OPTIMIZING MULTIPLEX PCR FOR A SET OF MALAY ANCESTRY INFORMATIVE MARKER SINGLE NUCLEOTIDE POLYMORPHISMS AND PRELIMINARY ANALYSIS OF GENOTYPES BETWEEN MALAY AND NON-MALAY POPULATION

By

CHENG YI-TING

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Pathology (Medical Genetics)

2021

ACKNOWLEDGEMENT

It is my great pleasure to acknowledge all the people who has been involved in this project. I would like to extend my sincere and heartfelt gratitude to all the personages involved in this endeavour, whose help, guidance, and cooperation, without whom I would not have made headway in this project.

I would like to thank Professor Zilfalil bin Alwi, whose great passion, guidance, assistance and encouragement gave this project shape and form. I would also like to express my gratitude to Associate Professor Dr Sarina Sulong, who advised, supported and assisted me throughout this project. I also appreciate the help and advice from my other supervisor Dr Nazihah Mohd Yunus, who helped throughout these years.

I also give thanks to all the staff in the Human Genome Centre who has assisted me in the laboratory work, to PhD candidate Puan Nor Shafawati Ab Rajab, Dr Ahmad Aizat and to Dr Nur Fadhlina Musa for their help in the laboratory procedures and brainstorming during the optimization process.

To my father, who passed away during the period of completing this dissertation, I thank you for giving me life and raising me all these years. May you be watched over by the Lord Buddha in your journey to the next world. To all my family, thank you for your support during these tumultuous years.

I also thank Universiti Sains Malaysia for the grant provided for this study. Any omission in this brief acknowledgement does not mean lack of gratitude.

Table of Contents

ABSTRAK

Pengenalan

Inferens keturunan atau *inference of ancestry* merupakan satu bidang yang amat menarik dan pelbagai kaedah telah digunakan untuk memilih dan mengesahkan panel penanda bermaklumat keturunan (AIM). Salah satu daripada kaedah yang biasa digunakan adalah polimorfisme nukleotid tunggal (SNP). Yahya et al, 2017 telah membentuk satu panel yang mengandungi lebih kurang 200 SNP boleh memebezakan populasi Melayu dengan ketepatan melebihi 80%. Lima SNP telah dipilih daripada panel di atas untuk dijadikan ujian *PCR-multiplex*. Kajian rintis ini memerihalkan proses pengoptimuman reaksi rantai *polymerase* (PCR) dan keputusan genotip antara subjek Melayu dan bukan Melayu.

Metodologi

Satu panel SNP yang bermaklumatkan keturunan telah dipilih daripada pangkalan data *Malaysian Node of the Human Variome Project* dan *Singapore Genome Variation Project* dan dirujuk kepada data daripada fasa ketiga *International HapMap Project*. Lima SNP telah dipilih untuk kajian rintis ini iaitu rs197824, rs752625, rs4599414, rs12550668 dan rs4134376. Subjek-subjek dikenal pasti sebagai Melayu atau bukan Melayu berdasarkan maklumat-maklumat yang diperolehi dari sekurang-kurangnya 3 generasi keluarga mereka. Primer hadapan dan terbalik telah direkabentuk untuk setiap SNP dan rantai reaksi *polymerase* dioptimumankan untuk setiap SNP secara *singleplex*. Seterusnya penjujukan DNA dilakukan untuk menentukan genotip di SNP yang disasarkan.

Keputusan dan Perbincangan

Empat SNP berjaya dioptimumkan dan penjujukan DNA telah dilakukan untuk 10 orang subjek. Genotip di antara populasi Melayu dan bukan Melayu tidak menunjukan perbezaan yang ketara.

Tetapi frekuensi alel untuk SNP rs752625 menunjukkan perbezaan yang ketara manakala SNP rs4134376 menunjukkan 3 alel pada lokus tersebut. Persamaan yang wujud pada alel ini mungkin disebabkan oleh sejarah pencampuran di antara populasi di Semenanjung Melayu dengan populasi yang lain, jarak geografi yang dekat, dan ketiadaan populasi rujukan yang relevan. Untuk pengesahan lanjut tentang ketepatan (*accuracy*) panel SNP ini, lebih banyak SNP dan subjek perlu dikaji dan proses genotip boleh dilaksanakan dengan reaksi *single base extension* untuk mengkaji pelbagai SNP secara serentak.

ABSTRACT

Introduction

Inference of ancestry is of great interest to many and various methods have been developed in selecting and validating panels of ancestry informative markers (AIMs). One of the ancestry informative marker that is commonly used is single nucleotide polymorphisms (SNPs). Yahya et al, 2017 had determined that a panel of approximately 200 SNPs can distinguish the Malay population with an accuracy of more than 80%. Five SNPs were chosen from the above panel of SNPs to be developed into a PCR-multiplex assay. This pilot study describes the PCR optimization process and the genotyping results of the Malay and non-Malay subjects.

Methodology

Ancestry informative marker SNP panels were selected from the genotyping databases of the Malaysian Node of the Human Variome Project and Singapore Genome Variation Project and referenced against the International HapMap Project Phase 3. Five SNPs were chosen for the pilot study (rs197824, rs752625, rs4599414, rs12550668 and rs4134376). The subjects were participants who identified themselves as Malay and non-Malay for at least 3 generations. Forward and reverse primers were designed for each SNP and the polymerase chain reaction (PCR) for each SNP were optimized in singleplex. The PCR products were then sequenced to determine the allele at the target SNP.

Results and Discussion

Four SNPs were successfully optimized and later genotyped for 10 subjects. The difference in the genotypes of the Malay and non-Malay populations were found to be statistically insignificant, however, there is a significant difference in the allele frequency for rs752625. One SNP rs4134376 was tri-allelic. The similarities may arise from several factors, including

the history of admixture in these populations that have occurred, the close geographical distance, and the absence of appropriate reference population. To validate the panel further, more SNP and more subjects should be involved, and genotyping could be done with single base extension reaction in multiplex to assay multiple SNPs simultaneously.

CHAPTER 1: 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 have 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 AISNP 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 population estimate of the Department of Statistics Malaysia in 2020, the population of Malaysia comprises Bumiputera 69.6%, Chinese 22.6% and Indians 6.8%.

In Malaysia, though Malays have a formal political definition (Article 160(2), Constitution of Malaysia), 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 descendants of 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 aborigines (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.

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

CHAPTER 2: OBJECTIVES OF THE STUDY

General Objective

To optimize multiplex PCR for relevant SNPs chosen from Malay ancestry informative marker SNPs.

Specific Objective

1. To select Malay ancestry informative marker SNP from the ancestry informative marker SNP panel reported by Yahya et al, 2017.
2. To design primers for the target SNPs and optimize the polymerase chain reaction conditions for each SNP
3. To validate the successful amplification of the target SNPs under optimized conditions via sequencing
4. To perform preliminary analysis of the genotypes of Malay and non-Malay subjects at the target SNP loci.
5. To combine the 5 singleplex reactions to develop a multiplex PCR reaction.

CHAPTER 3: MANUSCRIPT

TITLE

Malay Ancestry Informative Marker Single Nucleotide Polymorphism Panel: The First Step to Developing an Ancestry-Informative Marker SNP array for the Malaysian Malay Population

List of Authors:

YI-TING CHENG[1], SHARIFAH-NANY RAHAYU-KARMILLA SYED-HASSAN[1], AZIAN HARUN[2], NAZIHAH MOHD YUNUS[1], SARINA SULONG[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.

Corresponding Author

PROFESSOR DR ZILFALIL BIN ALWI

Human Genome Centre,

School of Medical Sciences, Universiti Sains Malaysia,

16150 Kota Bharu, Kelantan,

Malaysia.

Email: zilfalil@usm.my

SUMMARY

Inference of genetic ancestry has been of great interest in many fields and one of the markers in these analyses is ancestry informative marker single nucleotide polymorphisms (AIM SNPs). The Malay population is an ethnic group located mainly in South East Asia and comprises the largest ethnicity in Malaysia. To determine Malay ancestry, Yahya et al, 2017 selected 37,487 SNPs from the genotyping data collected by the Malaysian Node of the Human Variome Project and Singapore Genome Variation Project and referenced them against the data from the International HapMap Project Phase 3. The SNPs determined to be informative for ancestry were compiled into AIM SNP panels, and from these a few SNPs were selected for optimization in preparation for single base extension reaction multiplexing. This pilot study gives a brief account of the optimization process and the genotyping data for subjects of Malay ancestry and compared the results against those of non-Malay ancestry. 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 panel for Malay population.

Keywords:

Ancestry, Malay, single nucleotide polymorphisms

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 et al., 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 have 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 et al., 2005; Frazer et al., 2007; Gibbs et al., 2003; Pemberton et al., 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 marker 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 et al., 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 of Malaysia), this definition does not take into full account their Austronesian origins, the different ethnic sub-groups and their genetic ancestry (Norhalifah et al., 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 (Lim et al., 2014).

This pilot study aims to optimize a multiplex PCR assay to demonstrate and validate some of the chosen AIM SNPs from Yahya et al. (2017), by genotyping ten participants of self-reported

Malay ancestry and non-Malay ancestry, to detect any significant differences between the genotyping results of the two populations.

MATERIALS AND METHODS

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 I_n, and PCAIM, and pairwise F_{ST}. 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 90%, though the number of SNPs required to achieve that accuracy differ with method. While a set of 157 SNPs with the highest F_{ST} has an accuracy of more than 80%, the number of SNPs to achieve the same accuracy using I_n and PCAIMs methods respectively were 200. Among these SNPs, one SNP was shared.

To validate the AIM SNP panels selected, 5 SNPs were chosen, with three SNPs chosen at random while one SNP was shared between I_n and PCAIM results (rs12550668), and another shared between PCAIM and F_{ST} results (rs4134376). There were no shared SNP between I_n and F_{ST} 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 Center for Biotechnology Information (NCBI), thus eliminating the possibility of selective pressures on the frequency of the allele.

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	I_n	Chr4:7461577	Intron variant in SORCS2
Rs752625	PCAIM	Chr4:169050766	None
Rs1978241	F_{ST}	Chr17:61510491	None
Rs12550668	I_n and PCAIM	Chr8:11722445	Intron variant in GATA4
Rs4134376	PCAIM and F_{ST}	Chr15:85885026	Intron variant in LOC105370953

Study subjects

This is a cross-sectional study where the ten participants were of Malay and Non-Malay descent. 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. All participants are healthy adults more than 18 years of age, with no congenital malformations, major birth defects or chronic diseases, and be of either Malay or non-Malay descent for at least 3 generations. 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. Those with admixed ancestry were excluded. 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).

Molecular Analysis

Forward and reverse primers were designed for each SNP. 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. A primer set for a specific target SNP will produce amplicons of a certain size, while another primer set for another target SNP will produce amplicons of a different size. Thus, each primer set produced amplicons of a specific size that corresponds to a target SNP. The sizes of these amplicons range from 178 to 557 bp, which can be differentiated 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 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