ANCESTRY INFORMATIVE MARKERS SINGLE NUCLEOTIDE POLYMORPHISMS PANEL FOR ANCESTRY ESTIMATION IN THE MALAY POPULATION

PADILLAH BINTI YAHYA

UNIVERSITI SAINS MALAYSIA

2021

ANCESTRY INFORMATIVE MARKERS SINGLE NUCLEOTIDE POLYMORPHISMS PANEL FOR ANCESTRY ESTIMATION IN THE MALAY POPULATION

by

PADILLAH BINTI YAHYA

Thesis submitted in fulfilment of the requirements

for the Degree of

Doctor of Philosophy

August 2021

ACKNOWLEDGEMENTS

In the name of Allah S.W.T, the Most Gracious and the Most Merciful. All praises to Allah, for giving me strength, patience, perseverance and optimism throughout my postgraduate journey. My biggest gratitude to Him, with His permission this thesis has finally been completed.

My deepest appreciation to my main supervisor, Prof. Dr. Zilfalil Alwi, and my co-supervisors, Associate Prof. Dr. Sarina Sulong and Associate Prof. Dr. Azian Harun for their guidance, support and kind supervision throughout my study. My special thanks to Dr. Sissades Tongsima and his team members, Pongsakorn Wangkumhang, Alisa Wilantho, and Chumpol Ngamphiw at the National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand Science Park, Pathum Thani Thailand, for their time, guidance, lessons and technical supports in completion this study. Special thanks also to W. Nur Hatin W. Isa, Nur Shafawati Ab Rajab and Nurfazreen Mohd Nasir who have provided me all the SNPs genotype data used in this study and help me a lot in data analyses. Thanks to my great and helpful friends, Nurul Fatihah Azman and Diana Rashid and all students and staff of MyHVP, USM for helping me throughout my study. I am deeply thankful to my families and my best friend, Juriah Mohamed, for standing by my side when times get hard. I couldn't have achieved this without your support. Finally, I would like to thanks Public Service Department Malaysia (JPA) for the HLP scholarship, USM for providing all the facilities to carry out my study and Universiti Sains Malaysia Apex Grant: 1002/PPSP/910343 and NTU Grant (Muhammed Ariff Research Grant (MAS): 304.PPSP.6150148.N119 for supporting this study.

TABLE OF CONTENTS

CHAPTER 3-MATERIALS AND METHODS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS AND ABBREVIATIONS

%	: Percentage
+	: Plus
±	: Plus minus
®	: Registered Sign
Σ	: Sum of
≤	: Less than and equal to
≥	: More than and equal to
>	: More than
<	: Less than
~	: Approximately
β	: Beta
α	: Alfa
δ	: Absolute allele frequency differences
λ^2	: Chi-squared test
r^2	: Regression value
K	: Number of cluster
P	: Allele frequencies
Q	: ancestral component or membership coefficient
$Pr(C_k/d)$	: Posterior probabilities of data
A	: Adenine
aAIMs	: Ancient ancestry informative markers
ABI	: Applied Biosystems
AIM	: Ancestry informative marker

ANOVA : Analysis of variance

ASW : Africa America

AUC : Area under curve

bp : Basepair

C : Cytosine

CEU : European

CHB : Han Chinese Beijing

CHD : Chinese Colorado

CN-JN : Chinese Jinuo

CN-WA : Chinese Wa

dbSNP : Database Single Nucleotide Polymorphism

ddNTPs : Dideoxynucleotides triphosphates

DNA : Deoxyribonucleic Acid

EM : Expectation maximization

F_{ST} : Allele frequency differences between populations

FN : False negative

FP : False positive

FRET : Fluorescence resonance energy transfer

G : Guanine

GIH : Gujarati Indian

GWAS : Genome-wide association studies

HGDP-CEPH: Human Genome Diversity Cell Line Panel

HLA : Human Leukocyte Antigen

HMM : Hidden markov model

HNB : Hidden Naïve Bayes

H.Pylori : Helicobacter pylori

HVR : Hyper variable region

HWE : Hardy Weinberg Equilibrium

IBS : Identical-by state

IBD : Identical-by descent

I_n : Informativeness for assignment

ID-JV : Indonesian Java

ID-ML : Indonesian Malay

ID-TR : Indonesian Toraja

IN-DR : Indian Telugu

IN-WL : Indian Marathi

INDEL : Insertion deletion

ipPCA : Iterative pruning principal component analysis

JPT : Japanese

KNN : K-nearest neighbor

LAMP : Local ancestry in admixed population

LD : Linkage disequilibrium

LSBL : Locus specific branch length

LWK : Luhya

MAF : Minor allele frequency

MAL-CH : Chinese Malaysia

MAL-IN : Indian Malaysia

MAL-MY : Melayu Malaysia

MC1R : Melanocortin 1 reseptor

MCMC : Markov Chain Monte Carlo

MDS : Multidimensional scale

MEX : Mexican

MHC : Major histocompatibility complex

MHMM : Markov-hidden markov model

MKK : Maasai

mtDNA : Mitochondrial DNA

MY-BG : Melayu Bugis

MY-BJ : Melayu Banjar

MY-CH : Melayu Champa

MY-JH : Jahai

MY-JV : Melayu Java

MY-KD : Melayu Kedah

MY-KN : Melayu Kelantan

MY-KS : Kensui

MY-MN : Melayu Minang

MY-PT : Melayu Patani

MY-TM : Temuan

MyHVP : Malaysian node of the human variome project

NB : Naïve Bayes

NCBI : National Center for Biotechnology Information

PCA : Principal Component Analysis

PCAIMs : PCA-correlated ancestry informative markers

RFLP : Restriction fragment length polymorphisms

RMSE : Root mean squared error

ROC : Receiver operating characteristic

SBE : Single-base extension

SGVP : Singapore Genome Variation Project

sgvp-chs : Singapore Chinese

sgvp-ins : Singapore Indian

sgvp-mas : Singapore Melayu

SINEs : Short interspersed nuclear elements

SP : Sub-population

STRs : Short Tandem Repeats

SNPs : Single Nucleotide Polymorphisms

SVD : Singular value decomposition

T : Thymine

TB : Tuberculosis

TP : True positive rate

TSI : Toscani

WEKA : Waikato Environment for Knowledge Analysis

YRI : Yoruba

LIST OF APPENDICES

PENANDA MAKLUMAT KETURUNAN POLIMORFISME NUKLEOTIDA TUNGGAL UNTUK ANGGARAN KETURUNAN DALAM POPULASI MELAYU

ABSTRAK

Penanda maklumat keturunan (AIM) dapat digunakan untuk menyimpulkan (inferens) keturunan seseorang individu bagi meminimumkan ketidaktepatan maklumat keturunan yang dilaporkan sendiri yang digunakan pada masa ini dalam penyelidikan bioperubatan. Dalam kajian ini, tiga kaedah digunakan dalam membina panel SNP AIM Melayu, iaitu analisis *iterative pruning principal component* yang digabungkan dengan *pairwise* F_{ST} (*ipPCA-*F_{ST}), *informativeness for assignment* (I_n) dan *PCA-correlated SNPs* (*PCAIMs*). Dua set data genotip SNP Melayu yang diperolehi daripada Projek Variome Manusia Malaysia (MyHVP) telah digunakan untuk mengekstrak panel SNP AIM Melayu iaitu 135 genotip SNP Melayu yang dihasilkan melalui platform Affymetrix GeneChip Mapping Xba 50K array dan 76 genotip SNP Melayu yang dihasilkan melalui platform Affymetrix SNP-6 array dan SNP OMNI2.5 Illumina array. Tambahan 89 lagi genotip SNP Melayu yang dihasilkan melalui platform Affymetrix SNP-6 array diperolehi dari Projek Variasi Genom Singapura (SGVP). Pangkalan data SNP Pan-Asian digunakan sebagai populasi rujukan untuk pangkalan data genotip SNP Melayu pertama manakala pangkalan data HapMap Fasa 3 dan pangkalan data SGVP digunakan sebagai populasi rujukan untuk pangkalan data genotip SNP Melayu kedua. Ketepatan setiap panel SNP AIM Melayu yang dihasilkan dinilai dengan menggunakan *machine learning* "*ancestry-predictive model*" yang dibina dengan menggunakan WEKA, platform *machine learning* komprehensif yang ditulis dalam Java. Seterusnya, corak genetik bangsa Melayu

dianalisis mengunakan program ADMIXTURE berdasarkan kepada panel SNP AIM tersebut. Hasil analisis menunjukkan bahawa model SNP AIM 144, 299 dan 433 yang dipilih dari data Affymetrix 50K SNP menggunakan kaedah *ipPCA-*F_{ST}, mengklasifikasikan individu Melayu masing-masing, dengan ketepatan 76.5%, 70.6% dan 82.4%. Ketepatan meningkat kepada 88.2% menggunakan model SNP AIM 1772 dan 3145. Model SNP AIM 250 dan 2000 yang dipilih menggunakan kaedah I_n, berjaya mengklasifikasikan individu Melayu dengan ketepatan masing-masing, 89.8% dan 96.1%. Walaubagaimanapun, ketepatan sedikit lebih rendah, masing-masing 78% dan 92.1% untuk bilangan SNP yang sama yang dipilih dengan kaedah *PCAIM*. Analisis ADMIXTURE menunjukkan corak genetik populasi Melayu dapat dibezakan dengan jelas dari populasi rujukan menggunakan panel SNP AIM 1772 dan 3145 yang dipilih oleh *ipPCA-*F_{ST} dan 2000 SNP yang dipilih oleh I_n dan *PCAIMs*. Panel SNP AIM 101, 157 dan 294 yang dipilih dari data Affymetrix SNP-6 menggunakan kaedah *ipPCA-*F_{ST}, menunjukkan ketepatan klasifikasi masing-masing 88.8%, 94.4% dan 96.9%. Ketepatan meningkat kepada 100%. menggunakan panel SNP AIM 1250 dan 2240. Model 100, 200 dan 2000 SNPs dipilih menggunakan I_n, menunjukkan ketepatan klasifikasi masing-masing 67.5%, 80% dan 100%. *PCAIM* menunjukkan ketepatan 68.8%, 81.9% dan 99.4%, masing-masing untuk bilangan panel SNP AIM yang sama. Corak genetik populasi Melayu dapat dibezakan dengan populasi dunia lain yang digunakan dalam kajian ini menggunakan panel SNP AIM 1250 dan 2240 yang dipilih menggunakan kaedah *ipPCA-*F_{ST} dan 2000 AIM SNP yang dipilih menggunakan kaedah I_n dan *PCAIM*.

ANCESTRY INFORMATIVE MARKERS SINGLE NUCLEOTIDE POLYMORPHISMS PANEL FOR ANCESTRY ESTIMATION IN THE MALAY POPULATION

ABSTRACT

Ancestry-informative markers (AIMs) can be used to infer an individual's ancestry to minimize the inaccuracy of self-reported ethnicity in biomedical research. The AIM-SNP panels for the Malay population were developed using three methods in this study: iterative pruning principal component analysis (ipPCA) combined with pairwise F_{ST} (ipPCA-F_{ST}), informativeness for assignment (I_n), and PCA-correlated SNPs (or PCA-informative markers; PCAIMs). The Malay AIM-SNP panels were designed using two sets of Malay SNP genotype datasets stored in SNP arrays hosted by the Malaysian Node of the Human Variome Project (MyHVP). The first dataset contained 135 Malay SNPs genotypes generated from Affymetrix GeneChip Mapping Xba 50K array platform. The second dataset contained 76 Malay SNPs genotypes generated from Affymetrix SNP-6 array and OMNI2.5 Illumina SNP array platforms. In addition, 89 Malay SNP genotypes from the Singapore Genome Variation Project (SGVP) using the Affymetrix SNP-6 array platform were also included in the second set of the Malay SNP genotype datasets. The Pan-Asian SNP dataset was used as a reference population for the first Malay SNP genotype dataset, whereas the International HapMap Phase 3 project and SGVP datasets served as the reference populations for the second Malay SNP genotype datasets. The accuracy of each resulting Malay AIM-SNP panel was evaluated using machine learning "ancestry-predictive model" constructed using WEKA, a comprehensive machine learning

platform written in Java. The ADMIXTURE program was used to explore the genetic pattern of Malays based on the selected AIM-SNP panels. The results showed that models with 144, 299, and 433 AIM-SNP panels selected from the Affymetrix 50K SNPs dataset using the ipPCA-F_{ST} method, correctly classified Malay individuals with an accuracy of 76.5, 70.6, and 82.4%, respectively. The accuracy further increased to 88.2% when using models with 1772 and 3145 AIM-SNP panels. Models with 250 and 2000 SNPs ranked by I_n, correctly classified Malay individuals with an accuracy of 89.8 and 96.1%, respectively. However, the accuracy was slightly lower, 78 and 92.1%, respectively, for the same number of SNPs selected by the PCAIM method. ADMIXTURE analysis showed that the genetic structure of the Malay population can be distinctly differentiated from the reference populations using 1772 and 3145 AIM-SNP panels selected by ipPCA-F_{ST}, and 2000 SNPs selected by I_n and PCAIM. Models with 101, 157, and 294 AIM-SNP panels selected from the Affymetrix SNP-6 dataset using the ipPCA-F_{ST} method demonstrated classification accuracy of 88.8, 94.4, and 96.9%, respectively. Remarkable results were obtained using 1250 and 2240 AIM-SNP panels, where the accuracy increased to 100%. Models with 100, 200, and 2000 ranked by I_n, correctly classified Malay individuals with an accuracy of 67.5, 80, and 100%, respectively. For the same number of AIM-SNP panels, the PCAIM showed an accuracy of 68.8, 81.9, and 99.4%, respectively. The genetic structure of the Malay population can be differentiated from the other world populations used in this study using the 1250 and 2240 AIM-SNP panels selected by ipPCA-F_{ST} and 2000 AIM-SNP panels selected by I_n and PCAIM.

CHAPTER 1

INTRODUCTION

1.1 Research background

More than 100,000 years ago, modern humans were thought to have migrated from Africa to other parts of the world (Cavalli-Sforza, 2007). They formed local communities influenced by their surrounding environment and tended to mate in proximity, contributing to genetic drift and natural selection (Cavalli-Sforza, 2007). Mutations have also arisen, resulting in biological differences, though they retain some of their ancestor's genomic information. The migration, genetic drift, mutation and natural selection operated in parallel with demographic and historical events resulted in variants that are rare in some population but not in others which are likely arisen recently and contributed to the population differences (Cavalli-Sforza, 2007, Paschou et al., 2010). The differences of the genetic patterns were portrayed in their genetic ancestry and population structure carried in the genome of each individual (Cavalli-Sforza, 2007, Paschou et al., 2010).

Microsatellite markers were used to examine patterns of human genetic variation and population genetic structure across the entire genome (Paschou et al., 2010). Studies of population genetic structure based on Single Nucleotide Polymorphisms (SNPs) genome-wide data have successfully revealed the clines of genetic diversity around the world, especially with the advent of modern technologies and the realization of the HapMap project (Paschou et al., 2010). More recent studies of genetic ancestry to infer individual membership down to a population within a continent has attracted

considerable attention because of their value in biomedical, population genetics, anthropological, and forensic applications (Bryc et al., 2015, Byun et al., 2017, Das and Upadhyai, 2018, Vongpaisarnsin et al., 2017, Zeng et al., 2016).

To reveal the genetic variation within and among populations, the genetic distance between them can be measured by calculating the frequencies of the variant allele. Wright's F-statistics (F_{ST}) is commonly used to measure genetic differentiation. Small F_{ST} revealed similar allele frequencies, whereas large F_{ST} indicated that allele frequencies within each population differed (Holsinger and Weir, 2009). This variant/allele can be chosen as a candidate marker to infer ancestry for an individual of that population, and it is called ancestry-informative markers (AIMs).

AIMs are DNA markers that show high allele frequency differences between populations from different geographic regions; thus, this marker could infer an individual's biogeographic ancestry (Kayser and de Knijff, 2011). AIMs can be found on any DNA polymorphisms such as short tandem repeats (STRs), Alu elements, insertion-deletion polymorphisms (INDELs) or single nucleotide polymorphisms (SNPs) (Algee-Hewitt et al., 2016, Esposito et al., 2018, Gómez-Pérez et al., 2010, Inácio et al., 2016). Both haploid and diploid genetic markers can be used to study genetic ancestry or biogeographic ancestry. Mitochondrial DNA (mtDNA) and Y-chromosome polymorphisms are the two haploid markers that have been frequently used to study biogeography ancestry (Chaitanya et al., 2014, Dulik et al., 2012, Salas et al., 2006). However, mtDNA (maternal lineage) and Y-chromosome (paternal lineage) markers do not provide comprehensive information on individual ancestry due to their haplotype nature. Comparatively, autosomal markers can provide more

information about an individual's genetic ancestry because they represent a much greater proportion of genome history (both maternal and paternal ancestry) (both maternal and paternal ancestry) (Royal et al., 2010).

Autosomal markers commonly studied for genetic ancestry were STRs (Kutanan et al., 2014, Nunez et al., 2010, Phillips et al., 2013a), Alu elements (Gómez-Pérez et al., 2010, Hormozdiari et al., 2011, Krishnaveni and Prabhakaran, 2015), INDELs (Moriot et al., 2018, Pereira et al., 2012, Tao et al., 2019, Zaumsegel et al., 2013) and SNPs (Fondevila et al., 2013, Hwa et al., 2017, Kidd et al., 2014, Poetsch et al., 2013). Nonetheless, SNPs was the marker of choice for the ancestry studies due to its stability, abundance in human genome, demonstrated pronounced frequency variation among populations and thousands of SNPs can be assayed simultaneously using high throughput platform such as microarray chips (Royal et al., 2010). Furthermore, autosomal SNPs are almost entirely used to estimate genetic ancestry in epidemiological applications.

Single base pair substitution or SNPs play an important role to the variation among individuals including susceptibility to disease and reactions to drug (Kruglyak and Nickerson, 2001). SNPs play an important role in an individual susceptibility to most diseases and drugs metabolism and it is also said to be directly involved in determining the phenotype of an individual such as eye color, hair and skin; facial morphology and height (Butler, 2012). Millions of SNPs have been identified and made available at dbSNP homepage at NCBI and HapMap, however a small subset of SNPs (10-100s) should be enough to accurately infer an individual ancestry (Fondevila et al., 2013, Gettings et al., 2014, Sampson et al., 2011). AIM-SNPs are a small set of informative

SNPs. The advantages of AIM-SNPs over a random set of autosomal SNP markers include its ability to offer increased power for ancestry inference and to define admixed population (to determine the relative percentages of descendants). Simultaneously, a smaller set of markers can reduce genotyping costs while increasing throughput (Royal et al., 2010).

Several approaches have been used by researchers to select SNPs for ancestry studies. Rosenberg et al., (2003) suggested that SNPs can be ranked individually based on their ability to distinguish ancestry by calculating the F_{ST} value, allele frequency and the informativeness for assignment (I_n). I_n is the measure of information provided by multiallelic markers about individual ancestry (Rosenberg et al., 2003). Paschou et al., (2010) chose SNPs that are strong contributors to the principal component analysis (PCA) or PCA-correlated SNPs (PCAIMs). PCA is a multivariate analysis that provides a new coordinate system (Kayser and de Knijff, 2011). Kersbergen et al., (2009) used a model-based clustering approach STRUCTURE program to estimate genetic diversity among multiple groups of individuals and then used the pairwise F_{ST} ranking procedure to identify AIM-SNPs.

Iterative pruning PCA (ipPCA) is another algorithm suggested for searching AIM-SNPs. This algorithm assigned individuals to sub-populations and calculated the total number of sub-populations present, and the STRUCTURE program was then used to select the appropriate AIM-SNPs (Intarapanich et al., 2009). Galanter et al., (2012) used Locus Specific Branch Length (LSBL) approach to discover AIM-SNPs. Based on F_{ST} values, LSBL is a measure of population structure in one population sample relative to two other population samples. The AIM-SNPs were selected based on the

highest LSBL for that population. The selected AIM-SNPs were tested for linkage disequilibrium (LD), physical distance, and heterogeneity. The AIMs were excluded if the markers were in LD or the alleles showed significant allele frequency heterogeneity between the samples representing each ancestral group (Galanter et al., 2012).

ADMIXTURE program is another approach to observe the genetic structure of studied groups/clusters while simultaneously selecting AIM-SNPs (Vongpaisarnsin et al., 2015). Other approaches include selecting SNPs from databases that exhibit marked allele frequency differences between populations (high F_{ST} value) and strongest contributors to the PCA (Harrison et al., 2008, Phillips et al., 2013b, Sampson et al., 2011); exploring existing SNP panels with hundreds of genetic markers using commercially available SNP genotyping arrays (Kidd et al., 2014) and selecting SNPs with genes involved in melanin syntheses such as MC1R, OCA2, ASIP, or SLC45A2 (Gettings et al., 2014, Poetsch et al., 2013, Soejima and Koda, 2007).

Individual AIM-SNPs can then be genotyped using common SNP genotyping platforms such as allelic-specific hybridization, primer extension, oligonucleotide ligation, or invasive cleavage (Sobrino et al., 2005). Further, allelic products of these methods can be detected with several detection systems such as florescence-electrophoresis (Bouakaze et al., 2009, Fondevila et al., 2013, Mosquera-Miguel et al., 2009, Poetsch et al., 2013), fluorescence resonance energy transfer (FRET) (Lareu et al., 2001, Nicklas and Buel, 2008), fluorescence arrays (Divne and Allen, 2005, Zeng et al., 2012), and mass spectrometry (Li et al., 1999, Shi et al., 2011). Primer extension combined with the florescence-electrophoresis allelic detection method, also known

as mini-sequencing technology, is one of the most suitable methods for analyzing a small number of AIM-SNPs. This is because most laboratories have an automatic capillary electrophoresis instrument, which is also used for STR genotyping. To facilitate the detection of the SNP allelic products, the commercialized multiplex single-base extension reaction SNaPshot® kit (Applied Biosystems, USA), which utilizes fluorescent ddNTPs, is also available on the market (Daniel et al., 2009, Phillips et al., 2013a, Rogalla et al., 2015, Wei et al., 2016).

High throughput technology such as microarray is more powerful for analyzing hundreds or thousands of AIM-SNPs simultaneously. Keating et al., (2013) developed the Identitas v1 Forensic Chip, a diagnostic tool comprising of 201,173 genome-wide autosomal, X-chromosomal, Y-chromosomal, and mitochondrial SNPs for simultaneously inferring biogeographic ancestry, appearance, relatedness, and gender. The chip, which was manufactured by Illumina, uses the well-established Infinium technology (Keating et al., 2013). Galanter et al., (2012) genotyped 446 AIM-SNPs using both Affymetrix and Illumina platforms, and Krjutskov et al., (2009) used a microarray platform to analyze a 124-plex SNP comprising of 49 mtDNA SNPs, 29 Y-chromosomal SNPs, and 46 autosomal SNPs (Krjutskov et al., 2009).

According to previous studies, certain diseases are more prevalent in one ethnic group than others, such as hypertension, end-stage renal disease, tuberculosis, lung function, and prostate cancer (Daya et al., 2014, Menezes et al., 2015, Royal et al., 2010). Cappetta et al., (2015) studied the effect of genetic ancestry on leukocyte global DNA methylation in cancer patients and suggested that genetic ancestry should be considered as a modifying factor in epigenetic association studies, especially in

admixed populations. Thus, understanding ethnicity/ancestry and the substructure is vital for properly designing case-control association studies and identifying disease predisposing alleles that may differ across ethnic groups (Cappetta et al., 2015, Cavalli-Sforza, 2007, Royal et al., 2010, Tishkoff and Kidd, 2004). The current practice of using the self-identified ethnicity/ancestry approach in association diseases or medical genetics studies may result in false-positive or false-negative results, especially in studies of admixed populations, because this approach cannot account for the percentage of admixture in admixed cases (Liu et al., 2013b). Furthermore, understanding one's ancestry background can help in the proper diagnosis and subsequent treatment of diseases.

1.2 Problem statement

Self-reported ethnicity has limitations when conducting genetic studies (Al-Alem et al., 2014, Al-Naamani et al., 2017, Mersha and Abebe, 2015). A study participant's ethnicity is frequently misreported. This contributes to the spurious association with false-positive or false-negative results. Self-reported ethnicity errors may occur when subjects are unaware of their true ethnicity, only know their recent ancestors, or identify with one ethnic group despite their admixed background. Therefore, the accuracy of biomedical studies will be affected, and the study will fail to provide novel insights into variation in disease susceptibility and adverse drug reactions in the studied population or individual. As a result, analyzing the human genome can provide information about an individual's ancestry, especially AIM-SNPs, which can be used to describe actual genetic variation in studied individuals or populations.

Many studies on AIM-SNPs have been carried out in various populations worldwide, including the Brazilian population (Lins et al., 2010), the United States population (African American, East Asian, European American, and Hispanic American/Native American) (Getting et al., 2014), European-descendent populations (Huckins et al., 2014), the Han Chinese population (Qin et al., 2014), the Thai population (Vongpaisarnsin et al., 2015), Australia and Pacific region populations (Santos et al., 2016) and Singapore population (Ramani et al., 2017). To date, however, no study has been carried out to identify Malay AIM-SNPs for the Malay population. The designing of Malay AIM-SNP panels can be used to ascertain Malay ancestry and reveal the extent of admixture in Malay individual subjects. This information may improve the accuracy and precision of biomedical and pharmacological studies carried out in the Malay population.

1.3 Research justifications

Genetic epidemiology shows that many Mendelian diseases are concentrated in a few, usually small social or ethnic groups, especially for the rarer diseases (Cavalli-Sforza, 2007). Ethnicity/ancestry of individuals has also been proven to affect their response to some specific therapeutic agents or drugs (Al-Naamani et al., 2017). The study of biogeographical ancestry using AIM-SNPs is an important tool in the human medical research for gaining a better understanding of the ancestry-associated variations in human diseases without denying the contribution of other external factors such as the socio-economic, dietary and environmental background. Moreover, as the world becomes multi-ethnic, mixed marriage is becoming more common in some populations, making it difficult to assign a single ethnicity to an individual.

Consequently, understanding one's ancestry is a vital step in ensuring the proper designing of biomedical studies.

The development of AIM-SNP panels will help to address this problem, and the genetic admixture of an admixed individual will be fully discovered. This genomic knowledge will also be extremely useful in rapidly extending our knowledge of Mendelian diseases and individual responses to drugs that may be available for a specific disease. The accuracy of disease association studies can be enhanced, and false-positive and false-negative associations can be avoided. Individual genetic ancestry estimation can also bring us closer to more personalized/individualized genetic-based medicine.

In Malaysia, the Malays are the main ethnic group residing mostly in Peninsular Malaysia and some in Sabah and Sarawak. Malays are said to be an admixed population and demonstrated heterogeneous genetic structure (Hatin et al., 2011, Hatin et al., 2014, Deng et al., 2014, Deng et al., 2015, Hoh et al., 2015). The availability of Malay SNPs database hosted by the Malaysian Node of the Human Variome Project (MyHVP) and the Singapore Genome Variation Project (SGVP) enable the selection of AIM SNPs for the Malay population.

The Malay AIM-SNP panels can be used by the medical community to examine their subjects (who self-identified themselves as Malay) before recruiting them in their disease association studies, particularly for research involving ancestry/race as a vital contributor to the severity of the diseases. Pharmacology researchers may use these Malay AIM-SNPs to confirm their Malay ancestry subjects to gain an accurate

response to their drug research testing. These Malay AIM-SNPs are also useful in forensic investigations. The unknown samples found at the crime scene could be subjected to ancestry screening tests using these AIM-SNP panels, facilitating police investigations.

The MyHVP Malay SNP databases will be used in this study to select suitable SNPs that will be used to develop the Malay AIM-SNP panels using three approaches: pairwise F_{ST} (combined with ipPCA), I_n and PCA-correlated SNPs (or PCAIMs). It is hypothesized that hundreds of SNPs from these databases will be selected as AIM candidates for Malay ancestry. The accuracy of each resulting Malay AIM panel will be evaluated using a machine learning "ancestry-predictive model" constructed using WEKA, a comprehensive machine learning platform written in Java, and the genetic pattern based on the selected AIM-SNPs will be evaluated using the ADMIXTURE program. Consequently, the selected Malay AIM-SNPs can differentiate them from their closely related genetic ancestry counterpart and other world populations, and they can be used to describe admixture in the Malay population.

1.4 Objectives

1.4.1 General Objective

To select a subset of ancestry informative marker (AIM) SNPs from the Genome-wide Affymetrix GeneChip Mapping Xba 50 K array and Genome-wide Affymetrix SNP-6 array of the Malay SNPs datasets to design a panel of AIM SNPs that can estimate the ancestry of Malay population.

1.4.2 Specific Objectives

1. To identify a subset of autosomal AIM-SNPs from Malay SNP datasets, a genome-wide Affymetrix GeneChip Mapping Xba 50 K array with Pan-Asian SNP datasets, as reference samples.
2. To identify a subset of autosomal AIM-SNPs from Malay SNPs datasets, Genome-wide Affymetrix SNP-6 array platform with SGVP, and International HapMap Phase 3 project (HapMap) datasets as reference samples.
3. To compare the performance of the three AIM-SNP identification techniques: ipPCA combined with pairwise F_{ST} (ipPCA-F_{ST}), I_n and PCA-correlated SNPs (or PCAIMs).
4. To evaluate the genetic structure of the Malay population using the ipPCA and the ADMIXTURE program, as well as to validate the accuracy of the selected Malay AIM-SNP panels using ADMIXTURE and WEKA suite as the ancestry-predictive model.
5. To identify common SNPs between AIM-SNPs obtained from the Affymetrix GeneChip Mapping Xba 50 K Array and the Affymetrix SNP-6 array.

CHAPTER 2

LITERATURE REVIEW

2.1 Ancestry

Ethnicity, race and ancestry are three terms that used interchangeably to describe a group of persons that show different appearance, comes from different background and geographic origins and practising different cultural and social. However, there are differences between the terms race, ethnic and ancestry. Race is a term that was used during the slaving era as a way to differentiate people based on their appearance such as skin colour, feature, social, geographic origin and cultural practice (National Human Genome Research Institute, 2005, Tishkoff and Kidd, 2004). Race is often equated with *continental ancestry* which assumes the existence of the five races which correspond well to geographic regions that are Africa, Europe, East Asia, Oceania and the Americas (Tishkoff and Kidd, 2004, Royal et al., 2010). Race was replaced by ethnicity in 20th centuries because of this term is more appropriate to differentiate peoples come from different biogeographic origin (National Human Genome Research Institute, 2005).

Ethnicity emphasized more on characteristic to group people such as cultural, languages, religion, custom, dress, diet, as well as their historical or territorial identity hence not just their genetic ancestry. However by using the ethnicity to group people it may suffers several shortcomings especially ethnic group contains subgroups for example, “Hispanic” ethnic group has subgroups Cuban Americans, Mexican Americans and Puerto Ricans and the Malay ethnic group has various subgroups such

as Melayu Bugis, Melayu Banjar, Melayu Minang and Melayu Kelantan. Combining these groups into a single category does not result in a better understanding of these groups (National Human Genome Research Institute, 2005, Tishkoff and Kidd, 2004). Ethnicity and racial which both used to cluster population by broad geographic origin is not adequate to represent genetic variation in human (Tishkoff and Kidd, 2004). Categorizing people in term of ancestry is more appropriate since it may recognize a single predominant source or multiple sources (National Human Genome Research Institute, 2005).

Self-identified race or ethnicity to identify individual in biomedical studies is not sufficient (Mersha and Abebe, 2015). Knowledge of ancestry can be important clinically and in biomedical studies due to the differential distribution of normal genetic variation and of genetic variation affecting diseases which were resulted from the genetic drift, natural selection, mutation and migration processes (Tishkoff and Kidd, 2004). Knowing the genetic ancestry of the subject in biomedical research studies will be important to avoid spurious results in biomedical studies related to population stratification or may be used to map susceptibility variants that might be differentially distributed with respect to ancestry (Royal et al., 2010).

According to US Census Bureau, ancestry is "*a person's ethnic origin, the person's parents or descent or the 'roots' where the person comes from*" (Royal et al., 2010). Ancestry may be referred to a group of persons who are geographically different (biogeographic ancestry) for example Asian, sub-Saharan African, African American, European; geopolitically different for example Vietnamese, Zambian, or Norwegian

or culturally different for example Brahmin, Lemba or Apache (National Human Genome Research Institute, 2005).

We are somewhat related to each other by varying degree because of our common origin as a species (Royal et al., 2010). Mutation and recombination process contributed to differences in an individual's genome and various segments of an individual's genome may have ancestral histories that trace to different populations (Royal et al., 2010). Ancestry of an individual can be inferred from the genetic data based on the differences of the allele frequencies of the loci geographically. So ancestry can be used as alternative of the race and ethnicity in genetic research to describe a group of people and to study the population histories and relationships. The inferring of ancestry from the genetic data can be used in genealogical, anthropological and epidemiological research (Royal et al., 2010). Self-identified race or ethnicity which is commonly used in pharmacological and biomedical studies can be accompanied by the information of the genetic ancestral and considering together the education, environmental, diet, socioeconomic and culture factors which may contributed to the health disparities (Mersha and Abebe, 2015, Royal et al., 2010).

2.2 Genetic markers used in ancestry studies

The study of individual ancestry is useful in biomedical studies especially in the era of personalized medicine. Studies carried out by Mersha and Abebe (2015), showed limitation of self-reported race or ethnicity in biomedical research. Phenotype-based race information often disagrees with the genetic ancestry inferred using ancestry informative markers based on genetic or genomic data (Mersha and Abebe, 2015).

Estimation of the ancestry using genetic markers is not impossible by using the allele frequency differences between populations (Tishkoff and Kidd, 2004). Markers that demonstrate large differences in allele frequencies are good candidates to be used as a tool to estimate genetic ancestry (Barnes, 2010, Tishkoff and Kidd, 2004). Genetic markers such as Single Nucleotide Polymorphisms (SNPs), Y chromosomal haplotypes, mitochondrial DNA (mtDNA) haplotypes, X chromosomal, Short Tandem Repeats (STRs), Insertion deletion polymorphisms (INDEL), as well as Alu elements and Human Leukocyte Antigen polymorphisms (HLA) have been used to study the ancestry of the world populations (Mersha and Abebe, 2015).

Most of the ancestry studies have been carried out using haploid markers (mtDNA and Y chromosomal) (Cai et al., 2011, Corach et al., 2010, Dulik et al., 2012, Nunez et al., 2010) or multiple unlinked autosomal markers which are diploid where priory selected to be ancestry informative (Bryc et al., 2015, Gettings et al., 2014, Nakaoka et al., 2013). The selection of genetic markers in ancestry studies mostly rely on the purposes of the study where in genealogical and anthropological study haploid markers may be preferred whereas for the epidemiological research, it rely mostly on allele frequencies of the autosomal SNPs. Genetic epidemiologists employ the methods of ancestry inference using autosomal SNPs to identify genetic associations with diseases either to control for statistical biases related to population stratification among cases and controls or as a strategy to map susceptibility variants that might be differentially distributed with respect to ancestry (Royal et al., 2010).

Recently most of the studies of ancestry concentrated on the SNP marker on autosomal DNA. This is due to the haploid marker lack recombinant and autosomal give more information regarding the paternal and maternal ancestry which produce more accurate data. Moreover the studies of SNP data on autosomal DNA are more benefited to the epidemiological research especially the studies of relationships of the diseases susceptibility and protected SNPs.

On the other hand, combination of both haploid and diploid markers analysis might reveal more information about the genetic history of studied populations. Studies carried out by (Simonson et al., 2011) combined mitochondrial, Y-chromosomes and SNPs on autosomal chromosome to show that the Iban population from the Malaysian state of Sarawak exhibited genetic similarity with Indonesian and mainland Southeast Asian populations. Nunez et al., (2010) analysed the mitochondrial control region, Y-chromosome STRs and autosomal STRs to ascertain the origin of the Nicaraguan ancestors whereas Corach et al., (2010) used selected SNP marker on autosomal chromosome, mitochondrial DNA and Y-chromosome to study the genetic admixture of Argentineans.

2.2.1 Single Nucleotide Polymorphisms (SNPs)

Single Nucleotide Polymorphisms (SNPs) is a bi-allelic genetic marker. It refers to a single base sequence variation at a particular point in the genome and can be found abundantly throughout the human genome with a frequency of about one in 1,000 bp (Brookes, 1999). SNP markers are mostly bi-allelic markers which usually have two alleles per marker; A/G, C/T, A/T, T/G, C/G or A/C (Butler, 2012). SNPs can be found

in coding regions of gene, non-coding regions or in the intergenic regions (Syvänen, 2001).

SNPs found in coding region can be further divided into two types; the synonymous and non-synonymous SNPs. The synonymous SNPs do not affect the protein sequence whereas the non-synonymous SNPs alter the function or structure of the encoded proteins resulted in the recessively or dominantly inherited monogenic disorder (Syvänen, 2001). SNPs is the simplest form of DNA variation among individuals but yet known to be a very important genetic marker which responsible to the phenotypic differences between individuals. This bi-allelic marker plays an important role to the differences of individual's drug responses as well as the progression and development of many genetic diseases and said to be directly involved in determining the phenotype of an individual such as eye colour, hair and skin; facial morphology and height (Butler, 2012). SNP has been extensively studied not only for the biomedical research, but also in the field of anthropological and forensic research.

SNPs have been the marker of choice in biomedical research due to its vital contributions to the function of the regulation and expression of a protein. The studies of SNPs not only can help us to predict the response of individual to various type of drugs or environmental toxin, and risk of developing particular diseases but also to track the inheritance of diseases genes within families (Kim and Misra, 2007). Case-control association studies are the most common application of SNPs in biomedical studies. Large SNPs genotyping data of both the patient and healthy control groups which contributes to the changes in cellular biological processes inducing diseased states usually studied in case-control association studies utilizing SNPs (Kim and

Misra, 2007). The establishment of the relationship between a genotype and a phenotype is based on the comparison of the differences of genotypes for all phenotypic characteristic demonstrated by the groups being studied (Kim and Misra, 2007). The information can be used to characterise the susceptibility genes associated with a disease, hence the encoded protein can be determined for prevention or treatment of the disease (Kim and Misra, 2007).

Pharmacogenomics studies are another discipline that utilizes SNPs as a tool to study the effects of genetic polymorphisms on drug response (Kim and Misra, 2007). This study is becoming popular due to the strong demand of the personalized medication. In pharmacogenomics study where SNPs are utilized as the markers, the aim is to elucidate effects of genetic polymorphisms on drug responses. Patients who have been administered a specific drug are the targeted groups and a large scale SNPs genotyping data is needed to ensure the accuracy and the effectiveness of the study (Kim and Misra, 2007).

In Malaysia, the study of SNPs related to *Helicobacter pylori* (*H.pylori*) and Thalassemia are amongst two common diseases that intensively studied in biomedical research. This is due to the significant differences of *H.pylori* infection prevalence rates among major ethnic groups in Malaysia (ie Malays, Chinese and Indians). The highest infected was observed among Indians adults whereas Malays exhibited low prevalence of *H.pylori* (Goh, 2018, Kumar et al., 2015, Lee et al., 2013, Sasidharan et al., 2011). On the other hand, Malays demonstrated high prevalence of Thalassemia (HbE β-thalassemia) compared to the Chinese and Indians populations (George, 2013).

H.pylori is a major gastric bacterial pathogen which has been said to be distributed through the routes of human migration resulted in the division of six ancestral populations; three from Africa, two from Asia and one from Europe (Tay et al., 2009). Study carried out by He et al. (2015) identified several SNPs related to gene that involved in the process of gastric carcinogenesis. Three genes; PGC, PTPN11, and IL1B said to be associated with the susceptibility to gastric carcinogenesis (He et al., 2015). They found the interactions of the SNPs of PGC (rs6912200 and rs4711690), PTPN11 (rs12229892) and IL1B (rs1143623) modified the risks of gastric cancer.

Thalassaemias are autosomal recessive disorders which are caused by the defective synthesis of the globin chain or faulty synthesis of haemoglobin (Yatim et al., 2014). Two common types of thalassemia are α and β-thalassemia, which resulted from the defective synthesis of alpha and beta chains respectively (Yatim et al., 2014). Common type of thalassemia observed in Malays is HbE β-thalassemia (George, 2013). Nuinoon et al., (2010) reported three SNPs that have high association with β-thalassemia; SNPs of gene HBBP1 (rs2071348), gene HBS1L-MYB (rs9376092) and gene BCL11A (rs766432). Recent studies carried out by Cyrus et al., (2017) revealed another six SNPs located on chromosome 6 related to gene HBS1L-MYB (rs9376090, rs9399137, rs4895441, rs9389269, rs9402686, rs9494142, rs9376090) which has high association with the severity of the β-thalassemia.

2.2.1(a) SNPs as ancestry informative marker

SNPs is a useful biological marker in the anthropological genetics research to study the variations among different groups of humans, reconstruct evolutionary history, human physical traits and it can reveal the history of modern human migration and their adaptation to different environments. SNPs have played an important role in genetic anthropological studies because this polymorphism are believed to be stable and not deleterious to organisms and can be population specific. Most SNPs are located in non-coding regions of the genome hence not known to influence phenotype of an individual and can be used in evolutionary studies. Numerous studies of SNPs marker on mitochondrial DNA (mtDNA) and Y-chromosomal as well as autosomal have been reported (Bryc et al., 2015, Dulik et al., 2012, Elhaik et al., 2013, Kivisild, 2015). Mitochondrial and Y-chromosomal uni-parentally markers have been predominantly used to study human migration in past decades; however the trend has recently shifted to autosomal SNPs. This is due to the breakthrough of the whole genome autosomal SNPs research and the development of miniaturized and automated procedure for analysing thousands of SNPs simultaneously (Kundu and Ghosh, 2015).

Autosomal SNPs marker has been widely used in the study of human migration and tracing the ancestors of human populations. This marker utilizes DNA from the 22 pairs of autosomal chromosomes contributed by both parents; hence more information regarding the history of the ancestors can be obtained compared to the haploid marker such as mitochondrial DNA and Y-chromosomal DNA. Numerous studies have been carried out on autosomal SNPs to infer ancestry across diploid genome (Hou et al., 2014, Huckins et al., 2014, Hwa et al., 2017, Galanter et al., 2012, Kersbergen et al., 2009, Rogalla et al., 2015, Phillips et al., 2013b, Sampson et al., 2011, Santos et al.,

2016, Santos et al., 2011, Vongpaisarnsin et al., 2015, Vongpaisarnsin et al., 2017, Wei et al., 2016).

SNPs distributed throughout the human genome that occur at very different frequencies in different world populations are good candidates as ancestry informative markers (Budowle and Daal, 2008). The use of SNPs as ancestry-informative marker has been numerously published recently (Hwa et al., 2017, Das and Upadhyai, 2018, Esposito et al., 2018, Setser et al., 2020, Vongpaisarnsin et al., 2017). Yang et al., (2005) have identified 199 ancestry informative markers which were distributed throughout the human genome. The SNPs were selected based on allele frequency differences in the ABI database comprising USA Caucasian, African American, Chinese and Japanese (Yang et al., 2005). Using the ancestry informative markers, they successfully demonstrated that both continental and sub-continental populations can be readily distinguished. Furthermore, the contribution of the putative parental population can be examined in admixed population using the SNPs ancestry informative markers (Yang et al., 2005).

Kidd et al., (2014) have developed a panel of 55 highly informative SNPs. The panel has been used to analyse 73 world populations and said to be very robust and efficient to provide excellent information on ancestry especially for forensic application (Kidd et al., 2014). In developing the SNPs ancestry panel, they used several sources of SNPs databases including the ABI databases, the HGDP-CEPH SNPs databases and their own laboratory databases. They successfully identified the largest pairwise allele frequencies differences between the studied populations to develop the panel of SNPs for ancestry inference. The set of 55 ancestry SNPs is opened for improvement because

not all populations were represented and tested and the SNPs selected were less efficient in estimating the admixed populations (Kidd et al., 2014). Subsequently, in 2017 they have added another 14 reference populations allele frequency to enhance the use of the 55 ancestry informative SNPs (Pakstis et al., 2017). The allele frequencies of all 55 SNPs for a total of 139 population samples are available publicly and have been incorporated in commercial kits by ThermoFisher Scientific and Illumina (Pakstis et al., 2017).

Recent studies carried out by Esposito et al. (2018) revealed the usefulness of a panel of ancient ancestry informative markers (aAIMs) in identifying fine-scale ancient population structure in Eurasians. They utilized more than 150 thousand autosomal SNPs from 302 ancient genome classified to 21 populations recovered from Europe, the Middle East and North Eurasia (Esposito et al., 2018). They demonstrated that principal component analysis (PCA)-based approach outperforms other methods such as the Infocalc and the Wright's F_{ST} in capturing ancient population structure and identifying admixed individuals. Their finding can also be used to improve the accuracy of genetic studies utilizing ancient DNA. On the other hand, Das and Upadhyai, (2018) have approved the robustness and efficiency of autosomal ancestry informative SNPs in analysing the fine genetic structure of the highly admixed population of South Asian genetic origins. Comparison of the three methods; Infocalc, F_{ST} and the Smart PCA based on their whole genome data of Indian subcontinent, shows that the Infocalc method gave the best results compared to the Smart PCA and F_{ST}.

2.2.1(b) SNPs genotyping

SNPs genotyping protocol can be divided into two main parts; the biochemical reaction and the detection procedures (Chen and Sullivan, 2003). The biochemical reaction is basically the determination of the allele-specific products of the SNPs (Kim and Misra, 2007). Reviewed carried out by Sobrino et al., (2005) have listed a number of SNP genotyping chemistries such as allelic-specific hybridization, primer extension, oligonucleotide ligation and invasive cleavage. Primer extension involves the incorporation of nucleotides to the DNA template using specific enzyme to discriminate the SNPs alleles. A primer will be designed to anneal to the 3' end of the DNA template (SNPs) and nucleotides will be added to the template by the polymerase enzyme (Kim and Misra, 2007, Sobrino et al., 2005).

The allelic-specific hybridization involves the use of differences in thermal stability of the double-stranded DNA to discriminate the SNPs allele (Kim and Misra, 2007). Target-probe pairs must perfectly complementary to each other thus the effectiveness of the hybridization likely depending on the length and sequence of the probe, location of the SNPs and hybridization condition. This approach is suitable for the high throughput microarray platforms such as incorporated in the GeneChip® array technology (Affymetrix, CA) (Kim and Misra, 2007).

Oligonucleotide ligation is a technique where ligase enzymes are used to discriminate the SNPs allele. In these approach three oligonucleotides probes are involved, where the first two oligonucleotides are hybridized to the single stranded DNA template, adjacent to each other. Subsequently, the third probe binds to the template adjacent to the SNP immediately next to the allele-specific probe. The ligation product is detected

by various methods (Kim and Misra, 2007). Invasive cleavage involves the cleaving of the targeted DNA sequence by restriction enzyme and the product can be detected using gel electrophoresis. Invander® assay has adopted this technique by using two allele-specific probes attached with two types of dye at either end, the reporter (R) and the quencher (Q) and a common invader probe. The products can be detected using florescence analysis.

Allelic product of those methods can be detected with several detection systems such as florescence-electrophoresis, fluorescence resonance energy transfer (FRET), fluorescence polarization, fluorescence arrays, mass spectrometry and luminescence (Fondevila et al., 2013, Nicklas and Buel, 2008, Poetsch et al., 2013, Shi et al., 2011, Zeng et al., 2012). The florescence-electrophoresis detection method is perhaps the most accessible method due to the availability of the technique in most of the forensic laboratory around the world. The high throughput technology such as microarray (for example Affymetrix and Illumina platforms) is more powerful for analysing hundreds or thousands of SNPs simultaneously which had been used in many ancestry studies (Galanter et al., 2012, Keating et al., 2013, Krjutskov et al., 2009).

2.3 Ancestry informative markers (AIMs)

Historically, human dispersed from East Africa to other part of the world more than 100,000 years ago resulted in the genetic diversity of modern human due to adaptation to new environment, geographical and climate change (Cavalli-Sforza, 2007). The migration contributed to the genetic variation because of the founder populations usually carried a portion of their most immediate ancestral population and at the same