., & Roewer, L. (2012). SNaPshot® minisequencing analysis of multiple ancestryinformative Y-SNPs using capillary electrophoresis DNA Electrophoresis Protocols for Forensic Genetics (pp. 127-140): Springer.
14. Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., . . . Shen, Y. (2003). The international HapMap project.
15. Gungwu, W. (1958). THE NANHAI TRADE: A Study of the Early History of Chinese Trade in the South China Sea. Journal of the Malayan Branch of the Royal Asiatic Society, 31(2 (182)), 1-135.
16. Hatin, W. I., Nur-Shafawati, A. R., Etemad, A., Jin, W., Qin, P., Xu, S., . . . Consortium, H. P.-A. S. (2014). A genome wide pattern of population structure and admixture in peninsular Malaysia Malays. Hugo J, 8(1), 5. doi:10.1186/s11568-014-0005-z
17. Hatin, W. I., Nur-Shafawati, A. R., Zahri, M. K., Xu, S., Jin, L., Tan, S. G., . . . Consortium, H. P.-A.S. (2011). Population genetic structure of peninsular Malaysia Malay sub-ethnic groups. PloS one, 6(4), e18312. doi:10.1371/journal.pone.0018312
18. Hoh, B. P., Deng, L., Julia-Ashazila, M. J., Zuraihan, Z., Nur-Hasnah, M., Nur-Shafawati, A. R., . . .Xu, S. (2015). Fine-scale population structure of Malays in Peninsular Malaysia and Singapore and implications for association studies. Hum Genomics, 9, 16. doi:10.1186/s40246-015-0039-x
19. Jung, J. Y., Kang, P. W., Kim, E., Chacon, D., Beck, D., & McNevin, D. (2019). Ancestry
20. informative markers (AIMs) for Korean and other East Asian and South East Asian
21. populations. Int J Legal Med, 133(6), 1711-1719. doi:10.1007/s00414-019-02129-7
22. Kidd, K. K., Pakstis, A. J., Speed, W. C., Grigorenko, E. L., Kajuna, S. L., Karoma, N. J., . . . Odunsi,A. (2006). Developing a SNP panel for forensic identification of individuals. Forensic science international, 164(1), 20-32.
23.
24. Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., . . . Kidd, J. R. (2014). Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet, 10, 23-32. doi:10.1016/j.fsigen.2014.01.002
25. Li, C.-X., Pakstis, A. J., Jiang, L., Wei, Y.-L., Sun, Q.-F., Wu, H., . . . Kidd, J. R. (2016). A panel of 74 AISNPs: improved ancestry inference within Eastern Asia. Forensic Science
26. International: Genetics, 23, 101-110.
27. Lim, C. H., Zain, S. M., Reynolds, G. P., Zain, M. A., Roffeei, S. N., Zainal, N. Z., . . . Mohamed, Z. (2014). Genetic association of LMAN2L gene in schizophrenia and bipolar disorder and its interaction with ANK3 gene polymorphism. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 54, 157-162.
28. Lundsberg, B., Johansen, P., Børsting, C., Morling, N., & Consortium, E.-N. (2013). Development and optimisation of five multiplex assays with 115 of the AIM SNPs from the EUROFORGEN AIMs set on the Sequenom® MassARRAY® system. Forensic Science International: Genetics Supplement Series, 4(1), e182-e183.
29. Maran, S., Lee, Y. Y., Xu, S., Rajab, N.-S., Hasan, N., Aziz, S. H. S. A., . . . Zilfalil, B. A. (2013). Gastric precancerous lesions are associated with gene variants in Helicobacter pylorisusceptible ethnic Malays. World Journal of Gastroenterology: WJG, 19(23), 3615.

30. Nakanishi, H., Pereira, V., Børsting, C., Yamamoto, T., Tvedebrink, T., Hara, M., . . . Morling, N.(2018). Analysis of mainland Japanese and Okinawan Japanese populations using the
31. precision ID Ancestry Panel. Forensic Science International: Genetics, 33, 106-109.
32. Ngamphiw, C., Assawamakin, A., Xu, S., Shaw, P. J., Yang, J. O., Ghang, H., . . . Consortium, H. P.-A. S. (2011). PanSNPdb: the Pan-Asian SNP genotyping database. PloS one, 6(6), e21451.
33. Norhalifah, H. K., Syaza, F. H., Chambers, G. K., & Edinur, H. A. (2016). The genetic history of Peninsular Malaysia. Gene, 586(1), 129-135. doi:10.1016/j.gene.2016.04.008
34. Pardo-Seco, J., Martinón-Torres, F., & Salas, A. (2014). Evaluating the accuracy of AIM panels at quantifying genome ancestry. BMC Genomics, 15(1), 543. doi:10.1186/1471-2164-15-543
35. Pemberton, T. J., Wang, C., Li, J. Z., & Rosenberg, N. A. (2010). Inference of unexpected genetic relatedness among individuals in HapMap Phase III. Am J Hum Genet, 87(4), 457-464. doi:10.1016/j.ajhg.2010.08.014
36. Pereira, F., Guimaraes, R. M., Lucidi, A. R., Brum, D. G., Paiva, C. L. A., & Alvarenga, R. M. P. (2019). A systematic literature review on the European, African and Amerindian genetic ancestry components on Brazilian health outcomes. Sci Rep, 9(1), 8874. doi:10.1038/s41598-019-45081-7
37. Pereira, V., Mogensen, H. S., Borsting, C., & Morling, N. (2017). Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165
38. ancestral informative markers. Forensic Sci Int Genet, 28, 138-145. doi:10.1016/j.fsigen.2017.02.013
39. Phillips, C. (2015). Forensic genetic analysis of bio-geographical ancestry. Forensic Sci Int Genet, 18, 49-65. doi:10.1016/j.fsigen.2015.05.012
40. Phillips, C., Amigo, J., Tillmar, A., Peck, M., de la Puente, M., Ruiz-Ramírez, J., . . . Parsons, T.(2020). A compilation of tri-allelic SNPs from 1000 Genomes and use of the most
41. polymorphic loci for a large-scale human identification panel. Forensic Science International:Genetics, 46, 102232.
42. Phillips, C., Aradas, A. F., Kriegel, A., Fondevila, M., Bulbul, O., Santos, C., . . . Schneider, P. (2013). Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries. Forensic Science International: Genetics, 7(3), 359-366.
43. Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., . . . Fondevila, M. (2014). Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. Forensic Science International: Genetics, 11, 13-25.
44. Pritchard, J. K., & Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. The American Journal of Human Genetics, 65(1), 220-228.
45. Salleh, M. Z., Teh, L. K., Lee, L. S., Ismet, R. I., Patowary, A., Joshi, K., . . . Hamzah, A. S. (2013).Systematic pharmacogenomics analysis of a Malay whole genome: proof of concept for personalized medicine. PloS one, 8(8), e71554.
46. Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R. A., Burchard, E. G., . . . Via, M. (2016). Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacificregion. Forensic Science International: Genetics, 20, 71-80.
47. Sobrino, B., Brión, M., & Carracedo, A. (2005). SNPs in forensic genetics: a review on SNP typing methodologies. Forensic science international, 154(2-3), 181-194.
48. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nature Reviews Genetics, 20(8), 467-484.
49. Teo, Y. Y., Sim, X., Ong, R. T., Tan, A. K., Chen, J., Tantoso, E., . . . Chia, K. S. (2009). Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res, 19(11), 2154-2162. doi:10.1101/gr.095000.109
50. Wei, Y. L., Qin, C. J., Liu, H. B., Jia, J., Hu, L., & Li, C. X. (2014). Validation of 58 autosomal
51. individual identification SNPs in three Chinese populations. Croat Med J, 55(1), 10-13.
52. doi:10.3325/cmj.2014.55.10
53. Xavier, C., & Parson, W. (2017). Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit–MPS forensic application for the MiSeq FGx™ benchtop sequencer. Forensic Science International: Genetics, 28, 188-194.
54. Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., . . . Pan, X. (2009). Genomic dissection of
55. population substructure of Han Chinese and its implication in association studies. The
56. American Journal of Human Genetics, 85(6), 762-774.
57. Yahya, P., Sulong, S., Harun, A., Wan Isa, H., Ab Rajab, N. S., Wangkumhang, P., . . . Zilfalil, B. A.(2017). Analysis of the genetic structure of the Malay population: Ancestry-informative marker SNPs in the Malay of Peninsular Malaysia. Forensic Sci Int Genet, 30, 152-159.doi:10.1016/j.fsigen.2017.07.005
58. Yamaguchi-Kabata, Y., Nakazono, K., Takahashi, A., Saito, S., Hosono, N., Kubo, M., . . . Kamatani, N. (2008). Japanese population structure, based on SNP genotypes from 7003 individualscompared to other ethnic groups: effects on population-based association studies. TheAmerican Journal of Human Genetics, 83(4), 445-456.
59. Zhao, S., Shi, C.-M., Ma, L., Liu, Q., Liu, Y., Wu, F., . . . Chen, H. (2019). AIM-SNPtag: A
60. computationally efficient approach for developing ancestry-informative SNP panels. Forensic Science International: Genetics, 38, 245-253.