GENOME-WIDE SNP MICROARRAY ANALYSIS AMONG MALAY SUB-ETHNIC GROUPS IN PENINSULAR MALAYSIA

NUR SHAFAWATI BINTI AB RAJAB

UNIVERSITI SAINS MALAYSIA

2013

GENOME-WIDE SNP MICROARRAY ANALYSIS AMONG MALAY SUB-ETHNIC GROUPS IN PENINSULAR MALAYSIA

by

NUR SHAFAWATI BINTI AB RAJAB

Thesis submitted in fulfillment of the requirements for the degree of
Master of Science

July 2013

ACKNOWLEDGMENTS

First of all, I am grateful to Allah s.w.t for the healthiness and opportunity to complete my study. I would like to give my appreciation to Prof Zilfalil Bin Alwi who offered me a great opportunity to gain experiences and knowledge in applying this study. Many thanks to Ministry of Higher Education (MOHE) for the research funding of FRGS grant (203/PPSP/6170025) and and also to short term grant Universiti Sains Malaysia (304/PPSP/61311034).

My deep appreciation also goes to all the villagers, nurses and head villagers from Kelantan, Kedah, Johor, Negeri Sembilan and Perak who involved directly or indirectly with the process of collecting the samples succeed. An appreciation also goes to Dr Pornprot Limprasert and his students in providing of Pattani Malay samples from Narathiwat, Thailand. Thank you to Universiti Sains Malaysia for accepting me as postgraduate student. Many thanks to UKM Medical Molecular Biology Institute (UMBI) and Matrix Analytical Sdn Bhd, Malaysia for allowing me to do part of the genotyping works in their lab.

Special appreciation to Human Genome Centre that provide facilities in doing most of the lab works including staff and other students who encouraged me in advising and technical support in most of the works. Especially Arfah, Marini, Hanani, Roslina, Aishah, Wati, Sathiya, Yan Yan, Marjanu, Fazreen and many others. Special thanks also goes to Hatin, Prof Shuhua Xu and Prof Andrew Crosby as giving me the beauty essence of knowledge in crucial analysis part.

Full appreciations to my beloved family, especially to mama, Sharipah Bt Jusoh and abah, Ab Rajab Bin Mahmood who support and motivates me from the beginning. Not forgetting to all my siblings especially Ayu and Yaya.

Last but not least, a person who deserves my deepest thanks and appreciation for his continued support and motivation during the study especially in my final writing process of this dissertation; my beloved husband, Shazwan Bin Shamsuddin.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PLATES

LIST OF GRAPHS

LIST OF ABBREVIATIONS

A:	Adenine
A260/A280 :	ratio of 260 absorbance over 280 absorbance
Bp:	Base pair
ddH2O:	deionized distilled water
DGGE:	denaturing gradient gel electrophoresis
DHPLC:	denaturing high performance liquid chromatography
DNA:	deoxyribonucleic acid
dNTPs:	dinucleotide triphosphates
dsDNA:	double strand deoxyribonucleic acid
EDTA:	Ethylene diamine tetra acetic
g:	gram
HLA:	Human leukocye antigen
HWE:	Hardy-Weinberg equilibrium
kb:	kilo base
M:	Molar
MAF:	Minor Allele Frequency
mg/ml:	miligram per mililiter
MgCl2 :	magnesium chloride
ml/min:	mililiter per minute
mm:	milimeter
mM:	millimolar
mtDNA:	Mitochondrial DNA
n:	Number of individuals

NaCl:	Sodium chloride
OD:	optical density
PAGE :	polyacrylamide gel electrophoresis
PCR :	polymerase chain reaction
RFLP:	Restriction fragment length polymorphism
RNA:	ribonucleic acid
rpm :	round per minute
SD :	standard deviation
SNP :	single nucleotide polymorphism
SSCP :	single strand conformational polymorphism
ssDNA :	single strand deoxyribonucleic acid
SSOP:	Sequence specific oligonucleotide probe
SSP:	Sequence specific primer
SSPE:	Saline Sodium Phosphate EDTA
Taq:	*Thermophilus aquaticus*
TBE :	Tris-borate-ethylene
TE buffer:	Tris-EDTA buffer
Tris HCl:	Tris-hydrochloric acid
U:	Unit
UV:	ultra-violet
VNTRs:	Variable number of tandem repeats
Xba I:	Restriction enzyme of an E.coli strain that carries the *xbaI* gene from *Xanthomonas badrii*
Y-STRs:	Y-chromosome short tandem repeats

Mg:	microgram
µl:	microliter

A GLANCE OF THE STUDY

Chapter 1 Introduction	This chapter provides information on related introduction covered the association's story for this project. Start with basic knowledge of biological introduction and continue to population genetics study. Then, provide information on Single Nucleotide Polymorphism (SNP), brief information with the old tools to genotype SNPs. Continue with recent advent of Microarray SNP genotyping that was used for this study and end with the analysis part of genotype data with the available software to be use later. Also, the most important is the samples involved in this study which 7 Malay sub-ethnic groups from Peninsular Malaysia and one Malay sub-ethnic group; Pattani malay from southern Thailand. Objectives of the study were provided in the last part of the chapter.
Chapter 2 Material & methods	Overall, this chapter mainly described the complete methods step by step in order to achieve the objectives of the study. Started with collection samples, followed lab works which mainly based on Affymetrix Human Genome 50K *xba*I, extraction of SNP genotyped data and finally continue with steps of analyzing the abundant data with various software.
Chapter 3 Results	In this chapter, all the results obtained from the analysis parts were extracted and presented. Every result was shown to differentiate the 8 Malays sub-ethnic groups by genetic variations which are allele frequencies, Linkage Disequilibrium (LD) and Tag SNPs respectively. However, filtrations of clean SNPs were needed to run the analysis.
Chapter 4 Discussion	All related information was discussed in this chapter regarding all results adapted in the last chapter. This chapter also provides the list of SNP identification of every Malays by their selected Tag SNP. Interestingly, this chapter also discuss on 6 related genes that has been found to have special association with these Malays. Limitation, problem facing and the future prospect of this study also been discussed.
Chapter 5 Conclusion	Overall conclusion for this study was made in this chapter including the use of SNP microarray, the beauty of different Malays in Peninsular Malaysia and the importance of knowing their genetic differentiation in order to search correlation with diseases or health.

ANALISA MICROARRAY SNP PADA GENOM ANTARA KUMPULAN SUB-ETNIK MELAYU DI SEMENANJUNG MALAYSIA

ABSTRAK

Kebangkitan dalam penggunaan teknologi canggih di dalam bidang genetic telah banyak mempengaruhi serta menaiktaraf kemajuan dalam genetik populasi manusia. Antaranya Mikroatur nukleotid pollimorfisme tunggal (SNP) yang membolehkan pengliputan SNP yang sangat besar dalam genom manusia. Kemudahan mikroatur ini telah digunakan bagi kajian ini untuk mencari serta mendapatkan perbezaan genetik di kalangan etnik melayu di semenanjung Malaysia. Etnik melayu di semenanjung Malaysia terdiri daripada beberapa kumpulan sub-etnik melayu yang berbeza dalam pelbagai faktor antaranya bahasa, sejarah perpindahan ke Malaysia, tempat asal, adat serta kehidupan sosial harian. Seramai 135 orang melayu terlibat dalam kajian ini yang terdiri daripada sub-etnik Melayu Kelantan, Melayu Minang, Melayu Jawa, Melayu Bugis, Melayu Kedah, Melayu Champa, Melayu Banjar serta Melayu patani.

Daripada kajian yang dijalankan, lebih daripada 50000 SNP berjaya digenotipkan. Hasil kajian mendapati sememangnya terdapat perbezaan frekuensi alel antara etnik melayu ini dapat menjelaskan perbezaan mereka. Di samping itu, kajian ini ingin mandalami perbezaan kumpulan sub-etnik melayu yang terlibat menggunakan analisis hubungan ketaksamaan (LD) SNP, Haplotip dan Tag SNP yang terdapat pada 3 kromosom yang menunjukkan jarak genetik (genetic

distance) yang paling jauh. Seterusnya, SNP pengenalan untuk setiap sub-etnik ini dapat dihasilkan menggunakan Tag SNP yang terpilih. Selain itu, kajian kaitan dengan gen yang terlibat juga dapat diterokai. Terdapat 31 SNP yang terlibat dalam penemuan blok LD yang kuat bagi mencari identiti SNP setiap kumpulan etnik melayu ini menggunakan Tag SNP yang terpilih. Hasil akhir kajian ini ialah penemuan identiti SNP bagi setiap sub-etnik melayu ini selain daripada Melayu Champa yang tidak mempunyai blok LD yang kuat untuk ditafsirkan. Selain itu, terdapat 6 gen yang menarik yang boleh dikaitkan dengan sub-etnik melayu ini iaitu FRYL, SGCB, LIG1, LSM14A, LARGE serta FAM118A. Bagaimanapun, kajian yang lebih mendalam perlu dilakukan untuk memastikan penemuan ini.

GENOME-WIDE SNP MICROARRAY ANALYSIS AMONG MALAY SUB-ETHNIC GROUPS IN PENINSULAR MALAYSIA

ABSTRACT

The use of advanced technology in the field of genetic had influenced and upgraded the dicipline and had leds to a lot of advances in the genetics of human populations. Among them, microarray of single nucleotide polymorphism (SNP) allows large coverage of the human genome. SNP microarray was used for this study to find and characterize genetic differences among Malays sub-ethnic groups in Peninsular Malaysia. The Malay sub-ethnic groups of Peninsular Malaysia consist of several sub-groups that differ in a variety of factors including language, history of migration to Malaysia, origins, customs and daily social life. One hundred and thirty five Malays participated in this study and consisted of Kelantan Malay, Minang Malay, Javanese Malay, Bugis Malay, Kedah Malay Champa Malay, Pattani Malay and Banjar Malay.

From our study, more than 50,000 SNPs were successfully genotyped. The study found that there is indeed allele frequency differences among the Malay sub-ethnic groups which absolutely show their differences. In addition, this study goes deep into Malay differences by analyzing their differences of Linkage disequilibrium (LD), haplotype and tag SNPs on three selected chromosomes that showed the highest genetic distances. More on, SNP identification for each sub-ethnic group can be produced using tag SNPs. This study further investigated the related genes which were identified. There were 31 SNPs involved in the

discovery of a strong LD block which could identity each of sub-ethnic Malay based on selected tag SNPs. The end result of this study is the discovery of the SNP identity for each sub-ethnic Malay group apart from Champa Malays which did not have a strong LD block to be interpreted. In addition, there were six genes of interest that could be attributed to Malay sub-ethnic groups, namely FRYL, SGCB, LIG1, LSM14A, LARGE and FAM118A genes. However, further investigations need to be done to confirm these findings.

CHAPTER 1

INTRODUCTION

1.1 DNA as a basic unit of living organism

Deoxyribonucleic Acid (DNA) is a nucleic acid molecule that contains the genetic instructions used in the development and functioning of all known living organisms. Chemically, the DNA is a long polymer of four simple units called nucleotides or bases which are adenine, cytosine, guanine, and thymine, abbreviated as A, C, G, and T, respectively (Lewis, 2007; Reece, 2004)). It is organized in separate linear molecules called chromosomes (**Figure 1.1**). The human genome contains approximately three point bases from three billion bases divided into twenty pairs of autosomal chromosomes and two sex chromosomes. The chromosome range in lenght from about 50 to 250 million bp which each of them contains many genes; approximately around 30,000 genes.

Most sexually reproducing organisms are diploid which have a duplicate set of genetic material consisting of paired chromosomes, one from each parent. Such paired chromosomes, called homologous, are essentially identical which contains same genes in the same order, but having small differences in the DNA sequences originating from the variability present in the population (Lewis, 2007).

1.2 Genome variation

The uniqueness of human being are due to the environment factors and hence natural selection forces that cause the differences to their DNA sequence. This situation gives meaning to genome variation. In fact no two people are genetically identical. Each people have 0.1 percent DNA sequence or three million different from their entire genome. However, more closely related two people are, the more similar their genomes but more distant related two people, more difference their genomes sequence. In other words, people from same populations will have more similar genetic variation than the people from difference populations. Genomic variations in the human DNA sequences can affect the variation in human traits, the development of diseases and an individual's response to drugs, infections and vaccines (Nakamura *et al,* 2009). This interconnection had invited many projects to do research on human genome variation mapping.

One of the project is the International HapMap Project (2005). This project involves six countries known as Japan, United Kingdom, Canada, China, Nigeria and United States. Their goal is to compare the genetic sequences of different individuals from these six countries population in chromosomal regions and to make the data collected are freely available online from their website which is http://www.hapmap.org. Another aim of this project is to discover the genetic associations with diseases.

Furthermore, another project that is keen on genome variation is Human Variome Project Consortium, launched in 2006. Their main objective is to collect, characterize and share all the genetic variations that effecting diseases. Next, the human global health can be improved and enhanced by standardization and systemic program on managing all the population health. (Cotton *et at*, 2007).

Malaysia is also not left behind in the form of human variation mapping. The project is known as 1Malaysia Human Genome Variation Consortium which is a Malaysian node of Human Variome Project. This consortium was launched on October, 2010. The main objective of this project is to create the Malaysian genome variation map and to study its implications on many fields regarding human health, human history, and the applications of human social life also the country's migration history.

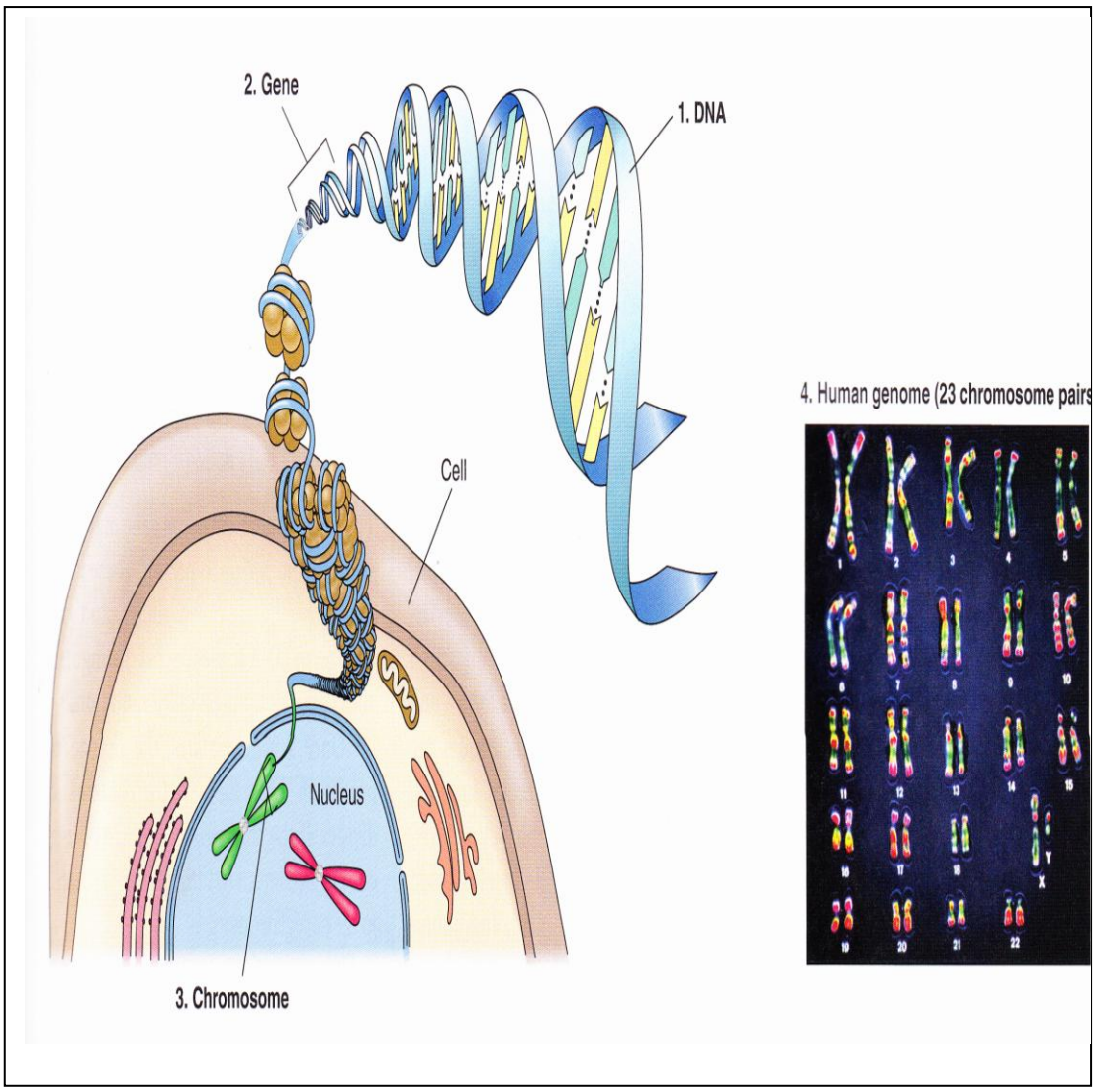

Figure 1.1: The illustration of cell, chromosome, gene and DNA (Lewis, 2007).

1.3 Genetic Polymorphism

Genetic polymorphism is the result of the sequence changes that occur in DNA and inherits from one generation to the next. It is generally defined as an occurrence of more than one percent in a population. The study of such inheritable genetic polymorphic markers provides an understanding of the human population history (Marth *et al.*, 2004). Examples of genetic polymorphisms that have been widely discussed are Single Nucleotide Polymorphisms (SNPs), sequence repeats, deletions insertions and also recombination (Smith, 2002).

1.3.1 Single Nucleotide Polymorphism (SNP)

The single-nucleotide polymorphisms (SNPs) are the most common polymorphic marker and abundant form of DNA variation in the human genome. It occurs exactly once in the human evolution (Chang *et al*., 2006). It occurs when a single nucleotide (A/T/C/G) in the genome sequence is altered, example "A**A**CTAT to A**T**CTAT" (Brookes, 1999). An individual with homologous chromosomes are almost identical with some small differences. Mainly, most differences occur at SNPs, although other types of polymorphisms occur (e.g. microsatellites, rearrangements, and copy number variations). An individual is said to be homozygous for a SNP if the same allele is present at both homologous chromosomes, and heterozygous otherwise (**Figure 1.2**).

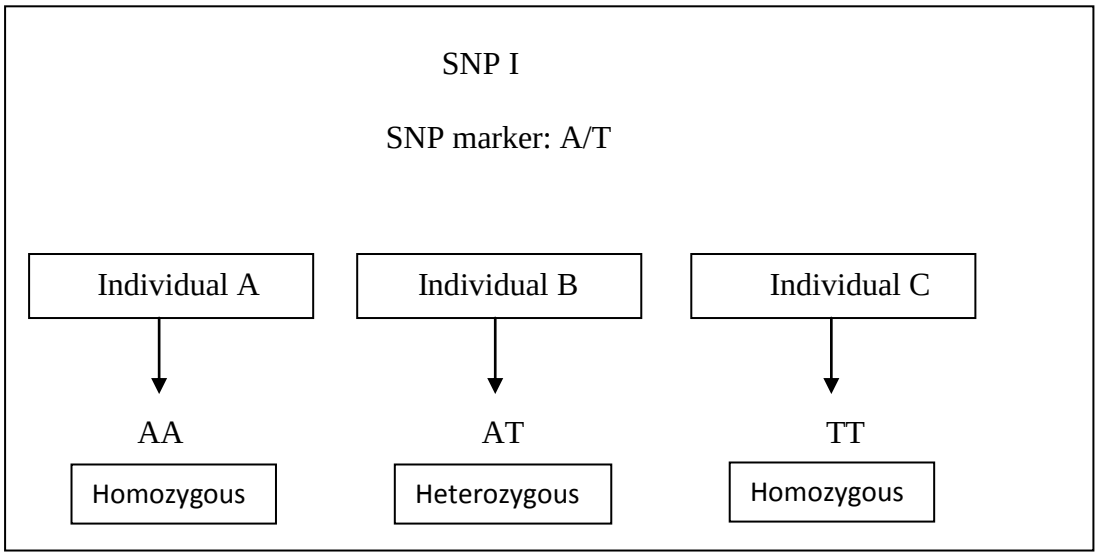

Figure 1.2. The diagram shows the example of homozygous and heterozygous SNP at one locus (SNP I as an example). Individual A and C show homozygous SNP (AA and TT). Homozygous SNP is an individual with the same allele at particular genomic locus. Individual B shows heterozygous SNP (AT). Heterozygous is an individual with different alleles at particular genomic region.

The majority of SNPs are located in non-coding parts of the genome and does not affect the sequence of encode proteins. These markers can be used for accessing the degree of genetic diversity in a population, in evolutionary studies and in forensics research. The SNPs resulted in altered protein transcription can be used for diagnosis of hereditary disorders or prediction of drug response in pharmacogenetic diagnosis.

According to the International SNP Map Working Group (The international HapMap consortium, 2007) a map of 3.1 million SNPs is distributed throughout the human genome. It has been estimated to be on average of one SNP per 1000 to 2000 bases that there are seven million common SNPs with a minor allele frequency (MAF) > 0.05. Hinds *et al* (2005) reported that most common SNPs are found in most of the populations studied which is American of European, African and Asian ancestry. However, the frequencies of the alleles are varying considerably between these populations.

1.4 Population genetics

Population genetics is the study of the frequency of occurrence of an allele within and between populations. The science of population genetics deals with Mendel's Laws and other genetic principles as they affect the entire population of an organism. Basic understanding of population genetic is essential and useful in medicine, law, biotechnology, molecular biology, cell biology, evolutionary biology, natural history, sociology and anthropology (Wolinsky, 2008). It includes the study of various forces that resulted in evolutionary changes of human through time such as genetic drift, mutation, natural selection, and human migration.

1.4.1 Hardy-Weinberg Equilibrium (HWE)

Hardy-Weinberg principle (HWP) proposed by the most fundamental and important law in population genetics is the. It was worked out by two scientist; Godfrey H. Hardy and Wilhem Weinberg on 1908. The principle was used to predict how gene frequencies can be inherited from generation to generation given a specific set of assumptions. This test involves comparing the observed and expected genotype frequencies for population studied. When a population is in Hardy-Weinberg equilibrium (HWE) for a given locus, it means that there is random mating, no selection, no mutation, no gene flow and a population large enough to avoid the random effects of genetic drift (Mao *et al.*, 2010).

Population genetics was used to determine how reproductively isolated between populations. Nevertheless, if differences occur in selection operation on a locus, the difference may be due to the selection acting differently on the different populations rather than the result of isolation in reproductive between the populations. Thus, if allele frequencies are compare between two populations, fisrt step we need to determine whether each population is in Hardy Weinberg equilibrium to ensure that the difference in allele frequencies of the two populations is due to reproductive isolation.

The test of HWE can be measured using the "goodness of fit" or chi-squared test (x^2). Mathematically, the chi-squared test is represented as:

$\chi^2 = \Sigma$ [(observed value – expected value)2 / expected value]

1.5 SNP genotyping with Microarray

Various methods can be used to screen SNPs. This include restriction fragment length polymorphism (RFLP) analysis, allele specific oligonucleotide hybridization, oligonucleotide ligation assay, single stranded conformation polymorphism (SSCP), allele specific primer PCR, analysis using beacons, TaqMan, invader method, mass spectrometry, pyrosequencing, analysis using molecular inversion probes, denaturing gradient gel electrophoresis (DGGE) and denaturing high performance liquid chromatography (dHPLC) (Twyman *et al*,. 2005; Gut., 2001; Ye *et al*., 2001; Wang *et al*, 2007).

Genotyping able characterized to these human variations. One of the new methods in this new era for screening the SNPs is Microarray genotyping (Hanai *et al*, 2006; Simpson *et al*, 2005; Gibson, 2006). The information provides clues about the evolutionary history of human genome. To study the evolutionary relationships between humans, various methods can be employed to estimate the time of their divergence from a common ancestor.

1.5.1 What is microarray

Microarray is a high throughput technology that allows detection of thousands of genes or targeted variations simultaneously. It allows scientist to easily detect and measure the expression of thousands of genes or markers at one time (Jenkins and Gibson, 2002). Microarray was designed and developed in 1990's by groups of scientist and engineers. It is modifications of multiple methods of molecular laboratories designed before. Most of basic and common molecular tools was used and applied as protocols in microarray methods involving digestion, polymerase chain reaction (PCR), fragmentation, labelling, targeting and hybridyzation (Metspalu, 2005).

Microarray can be use for detection of DNA or RNA level. It involves mainly in the detection of changes in gene expression levels, detection of genomic gains and losses and detection of mutations in DNA (Hoheisel, 2006).

1.5.2 Application of microarray

Since the development and existent of microarray technology in genetic studies, the uses and application of microarray has raised by year. The use of microarray can be applied for identification of complex genomic studies. The technology was used to look for genes related to human diseases (Craig and Stephan, 2005). This whole-genome study produced individual genotypes and found related genes with many diseases.

Microarray also has been applied in pharmacogenomics especially in the drug discovery. Drug Pharmacogenomics is the study of correlations between therapeutic responses to drugs and patients profile in genetics (Liljedahl *et al*, 2003). Genes from a diseased and a normal cell will help the identification of the biochemical constitution of the proteins synthesized by the diseased genes from comparative analysis. The information can be use to synthesize drugs which combat with these proteins and reduce their effect. Microarray was extensively used in characterize the human genetic variations (Gunderson *et al*, 2005). This phenomenon gave lights to many researchers globally in producing genotype profile and differences among ethnics or populations.

In this study, Affymetrix 50K GeneChip platform (**Figure 1.3**) which provide more than 50,000 SNPs in a single chip, was chosen to screen the Malay sub-ethnic populations. The application of the microarray enables and promises simultaneous genome-wide screening and much information can be obtained in a relatively short time (Syvanen *et al.*, 2005; (Craig and Stephan, 2005). It features typically less than 200 microns in diameter with thousands of spots and is usually printed onto a coated glass microscope slide or chip. Microarray was originally designed for the detection of differences between samples and is ideally suitable for high-throughput studies of natural variation (Gunderson *et al*, 2005).

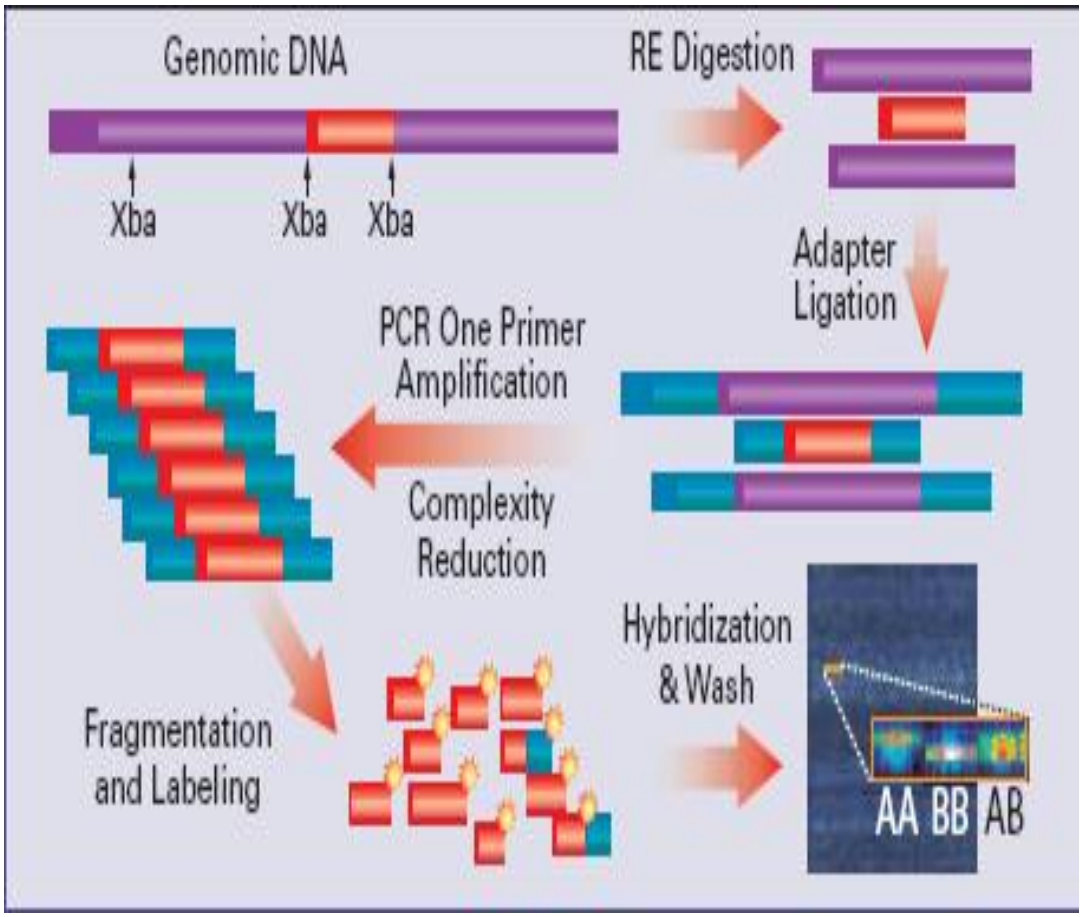

Figure 1.3: SNP Genotyping mapping assay overview (Affymetrix 100K Manual Assay)

1.6 Analysis of SNP genotyping data

Population genetics studies using microarray SNP genotyping involved several steps that must be followed: accessing, decomposition, generating, and analyzing large volumes of data which often take time consuming to be implemented. The most important and crucial step in microarray SNP genotyping is the analysis of the microarrays data. This challenging part gives bombastic and crucial task to the researcher to analyze the large data (Stokes *et al*, 2007; Metspalu, 2005; Liang et al, 2008). The SNP data generated via difference software depends on the objectives of the study. The third party software on population genetics varies either freely available online or provided by private companies. The software also differs among them according to the algorithms of the task chosen.

1.6.1 PEAS V1.0 (a package for elementary analysis of SNP data) software

Freely available downloaded software, PEAS software was created and developed by Xu *et al* (2010) with the aim to facilitate analyses on population genetics and molecular phylogenetics studies. One of the function in this package is the creation of input file for other packages such as Haploview (Barrett *et al.*, 2005), STRUCTURE (Pritchard *et al.*, 2000), Arlequin (Schneider *et al*. 2000), LDhat (McVean *et al.*, 2004), PLINK (Purcell *et al.*, 2007), MEGA (Kumar *et al*. 2004), PHYLIP (Felsenstein, 1989), PHASE (Stephens *et al*. 2001) and fastPHASE (Stephens & Donnelly, 2003).

The PEAS software (Xu *et al.*, 2010) provides many other functions which involve in population genetics and molecular phylogenetics studies such as, basic statistics, data filtering, individual and population distance and many more.

1.6.2 Haploview version 3.32 software

The International HapMap Project conducting large-scale surveys of human genetic variation which resulted in an increas volumes of single-nucleotide polymorphism (SNP) genotyping data that have produced delightful opportunities for association studies. However, they have stimulated the difficulty of treating and analyzing such dense data created. Another freely available downloaded software; Haploview is an aesy-to-use program created by Barrett *et al* (2005) which developed in Mark Daly's lab at the Broad Institute of MIT and Harvard.

Haploview has several features that are useful throughout different phases of association studies. In this study, the Haploview software is used to analyse the SNP genotype data in order to get the SNP haplotype, linkage disequilibrium (LD) and Taq SNP for the populations involved in this study.

1.6.3 Haplotype

The sequence of alleles in contiguous SNP positions along a chromosomal region is called a haplotype (**Figure 1.4**). The number of haplotypes in the human genome is estimated to be about 200,000 compared to the number of SNPs, which is around 10 times higher (Metspalu, 2005). Recombination is the major source for variation of haplotypes in the population. Generally, recombination occurs when a strand of DNA breaks and joined at the end of another DNA molecule. In this context recombination occurs as chromosomal crossover between paired homologous chromosomes. Sex cells, a process of mixing of two chromosomes in each parent occurs naturally in meiosis that caused the chromosomes in human cells come in pairs.

However, the section of the ancestral chromosomes are shifted through frequent recombination events but more or less of the sections still occur at some regions of DNA sequences and shared by multiple populations. This is due to no recombination occurrs at those regions. Hence, this is the sections of haplotype that allow for gene inspections involved in diseases and other important traits.

As a young species, most of the variation in any present-day human population comes from the variation present in the ancestral human population. Also, as humans migrated, they carried part but not all of the genetic variation that existed in the ancestral population. As a result, the haplotypes of recent human populations tend to be subsets of the haplotypes ancestral. In addition, the haplotypes in recent populations tend to be longer than in ancestral populations, because the ancestral

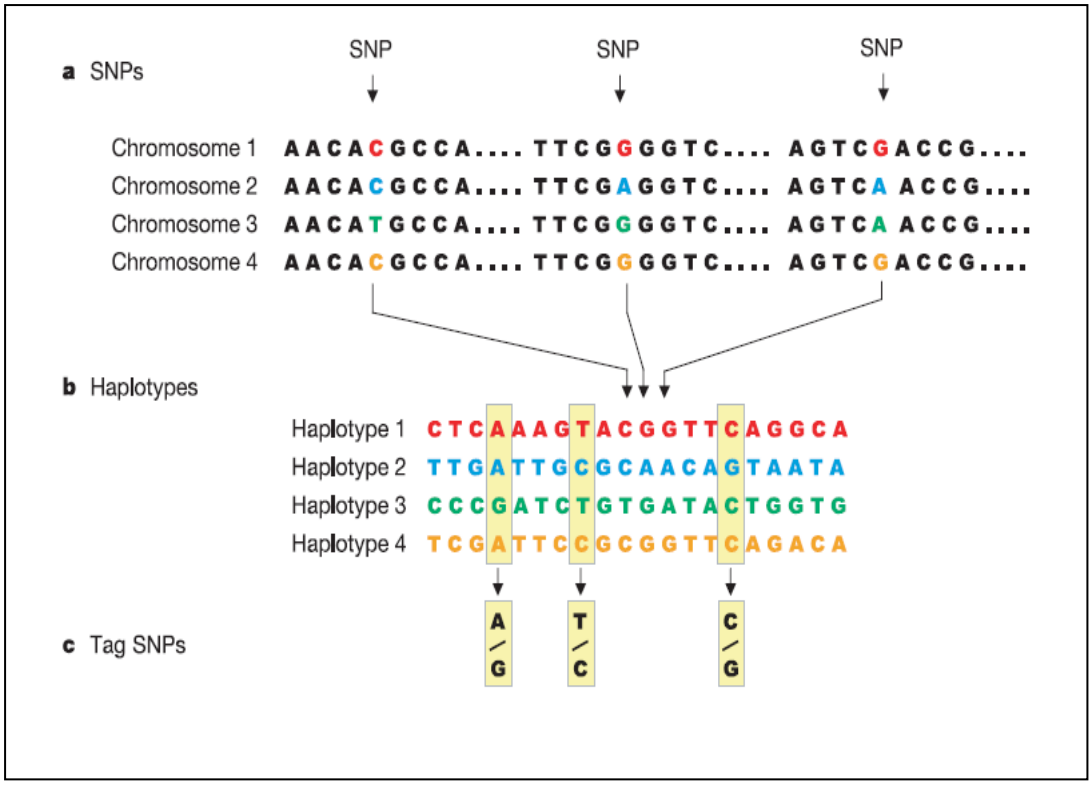

Figure 1.4: SNPs, haplotype, and Tag SNPs. (a) A population shown with four chromosomes. Three positions from the choromosomes differs by one nucleotide (indicated by an arrow) indicating the SNPs. (b) the contiguous sequences of SNPs from each of four chromosomes shows the haplotypes. (c) The selected SNPs from the each haplotype show the formation of Tag SNP. (The International HapMap Consortium., (2003).

have been larger through much of recent population history and recombination has had more time there to break up haplotypes.

Using Haploview software, the haplotypes are estimated using an accelerated EM algorithm similar to the partition/ligation method described by Qin *et al.*, (2002). This creates highly accurate population frequency estimates of the phased haplotypes based on the maximum likelihood as determined from the unphased input. The haplotype shows each haplotype in a block with its population frequency and their connections between blocks. A value of multiallelic D' is shown in the crossing areas that represents the level of recombination between the two blocks.

1.6.4 Linkage Disequilibrium (LD)

In population genetics, linkage disequilibrium is the measurements of non-random associated alleles at two loci based on the expectations relative to allele frequencies in a population. The LD can be estimated by the parameters of *D'* and r^2. Haploview software defined a population in LD by the structure of haplotype blocks using solid spine (Tulio *et al.*). From Haploview software, the default algorithm is taken from Gabriel *et al.* (2002). The result can be differentiated as "strong LD", "inconclusive" or "strong recombination" or “low LD” if ninety five percent confidence bounds on D prime are generated. Creation of a block was designed if 95% of informative comparisons are "strong LD" which ignores markers with MAF < 0.05.

A strong LD between variants and neighboring SNPs is required for associative detection of the genes responsible for the disease or population characteristics. By mapping the LD throughout the whole genome, SNPs in haplotypes can be related to genes that predispose individuals to common multifactorial disorders. Currently, the distribution of LD in different Malay populations is still unknown, but with the development of high-throughput genotyping technologies it will be possible to create a genome-wide LD map among Malays.

1.6.5 Taq SNP

A tag SNP is a selected SNP in a region of the genome with high linkage disequilibrium (**Figure 1.4).** The tag SNP help researcher to identify special unique of SNPs from haplotype that can be the identity details of the haplotype (Kruglyak, 1999). Hence, by testing an individual's tag SNP, the researcher will be able to identify the whole haplotype involved in that individual. The Hapmap project estimates that the tag SNP is about 300,000 to 600,000, which is far fewer than the 10 million common SNPs.

The Haploview software measures the tag SNP using the implementation of Paul de Bakker's Tagger tag SNP selection algorithm. The Haploview's Tagger operates by pairwise or aggressive mode. Both case actualize by choose a minimal set of markers such that all alleles to be captured are correlated at an $r^2 >= 0.8$.

1.7 The Malay people

The Malays are a diverse group of Austronesian people inhabiting Southeast Asia region. They constitute the dominant ethnic group in Malaysia, Indonesia, Brunei, the minority in South Philippines, the Pattani region of Thailand, East Timor, and Singapore (Tryon, 2006).

One of theory about the origin of malays in Peninsular Malaysia from Hussin *et el* (2004) is the malays were migrated from Yunnan since 1500 before century and other theory from Coomaraswamy (1985) said that the Malay is a mixed of Proto Malays and Deutero Malays. The Proto malays is a primitive people in Peninsular Malaysia and was migrated into rural and mountainous area after the migration of Deutero Malays from Mongolia.

The present-day Malays is the intermarriages between Deutero malays and traders of the ancient trade from many countries; Indian, Arab, Chinese, Sumatram, Javanese and Siamese (Hatin *et al.*, 2011). These intermarriages produce various recent Deutero-Malays in Peninsular Malaysia due to their geographical origin (ancestral) and their social-linguistic practice. Several groups of Malays in Peninsular Malaysia have been identified, such as Melayu Kelantan, Minang, Riau, Jawa, Acheh, Bugis, Rawa, Banjar, Champa, Pattani, Kedah and Jambi (Hoh *et al*, 2008).

According to the federal constitution ("Federation of Malaya" Part XII 124(3)(b) Federal Citizenship, Acquisition of Federal Citizenship by operation of law), malay people is a people who consider Islam as their religion, and use malay language as language of communication and conforms to the Malay customs.

Historically, these Malays sub-group came to Malaysia from different routes. According to Hussin *et al.* (2004) and Roux, (1998), Kelantan malays is a decendent from Langkasuka kingdom which arose approximately in 100 BCE to 7th century CE and then known as Pattani Kingdom. This kingdom involved most of Malays from northern Peninsular Malaysia which are Kelantan malay, Kedah malay, Pattani malay, and Terengganu malay (Omar Din, 2011). However, the existence of Champa Malay to the Peninsular Malaysia was very recent among other Malays as they exodus Cambodia in 1975 as the government falls to communist. These refugee sub-ethnic malay also known as "boat people" arrived Peninsular Malaysia with hope to resettlements (Tze-Ken, 2008)

The malays in Peninsular Malaysia also descendent from Indonesia Archipelago since the nineteenth centuries (Mohd Jali, 2003). The Banjar malays was from Kalimantan, the Jawa and Bawean malays from Java and others from Sumatera such as malays of Minangkabau, Batak, Rawa, Acheh and mandailing (Sainuddin, 2003).

The Bugis malay was migrated from Sulawesi since seventeenth century. Their settlements to malay peninsula particularly Johor and Selangor is actually to

escape from political issues and mainly because of the Bugis malay itself love to travel (Omar *et al*., 2009). The Minang malays migrated to Negeri Sembilan in the early 14th century after the fall of the sultanate of Malacca.

The migration of Jawa malays from Java into Johor happened after the goverments open the district in the state for immigrants (Sainuddin, 2003). The Jawa malays also settled in Selangor but the groups was not as big as in Johor. However, the Banjar malays is more diverges which they settled to three states in Peninsular Malaysia which is Perak, Selangor and Johor (Mohd Jali, 2003).

To date, the International HapMap Project includes more than 4 million SNPs in 270 individuals from four geographically diverse populations (Nigeria, Utah, Japan and China) and is one of the most important sources of information regarding the variation in the human genome but this database did not include our Malay population. In addition, populations from the South East Asia region, which is well known to have the largest record of human migration outside Africa, are yet to be studied.

The report on Malays group in Peninsular Malaysia on SNP genotyping is too little according to publication. The only SNP study on population genetics was done by Hatin *et al*., (2011) and The HUGO Pan Asian SNP Consortium *et al* (2009). Hatin *et al.*, (2011) discussed on genetic structure between malays in peninsular Malaysia and other population groups mainly in Asia. It shows the differences in genetic structure among the malays studied (Kelantan malays, Minang malays, Jawa malays and Bugis malays) based on Fst calculation by

neighbor-joining tree. This is would be due to the differences of their geographical and also admixtures (Hatin *et al*., 2011).

The HUGO Pan Asian SNP Consortium *et al* (2009) studied only two Malay sub-groups which are Kelantan malay and Minang malay. Seventy three populations involve in this study from south east asia and east asia including the two malays. This paper showed correlations and relatedness of genetics ancestry with linguistic groups and geography. From the result of a hypothetical most recent common ancestor (MRCA), the study shows the most ancestral from both Malays was Kelantan malays.

Other study on malays sub-group in peninsular Malaysia was done using HLA study (Edinur *et al,* 2009), mtDNA study (Haslindawaty *et al*, 2010), and short tandem repeats (STR) on Y chromosome (Hoh *et al*, 2008). Only Hatin *et al* (2011) studied the genetic structure of four Malays using SNP as the markers. The Malay sub-ethnic groups involved in this study were Kelantan Malays, minang Malays, Jawa Malays and Bugis malay. Most of the studies showed the differences in genetic variations among the malays. However, the study on genetic variations among Malays with SNPs by LD and tag SNP were not detected and limited populations involved. The lack of information on SNP study of LD, haplotype and Tag SNP on malays population in peninsular Malaysia has attract our study to enlarge and extent the information on the malays sub-group using SNPs.

The aim of this project is to determine the common genome variation among the Malay sub-populations using SNP microarray genotyping. The variations will concentrate on the haplotype map, linkage disequilibrium and Taq SNPs among the malays sub-groups in Peninsular Malaysia. The large amount of data generated will be useful in the mapping of human phenotypes and complex diseases in South East Asia.

At the end of the study, the result produced will provide an insight into the special characterization and variation as well as the unique differences between Malay sub-groups and will contribute to the creation of Malaysian SNP database. On top of that, it provides detail information for the complex disease mapping and thus leads to a better understanding of the complex causes of many common diseases in human.

1.8 Benefits and outcomes of this study

1.8.1 New technology enables new science

This study includes the eight Malay sub-groups. This new SNP microarray technology allows us to screen more than 50,000 SNPs simultaneously and this is made possible by combining semi-conductor manufacturing technology with chemistry to put millions of different strands of DNA on a single glass chip the size of a thumbnail. This new technology enables us to do new science although with less developed research infrastructures.

1.8.2 Establishment of the application for screening SNPs for population genetics

Microarray has become the most popular method in detecting multiple SNPs in one chip. The establishment of this method helps in studying of the whole genome screening assay for population genetics. The idea of high-density SNP map of the genome using microarray will be useful for the mapping of the genes involved in complex disorders, via association studies. Human genetic variation research determines how variation among individuals or groups contributes to the health status of that individual or group and to understand the pattern of diversity in order to accelerate the search for the genetic cause of human disease.