BIOINFORMATICS: INFERENCE OF POPULATION GENETIC STRUCTURE OF PENINSULAR MALAYSIA MALAY SUB-ETHNIC GROUPS USING SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs) MICROARRAY MULTILOCUS GENOTYPE DATA

by

W. NUR HATIN BINTI W. ISA

Thesis submitted in fulfillment of the requirements

for the degree of

Master of Science

UNIVERSITI SAINS MALAYSIA

MAY 2012

BIOINFORMATIK: KESIMPULAN STRUKTUR GENETIK POPULASI BAGI KUMPULAN SUB-ETNIK MELAYU DI SEMENANJUNG MALAYSIA MENGGUNAKAN DATA MIKROARRAY GENOTIP PELBAGAI LOKUS POLIMORFISME NUKLEOTIDA TUNGGAL (SNPs)

oleh

W. NUR HATIN BINTI W. ISA

Tesis yang diserahkan untuk

memenuhi keperluan bagi

Ijazah Sarjana Sains

UNIVERSITI SAINS MALAYSIA

MEI 2012

ACKNOWLEDGEMENTS

First and foremost my biggest gratitude to Allah S.W.T, the most gracious and most merciful, for giving me strength, patience and healthiness to finish this thesis within the stipulated time. My deep appreciation to Associate Professor Dr. Zilfalil Bin Alwi (main-supervisor), Dr. Mohammed Rizman Bin Idid (co-supervisor), Professor Felix Li Jin and Professor Shuhua Xu for giving me so much of guidance, precious time, lessons, comments, and technical supports in completion this study. Thank you to Ministry of Higher Education (MOHE) for the research funding of FRGS grant (203/PPSP/6170025) and Universiti Sains Malaysia (USM) Fellowship Scheme for the scholarship in finishing the study. I also acknowledged the contributions by the other members of this study group from the School of Health Sciences and the School of Dental Sciences, USM. Many thanks to the UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia and Matrix Analytical Sdn. Bhd., Malaysia for allowing us to use their laboratory facilities. Not to forget, special thanks to my wonderful research fellow; Nur Shafawati Ab Rajab who has done a great job in genotyping the samples and for all the supports. Also to other fantastic friends in the lab such as Arffah, Marjanu, Mareen, Alia, Tasya, Dayah, Hasnah, Huda, Kila, Fizah, Ina, Siti, Sathiya, Aizat, Zaki, Syibli, Amin, Abg. Nizam, Rani, Mr. Chia and many others that I couldn't mention here due to limited space, thank you for all the cheerful times and your help. Last but not least, my gratitude to my family, especially my beloved parent, Wan Isa Bin Wan Ali and Samsiah Bt Embong, who were always there for me and understand my 'turbulence' during the writing of this thesis. May Allah repay all your kindness and bless us all. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figures **Page**

Figures **Page**

Figures **Page**

LIST OF APPENDICES

LIST OF ABBREVIATIONS

arp	: input file format for Arlequin software
BCE	: before current era/before common era
BMD	: between marker distance
BRLMM	: Bayesian Robust Linear Model with Mahalanobis distance algorithm
CE	: current era/common era/calendar era
CEL	: cell intensity values of the individual probes
CHP	: information of the probe sets data file, including intensity values
CN-JN	: Chinese *Jinuo*
CN-WA	: Chinese *Wa*
DA	: matrix distance
DAT	: raw image data file (pixel intensity values)
DC	: chord distance
DM	: Dynamic Model Mapping Analysis algorithm
DMS	: degree-minute-second
DNA	: deoxyribonucleic acid
drawparams	: parameters file for *distruct* program
DTT	: data transfer tool
Fst	: fixation index statistic
HWD	: Hardy-Weinberg disequilibrium
HWE	: Hardy-Weinberg equilibrium
HWP	: Hardy-Weinberg principle

IBD : Isolation-by-distance

ID-JV : Indonesian *Jawa*

ID-ML : Indonesian *Melayu*

ID-TR : Indonesian *Toraja*

indfile : individual *Q* matrix file

IN-DR : Indian *Telugu*

indvq : permuted individual *Q* matrix file

inp : input file format for PEAS software

inpfile : input file

IN-WL : Indian *Marathi*

kb : kilo base

LD : linkage disequilibrium

LE : linkage equilibrium

MAF : minor allele frequency

MCMC : Markov Chain Monte Carlo

MDS : multi-dimensional scale

miscfile : miscellaneous output file

ML : Maximum likelihood algorithm

M_l : strict consensus method

MRCA : most recent common ancestor

MS-DOS : Microsoft disk operating system

MSG : mean square errors for loci within populations

MSP : mean square errors for loci between populations

mtDNA : mitochondria DNA

MY-BG : *Melayu Bugis*

MY-JH : Malaysian *Semang Jahai*

MY-JV : *Melayu Jawa*

MY-KD : *Melayu Kedah*

MY-KN : *Melayu Kelantan*

MY-KS : Malaysian *Semang Kensiu*

MY-MN : *Melayu Minang*

MY-TM : Malaysian Proto-Malay *Temuan*

NJ : Neighbor joining algorithm

outfile : output file

outtree : Newick notation tree file

paramfile : parameter file for CLUMPP program

perm : file containing the list of allowed colors for the permuted clusters

popfile : population *Q* matrix file

popq : permuted population *Q* matrix file

RNA : ribonucleic acid

RPT : report file

RS ID : reference SNP ID

RSQ : r-square correlation

S : sampling SNPs dataset for STRUCTURE analysis

SD : standard distance

SNPs : single nucleotide polymorphisms

SSC : symmetric similarity coefficient

TH-PT : Thai *Pattani*

TSC ID : The SNP Consortium ID

XLS : Microsoft Excel file format

XML : extensible markup language

Y-DNA : Y chromosome DNA

YRI : *Yoruba*

LIST OF SYMBOLS

°	: degree
'	: minute
"	: second
<	: less than
>	: more than
~	: more or less
θ	: population mutation rate/moment estimator
σ	: population standard deviations
μ	: mutation rate per locus per generation
α	: alpha parameter
e	: margin of error or desired level of precision
K	: cluster
Ln(Pr)	: logarithm value of posterior probabilities
N_e	: effective population size
Pr (X|K)	: posterior probabilities of data
Q	: ancestral component or membership coefficient
r	: correlation coefficient
R^2	: linear regression
X	: given genotype data
χ^2	: Pearson's chi-square test of goodness of fit

BIOINFORMATIK: KESIMPULAN STRUKTUR GENETIK POPULASI BAGI KUMPULAN SUB-ETNIK MELAYU DI SEMENANJUNG MALAYSIA MENGGUNAKAN DATA MIKROARRAY GENOTIP PELBAGAI LOKUS POLIMORFISME NUKLEOTIDA TUNGGAL (SNPs)

ABSTRAK

Di Semenanjung Malaysia, orang Melayu terdiri daripada pelbagai kumpulan sub-etnik yang dipercayai mempunyai susur galur keturunan yang berbeza berdasarkan penghijrahan mereka berabad yang lalu. Pengetahuan struktur genetik populasi Melayu bukan sahaja penting bagi reka bentuk kajian bioperubatan yang sepatutnya, tetapi juga membantu dalam memahami sejarah demografi yang berkaitan penghijrahan dan percampuran mereka. Kajian ini adalah untuk mengenal pasti dan menganalisis struktur genetik dan pekali komponen keturunan bagi lima kumpulan sub-etnik Melayu iaitu *Melayu Bugis*, *Melayu Jawa*, *Melayu Minang*, *Melayu Kedah* dan *Melayu Kelantan*. Data genotip pelbagai lokus polimorfisme nukleotida tunggal (SNPs) untuk semua kumpulan sub-etnik Melayu dijana menggunakan cip Affymetrix 50K Array. Bagi menguatkan kesimpulan kajian, 12 populasi lain dari Thailand, Indonesia, China, India, Afrika dan sub-kumpulan Orang Asli di Semenanjung Malaysia yang diperolehi dari Pangkalan Data Pan Asian SNP Initiative (PASNPI) dan juga projek antarabangsa HapMap telah dimasukkan dalam analisis. Data 54,794 autosomal SNPs yang telah di genotip bagi setiap 472 individu daripada 17 populasi telah dianalisis menggunakan dua pendekatan analisis statistik untuk kelompok genetik, iaitu kaedah data jarak dan

kaedah kelompok berasaskan model. Kaedah berasaskan data jarak telah berjaya mengesan sekurang-kurangnya tiga kelompok genetik, dengan kemungkinan percampuran genetik dalam Melayu. *Melayu Bugis* dan *Melayu Minang* mempunyai hubungan genetik yang sangat rapat dengan populasi Indonesia, manakala *Melayu Jawa* mempunyai hubungan genetik dengan Proto-Melayu dan Cina, menunjukkan asal-usul keturunan yang sama. *Melayu Kedah* dan *Melayu Kelantan* berbeza secara genetiknya daripada kumpulan-kumpulan sub-etnik Melayu yang lain, tetapi dekat dengan populasi *Pattani* dari Thailand. Kaedah berasaskan model telah mengenal pasti campuran dan pekali komponen keturunan bagi setiap orang Melayu. Orang Melayu berkongsi keturunan dengan orang Indonesia dan Thai. Garis keturunan mereka telah dikesan mempunyai perkongsian keturunan dengan Proto-Melayu dan Cina. Walau bagaimanapun, *Melayu Minang*, *Melayu Kedah* dan *Melayu Kelantan* mempunyai komponen keturunan India yang besar dalam genom mereka, berbanding dengan Melayu yang lain. *Melayu Jawa* mempunyai komponen Cina dengan perkadaran tertinggi dalam genom mereka, manakala genom *Melayu Bugis* hampir tidak ada percampuran dengan komponen-komponen yang lain. Keputusan ini telah menyediakan maklumat mengenai perbezaan genetik antara kumpulan sub-etnik Melayu dan maklumat yang berharga bagi mengenal pasti asal-usul orang Melayu di Semenanjung Malaysia.

BIOINFORMATICS: INFERENCE OF POPULATION GENETIC STRUCTURE OF PENINSULAR MALAYSIA MALAY SUB-ETHNIC GROUPS USING SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs) MICROARRAY MULTILOCUS GENOTYPE DATA

ABSTRACT

In Peninsular Malaysia, the Malays consist of various sub-ethnic groups which are believed to have different ancestral origins based on their migrations centuries ago. Knowledge of Malay population genetic structure is not just important for a proper design of biomedical studies, but also helpful in understanding their demographic histories of migration and admixture. This study was conducted with the aim of identifying and analyzing the genetic structures and ancestral membership coefficients of five Malay sub-ethnic groups namely *Melayu Bugis*, *Melayu Jawa*, *Melayu Minang*, *Melayu Kedah*, and *Melayu Kelantan*. The multilocus genotype data of single nucleotide polymorphisms (SNPs) for all of the Malay sub-ethnic groups were generated using Affymetrix 50K Array chip. For powerful inference, 12 other study populations from Thailand, Indonesia, China, India, Africa and *Orang Asli* sub-groups in Malay Peninsula, obtained from the Pan Asian SNP Initiative (PASNPI) and International HapMap project database, were included in the analysis. The genotyped data of 54,794 autosomal SNPs for each of 472 individuals from the 17 populations were analyzed by two statistical analyses approaches for genetic clustering, which are distance-based and model-based clustering methods. The distance-based method has

successfully detected at least three genetic clusters, implying probable admixtures within the Malays. *Melayu Bugis* and *Melayu Minang* have a very close genetic relationship with Indonesian populations, whilst *Melayu Jawa* have interestingly close genetic relationship with Proto-Malays *Temuan* and Chinese, indicating a common ancestral origin. *Melayu Kedah* and *Melayu Kelantan* are genetically different from the other Malay sub-ethnic groups, but close to Thai *Pattani*. The model-based method has uncovered the level of admixture and ancestral membership coefficient for each of the Malays. Apparently, Malays shared a common ancestor with the Indonesians and Thais. The ancestry lines of Malays, Indonesians and Thais were traced back to have shared a common ancestor with the Proto-Malays and Chinese. Nevertheless, *Melayu Kedah* and *Melayu Kelantan* have substantial Indian ancestral component in their genome, as well as the *Melayu Minang*, relative to the other Malays. The *Melayu Jawa* had the highest proportions of Chinese ancestral component in their genome, whilst the *Melayu Bugis* had almost no admixture in their genome. These results contribute significantly on the genetic differentiation between the Malays sub-ethnic groups and provided valuable insight into the origins of the Malays in the Malay Peninsula.

CHAPTER 1

INTRODUCTION & LITERATURE REVIEW

1.1 Malays

1.1.1 Definition of Malays

Malays (*Melayu*) are an ethnic group who speak Malayo-Polynesian language which is a member of the Austronesian family (Bellwood, 1997; Omar, 2004). They belong to the Southern Mongoloid group of races and predominantly inhabit the Malay Peninsula, the east coast of Sumatra, the coast of Borneo and the smaller islands between these locations (Bellwood, 1997). The Malay Peninsula is the region that comprises of southern Thailand, Peninsular Malaysia, and the island of Singapore.

In Malaysia, Malay is defined according to Article 160 of the Constitution of Malaysia Part XII 124(3)(b) Federal Citizenship, Acquisition of Federal Citizenship by operation of law, which states that Malays are Malaysian citizen born to a Malaysian citizen who professes the religion of Islam, habitually speaks the Malay language, conforms to Malay custom and is domiciled in Malaysia (Constitution of Malaysia).

1.1.2 Malays of Peninsular Malaysia

In Peninsular Malaysia, the Malays consist of various sub-ethnic groups which are believed to have different ancestral origins based on their migrations centuries ago (Paul, 1961). The Malay populations in the western (*Melayu Minang*) and southern parts (*Melayu Jawa* and *Melayu Bugis*) of the Peninsular Malaysia were believed to have had more historical and cultural links with the populations from the Indonesian archipelago compared to the Malay populations in north-eastern regions (*Melayu Kelantan* and *Melayu Kedah*) (Benjamin, 1986; Fix, 1995). The Malays at the northern part of peninsula had more historical connections with the civilizations in mainland Southeast Asia such as Thailand, Cambodia and Myanmar (Allan, 1998; Syukri, 2002; Stark, 2006).

According to the 2010 Malaysian population and housing census, the Malays form the majority of the population followed by Chinese as the second largest ethnic group while Indians comprise the third largest ethnic group (Jabatan Perangkaan Malaysia, 2010). The existence of Chinese and Indian in the Malay Peninsula with different timelines throughout the centuries brought varying degrees of cultural influences and genetics admixtures to the Malay populations. Substantial influx of Chinese and Indians were started only during the British colonial era to work as laborers in the tin mines and the plantation industry that were mainly concentrated on the west coast of peninsula (Marshall Cavendish, 2008). Prior to British colonization, Chinese and Indian traders had established strong trading links with the Malay Peninsula. These early contacts did

not cause large scale migration but intermarriage and integration between them and the Malays were common (Marshall Cavendish, 2008).

1.1.3 Relationship with aboriginal peoples of Peninsular Malaysia

The existence of indigenous *Orang Asli* (aboriginal peoples) populations in the Peninsular Malaysia has also raised questions as to whether they are associated with the first wave of human migration from Africa, or belong to the more recent events of Asian human evolution (Allen, 1879; Hill *et al*., 2006). And, to what extent they have contributed to the uniquely admixed gene pool of Malays (Bellwood, 1993).

Based on language, physical features and sociological differences, the *Orang Asli* groups are classified into three main tribal groups; *Semang*, *Senoi*, and Proto-Malay (Benjamin 1985; Bellwood, 1993; Fix 1995; Bulbeck, 1996). The *Semang* are usually found in the northern region of the peninsula, the *Senoi* in the central region, and the Proto-Malays in the southern region (Benjamin 1985; Oppenheimer, 1998).

The *Semang* who speak the Northern Aslian languages which are part of the Austro-Asiatic language family are rainforest hunter-gatherer nomadic population (Benjamin, 1985; Bellwood 1993). The *Semang* are of Australo-Melanesian affinity and share some common physical features with African pygmy populations, including short stature, tight frizzy hair and dark skin (Allen, 1879; Fix 1995). Researches done by Carey (1976) and Fix (1995) have explained much about the *Semang*. Their physical

characteristics associated them with the Philippine *Aeta*, Andaman Islanders, Melanesians, Tasmanians, as well as certain tropical Australian rainforest foragers into a Negrito group. They were assumed to have originated in Africa and spread throughout Southeast Asia before establishing settlements in the southwest Pacific (Tarling, 1999; Wells, 2002). The *Semang* are believed to be the earliest settlers and original coastal inhabitants of the Malay Peninsula but the arrival of newcomers forced this group further inland, resulting in them being isolated in forested hilly regions.

The *Senoi* are regarded as the second wave migration into Peninsular Malaysia after thousands of years of the *Semang* arrival. They are thought to have its origin in South Asia and associated with the forest foragers in South Asia and most mainland Australian Aborigines (Cole, 1945; Fix, 1995). Nonetheless, from some other studies (Bellwood, 1993; Nicholas, 1996; Baer 1999; Hill *et al*., 2006) the *Seno*i are believed to be a composite group, with many of their phenotypic and cultural features as well as their maternal lineages have been shared with the *Semang* since Hoabinhian times and about half of its could be traced back to an origin in Indochina within 7,000 years ago. They probably are descendants of Neolithic peoples such as *Ban Kao* at central and southern Thailand who have intermarried with indigenous groups and became the ancestors of the modern *Senoi*. This may explains the variably Negrito to Mongoloid appearance in *Senoi*, with lighter skin and wave hair. The *Senoi* are the cultivators and speak central Aslian languages in the Austro-Asiatic family (Benjamin, 1985).

The Proto-Malays, also known as Aboriginal Malays (*Melayu Asli*) are the farmer-traders that form the third wave migration into peninsula in about 4,000 years ago (Bellwood, 1997; Fix, 1995). Their arrival supposedly represented the first influx of Southern Mongoloid in Peninsular Malaysia with the light skin and straight hair (Carey, 1976; Fix 1995). They are Austronesian speakers apart from one tribe, (the *Semelai*) who speak Aslian and are similar in appearance to the Malays but of diverse origins. Some were seafarers and probably having entered the region by sea in recent centuries whilst others may have been living in the peninsula for thousands of years (Benjamin, 1985; Andaya, 2001; Hill *et al*., 2006).

In contrast, the present-day Malays of the Malay Peninsula are described as Deutero-Malays, the descendants of the Proto-Malays who had admixed with Siamese, Javanese, Sumatran, Indian, Thai, Arab and Chinese traders (Comas *et al.*, 1998). However according to Fix (1995), the original Deutero-Malays migrated through southern China (after the migration of the Proto-Malays) over 1500 years ago and their intermarriages with the Proto-Malays and traders of the ancient trade routes resulted in the diverse recent Deutero-Malay populations that became known currently as the Malays.

1.1.4 Early Malay kingdoms

After the arrival of aboriginal peoples into peninsula during prehistoric times, more subsequent expansions of peoples from adjacent places like Indochina and Sumatra, also from as far as India, China and Egypt were registered in Peninsular Malaysia. The

period was referred as proto-historic times when the Malay Peninsula was once a very strategic port and became the crossroads in maritime trades of the ancient age that connected Indochina and the Indonesian archipelago.

At the beginning of the first century, the trading activity between Southeast Asia and the two flanking regions of India and China began to expand. The early sea traders reached the coast of the peninsula and let the local porters transport their goods, using rafts, elephants and man powers to the opposite coast of the peninsula along the rivers that was known as early trans-peninsular route (Rahman, 1998; Jacq-Hergoualc'h and Hobson, 2002). The map of the early trans-peninsular route and the Spice Route is shown in **Figure 1.1**.

By the 5th century CE, several kingdoms and entrepôts appeared on the Malay Peninsula, not just bringing many benefits to the people of Peninsular Malaysia but the communities in the region evolved to form complex cultures with varying degrees of influence from India and China (Rahman, 1998; Marshall Cavendish, 2008). Moreover, the Indians had been conspicuous in the region very much earlier, since the period of the ancient Hindu Malay kingdoms which arose approximately in 100 BCE to 7th century CE such as *Chi Tu, Langkasuka* and *Kadaram* that controlled much of the northern Malay Peninsula (Arasaratnam, 1970). The Indian influxes continued to expand during the subsequent empires of *Srivijaya* and *Majapahit* (Paul, 1961). These early Malay states were heavily influenced by concepts of religion, government and arts that were brought by the Indians. Traces of this influence can still be found in Malay

culture despite the later influence of Islam in 16th century CE (Shuhaimi, 1984; Arasaratnam, 1970; Syukri, 2002).

The *Chi Tu* kingdom was the earliest port kingdom recorded in peninsula that was said to be founded approximately in 100 BCE by Mon-Khmer peoples from Funan, an ancient pre-Angkor Indianized kingdom located around the Mekong Delta. The *Chi Tu* kingdom which means the 'Red Earth Land' was mentioned in the 7th century CE of Sui Dynasty annals from China and summarized by Rahman (1998). The location of the kingdom is never been confirmed, but most likely the interior of the Peninsular Malaysian state of Kelantan (**Figure 1.1**) based on the time of sailing distance to Champa, an Indianized kingdom that controlled what is now southern and central Vietnam. It was a prominent inland kingdom, despite the majority of the several kingdoms appeared on the Malay Peninsula and other parts of Southeast Asia were situated on the coast.

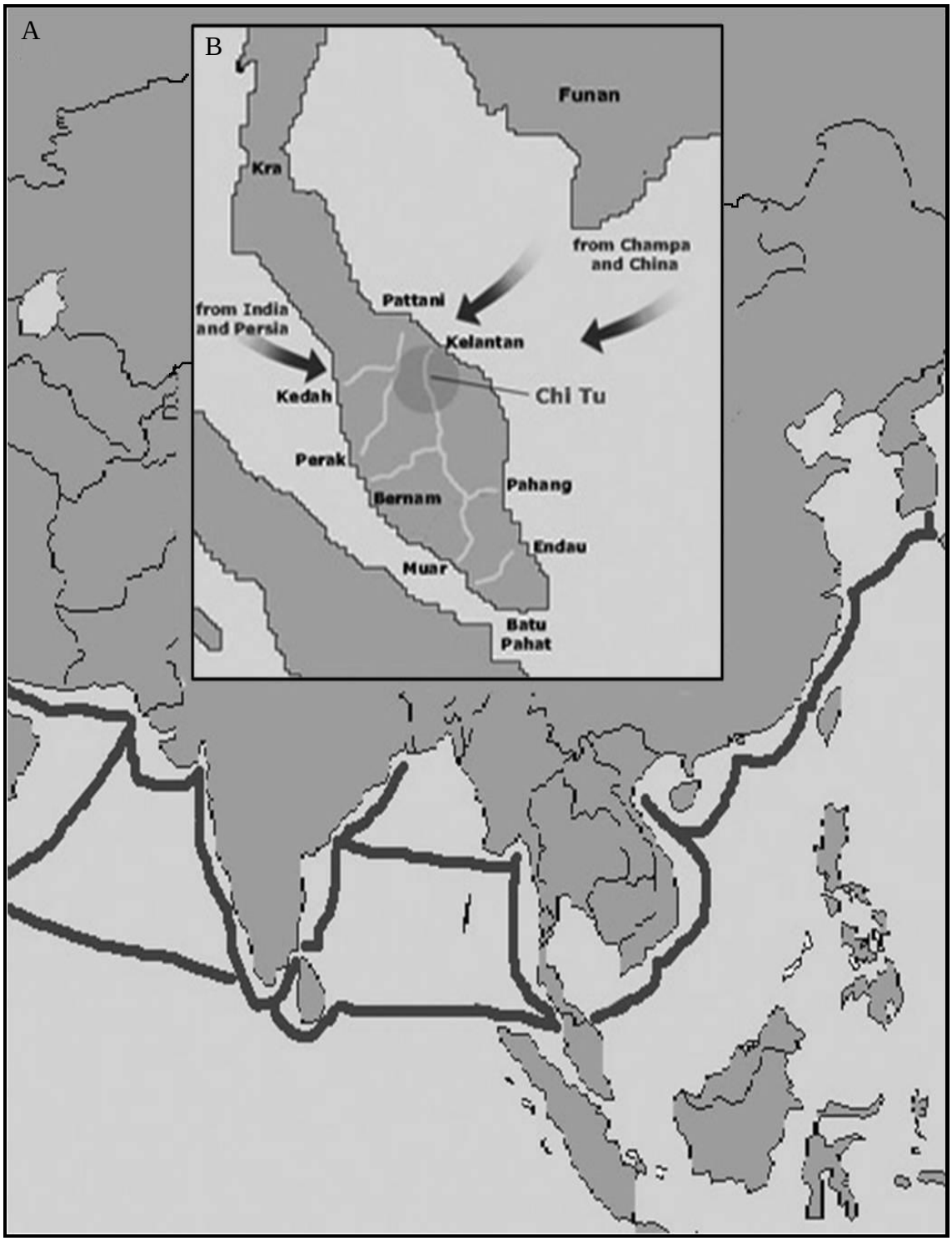

Figure 1.1: Map of Asia showing the ancient sea trade route. **A)** The black line shows the sea trade route with the Malay Peninsula as the crossroad of the maritime trade. **B)** The close-up map of Malay Peninsula shows the early trans-peninsular route in white line within the peninsula. Image was adapted from http://en.wikipedia.org/wiki/Early_history_of_Kedah.

In the early 2nd century CE, an ancient Hindu-Malay kingdom named *Langkasuka* was founded at Kedah by seafarers with Hindu faith from India. The kingdom later was moved to Pattani, on the east coast of the northern part of the Malay Peninsula that became the center of the earliest Malay kingdom of *Langkasuka* which the territory encompassed the states of Kelantan (Syukri, 2002). Meanwhile in Kedah, according to the Chinese and Indians literatures, as well as paper from Allen (1997) have cited that the trace of Indianization may have begun as early as 2nd century CE. Whereas the archeological evidences of Hindu-Buddhist kingdom from the excavation site at Bujang Valley have been dated to the 4th century CE and making it the oldest civilization in Peninsular Malaysia (Braddle, 1980, Lembah Bujang Archeological Museum, 2011). The Kedah Kingdom or *Kadaram* was founded around 630 CE by a Persian who embraced Hinduism. The Persian-Hinduism dynasty ended with the ninth king converted to Islam in 1136 CE (Shuhaimi, 1984; Othman, 1990).

In the 7th to 8th century CE, *Langkasuka* and *Kadaram* were believed to have succumbed to the *Srivijaya* but still flourished as long as they did not challenge their overlord (Rahman, 1998; Allen, 1998). The *Langkasuka* kingdom along with *Kadaram* are probably among the earliest kingdoms founded on the Malay Peninsula that later were arose as the powerful entrepôt in the region. *Kadaram* was also the first port in Southeast Asia used its sea route as an alternative route to China or Far East to replace the early trans-peninsular route, which connected Chinese and Far Eastern traders by land (Jacq-Hergoualc'h and Hobson, 2002). During these times, Indian traders and

priests travelled the maritime routes and brought with them Indian concepts of religion, government, and arts (Arasaratnam, 1970, Shuhaimi, 1984).

The *Srivijaya*, an ancient Hindu-Buddhist Malay empire that ruled some 600 years was centered in Palembang on the island of Sumatra (Allen, 1998). It influenced much of the Malay Archipelago with a reach spanning from Sumatra and Java to the kingdoms of the Malay Peninsula (*Chi Tu*, *Langkasuka* and *Kataha*) and also as far north as the southern part of Thailand (Munoz, 2006). The great influence of this Indianized empire affected much of Malay culture and at the same time, it also helped spread this culture throughout Sumatra, the Malay Peninsula, and western Borneo (Munoz, 2006). This empire started to rule from the 3rd century CE and remained a formidable sea power until the 13th century CE before being conquered by *Majapahit* (Paul, 1961). The *Majapahit* was an Indianized kingdom based in eastern Java from the 13th to around 16th century CE. It is considered to be one of the greatest and also the last of the major Hindu empires of the Malay Archipelago. Its influence extended far wider than the *Srivijaya* Empire did till the eastern part of Indonesia, including Sulawesi (Ricklefs, 1991, Wolters, 2008).

The proto-historic of Malaysia ended in the beginning of 15th century CE with the emergence of Malacca Sultanate. Based on Encyclopedia of Malaysia: Early History (1998), Malacca was established in 1402 CE by *Parameswara* from *Temasek* (now known as Singapore), a royal bloodline of *Srivijaya*. He embraced Islam in 1409 CE after marrying a princess from *Pasai* and the marriage has thrust most of his peoples to

embrace Islam. In the 15th to the early of 16th century CE, Malacca that encompassed most of modern day Peninsular Malaysia, Singapore and a great portion of eastern Sumatra thrived into the most important entrepôts in Southeast Asia and a hub of Islamic studies, spreading Islam to Malay Archipelago. During this period, the interactions between Malay Peninsula with Eastern and Western civilizations such as the Chinese Empire, Siamese, Gujerat, Arabs and Europeans that came and traded with Malacca, was very common. The trace of influences between civilizations whether in the architectures of the building or cultures, still can be seen in Malacca state until now.

1.2 Population genetics

Population genetics study is the analysis of changes in allele frequency and describes the variability over time in populations. It is the field that is closely linked to evolutionary study as the genetic composition is mostly affected to the evolutionary processes, such as natural selection, genetic drift, mutation, and gene flow. In genetics, the population is defined as a group of individuals who share a common gene pool and have the potential to interbreed. The gene pool is the complete set of unique alleles in a population. The statistical measures of population genetics can clarify the genetic structure and ancestry (Slatkin, 1987; Hartl and Clark, 2007).

Strong foundations of population genetics have been founded by Ronald A. Fisher, J.B.S. Haldane and Sewall Wright in 1920s and 1930s. It came into being from the need to integrate the re-discovered principles of Mendelian genetics in 1900 (originally

published by Gregor Johann Mendel in 1865, but overlooked by scientific community for 40 years) with the Darwin's theory, published in 1859. The Mendelian genetics is about primary principles involved in transmission of hereditary characteristics from parent to offspring, whereas Darwin has established the evolution theory of common ancestor for all life of modern species and that the process of natural selection was the major mechanism that caused the pattern of evolution. By exploring quantitatively the affection of natural selection and other evolutionary processes on Mendelian population's genetic composition over time, the full reconciliation of Mendelian and Darwinism was achieved. It was a pioneer model towards better understanding of evolution process and marked as the birth of modern population genetics (Samir, 2008).

1.2.1 The Hardy-Weinberg principle

The Hardy-Weinberg principle (HWP) is a fundamental law in population genetics about the relationship between allele and genotype frequencies in population. It was discovered by two scientists, Godfrey H. Hardy, an English mathematician, and Wilhelm Weinberg, a German physician in 1908. The mathematical equation is the simplest but the most crucial one as it is the standard against which to measure changes of allele frequency in a population (Crow, 1988). The principle states that the genotype frequencies and gene frequencies of a large, randomly mating population remain constant from generation to generation provided migration, mutation and selection do not take place. This situation is said to be in Hardy-Weinberg equilibrium (HWE).

The HWP has simplified the modeling of evolutionary change with the application of HWE to compute the entire distribution of genotype frequencies if given the relative frequencies of all the alleles at a single locus are constant (Samir, 2008). Nevertheless, in real world, the conditions of HWE are not exactly applicable to all of human populations because most of human do not mate at random. Deviation from HWP generally indicates that the population has undergone the evolution process such as inbreeding or presence of population structure. In disease association study, the deviation from HWP at particular markers may suggest an association between the marker and disease susceptibility.

The deviation can be measured using the "goodness of fit" or chi-squared test (χ^2). If the value of χ^2 is less than the 5% significance level of the degree of freedom, the population is still in Hardy-Weinberg proportions. Recently, for large amounts of alleles, a number of powerful algorithms of Markov Chain Monte Carlo (MCMC) methods have been introduced to assess the deviation from HWE which require a super power computer to perform the analysis (Wigginton *et al*., 2005).

1.2.2 Patterns of human genetic variation

The complete sequence of human genome and mapping of human genetic variation at thousands of loci by recent advances technologies together with the robustness of powerful statistical analysis in estimating population parameter have increased the ability to understand the human evolution. Although the difference between humans at

the molecular level is only 0.2-0.5% of the three billion nucleotides in the total genome (Levy *et al*., 2007), it is more than adequate to show uniqueness among individuals and give major impact to biomedical studies.

The pattern of genetic variations among modern humans reflects ancient human demographic history. The knowledge of these variations have provided scientist with much of information about the recent origin, human dispersals and human relatedness to one another (International HapMap Consortium, 2005; Conrad, 2006; Relethford, 2008). As the human genetic variation is under influenced of natural selection, gene flow, genetic drift, and mutation as well as recombination processes (Myers, 2005; Wu and Lin, 2006), the disease causing alleles may differ across diverse populations (Chakravarti, 2001; Reich and Lander, 2001) and may be geographically restricted due to the evolutionary processes (Pritchard, 2001; Pritchard and Cox, 2002). Therefore, the human genetic variations hold great potential towards better understanding of the differential susceptibility to disease, response to drugs and complex interaction of genetic and environment factors (Collins *et al*., 2003; Campbell and Tishkoff, 2008).

Researches on genetic data of autosomal variations, mitochondria DNA (mtDNA) and Y chromosome have shown that human populations geographically cluster at continental level (Rosenberg *et al*., 2002; Cavalli-Sforza and Feldman, 2003; Bamshad *et al*., 2004). In general, about 75-85% of genetic variation is within local populations, about 7-12% is between local populations within the same continent and another 8-13% of variation occurs between large groups living on different continents (Lewontin, 1972;

Jorde *et al*., 2000; Hinds *et al*., 2005). Moreover, the global patterns of human genetic variation from worldwide populations has been described as consistent with Isolation-by-distance (IBD) pattern and have supported the Out of Africa model, with the human diaspora originated in East Africa (Bowcock *et al*., 1994; Prugnolle *et al*., 2005; Ramachandran *et al*., 2005; Gonder *et al*., 2007; Li *et al*., 2008).

Based on genetic data integrated with insights from geography, ecology, archaeology, physical anthropology, and linguistics, the Out of Africa model illustrates a recent common ancestor for all human populations that evolved in Africa ~200,000 years ago (Cavalli-Sforza *et al*., 1994). According to the model, the initial population migrated out of Africa ~100,000 years ago with the subsequent expansions spreading and diversified throughout the world. The expansion into Australo-Melanesia was believed to happen ~60,000 years ago, through India and Southeast Asia (The HUGO-PASNP Consortium, 2009). Studies on mtDNA have shown that India has the second highest genetic diversity in the world as it was the first major settling point for modern humans after the initial migration from Africa (Thangaraj *et al*., 2006; Maji *et al*., 2009; Thangaraj *et al*., 2009). Map of the migration of modern humans out of Africa based on mtDNA is shown in **Figure 1.2**.

Analogous to the model is the analyses of various genetic data that have suggested the serial founder effects with decreasing of heterozygosity and slope of the ancestral allele frequency spectrum along with increasing of linkage disequilibrium (LD) in populations away from Africa (Tishkoff and Verrelli, 2003; Jakobsson *et al*., 2008; Li *et al*., 2008;

Deshpande *et al*., 2009). The population bottleneck event at the time of initial migration out of Africa was showed by numerous studies of autosomal SNPs, microsatellites, X chromosome and mtDNA variations. The studies indicate that African have more heterozygosity and larger amounts of population-specific alleles than non-African that carry only a subset of the diversity from Africa (Quintana-Murci *et al*., 1999; Gabriel *et al*., 2002; Verrelli and Tishkoff, 2004; Hunley *et al*., 2009; Tishkoff *et al*., 2009).

The population bottleneck has reduced the population size and increased genetic drift that caused loss of genetic variations in the migrated population. This has lead to flatten spectrum of ancestral allele frequency and greater LD in population with increasing geographic distance from Africa (Reich *et al*., 2001; Kidd *et al*., 2004; Jakobsson *et al*., 2008; Li *et al*., 2008). Furthermore, the levels of LD which referred to the non-random association of alleles at two or more loci on the same or different chromosome may vary within geographic region as it is also influenced by selection, rates of mutation and recombination as well as population structure and admixture (Tishkoff and Williams, 2002; Plagnol and Wall, 2006).

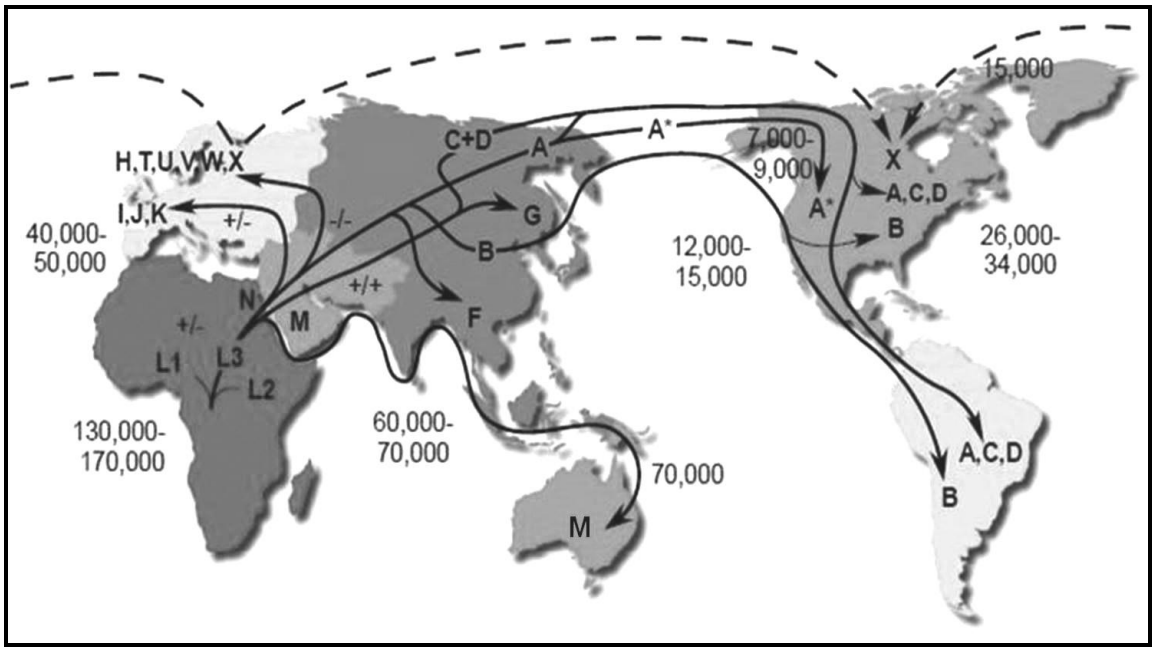

Figure 1.2: Map of the migration of modern humans out of Africa, based on Mitochondrial DNA (mtDNA). Capital letters represented mtDNA haplogroups. Numbers indicate years before present. Symbols of +/-, +/+ or -/- are the status of Dde I 10394 or Alu I 10397 markers and * for Rsa I 16329 marker. Image was taken from http://www.mitomap.org/pub/MITOMAP/MitomapFigures/WorldMigrations.pdf

1.2.3 Population genetic structure and ancestry

Genetic structure is the pattern of the genotype or allele frequency of individuals within a population. Whereas genetic ancestry is a genetic concept that describes the pattern of genome variation between populations that referred to the fraction of genome inherited from common ancestor (Kittles and Weiss, 2003).

Population structure or population stratification is the presence of subpopulations within a population that was caused by a systematic difference in allele frequencies. In the study of human population genetics, the genetic structure of a population, describes how the relative numbers of people with similar characteristics within a population affect its composition and characteristics. By studying the gene frequency dynamics for that particular group of people, the likely patterns of genetic variation in actual populations could be concluded and tested against empirical data. It will provide information on the change and distribution of allele frequency among populations (Crow and Kimura, 1972). The population structure analysis is important in understanding the human demographic and evolutionary history, forensics applications and medical applications mainly in genetic association studies to avoid false disease-risk allele associations (Rosenberg *et al*., 2002; Ray *et al*., 2005; Tang *et al*., 2005; Copper *et al*., 2008).

The population structure corresponds largely to geographic origin and ancestry (Jorde and Wooding, 2004). It is due to non-random mating between groups, usually limited to geographic region and localized endogamy. Individuals in the same region and speak

the same language within the same ethnicity tend to mate each other than with individuals in different ethnic groups. This assortative mating reduces gene flow and increases the genetic distance between the groups, resulting in genetic drift of allele frequencies in each group and the development of sub-structures in population. In the worldwide populations, the observed population structure can be largely explained by cumulative effect of random genetic drift at neutral loci (Li *et al*., 2008; DeGiorgio *et al*., 2009), although some of the genome may diverge due to local selection in adapting various habitats and climates as the modern human colonized the world (International HapMap Consortium, 2005).

Nowadays, advances in mobility, education and lifestyle have decreased the mating barriers and increased population structure with admixture of individuals from different ancestries for several generations (e.g. African-Americans and Latinos) (Rudan *et al*., 2008). The admixture reduces the average of genetic distance between the mixing populations and in this situation self-reported ancestry may be insufficient for assessing population stratification in disease studies, especially with recently admixed populations (Rosenberg *et al*., 2002; Tang *et al*., 2005). The genetic admixture along with recombination events could generate a new population with the genome of each individual is fragmented into shorter regions that originated from different ancestral populations. Thus, the inference of population structure in admixed populations is very challenging, especially in finding ancestral information of the admixed individuals (Sankararaman *et al*., 2008).

The identification of genetic ancestry in the admixed populations is crucial for admixture mapping technique as it can be used to identify loci in complex disease risk that may be at high frequency in one of the admixed populations (Zhu *et al*., 2004). The admixture mapping has greater statistical power than family-based linkage analysis and requires fewer markers than standard genetic-based association studies (McKeigue, 2005). The proportion of genetic ancestry of individual or population can be correctly estimated from various continental populations by analyses of large numbers of loci with proper evolutionary assumptions (Shriver *et al.*, 2003; Bamshad *et al.*, 2004). Several methods have been proposed to estimate the ancestry proportions in a population that incorporated with the likelihood function effects of genetic drift and other evolutionary forces such as coalescent-based approach, Bayesian, principle component analysis and many more (Wang, 2003; Price *et al*., 2006; Sankararaman *et al*., 2008; DeGiorgio *et al*., 2009).

1.2.4 Single nucleotide polymorphism (SNP)

A Single Nucleotide Polymorphism (SNP; pronounced "snips") is a variation of single nucleotide (A, C, G or T) in DNA sequences between individuals or paired chromosomes in an individual. SNPs are the most abundant variation throughout human genome with 90% of the total human variations constitute the common SNPs (Kruglyak and Nickerson, 2001; Reich *et al*., 2003; International HapMap Consortium, 2003) and the remaining 10% are known as the rare variants. In the latest release of NCBI dbSNPs Build 132, there are more than 10 million SNPs in the human species, of which 4.5

million are already validated. SNPs can occur in both coding and non-coding regions of genes with most of them are neutral or no known function but useful for pinpointing disease genes and some may affect function of genes or its expression (Jorde *et al*., 2004). The growing interest in identifying complex disease causing variants at the whole genome level has revealed the SNPs as powerful marker of choice for genome wide association studies (Tsuchihashi and Dracopoli, 2002).

At the beginning of new millennium, the analysis of SNPs on coalescent theory has been appropriately founded and developed in the field of population genetics (Kuhner *et al*., 2000; Nielson, 2000). Since then, the potential of SNPs in the study of population demographic and evolution history briskly supersede other classes of genetic markers (Lao *et al*., 2006; Norrgard and Schultz, 2008). SNPs can be the basis of evolutionary change as any observed variation of SNPs that favored by natural selection, can be followed over time and quantified (Barreiro *et al*., 2008). This is because SNPs are mutationally stable with a highly stable inheritance unlike repeat polymorphisms, which could cause bias estimation to inheritance analysis (Risch, 2000). Furthermore, relative ease of scoring for many SNPs loci is in line with the need of many unlinked loci to estimate population genetic parameters with statistical confidence using appropriate evolutionary assumptions (Nielson, 2000; Hare, 2001; Brumfield *et al*., 2003; Bamshad *et al*., 2004).

In the study of population ancestry, the used of SNPs multi-locus genotyped data has overpowered the sex-specific markers of mitochondrial DNA (mtDNA) and Y

chromosome (Y-DNA) from maternally and paternally inherited, respectively (Petkovski *et al*., 2003). Transmission of SNPs are sex-unspecific and have larger effective population size compared to mtDNA and Y-DNA that are haploid and lack of recombination (Falush *et al*., 2003; Templeton, 2007). As the genetic admixture becoming increasingly common within geographic region (Tishkoff and Kidd, 2004; Katarzyna *et al*., 2010), the SNPs have greater power of inferences than the mtDNA and Y-DNA.

1.3 High-throughput genotyping technology

High-throughput genotyping means the capability to characterize a very large number of SNPs (thousands or more) across the genome in individuals. The technology emerges in the past ten years and has rapidly evolved to provide relevance to human historical demography and the study on complex diseases (Kwok and Chen, 2003; Dalma-Weiszhausz *et al*., 2006). There are various types of high-throughput technology to screen the SNPs such as Matrix Assisted Laser Desorption/Ionization-Time of Flight (MALDI-TOF) mass spectrometry-based systems (Bray *et al*., 2001), TaqMan (Latif *et al*., 2001), Invader assay (Hsu *et al*., 2001), Pyrosequencing (Ronaghi, 2001) and many more. Some of these assays such as MALDI-TOF and Pyrosequencing could attain genotyping up to more than 10,000 SNPs a day, but the rest are lacking of potential for a high level of multiplexing (Tsuchihashi and Dracopoli, 2002).

Recently, a new advance feature of genotyping technology called SNP array has easily usurped other assays in the SNPs high-throughput field. The SNP array is a type of DNA microarray that consists of an arrayed series of thousands of microscopic spots. The spots are DNA oligonucleotides, known as probes on a surface of a glass or silicon chip. Each of the probes contains 10^{-12} moles of a specific DNA sequence and is properly designed from raw sequences and SNPs of multiple databases. The target SNPs are interrogated by multiple probes on the chip and each of the probes specifically hybridized to a particular SNP. The hybridization signal which is detected by a scanner system determines the genotyping call rate. The SNP array can be differ in aspect of accuracy, efficiency, and cost depending on its fabrication by the manufactured company as well as study design and analyzing methods (Mei *et al*., 2000; John *et al*., 2004; Hehir-Kwa *et al*., 2007).

The principal technology platform for this study is the Affymetrix GeneChip Mapping Xba 50K Array. A SNP array chip that simultaneously genotype and screen more than 50,000 SNPs loci in each individual. It is a genome-wide scan method that can maximize the efficient screening of the functional SNPs with respect to accuracy, speed and cost (Dalma-Weiszhausz *et al*., 2006; Zhou and Wong, 2007). Recent survey has proved that the Affymetrix GeneChip technology is the most preferred high-throughput technology for analyzing complex genetic information and other genomic application growth areas (Frost and Sullivan, 2011).

1.4 Types of data analysis

The analyses of SNPs microarrays data is the most important and challenging part which poses a large number of statistical problems (Hiekkalinna and Talikota, 2005; Stokes, 2007). Two statistical analysis approaches for genetic clustering, distance-based and model-based are used to conquer the challenge. Both distance-based and model-based approaches have their own limitations but both methods have great potentials in population structure specifically and generally in population genetics (Handley *et al*., 2007; Rodriguez-Ramilo *et al*., 2009). Therefore, to achieve the aims of this study, both distance-based and model-based clustering approaches were implemented in multiple analyses. The most prominent ones is phylogeny analysis as it can be applied on both of the approaches.

1.4.1 Distance-based clustering approach

The distance-based methods that are used in this analysis can detect fine-scale population structure of studied populations and are not computationally demanding compared to model-based approaches (Gao and Starmer, 2007; Li *et al*., 2010). Despite conceptually simple as necessitate almost similar statistics algorithm for many polymorphisms, the method has proven powerful for couples of population genetics studies. For instance, Young *et al*. (2005) has successfully identified susceptible SNPs to hypertension from five genes involved in blood pressure regulation in population samples of Human Genome Diversity Project-Centre d'etude du Polymorphisme