The genetic diversity and population history of indigenous peoples in Asia

Timothy Adrian anak Joseph Jinam

DOCTOR OF PHILOSOPHY

Department of Genetics,

School of Life Science,

The Graduate University for Advanced Studies (SOKENDAI)

2011

Acknowledgements

First and foremost I would like to thank my academic supervisor, Naruya Saitou for his expert advice and support throughout the duration of my time in Japan. I would also like to thank members of my progress committee: Toshiyuki Takano, Takashi Gojobori, Hiroshi Akashi, Nori Kurata and Ituro Inoue for their useful advice and constructive criticisms. I am also grateful to Katsushi Tokunaga and Nao Nishida from the University of Tokyo for their assistance and collaboration for this project.

I would also take this opportunity to thank Mark Stoneking and members of his group at the Max Planck Institute for Evolutionary Anthropology in Leipzig for being such graceful hosts and for their valuable input to my research. Thanks also to SOKENDAI for making it all happen and also to members of the Educational Support Team in NIG for your patience and support. I also thank friends and collaborators in Malaysia, especially Maude Phipps, Mahmood Ameen, Juli Edo and Hong Lih Chun.

I would also like to acknowledge the Ministry of Education, Culture, Sports, Science & Technology in Japan (MEXT) for giving me the opportunity to pursue my doctorate degree in Japan. I also extend my gratitude to SOKENDAI for providing the research grant (SOKENDAI Strategic Research Project) for me to conduct my research smoothly.

Thanks also to all members of Saitou Lab, past and present, for all your valuable support and friendship. Finally I wish to thank my family and friends (wherever they may be) for their endless support and encouragement. Thank you and *terima kasih*.

TABLE OF CONTENTS

List of figures

List of tables

Abbreviations

bp:	Base pairs
BSP:	Bayesian Skyline Plot
ML:	Maximum-Likelihood
mtDNA:	Mitochondrial DNA
NJ:	Neighbor-Joining
PCA:	Principal Component Analysis
PCR:	Polymerase Chain Reaction
SNP:	Single Nucleotide Polymorhism
TMRCA:	Time to the Most Recent Common Ancestor
YBP:	Years Before Present

ABSTRACT

Asia is home to many diverse human populations and has been of much interest to anthropologists and geneticists alike. The focus of this thesis is the genetic diversity and migration histories of indigenous populations from Southeast Asia and the Japanese Archipelago. Based on archaeological and linguistic data, the origins of Southeast Asians can be traced back to two major migrations; the ancient 'Out of Africa' migration circa 50,000 years before present (YBP) and the relatively recent 'Out of Taiwan' expansion of Austronesian agriculturalists approximately 5,000 YBP. In Malaysia and the Philippines, there are indigenous groups called Negritos whose physical appearance is distinct from their neighboring populations and are thought to have originated from the 'Out of Africa' migration. The majority of other Southeast Asian groups are thought to have originated from the 'Out of Taiwan' migration. As for the peopling of the Japanese archipelago, it is characterized by two important periods: the Jomon period from 15,000 to 3,000 YBP; and the Yayoi period from 3,000 to 1,700 YBP. According to the dual-structure model, the current Japanese population was the result of admixture between Jomon populations who originated from Southeast Asia and the incoming Yayoi migrants from mainland Asia. Some minority groups in Hokkaido and the Ryukyu islands may be direct descendants of the Jomon peoples.

By analyzing approximately 50,000 genome-wide SNP data generated by me and other Pan-Asian SNP Consortium (PASNP) members, I investigated the genetic structure that may exist within indigenous groups of Malaysia and elucidated their relationship with other Southeast Asians. Using Principal Component Analysis (PCA) and STRUCTURE analysis, I found traces of recent and sustained admixture within the Negrito, Indian and Austronesian (Temuan, Bidayuh, Malay) groups. Comparisons with other Southeast Asians revealed that the Malaysian Negritos are distinct from the Philippine Negritos, putting doubt into their common origins as proposed by the ‘Out of Africa’ model. A closer look at the relationship between Austronesian populations revealed genetic substructure which mirrors geographical affinities, which may be explained by geographical isolation following the ‘Out of Taiwan’ expansion or alternatively there may be separate population movements involving other Austronesian groups. These observations demonstrate the impact of admixture on the genetic substructure of indigenous Southeast Asian groups and hints at a more complex migration history of the Negritos and Austronesians than the simple ‘Out of Africa’ and ‘Out of Taiwan’ models would suggest.

Next I conducted an analysis of complete mitochondrial DNA (mtDNA) sequences to test the plausibility, impact and timing of the migration models in indigenous Malaysian populations. I generated complete mtDNA sequences in 86 individuals from four indigenous Malaysian groups. In the Jehai (Negrito), one of the frequent haplogroups was R21 which is indigenous to West Malaysia and dates back to the Pleistocene (~40,000 YBP). The three Austronesian groups (Bidayuh, Selatar and Temuan) showed high frequencies of haplogroups N9a6, N9a6a, F1a’c, N21 and N22 which have mainland Asian origins around 30,000 to 10,000

YBP. Haplogroups associated with the ‘Out of Taiwan’ expansion were either found at very low frequencies or not detected at all in those three Austronesian groups. Principal Component Analysis distinguishes the Malaysian Negritos from the Austronesians and also shows a dichotomy between Austronesians from Sumatra and Java and those from Taiwan and Philippines. As with the SNP analysis, results from mtDNA showed no apparent link between the Negritos of West Malaysia and those from Andaman and Philippines, again putting in question their common origins from the ‘Out of Africa’ migration. Regarding the origin of Austronesians, our results show support for an ‘early train’ migration originating from Indochina or South China around 30,000 to 10,000 YBP which predates but does not rule out the subsequent ‘Out of Taiwan’ expansion.

Finally I conducted a study to find out the genetic structure in Japanese populations and to answer questions regarding which model of Japanese origins would be best supported by the genome-wide SNP data. I performed data analysis of close to 1 million genome-wide SNP genotypes generated using the Affymetrix 6.0 genechip in three Japanese populations: Hondo-Japanese, Ryukyuan and Ainu. Principal Component Analysis (PCA) plots showed that these three populations formed three distinct clusters, with greater genetic variation within individuals of the Ainu group, brought about by admixture with the mainland Japanese and possibly another population from Northeast Asia. Phylogenetic analysis revealed that the Ryukyuans and Ainu form a cluster with 100% bootstrap probability and comparisons with other

global populations showed that all three Japanese populations cluster with other North East Asians. Current results appear to support the common ancestry of Ainu & Ryukyuans, which is compatible with the dual-structure model. However, the close affinity of all three Japanese populations with other North East Asians put the idea of Jomon origins from Southeast Asia in doubt although not entirely ruled out.

In summary, my results demonstrate the influence of surrounding populations to the genetic diversity in indigenous Malaysian and Japanese populations which also contributes to the genetic substructure in these indigenous groups. The presence of admixed individuals has to be considered when designing sampling strategies for future population genetic studies as well as when conducting and interpreting results of association studies. Regarding the history and origins of Austronesians in Southeast Asia, results suggest an earlier movement originating from Indochina around 30,000 to 10,000 YBP which has more impact on the mtDNA diversity of indigenous Austronesians in West Malaysia and Borneo than the proposed 'Out of Taiwan' expansion around 5,000 YBP. As for the origins of the Japanese population, my data supports some aspects of the dual-structure model in that the Ainu and Ryukyuans have shared genetic ancestry and that the mainland Japanese are the result of admixture between ancestral Yayoi and Jomon peoples. However, our data does not indicate a Southeast Asian origin of Jomon peoples but shows a closer affinity to Northeast Asian populations.

CHAPTER 1

General introduction

1.1 Human population diversity in Asia

Asia is the world's largest and most populous continent, spanning 44 million square kilometers and includes countries from Turkey in the west to the Pacific Islands in the east. Such a vast continent naturally houses a multitude of human populations, each with their own language and culture. The scope of this thesis is narrowed down to the human populations in two areas of interest, namely Southeast Asia and East Asia.

Southeast Asia currently consists of 11 countries and can be classified geographically into mainland Southeast Asia (Thailand, Vietnam, Laos, Cambodia, Myanmar, West Malaysia) and island Southeast Asia (Indonesia, East Malaysia, Brunei, Singapore, Philippines, East Timor). Mainland Southeast Asia is also sometimes referred to as Indochina but in this thesis, Indochina will be used to refer to more limited region encompassing only Cambodia, Laos and Vietnam. The division between mainland and island Southeast Asia did not become apparent until after the Last Glacial Maximum around 20,000 years before present (YBP). Up to the Last Glacial Maximum, the current islands of Borneo, Sumatra and Java were joined with the Asian mainland in what is called Sundaland (Figure 1.1). It was separated from the Sahul landmass which was made up of Papua New Guinea and the continent of Australia, by a boundary called the Wallace line, named after Sir Alfred Russel Wallace (Glover and Bellwood 2004).

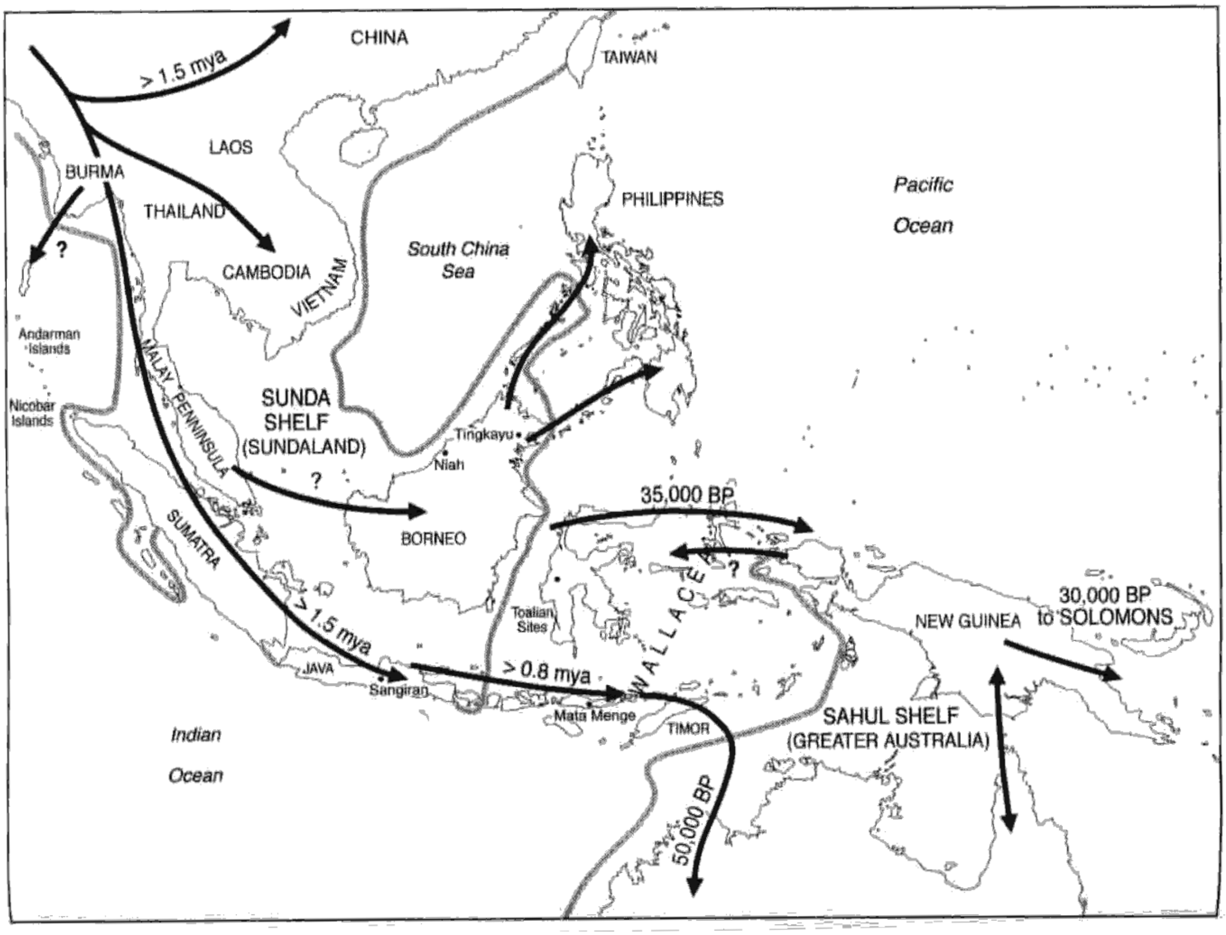


Figure 1.1: Migration routes into Southeast Asia based on archaeological data. Thick grey lines indicate the boundaries of the Sundaland and Sahul landmasses up to the Last Glacial Maximum (~20,000 YBP). Black arrows indicate proposed human movements during the Pleistocene period (>20,000 YBP) according to archaeological data. Image source: Glover and Bellwood, 2004.

The languages spoken in Southeast Asia can be classified into Tibeto-Burman, Tai-Kadai, Hmong-Mien, Austro-Asiatic and Austronesian. Of those, Austro-Asiatic and Austronesian are the two largest and most-spoken linguistic families in Southeast Asia. Austro-Asiatic languages are spoken mostly in Indochina and by some populations in India and West Malaysia whereas Austronesian languages are spoken not only in island Southeast Asia but also in Madagascar and the Pacific islands, making it the most widespread ethnolinguistic group in the world prior to the spread of the English language (Adelaar and Himmelmann 2005).

The populations in Southeast Asia are still dominated by groups who have a long association with the region despite a long history of contact with various civilizations and kingdoms from Arabia, India, China and even Europe. These majority groups such as the Thais, Khmers, Malays, Filipinos and Javanese traditionally practice agriculture (Bellwood 2005). Living amongst them are several indigenous minority groups such as the *Orang Asli* of West Malaysia, the Mlabri of Thailand or the Penan from Borneo who still practice hunter-gathering lifestyles even to this day. These minority groups are generally regarded to have been present in the region prior to the arrival of the agriculturalist societies. Of considerable interest to physical and cultural anthropologists are a group of hunter-gatherers collectively called Negritos. The term owes much to the fact that they exhibit physical features such as darker skin and frizzy hair that are distinct from the general surrounding population. Included in this blanket term are Negritos from Andaman Islands, West Malaysia and the Philippines (Carey 1976).

The Negritos in West Malaysia (also referred to as Semang) are included in a broader category called *Orang Asli*. The term *Orang Asli* translates to ‘Original Peoples’ in the Malay language and refers to about 18 distinct cultural-linguistic groups which are scattered throughout the rural and coastal regions of West Malaysia. Other subgroups include the Senoi and Proto-Malay (also referred to as Aboriginal Malay). The ancestors of the Senoi are thought to have originated from Indochina some 4,000 YBP and brought with them Neolithic cultures and introduced Austro-Asiatic languages to the Negritos (Bellwood 2005). The Proto-Malays and the various tribes from East Malaysia who speak Austronesian languages are generally associated with the Austronesian expansion from Taiwan (Blust 1995; Bellwood 2007).

East Asia is another subregion in Asia and includes China, Korea, Japan, Mongolia and Taiwan. In this thesis, I will mostly focus on the populations from the Japanese Archipelago. The Japanese Archipelago consists of the four main islands of Hokkaido, Honshu, Shikoku and Kyushu as well as the smaller cluster of Ryukyu Islands. During the glacial period (up to ~20,000 YBP) the three main islands of Honshu, Shikoku and Kyushu were connected by land bridges (Figure 1.2) whereas the gap between Hokkaido and Honshu was not deemed enough of an impediment to prevent movements of fauna or humans (Imamura 1996). The majority of the current Japanese (referred to as mainland Japanese or Hondo Japanese) are believed to be descendants of agriculturalists from the mainland but there also exists indigenous minority groups known as the Ryukyuans from the Ryukyu Islands and the Ainu from Hokkaido. The Ainu in particular exhibit physical features that are unique from the Hondo Japanese or even the Ryukyuans, leading to various theories and speculation regarding their origins. Unlike the Hondo Japanese who mainly practice agriculture, the Ainu have until recently maintained their traditional hunter-gathering lifestyles.

It is evident that even in these two sub regions of Asia there exists a rich diversity of human populations and naturally one would be curious regarding their origins and history. The following sections will attempt to make a brief introduction to the human histories in Asia from archaeological, linguistic and genetic perspectives.

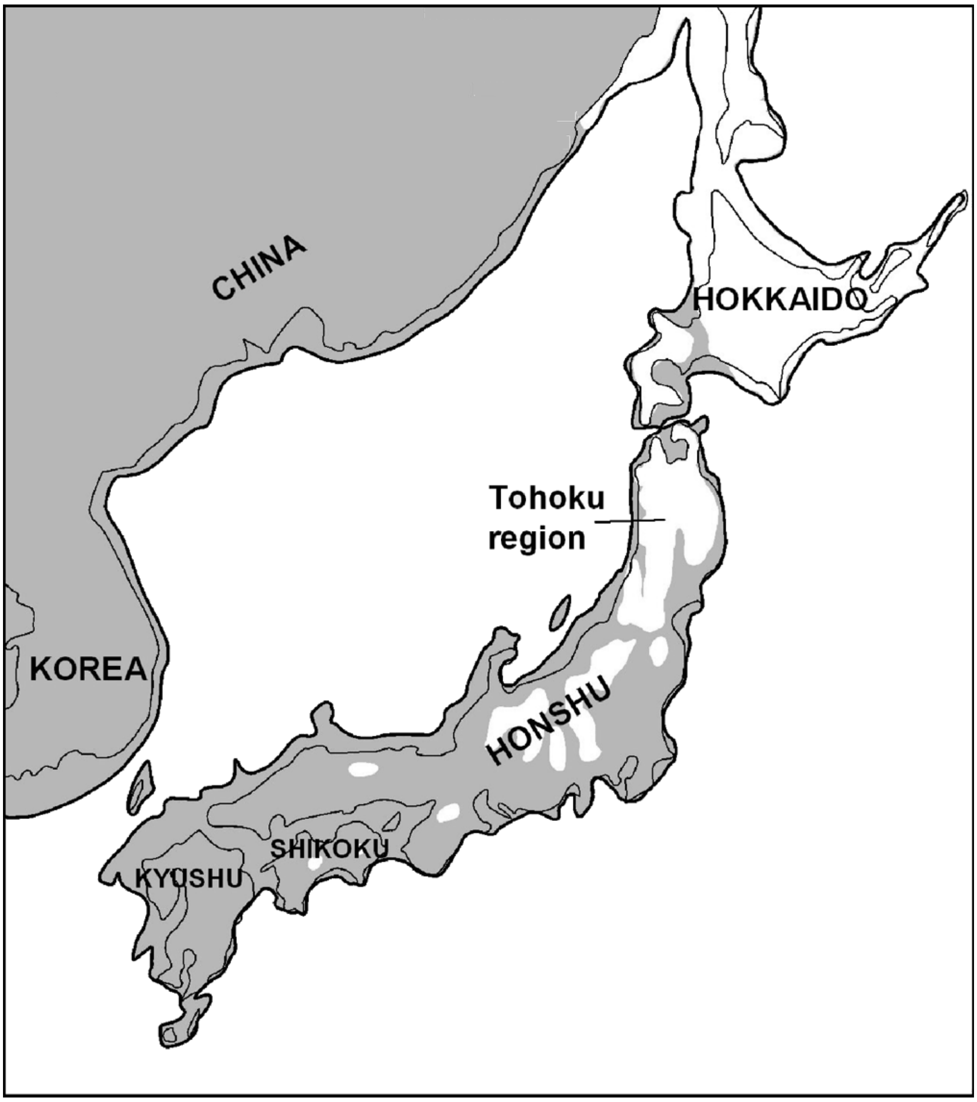


Figure 1.2: Geographical map showing the main islands in the Japanese Archipelago circa 20,000 YBP. Not shown are the Ryukyu Islands further south of the main islands. Thick black lines show the shoreline boundaries up to the Last Glacial Maximum (~20,000 YBP). Image modified from (Davison et al. 2005)

1.2 Human migrations into Asia: from archaeology, linguistics to genetics

With regards to the origins of anatomically modern humans, the generally accepted opinion is the Out of Africa model which posits the origin of *Homo sapiens* in Africa some 200,000 YBP and an eventual dispersal to other parts of the world, replacing the indigenous archaic humans. The earliest dispersal around 60,000 YBP was thought to have taken a southern, coastal route via India and Southeast Asia before eventually reaching Sahul. Dating of various archaeological specimens from Australia, New Guinea (Leavesley and Chappell 2004; O' Connell and Allen 2004) as well as from Borneo (Barker et al. 2007) was taken as supporting evidence for this early southern dispersal. This initial migration wave was thought to be responsible for the origins of the Negritos as well as Melanesians and Australian Aboriginals.

For the majority of Southeast Asian populations, their ancestry can be traced back to the so-called Austronesian migration, which was named as such because it accounts for the origins of most, if not all, populations who speak the Austronesian family of languages. As mentioned previously, the spread of Austronesian speakers ranged from Madagascar off the coast of Africa to the Asia Pacific islands. Such an expansive migration was driven by a great innovation at the time, which was agriculture, thought to have been developed in China (Bellwood 2005). The origin of this Austronesian migration was pinpointed back to Taiwan, based on language phylogenies which placed Austronesian languages spoken by the Taiwanese aboriginals at the root and all other subgroups are derived from this ancestral form (Diamond 1988; Gray and Jordan 2000; Adelaar and Himmelmann 2005). Archaeological and anthropological observations estimate the timing of this diaspora to the Neolithic or mid-Holocene period, approximately 5,000 to 7,000 YBP (Glover and Bellwood 2004; Bellwood 2007). Starting from Taiwan, these Neolithic agriculturalists spread south via the Philippines and forked westwards to island Southeast Asia and eastwards to the Asia Pacific islands.

Regarding the migration history of East Asia, in particular Japan, archaeological evidence points to the presence of anatomically modern humans around 40,000 YBP based on dating of the earliest stone tools found. The oldest human remains found in Japan were dated back to 30,000 YBP, corresponding to the upper Paleolithic period (Imamura 1996; Hudson 2006; Yoneda M, personal communication). The time period starting from 15,000 YBP marks the start of the Jomon period. The origins of the Jomon people are still debatable, but early archaeological studies suggest an origin in Southeast Asia based on similarities in cranial and dental morphologies from Jomon archaeological samples and extant Ainu/Ryukyuu with Southeast Asian populations (Hanihara 1991). The Jomon people settled most parts of the Japanese Archipelago until the emergence of agriculturalists during the Yayoi period approximately 3,000 YBP (Haruhari and Imamura 2004). Although the exact origins of the Yayoi people are still unknown, they are believed to have migrated to the Japanese Archipelago from the mainland via the Korean peninsula. Their interactions with the extant Jomon people formed the basis of several proposed models for the origins of the current Japanese population.

In general, population histories put forward by archaeological and linguistic data have been supported by genetic data, though in some cases some ambiguities remain. Early studies using ‘classical’ protein markers were consistent with the archaeological point of view regarding the origins of modern humans in Africa and the early southern route taken by the ancestors of Australian Aboriginals and Melanesians (Ruiz-Linarez et al. 1995; Cavalli-Sforza and Feldman 2003). The maternally inherited mitochondrial DNA (mtDNA) has also been used in support of this opinion, in which all human mtDNA lineages coalesce back to Africa 170,000 YBP (Ingman et al. 2000). The early southern dispersal via the coast was supported by mtDNA analysis whereby the basal M and N lineages found in India and indigenous populations in Southeast Asia date back to around 60,000 YBP (Macaulay et al. 2005; Majumder 2010). Recent advancements in Single Nucleotide Polymorphism (SNP) genotyping technology allows for the generation of up to hundreds of thousands of genome-wide SNP to be used for population genetics analysis, among other applications. One such study by Li et al. (2008) provided further support for the Out of Africa model by showing a serial reduction in SNP haplotype diversity with increasing distance from Africa. Another study by the Pan-Asian SNP Consortium (Abdulla et al. 2009) suggested a single entry into Asia followed by a south to north migration in a model which unites the histories of Southeast and East Asian populations. With regards to the Austronesian expansion, genetic data has been equivocal in the support of archaeological and linguistic studies. While some mtDNA analyses provided support for the Austronesian expansion from Taiwan (Trejaut et al. 2005; Tabbada et al. 2009), others proposed an origin in island Southeast Asia (Oppenheimer and Richards 2001). Although several Southeast Asian populations were included in the Pan-Asian SNP paper, no explicit inference was made regarding the Austronesian expansion.

As for the origins of the Japanese populations from a genetic standpoint, analyses using classical protein and blood group markers pointed to close affiliations between the Ainu and Ryukyuans suggesting a common origin during the Jomon period (Omoto 1983, 1995; Nei 1995; Omoto and Saitou 1997). However, none of those studies demonstrated any links between the Ainu and Ryukyuans with Southeast Asian populations, contradicting the idea proposed by archaeological and anthropological data. While the uniparentally inherited mtDNA and Y-chromosome markers were rather ambiguous regarding the origins of the Jomon people, they all conclude that the Hondo Japanese experienced considerable influence from the mainland, in line with the emergence of Yayoi agriculturalists. Recent analysis of genome-wide SNP mostly involved the Ryukyuans and Hondo Japanese (Yamaguchi-Kabata et al. 2008; Abdulla et al. 2009) but no such data was available for the Ainu.

1.3 General goals & organization of dissertation

Previous studies into the origins and genetic diversity of Asian populations have yielded equivocal results, with some conclusions supporting the earlier models proposed by archaeological and linguistic data while others offered differing views and possible alternative models regarding the topic. This thesis will focus on the indigenous populations from two subregions in Asia, namely Southeast Asia and East Asia and will attempt to provide insights to the migration histories and genetic diversities using genome-wide SNP and complete mtDNA sequence data.

Chapter 2 will focus on the analysis of 50,000 genome-wide SNP in the Malaysian population and other Southeast Asians with the aim to investigate in further detail the substructure that may exist within indigenous groups of Malaysia and to elucidate their relationship with other Southeast Asian populations. Chapter 3 will report the in-depth analysis of complete mtDNA sequences from four indigenous groups in Malaysia. By exploring the diversity of the maternally inherited mtDNA in these groups and comparing them with other populations within the Southeast Asian region, I hope to shed light on some questions regarding the timing and impact of the proposed human migrations in the region. Chapter 4 will reveal in closer detail the genetic substructure within Japanese populations using more than 500,000 genome-wide SNP with the aim to answer questions regarding which model of Japanese origins would be best supported by the genome-wide SNP data.

CHAPTER 2

Genetic substructure in Malaysian populations and relationships with other Southeast Asians estimated from genome-wide SNP data

2.1 Introduction

The populations in Malaysia typify the diversity found in the Southeast Asian region, with a mix of indigenous groups and those with ancestry tracing back to China and India. The majority of the Malaysian population consists of the indigenous Malay, Chinese and Indian populations and is also the case for Singapore. Although contact with traders from India and China dates back to at least the 15th century, it was only during the British colonial period, the late 19th century, which saw a massive influx of migrants predominantly from South India and South China (Andaya and Andaya 1984). Indigenous minority groups in Malaysia consist of the *Orang Asli* in the Malay Peninsula and various ethnic tribes in Borneo. Within this major group of *Orang Asli*, there exist some subgroups that are identified as Negritos who are thought to be descendants of the earliest migrants to the Southeast Asian region (Macaulay 2005; Hill et al. 2006). Other subgroups of the Orang Asli, the various tribes in Borneo as well as the Malays may have originated from the Austronesian expansion during the Neolithic period (Glover and Bellwood 2004; Bellwood 2005; Bellwood 2007).

Until recently, not much was known regarding the population genetic structure and diversity of these various indigenous groups, given their fairly reclusive nature. However, that trend has been changing with studies reported using mitochondrial DNA (Macaulay 2005; Hill et al. 2006; Hill et al. 2007), Y-chromosomal markers (Chang et al. 2009) and autosomal genetic markers (Dhaliwal et al. 2010; Jinam et al. 2010). Most recently, genome-wide SNP analyses on Asian populations, including several indigenous Malaysian groups have been reported (Abdulla et al. 2009). However, detailed analysis of Southeast Asian populations was not reported explicitly in that study.

Thus, the aims of this chapter are to elucidate the detailed substructure that may exist within indigenous groups of Malaysia which include the Negritos and Austronesians in this study and to infer their relationships with other Southeast Asian populations.

2.2 Materials & Methods

2.2.1 Ethical approval & data retrieval

DNA samples from four indigenous Malaysian populations (Jehai, Kensiu, Temuan and Bidayuh) were collected as part of my Master's project (Jinam 2007). I performed the SNP genotyping experiments for those samples at the Genome Institute of Singapore under the guidance of Professor Mark Seielstad as part of the Pan-Asian SNP (PASNP) project (Abdulla et al. 2009). All samples were genotyped using the Affymetrix Xba 50k Genechip microarray and the workflow is shown in Figure 2.1. Briefly, 50ng of DNA for each sample was digested using *Xba I* restriction enzyme. The digested DNA fragments were then ligated with adapter molecules which would act as annealing sites for the primers in the subsequent PCR step. The PCR products were then pooled, purified and then fragmented to yield ~50bp fragments. The fragments were then labeled and hybridized onto the genechip array which contains oligonucleotide probes. Following hybridization, the genechip was then washed, stained and finally scanned to obtain the raw intensity files which were then converted to 58,960 genotype calls per sample.

Genotype data from other Southeast Asian populations were also retrieved from the PASNP database (http://www4a.biotec.or.th/PASNP). Those samples were also genotyped using the same Affymetrix Xba 50k Genechip microarray. Initial filtering of the data yielded 54,794

SNP genotypes. The list of the populations used in this analysis and their corresponding linguistic and geographical information are shown in Table 2.1 and Figure 2.1. This study involves the use of digital genotype data which was generated by me and other PASNP collaborators. It has been approved by the ethical committee board of the National Institute of Genetics, Mishima.

2.2.2 Data filtering and quality control of SNP genotypes

From an initial number of 54,794 SNP, further filtering was performed to exclude SNP with call rates of less than 95% and minor allele frequencies less than 0.1% loci using PLINK software (Purcell et al. 2007), resulting in a final tally of SNP used for subsequent analysis to be 51,585.

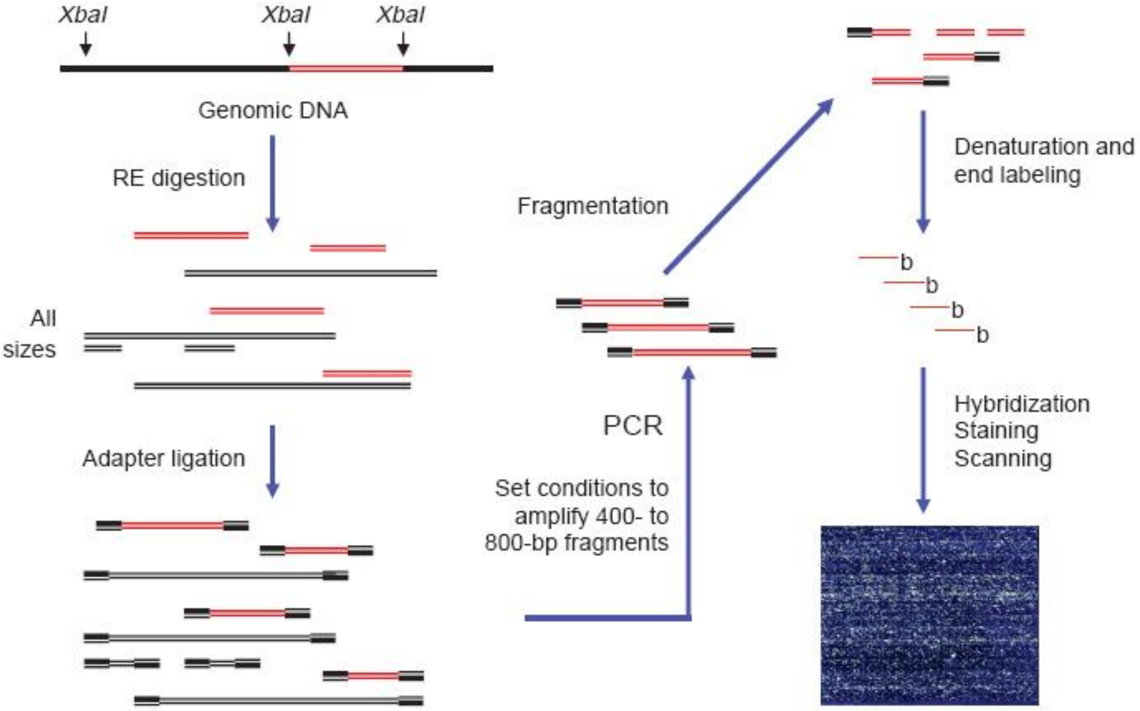


Figure 2.1: Overview of the experimental procedures for the GeneChip Mapping 50K Xba Assay (Image sourced from the Affymetrix product brochure).

Table 2.1: Pan-Asian SNP (PASNP) population information used in this study

Ethnicity	PASNP ID	Linguistic group	Geographical location	n
Jehai (Negrito)	MY-JH	Austro-Asiatic	West Malaysia	50
Kensiu (Negrito)	MY-KS	Austro-Asiatic	West Malaysia	30
Temuan	MY-TM	Austronesian	West Malaysia	49
Bidayuh	MY-BD	Austronesian	East Malaysia (Borneo)	50
Malay (KN)	MY-KN	Austronesian	West Malaysia	18
Malay (MN)	MY-MN	Austronesian	West Malaysia	20
Malay (SG)	SG-MY	Austronesian	Singapore	30
Chinese (SG)	SG-CH	Sino-Tibetian	Singapore	30
Indian (SG)	SG-ID	Dravidian	Singapore	30
Ami	AX-AM	Austronesian	Taiwan	10
Atayal	AX-AT	Austronesian	Taiwan	10
Melanesian	AX-ME	Papuan	Melanesia	5
Alorese	ID-AL	Austronesian	Indonesia (Nusa Tenggara)	19
Dayak	ID-DY	Austronesian	Indonesia (Borneo)	12
Javanese	ID-JA	Austronesian	Indonesia (Java)	34
Javanese	ID-JV	Austronesian	Indonesia (Java)	19
Batak Karo	ID-KR	Austronesian	Indonesia (Sumatra)	17
Lamaholot	ID-LA	Austronesian	Indonesia (Nusa Tenggara)	20
Lembata	ID-LE	Austronesian	Indonesia (Nusa Tenggara)	19
Malay	ID-ML	Austronesian	Indonesia (Sumatra)	12
Mentawai	ID-MT	Austronesian	Indonesia (Sumatra)	15
Manggarai	ID-RA	Austronesian	Indonesia (Nusa Tenggara)	17
Kambera	ID-SB	Austronesian	Indonesia (Nusa Tenggara)	20
Manggarai	ID-SO	Austronesian	Indonesia (Nusa Tenggara)	19
Sundanese	ID-SU	Austronesian	Indonesia (Java)	25
Batak	ID-TB	Austronesian	Indonesia (Sumatra)	20
Toraja	ID-TR	Austronesian	Indonesia (Sulawesi)	20
Agta (Negrito)	PI-AE	Austronesian	Philippines	8
Aeta (Negrito)	PI-AG	Austronesian	Philippines	8
Ati (Negrito)	PI-AT	Austronesian	Philippines	23
Iraya	PI-IR	Austronesian	Philippines	9

Manobo	PI-MA	Austronesian	Philippines	18
Mamanwa (Negrito)	PI-MW	Austronesian	Philippines	19
Urban	PI-UB	Austronesian	Philippines	20
Urban	PI-UI	Austronesian	Philippines	20
Urban	PI-UN	Austronesian	Philippines	19
Hmong	TH-HM	Hmong-Mien	Thailand	20
Karen	TH-KA	Sino-Tibetian	Thailand	20
Lawa	TH-LW	Austro-Asiatic	Thailand	19
Mlabri	TH-MA	Austro-Asiatic	Thailand	18
Mon	TH-MO	Austro-Asiatic	Thailand	19
Palong	TH-PL	Austro-Asiatic	Thailand	18
Plang	TH-PP	Austro-Asiatic	Thailand	18
Tai-Kern	TH-TK	Tai-Kadai	Thailand	18
Tai-Lue	TH-TL	Tai-Kadai	Thailand	20
H'Tin	TH-TN	Tai-Kadai	Thailand	18
Tai-Yuan	TH-TU	Tai-Kadai	Thailand	20
Tai-Yong	TH-TY	Tai-Kadai	Thailand	18
Yao	TH-YA	Hmong-Mien	Thailand	19

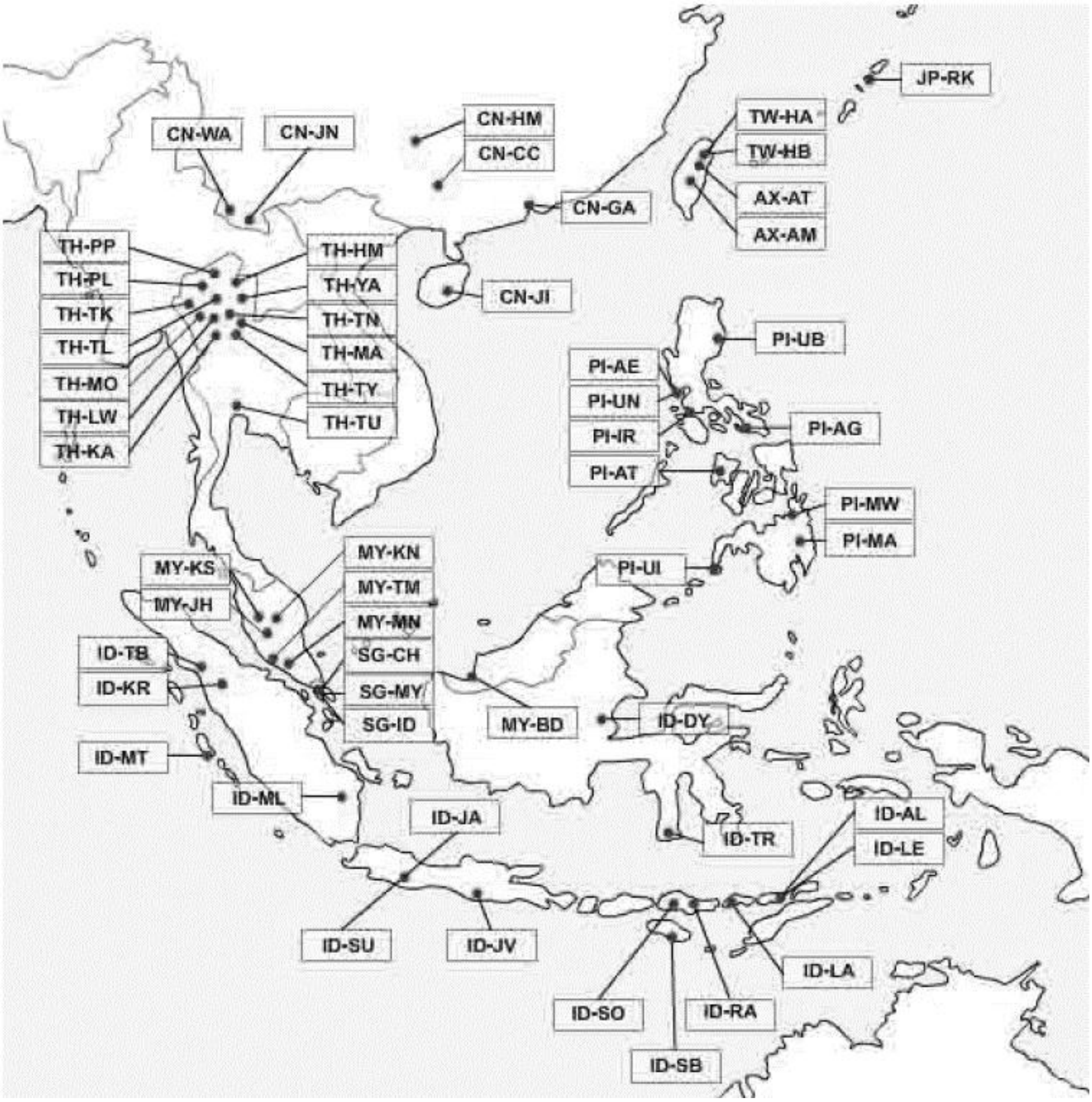


Figure 2.2: Geographical locations of PASNP populations from Southeast Asia (from Abdulla et al. 2009, supplementary material)

2.2.3 Data analysis

Genetic distances (Fst) between Malaysian populations were calculated for each SNP according to Weir and Cockerham (1984). Pairwise Fst distances between pairs of populations were obtained by averaging the Fst values for all SNPs. Principal Component Analysis (PCA) was done to assess relatedness between individuals using the *smartpca* program from the EIGENSOFT software package (Patterson, Price, and Reich 2006). In order to identify population structure and levels of admixture within individuals, a Bayesian clustering method implemented in the STRUCTURE software was used (Pritchard, Stephens, and Donnelly 2000). STRUCTURE assigns individuals based on their genotypes into a user-defined number of ancestral populations, denoted as k. Under the admixture model, individuals who are jointly assigned to two or more ancestry components are considered to be admixed. Burn-in length and number of repeats were both set to 10,000. A maximum-likelihood approach to identifying admixture and population structure as implemented in the *frappe* software (Tang et al. 2005) was also used.

2.3 Results

2.3.1 Genetic substructure in Malaysian populations

Pairwise Fst distances between populations are shown in Table 2.2. The greatest genetic distances were between the Negritos and Indians (0.06) whereas the distance among the three Malay groups was the lowest (0.01). The three Malay groups also had the lowest distance from the Indians (0.03) compared to an average 0.05 for other populations, and this may reflect their Indian ancestry as suggested in STRUCTURE analysis. Omitting recently admixed individuals from the Indian-SG, Malay-SG, Bidayuh and Negritos resulted in higher Fst distance measures (Table 2.3).

Table 2.2: Pairwise Fst distances between populations averaged over all SNP and using all individuals

	Jehai	**Kensiu**	**Temuan**	**Bidayuh**	**Malay (KN)**	**Malay (MN)**	**Malay (SG)**	**Chinese (SG)**
Kensiu	0.0264							
Temuan	0.0384	0.0498						
Bidayuh	0.0466	0.0599	0.0260					
Malay (KN)	0.0423	0.0503	0.0217	0.0274				
Malay (MN)	0.0460	0.0544	0.0248	0.0293	0.0167			
Malay (SG)	0.0371	0.0467	0.0164	0.0201	0.0111	0.0136		
Chinese (SG)	0.0480	0.0585	0.0243	0.0266	0.0202	0.0226	0.0153	
Indian (SG)	0.0620	0.0622	0.0497	0.0610	0.0313	0.0369	0.0340	0.0494

Table 2.3: Pairwise Fst distances between populations after removing admixed individuals from Jehai, Kensiu, Malay (SG), Bidayuh and Indian (SG) populations. Values that differ from Table 2.2 are indicated in red.

	Jehai	**Kensiu**	**Temuan**	**Bidayuh**	**Malay (KN)**	**Malay (MN)**	**Malay (SG)**	**Chinese (SG)**
Kensiu	0.0297							
Temuan	0.0490	0.0510						
Bidayuh	0.0585	0.0617	0.0264					
Malay (KN)	0.0509	0.0518	0.0217	0.0279				
Malay (MN)	0.0544	0.0558	0.0248	0.0298	0.0167			
Malay (SG)	0.0471	0.0492	0.0164	0.0200	0.0120	0.0144		
Chinese (SG)	0.0575	0.0599	0.0243	0.0272	0.0202	0.0226	0.0152	
Indian (SG)	0.0738	0.0712	0.0601	0.0730	0.0394	0.0452	0.0474	0.0607

The results of PCA for the first two principal components (PC) are shown in Figure 2.2, whereby the first principal component (x-axis) and the second principal component (y-axis) describe 2.6% and 2.3% of the variation between individuals, respectively. PC1 separates the Indians from the other populations whereas PC2 separates the two Negrito populations from others. As a result, three broad groupings were observed corresponding to the Negritos, Indians and the rest which consist of the Bidayuh, Temuan, Malays, indigenous Taiwanese and Chinese. Several Indian and Malay individuals in Singapore were observed to be in intermediate positions between their respective population clusters. This suggests that these individuals are hybrids of the two populations (Indians and Malays) and may be the result of fairly recent admixture. The Negritos also appeared to experience some varying degree of admixture, as seen by the way some individuals, especially in the Jehai, seem to form a gradient along PC2 heading towards the Malay/Temuan/Bidayuh/Chinese cluster in Figure 2.2. This ‘comet-like’ pattern could also be observed in the Temuan, Bidayuh, and Malay-MN when PCA was rerun after omitting the Negrito, Indian and indigenous Taiwanese individuals. PC1 in Figure 2.3(A) separates the Temuan and Bidayuh populations from the others while PC2 separates the Temuan and Bidayuh from the Malays and Chinese. The third PC (Fig. 2.3(B) shows some substructure within the Temuan population, whereas PC4 displays the ‘comet-like’ pattern in the Malay-MN.

To further investigate admixture and population structure in these populations, we performed Bayesian clustering implemented in the STRUCTURE software in which individuals were assigned into k number of clusters. Starting from k=2, the number of k is increased until the k value showing the greatest posterior probability was reached, in this case k=6. The results of STRUCTURE analysis from k=2 to k=6 is shown in Figure 2.4. Each individual is represented by a vertical bar and their respective ancestry components are indicated by different colors. Multiple ancestry clusters within an individual (multiple colors in a single vertical bar) signifies an admixed individual. At k=2, the two population subdivision corresponds to the Negritos

(Jehai, Kensiu) and the rest of the populations. As k is increased to 3, the population clusters observed were the Indians, Negritos and the rest. At k=4, there appeared to be a component shared mostly among the Chinese and indigenous Taiwanese. The new population cluster at k=5 corresponds to the Temuan and at k=6, the previous Negrito component was further split into the Jehai and Kensiu. At k=6, the six ancestry components correspond to the Jehai, Kensiu, Indian, Temuan, Bidayuh and Chinese populations. Admixture seems to be a predominant feature based on the STRUCTURE results at k=6.

Similar results were obtained using the *frappe* software which utilizes maximum-likelihood methods (Fig. 2.4), although the order at which new clusters were formed at k=5 and 6 was different. The ancestry component observed in the Austronesian-speaking populations was highest in the Bidayuh (74%) followed by the Temuan, Malays and was even present in the Chinese at roughly 20%. The ancestry component corresponding to the Chinese and indigenous Taiwanese was also observed in all other Austronesian groups. The three Malay groups have fairly similar ancestry components, with major contributions from the Chinese (45%), Austronesian (25%) and Indian (15%) components. Admixed Indian and Singapore Malay individuals identified from the PCA plots were confirmed in the STRUCTURE analysis.

In order to analyze more closely the ‘comet-like’ pattern observed in PCA analysis, the PC coordinates in the Negritos, Temuan and Bidayuh were plotted against ancestry components obtained from STRUCTURE analysis. For the Negritos, the PC2 coordinates from Figure 2.2 was plotted against the Negrito ancestry at k=5 from STRUCTURE analysis. For the Bidayuh and Temuan, the PC2 coordinates from Figure 2.3(A) was plotted against their respective Austronesian ancestry proportion at k=4 from the STRUCTURE analysis. All PC coordinates have been normalized to range from 0 to 1 so that 0 reflects coordinates closest to the Malay cluster and 1 reflects coordinates farthest from the Malay cluster in PCA plots. The results are shown in Figure 2.5. In all three cases, there was high correlation between the PC coordinates and the amount of admixture from STRUCTURE analysis.

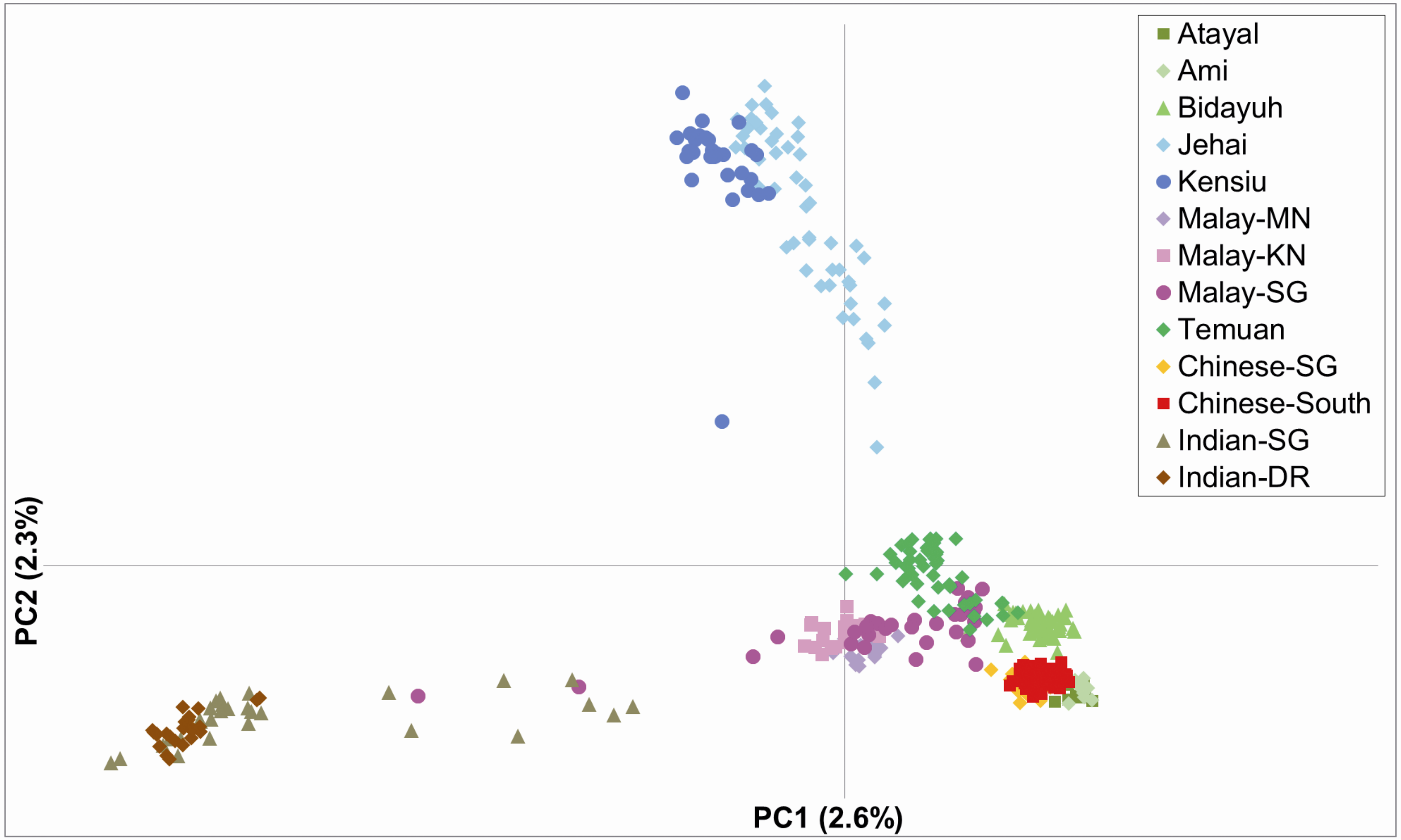


Figure 2.3: The first two Principal component (PC) analysis plots of individuals. Numbers in parentheses are percent of variation explained by the PC.

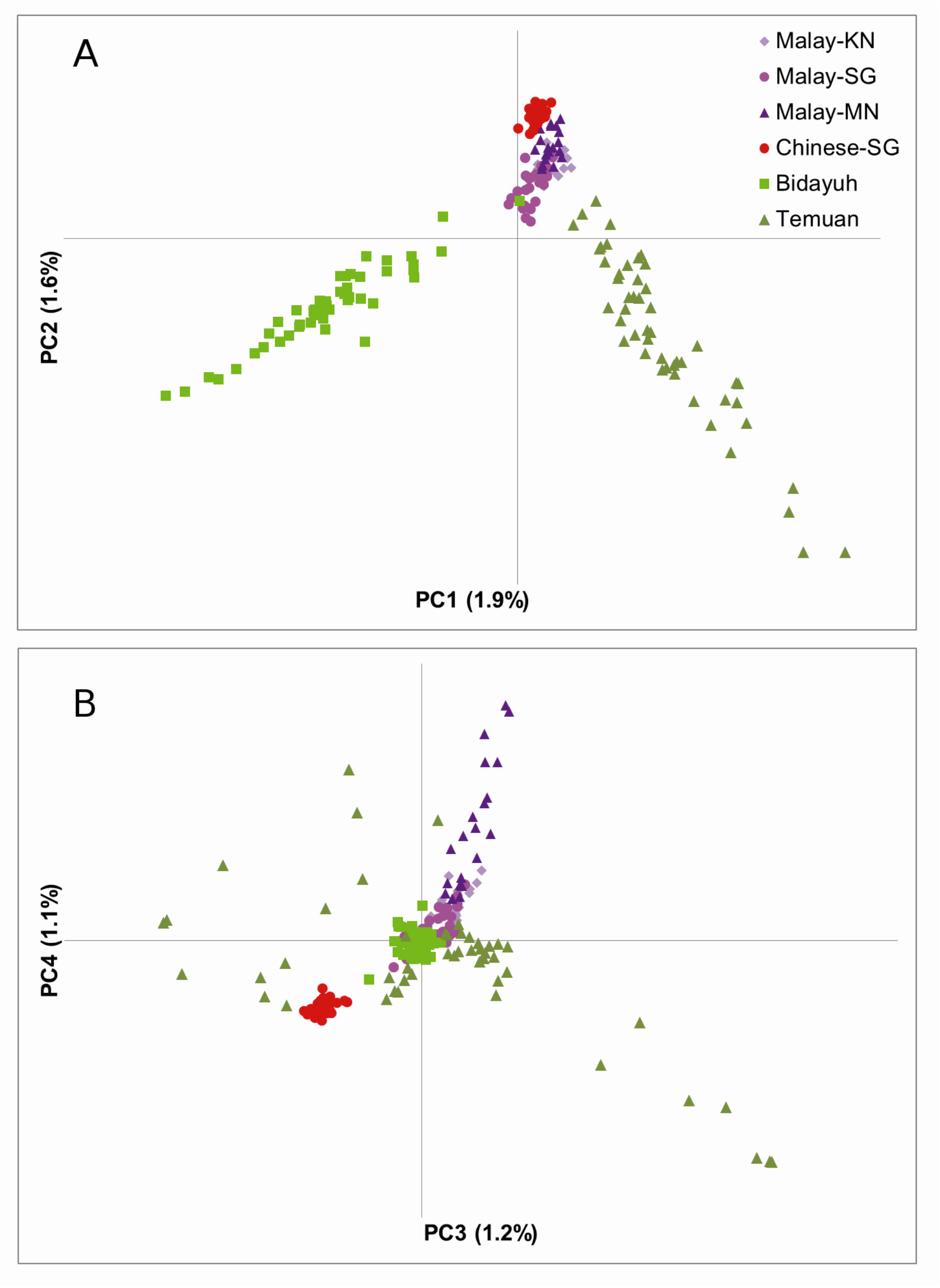


Figure 2.4: Principal component (PC) plots of individuals from Bidayuh, Temuan, Malay and Chinese groups. A) PC1 versus PC2 and B) PC3 versus PC4.

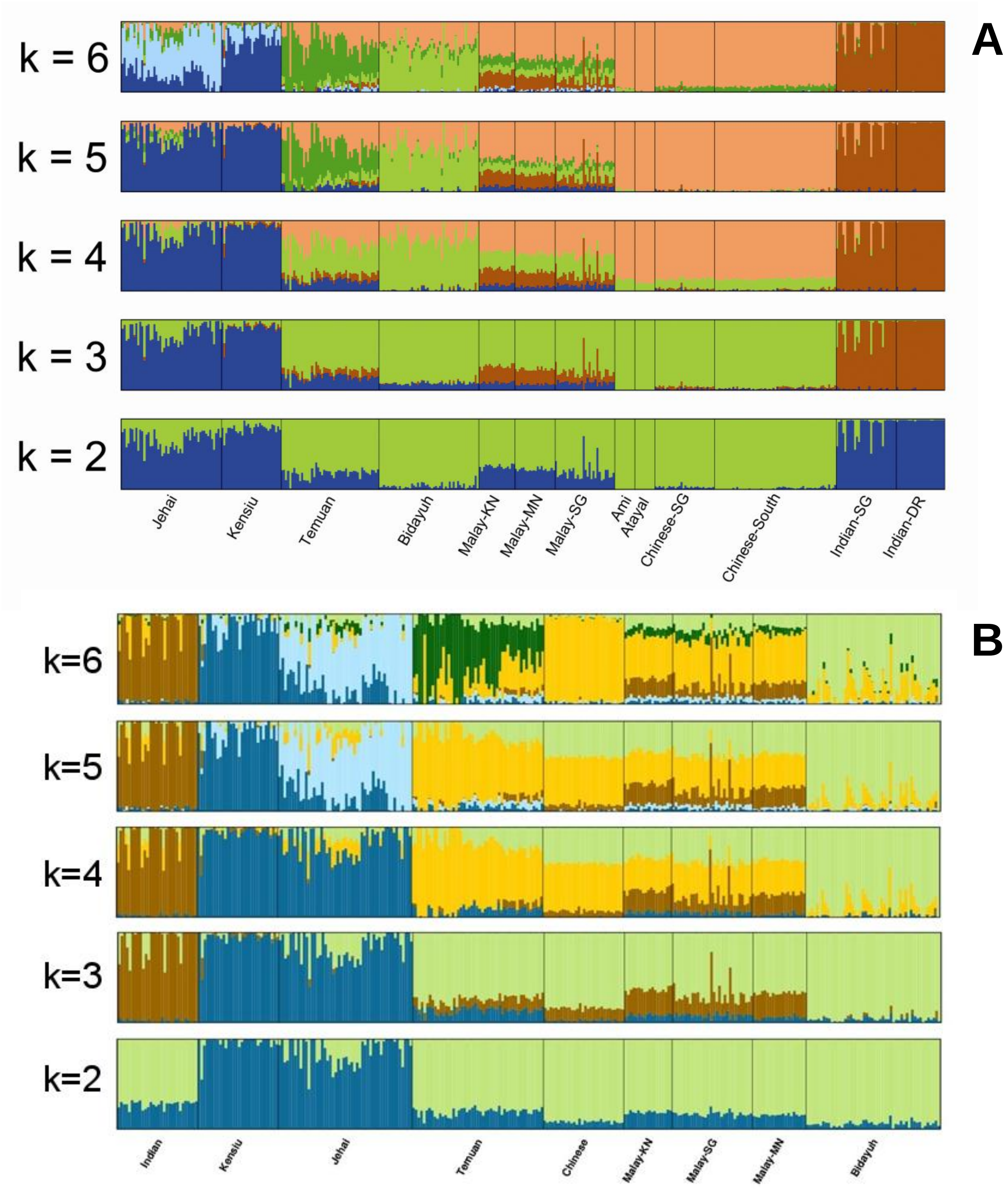


Figure 2.5: STRUCTURE and *frappe* output from k=2 to k=6. Each individual is represented by a vertical bar and the proportions of each cluster (k) are represented by different colors. Population labels are listed at the bottom. A) STRUCTURE results; B) *frappe* results

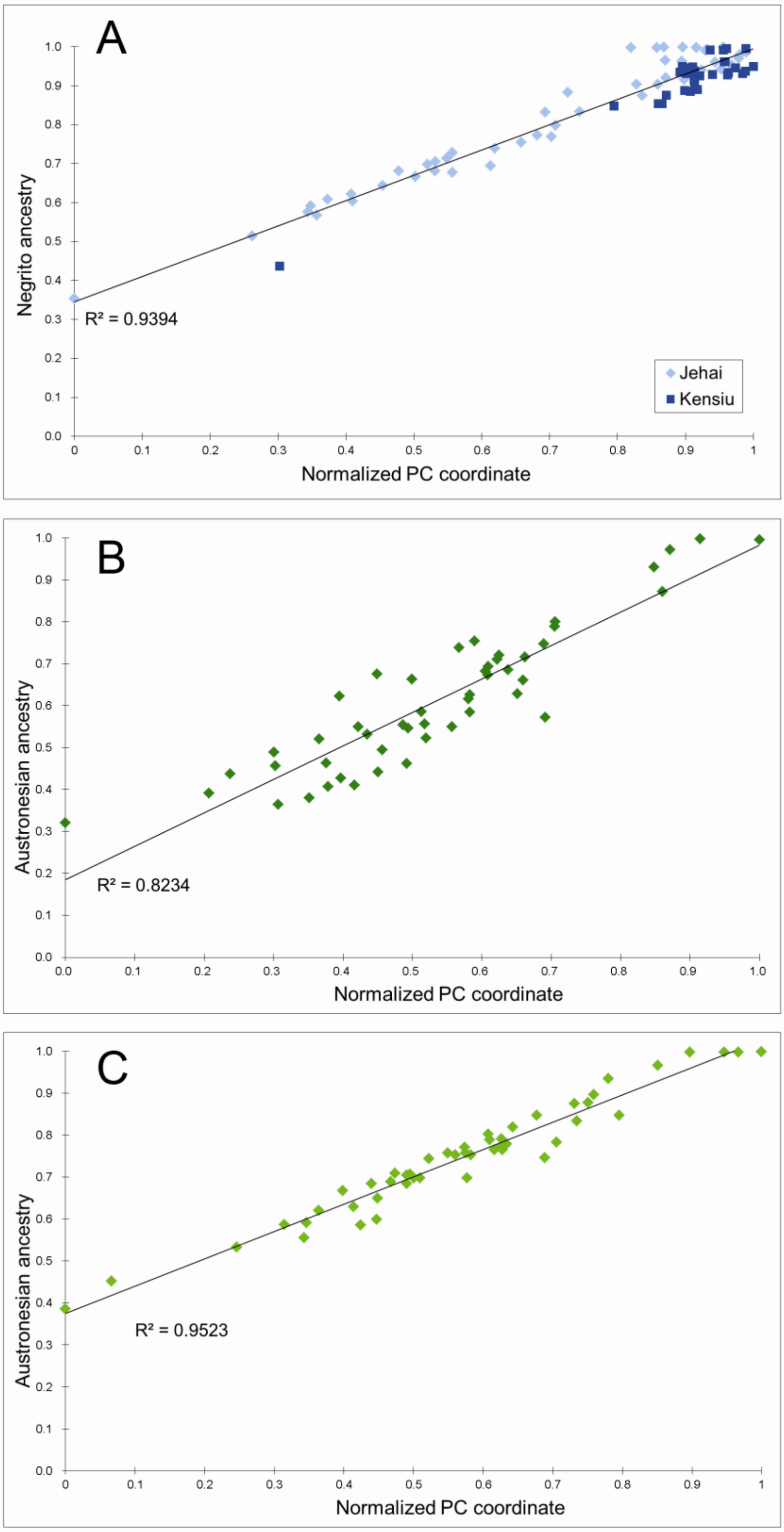


Figure 2.6: Correlation between Principal component values and STRUCTURE ancestry proportions. A) Negrito ancestry at k=5 vs. principal component (PC)2 coordinates in the Jehai and Kensiu. B) Austronesian ancestry at k=4 vs. PC2 coordinates in the Bidayuh from Figure 2.3. C) Austronesian ancestry at k=4 vs. PC2 coordinates in the Temuan from Figure 2.3. All PC coordinates have been normalized to range from 0 to 1 so that 0 reflects positions closest to the Malay cluster and 1 reflects positions farthest from the Malay cluster in PCA plots.

2.3.2 Relationships between Southeast Asian populations

To gauge the relationships between the indigenous populations in Malaysia with other Southeast Asian populations, PCA analysis was performed only on populations from Malaysia, Indonesia, Philippines and Thailand which are listed in Table 2.1. The resulting PCA plot is shown in Figure 2.6. In panel (A), the first PC separates the Malaysian Negritos from the rest while PC2 separates the Melanesians, Indonesians from Nusa Tenggara (Alor, Sumba, Flores islands) and Philippine Negritos from other populations. The majority of Austronesian-speakers with the exception of Indonesians from Nusa Tenggara and the Iraya and Manobo from the Philippines appeared to cluster closely with the Thai populations. Another exception was the Mlabri from Thailand, who appeared as an outlier and they are known to be a population isolate (Oota et al. 2005). To have a closer look at the relationship between Austronesians and Thai populations, PCA was performed after omitting the outlier populations mentioned above. The resulting PCA plot is shown in panel (B). PC1 appears to separate the populations according to an East-West division with the indigenous Taiwanese and Filipinos at one end (East) and the Thais, West Malaysians and Javanese at the other end (West). The second PC separates the Thais from the other Austronesians and the ‘comet-like’ pattern in the Temuan and Bidayuh were again observed, similar to Figure 2.3(A).

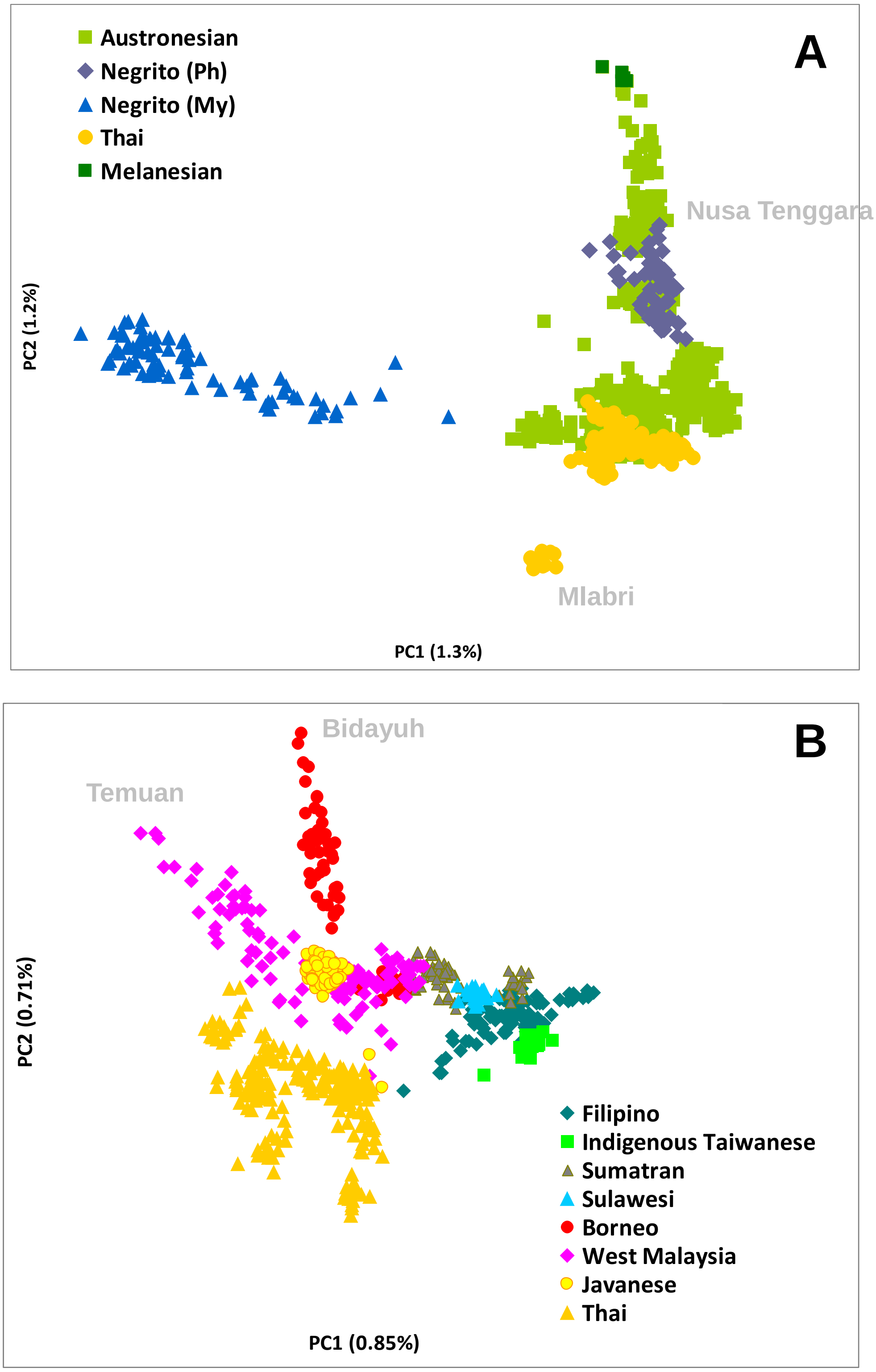


Figure 2.7: PCA plot of Southeast Asians. A) All individuals included. B) After excluding the following individuals: Negritos (Malaysian & Philippines), Melanesian, Indonesians (Nusa Tenggara), Thai (Mlabri)

2.4 Discussion

This report describes the population substructure and admixture within several indigenous populations as well as the relatively recent migrant populations in Malaysia and Singapore. STRUCTURE and PCA results revealed the presence of admixed individuals in some populations, most notably the admixture between the Singapore Malay and Indians. In this case the admixed individuals from both sides were easily identified by their intermediate positions between parental clusters in PCA plots as well as their ancestral proportions in STRUCTURE analysis. The amount of admixture in these individuals is most likely the result of recent admixture, given their population history.

With regards to the Malays, all three groups appeared very closely related based on PCA, STRUCTURE and Fst distances, despite coming from three separate geographical locations. This observation is consistent with a recent report by Hatin et al. (2011), who showed a close relationship between *Melayu Minang* and *Melayu Kelantan* (Malay-MN and Malay-KN in this study, respectively) in a population-based multidimensional scaling plot (their Fig. 2). On average, STRUCTURE ancestry components in the three Malay groups mainly consisted of Chinese ancestry (46%), Austronesian ancestry (30%) and Indian ancestry (18%). In the report by The HUGO Pan-Asian SNP Consortium (Abdulla et al. 2009) which included analysis of 73 Asian populations, three other populations (Malay, Batak and Batak Karo) from Sumatra, Indonesia also had roughly the same proportion of Indian ancestry with the Malays from Malaysia and Singapore but not in other Austronesian populations from nearby islands. This suggests that the current Malay populations in the Malay Peninsula may have originated from Sumatra and admixture with Indians may have occurred before the split. Alternatively, both populations in Sumatra and Malay Peninsula received the same amount of admixture. Linguistic, archaeological and historical evidences seem to indicate that Malays originated from Sumatran populations who themselves originated from an earlier Austronesian migration from

South China or Taiwan (Vlieland 1934; Andaya 2001) and results from classical biochemical markers suggested sustained contact and gene flow between Indians and Malays (Teng and Tan 1979).

The ‘comet-like’ pattern seen in the PCA plots of the Negritos, Bidayuh, Temuan and Malay-MN was also observed in other indigenous populations from Australia (McEvoy et al. 2010) and Latin America (Bryc et al. 2010). The observed pattern is likely to have been the result of continuous and sustained admixture with surrounding populations, stretching over several generations. While the source of admixture in those Australian aboriginals and Latin Americans included Caucasians, it does not seem the case for the Malaysian indigenous population, as the PCA plot did not reveal any admixture with Europeans (data not shown). Even though there has been continuous contact between Europeans and Malaysians since the 15th century, there has been no evidence of massive gene flow from Europeans to Malaysian populations such as those observed in some South American indigenous populations.

Although the indigenous populations in Malaysia have historically been isolated and had least contact with other groups, the admixture gradient implies that admixture has been an ongoing process. Although the source population for this admixture cannot be exactly determined, results suggest that it may be the Malay populations, given their continuous presence in the Malay Peninsula and neighboring islands. As for the Chinese and Indian populations in Singapore and by extension, Malaysia, they appeared to cluster closely with their respective ancestral populations from South India and South China, respectively, consistent with their recorded history (Vlieland 1934; Andaya and Andaya 1984).

PCA analysis of Southeast Asians using 50,000 genome-wide SNP reveals that the Malaysian Negritos are again unique compared to the rest of their surrounding populations. Interestingly, PC1 in Figure 2.6(A) clearly separates the two Negrito groups from West Malaysia and the Philippines and it appears that the Philippine Negritos are closer to their neighboring

Austronesian populations (Iraya and Manobo), probably as a result of more pronounced admixture compared to the Malaysian Negritos. A closer look at the relationship between Austronesian populations in Figure 2.6(B) implies a substructure akin to geographical division. Populations east of the Wallace line (Taiwan, Philippines, Sulawesi) appeared to have a closer genetic affinity to each other whereas populations from West Malaysia, Sumatra, Java and Borneo which formed the Sundaland landmass appeared closely related.

In summary, this chapter reveals in further detail the admixture and genetic substructure within Malaysian and their relationship with other Southeast Asian populations. It demonstrates that the indigenous groups have their own population substructure which is influenced by their surrounding populations. While categorizing individuals into an assumed panmictic population may still be practiced, the presence of admixed individuals should be considered, as their inclusion may affect genetic measures such as Fst as demonstrated here. The clustering patterns of individuals may shed some clues into their population migration histories, particularly in the indigenous groups whose origins have yet to unambiguously explained. More importantly, admixture and population structure within these populations should be taken into consideration, especially when conducting association studies as the presence of population stratification may lead to increased false positive results (Tian, Gregersen, and Seldin 2008; Yamaguchi-Kabata et al. 2008).

CHAPTER 3

Complete mitochondrial DNA analysis in indigenous Malaysian populations

3.1 Introduction

The focus of this chapter is the genetic diversity and population history of Southeast Asians from the point of view of mitochondrial DNA (mtDNA). The use of mtDNA as a genetic marker for population studies was popular due to such features including high copy number, high substitution rate and lack of recombination. Furthermore, the maternal mode of inheritance makes it possible to trace back lineages and infer histories relating to female migration.

The mtDNA is a circular molecule 16,569 base pairs (bp) in length. The first complete human mtDNA sequence was published by researchers at Cambridge University (Anderson et al 1981) but was later revised by Andrews et al. (1999) to produce the revised Cambridge Reference Sequence (rCRS). Mitochondrial haplogroups are defined by specific nucleotide substitutions or other features such as a 9-bp deletion on the mtDNA molecule. Haplogroups are labeled with a series of alphanumeric characters as shown in Figure 3.1. The three major haplogroups or macrohaplogroups are L, M and N and they display a characteristic geographical distribution in human populations. Macrohaplogroup L is mostly restricted to African populations whereas M and N are found mostly, but not restricted to, Asians and Europeans, respectively. The 917 bp non-coding region of the molecule is often referred to as D-loop or HyperVariable Region (HVR) and is often used for population genetic studies or even in forensics. This is because it is relatively short and easier to sequence but more importantly, it has a higher substitution rate compared to the coding-region of the molecule. Therefore, there are more substitutions in the D-loop region, making it useful for defining haplogroups and subsequently to discern populations and/or individuals. However, the variable rate of mutation in the D-loop region makes it not too suitable to infer the timing of evolutionary events.

There have been several previous studies using mtDNA sequences in Southeast Asian populations which attempted to address the Austronesian expansion. As mentioned in Chapter 1, the ‘Out of Taiwan’ model was largely supported by archaeological data and the linguistic phylogeny of Austronesian languages. It assumes a recent and rapid expansion from Taiwan to Polynesia with little or no admixture between the expanding and extant populations in what was called the ‘express train’ model (Diamond 1988; Gray and Jordan 2000). The model was later expanded to involve a series of pulses and pauses but remains fundamentally similar (Gray, Drummond, and Greenhill 2009). On the other side of the discussion is the ‘slow boat’ model, which posits an island Southeast Asian origin of Polynesians based on the age estimates of haplogroup B4a1a, also known as the Polynesian motif (Oppenheimer and Richards 2001). However, those previous studies mostly concentrated on the origins of the Austronesians from the Polynesian islands. In the Austronesians from island Southeast Asia, studies also came to inconsistent conclusions, with some reports supporting the ‘Out of Taiwan’ model (Trejaut et al. 2005; Tabbada et al. 2009) while others proposed earlier migrations from the Asian mainland during the late-Pleistocene to early-Holocene period (Hill et al. 2006; 2007).

A more comprehensive look into the mtDNA diversity in the indigenous groups in Malaysia, who include descendants of the earliest settlers of the region, should provide more insight into the origins and migration of humans in Southeast Asia. Under this current backdrop, an in-depth analysis of complete mtDNA sequences from four indigenous groups in Malaysia who represent the ancient migrants (Negritos) and subsequent migrants (Austronesians) to the Southeast Asian region was conducted. By exploring the diversity of mtDNA in these four groups and comparing them with other populations within the Southeast Asian region, we hope to shed light on some questions regarding their demographic and migration histories.

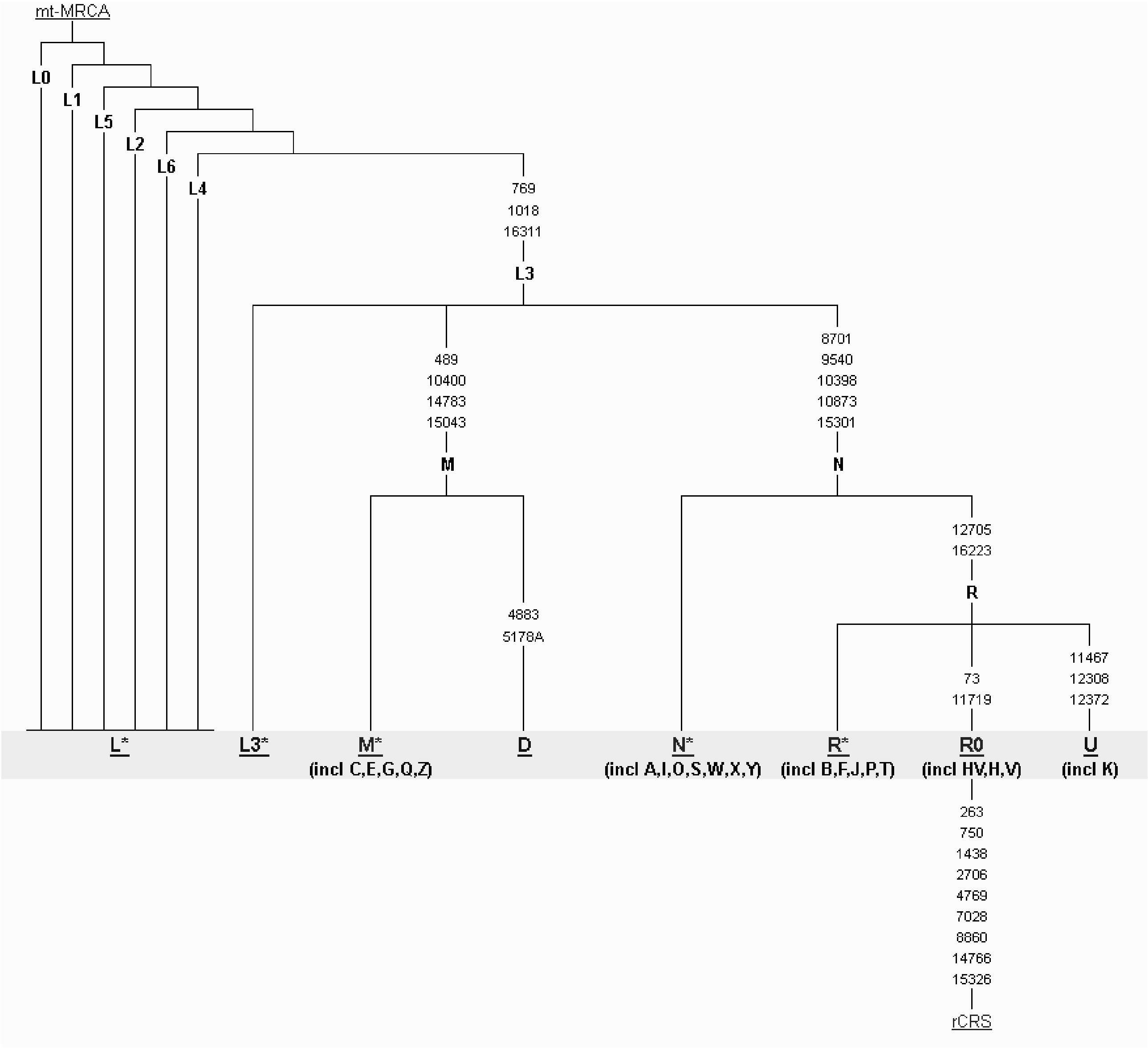


Figure 3.1: Phylogenetic tree of human mtDNA. Mitochondrial DNA haplogroups are shown as alphanumeric letters in bold. The numbers on the branches indicate the position of substitutions that define the haplogroup. rCRS stands for revised Cambridge Reference Sequence. Image source: http://www.phylotree.org/

3.2 Material & methods

3.2.1 Sample collection and ethical approval

In addition to samples from the Jehai, Temuan and Bidayuh which were previously collected as part of my Master's thesis (Jinam 2007), samples from an *Orang Asli* group from West Malaysia called the Seletar were also included. The new samples from the Seletar were collected on a sampling trip in Johor, West Malaysia with collaborators from the Monash University Sunway Campus and University of Malaya in July 2009. The Jehai, Temuan and Seletar represent the *Orang Asli* groups from West Malaysia while the Bidayuh are one of the indigenous groups from Borneo. The Jehai are further grouped as Negritos while the Temuan and Seletar are classified as Proto-Malays. Linguistically, the Jehai speak Austro-Asiatic languages whereas the Temuan, Seletar and Bidayuh speak Austronesian languages.

Additionally, mtDNA haplogroup frequencies in the Kensiu, another Negrito group from West Malaysia, were kindly provided by a colleague, Miss Hong Lih Chun from the Department of Molecular Medicine, University of Malaya (Hong LC, unpublished data). Haplogroup frequencies and complete mtDNA sequences in other populations were obtained from available literature. The list of populations used in this study is in Table 3.1 whereas their geographical locations are depicted in Figure 3.2. This study has been approved by the respective ethical committee boards of the National Institute of Genetics Mishima, University of Malaya and Monash University Sunway Campus.

Table 3.1: Population information used for mtDNA analyses

Population label	Ethnicity	Location	Data used[a]	References
1	Jehai	West Malaysia	1,2	This study
2	Temuan	West Malaysia	1,2	This study
3	Seletar	West Malaysia	1,2	This study
4	Bidayuh	Borneo	1,2	This study
5	Kensiu	West Malaysia	2	Hong LC, unpublished data
6	Batek	West Malaysia	1,2	Macaulay et al 2005; Hill et al 2006
7	Mendriq	West Malaysia	2	Hill et al., 2006
8	Temiar	West Malaysia	2	Hill et al., 2006
9	Semelai	West Malaysia	1,2	Macaulay et al 2005; Hill et al 2006
10	Jakun	West Malaysia	2	Hill et al., 2006
11	Malay	West Malaysia	2	Hill et al., 2007
12	Iban	Borneo	2	Simonson et al., 2011
13	Kadazan	Borneo	1	Soares et al, 2008
14	Alor	Indonesia	2	Hill et al., 2007
15	Ambon	Indonesia	2	Hill et al., 2007
16	Banjarmasin	Indonesia	2	Hill et al., 2007
17	Java	Indonesia	2	Hill et al., 2007
18	Bali	Indonesia	2	Hill et al., 2007
19	Lombok	Indonesia	2	Hill et al., 2007
20	Sumba	Indonesia	2	Hill et al., 2007
21	Sumatra	Indonesia	2	Hill et al., 2007
22	Sulawesi	Indonesia	2	Hill et al., 2007
23	Besemah	Indonesia	1,2	Gunnarsdottir et al, 2011
24	Semende	Indonesia	1,2	Gunnarsdottir et al, 2011
25	Manobo	Philippines	1,2	Gunnarsdottir et al, 2010
26	Mamanwa	Philippines	1,2	Gunnarsdottir et al, 2010
27	Surigaonon	Philippines	1,2	Gunnarsdottir et al, 2010
28	Filipino	Philippines	1,2	Hill et al., 2007; Tabada et al 2009
29	Indigenous Taiwanese	Taiwan	1,2	Trejaut et al, 2005; Hill et al., 2007; Soares et al 2008
30	Thai	Thailand	2	Hill et al, 2006
31	Vietnamese	Vietnam	1,2	Jin et al, 2009; Hill et al, 2006
32	Cham	Vietnam	1	Peng et al, 2010
33	South Chinese	China	1,2	Ingman et al, 2000; Kong et al., 2003; Kong et al., 2006; Hill et al., 2007; Kong et al., 2011

34	Great Andamanese	Andaman Islands	1	Thangaraj et al 2005, Barik et al, 2003
35	Onge	Andaman Islands	1	Thangaraj et al 2005, Barik et al, 2003
36	Nicobarese	Andaman Islands	1	Thangaraj et al 2005, Barik et al, 2003
37	Australian Aboriginal	Australia	1	Ingman et al, 2000; Ingman et al 2003; van Host Pelikan 2006
38	Papuan	Papua New Guinea	1	Ingman et al, 2000; Macaulay et al 2005
39	Melanesian	Melanesia	1	Freidlander et al 2007
40	Samoan	Polynesia	1	Ingman et al, 2000
41	African	Africa	1	Ingman et al, 2000

a- Data used: 1) Complete mtDNA sequences; 2) mtDNA haplogroup frequencies

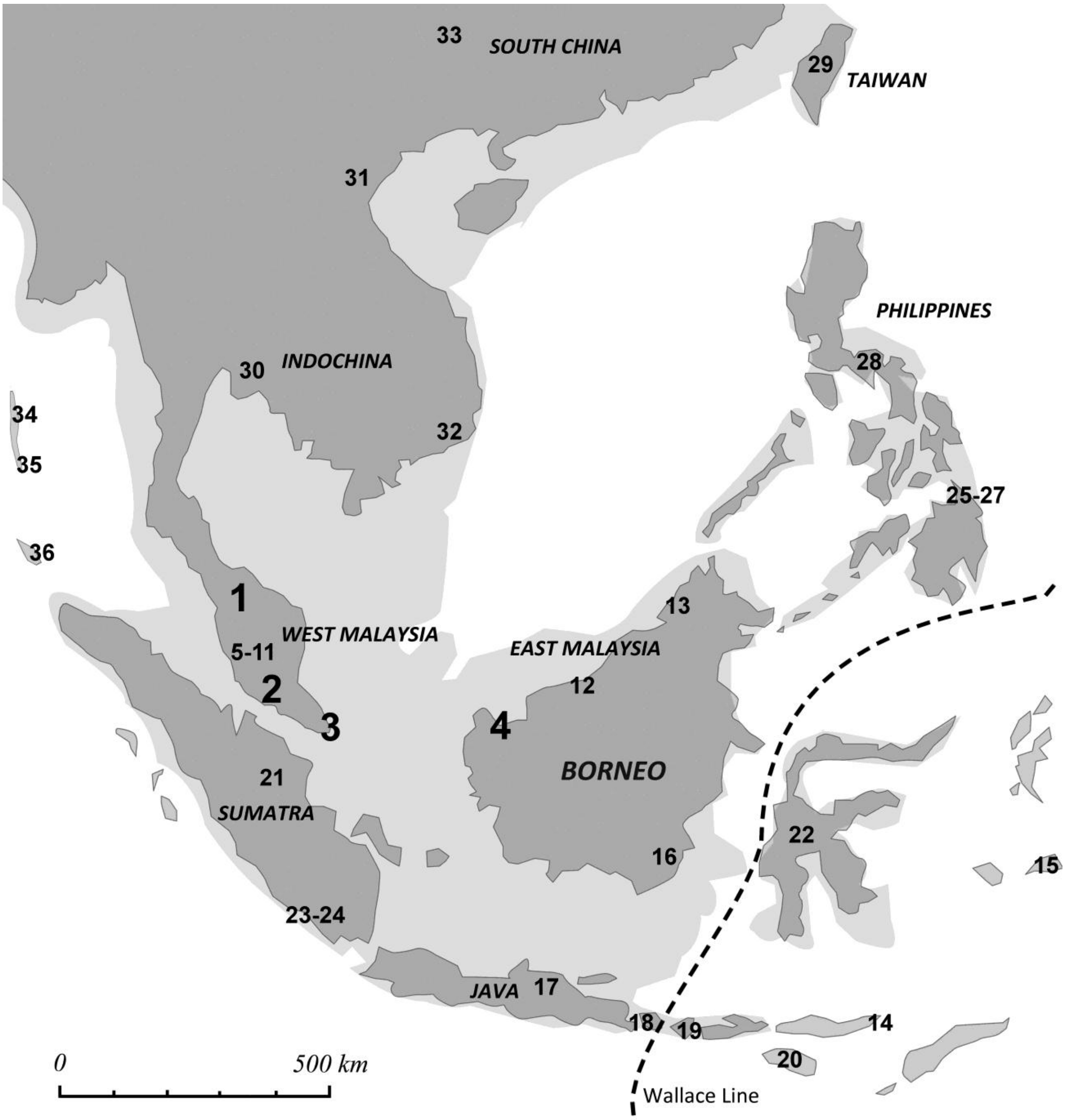


Figure 3.2: A map of Southeast Asia indicating geographical positions of populations used for analysis. Numbers indicate the locations of populations listed in Table 3.1. Not shown on the map are Australian Aboriginals, Papuans, Melanesians, Samoans and Africans. Areas shaded light gray indicate the extent of the landmass up to the Last Glacial Maximum following Hill et al (2007).

3.2.2 Complete mitochondrial DNA sequencing

Complete mtDNA sequencing was performed in a total of 68 samples using 11 pairs of PCR primers and 32 sequencing primers from (Torroni et al. 2001). The list of primer sequences are shown in the Appendices. A slight modification to the protocol involved optimizing annealing temperatures for all PCR reactions to 60°C, instead of 55ºC as reported in their paper. For details of gradient PCR reactions used for optimization, refer to the Appendices. PCR products were purified using ExoSAP-IT reagent before being subjected to sequencing reactions using the BigDye Terminator kit (Applied Biosystems). Capillary separation was performed on the ABI3130xl Genetic Analyzer (Applied Biosystems). For each sample, the resulting 32 traces were aligned to the revised Cambridge Reference Sequence (Genbank ID NC_012920.1) to obtain the consensus sequence.

In addition, complete mtDNA sequences in a total of 18 samples from the Jehai (1), Temuan (10) and Bidayuh (7) were generated using the Illumina Genome Analyser at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany under the guidance of Mark Stoneking. The details of the methodology were described previously (Gunnarsdottir et al. 2010). In total, complete mtDNA sequences were obtained from 86 individuals (24 Jehai, 18 Temuan, 21 Seletar and 23 Bidayuh).

3.2.3 Data analysis

Individuals were assigned to mtDNA haplogroups according to nomenclature found at http://www.phylotree.org (van Oven and Kayser 2009). Mitochondrial DNA haplogroup frequencies from the populations listed in Table 3.1 were used for Principal Component Analysis (PCA) using R software package (http://www.R-project.org).

Coding region sequences (nucleotide positions 577 to 16,023) were extracted from complete mtDNA sequences from available literature (Table 3.1) and were used to generate a Maximum-Likelihood (ML) tree using MEGA software version 5 (Tamura et al. 2011). Using the Tamura-Nei substitution model (Tamura and Nei 1993) with invariant sites and gamma-distributed rate, the ML tree was used to estimate time-depth of haplogroups. Age estimates were also done using a Bayesian MCMC method as implemented in BEAST software (Drummond and Rambaut 2007). The software generates phylogenetic trees using the same substitution model as described above, a normally distributed prior for the mutation rate with a mean of 1.26 x 10^{-8} following Mishmar et al. (2003) and a further assumption of constant population size. The trees were generated on a run of 40,000,000 steps, sampling every 4000 steps and the first 4,000,000 steps were regarded as burn-in. Bayesian Skyline Plots (BSP) were also generated for the Jehai, Temuan, Seletar and Bidayuh groups using the above parameters, but with a coalescent-based tree prior with a piecewise linear model.

3.3 Results

3.3.1 Summary statistics

The summary statistics regarding the variation of complete mtDNA sequences in the four Malaysian groups are shown in Table 3.2. The highest haplotype diversity was observed in the Temuan, followed by Jehai and Bidayuh. Interestingly, only five distinct mtDNA haplotypes were observed in the Seletar, and this was further reflected by the very low values of haplotype and nucleotide diversity in that group. Negative values for Tajima's D test were observed in the Bidayuh, Seletar and Temuan, suggesting a history of population expansion. However the p-values obtained did not indicate statistically significant deviation from expectation.

3.3.2 Haplogroup nomenclature and frequencies

All individuals were assigned to specific haplogroups belonging to M and N macrohaplogroups by following the nomenclature in www.phylotree.org as much as possible. A total of 23 haplogroups were observed, and the specific mutations that define those haplogroups are shown in Figures 3.3 and 3.4. In addition to haplogroup-defining mutations, all samples were observed to have additional mutations and in some cases, these additional mutations further differentiate populations. For example in Figure 3.4, additional mutations in haplogroup N9a6a are different between the Bidayuh, Jehai and Temuan. In most cases, all of the assignable haplogroups have been reported previously and all the samples in this study share the basal mutations with existing haplogroups. There are instances where some observed haplotypes share only the basal mutations with known haplogroups and additional mutations did not match any existing ones. Those haplotypes were therefore assigned to the closest basal haplogroup, for example B4a and F1a’c in Figure 3.4.

The haplogroup frequencies in the four populations studied as well as the Kensiu (Hong LC, unpublished data) are presented in Table 3.3. The most frequent haplogroup in the Bidayuh is N9a6a and the basal haplogroup F1a’c whereas in the Temuan, the most frequent haplogroups are M21a, N22 and N21 which are lineages that branched off directly from basal M and N haplogroups (Macaulay et al. 2005). The low nucleotide diversity observed in the Seletar was further demonstrated by the limited number of observed haplogroups and the very high frequency of one particular haplogroup, N9a6 at 71%. Haplogroups M21a and R21 are the most frequent in the Jehai and Kensiu, who are both Negrito groups. In short, the four indigenous Malaysian groups demonstrated some differences in their haplotype frequency distribution.

Table 3.2: Summary statistics for complete mtDNA sequences in four Malaysian groups

Statistics	Bidayuh	Jehai	Seletar	Temuan
No. sequences	23	24	21	18
No. haplotypes	13	11	5	14
Haplotype diversity ± SD	0.88±0.05	0.89±0.04	0.54±0.11	0.94±0.05
Mean no. pairwise difference	30.5	32.1	15.9	29.1
No. polymorphic sites	157	87	65	135
Nucleotide diversity	0.00184	0.00194	0.00096	0.00175
Tajima's D (p-value)	-1.14495 (NS)	1.4924 (NS)	-0.49193 (NS)	-1.09762 (NS)

NS – not significant

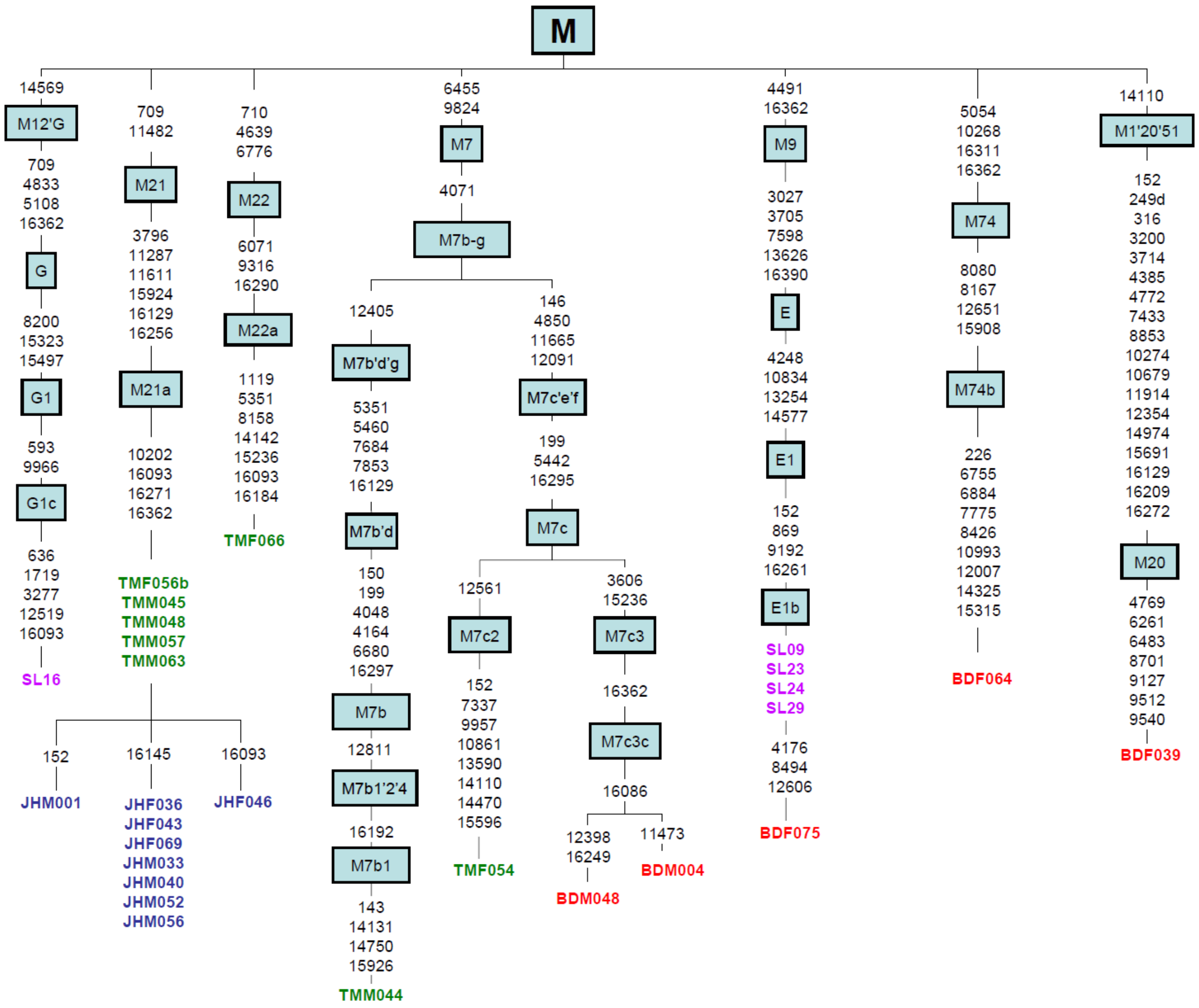


Figure 3.3: Haplogroup classification of complete mtDNA sequences from macrohaplogroup M. Numbers indicate nucleotide positions, d indicates deletion. The individual IDs are colored according to population: Jehai (blue); Temuan (green); Bidayuh (red); Seletar (pink).

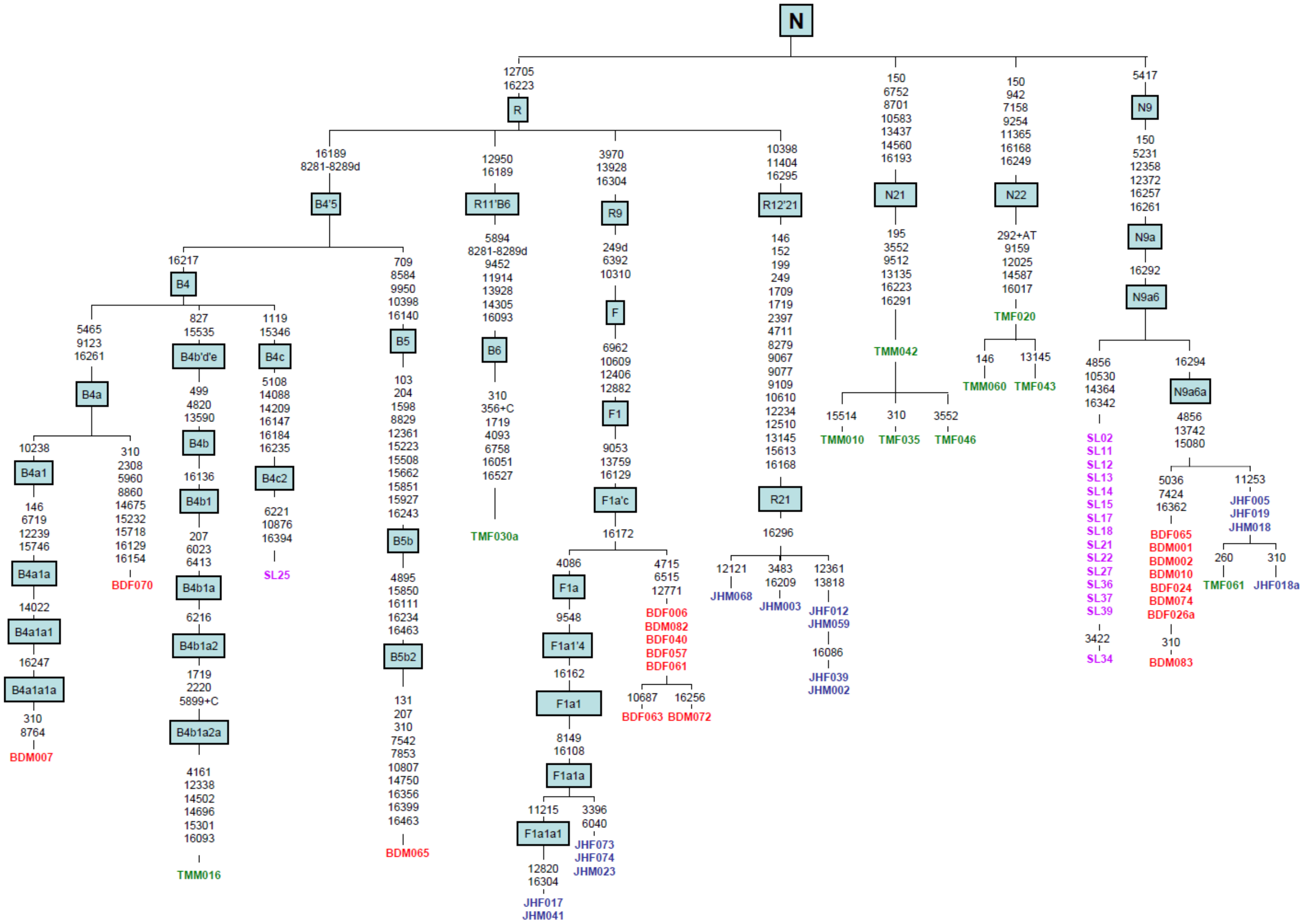


Figure 3.4: Haplogroup classification of complete mtDNA sequences from macrohaplogroup N. Numbers indicate nucleotide positions, d indicates deletion. The individual IDs are colored according to population: Jehai (blue); Temuan (green); Bidayuh (red); Seletar (pink).

Table 3.3: mtDNA haplogroup frequencies (%) in five indigenous Malaysian groups

Haplogroup	**Bidayuh**	**Jehai**	**Seletar**	**Temuan**	**Kensiu**[a]
G1c			4.8		
M20	4.3				
M74b	4.3				
M21a		37.5		27.8	43.2
M22a				5.6	
M7b1				5.6	
M7c2				5.6	
M7c3c	8.7				2.7
E1b	4.3		19.0		
B4a*	4.3				
B4a1a1a	4.3				
B4b1a2a				5.6	
B4c2			4.8		
B5a					2.7
B5b2	4.3				
B6				5.6	
F1a'c	30.4				
F1a1a		12.5			2.7
F1a1a1		8.3			
R21		25.0			43.2
N21				22.2	
N22				16.7	
N9a6			71.4		
N9a6a	34.8	16.7		5.6	5.4
Total individuals	***23***	***24***	***21***	***18***	***37***

[a] Unpublished data from Hong LC

3.3.3 Principal Component Analysis based on haplogroup frequencies

To further elucidate the relationship between the indigenous Malaysian populations and other surrounding populations, PCA was performed using haplogroup frequencies from selected populations from the literature (Table 3.1). The resulting PCA plot is shown in Figure 3.5. A clear division appears along the first Principal Component (PC) on the X-axis, which places the Negrito populations from West Malaysia (Jehai, Kensiu, Batek, Mendriq) at one end and most other Austronesian-speaking populations from Southeast Asia at the other end. On the second PC on the Y-axis there appears to be a geographical divide between the Austronesians groups, whereby populations from West Malaysia, Sumatra, and Java tend to cluster with mainland or continental groups (South Chinese, Thai, and Vietnamese). On the other hand, populations from Taiwan, Philippines and other Indonesian islands to the east (Sulawesi, Alor, Ambon) tend to group together. Thus it appears that PC1 represents a Negrito-Austronesian divide whereas PC2 corresponds to a continental-island division of Austronesian groups. However not all population affinities fall nicely into these two generalized trends.

The Temuan for example, are Austronesian speakers and are physically distinct from the Negritos but they clustered with them on PC1. This may be due to the high frequencies of haplogroup M21a in both the Temuan and Negritos. The three populations from Borneo (Bidayuh, Iban, Banjarmasin) also displayed somewhat irregular patterns. The Iban and Banjarmasin appeared closer to the continental and island clusters, respectively, whereas the Bidayuh clustered with two other Proto-Malay groups (Seletar and Semelai) from West Malaysia. These three groups (Bidayuh, Seletar and Semelai) appear to be intermediate between the two divides (Negrito-Austronesian and continental-island). Social practices may also have a significant bearing on mtDNA diversity, as shown by the Semende and Besemah from Sumatra. The Semende are a matrilocal group and are closer to the continental populations whereas the Besemah are patrilocal and are closer to island Southeast Asians. This suggests that the mtDNA diversity in the Besemah is shaped by female migrations from island Southeast Asia.

3.3.4 Phylogenetic analysis and age estimates of haplogroups

A Neighbor-Joining tree was constructed using coding-region sequences and was rooted using a chimpanzee mtDNA sequence (Genbank ID D38113). All haplogroup M and N lineages branched off the African L0 haplogroup and are shown separately in Figure 3.6. All taxa were found to cluster according to their respective haplogroups, further validating the haplogroup assignment in our samples. Using only a subset of sequences which represent haplogroups of interest, age estimates of those haplogroups were performed using ML and Bayesian methods. The estimated mean mutation rate of the mtDNA coding region according to the Bayesian MCMC analysis was 1.37 x 10^{-8} with a 95% Highest Posterior Density range of 1.01 x 10^{-8} to 2.35 x 10^{-8} substitutions per site per year. The ML tree was calibrated using the divergence time between African and non-African lineages of 170,000 YBP (Ingman et al. 2000), yielding a mutation rate of 1.39 x 10^{-8}. Ages estimates using the ML tree were based on the coalescence time of all mtDNA sequences that belong to the same haplogroup. The resulting ML tree showing ages of selected haplogroups is shown in Figure 3.7 while the rest of the age estimates (ML and Bayesian) are listed in Table 3.4.

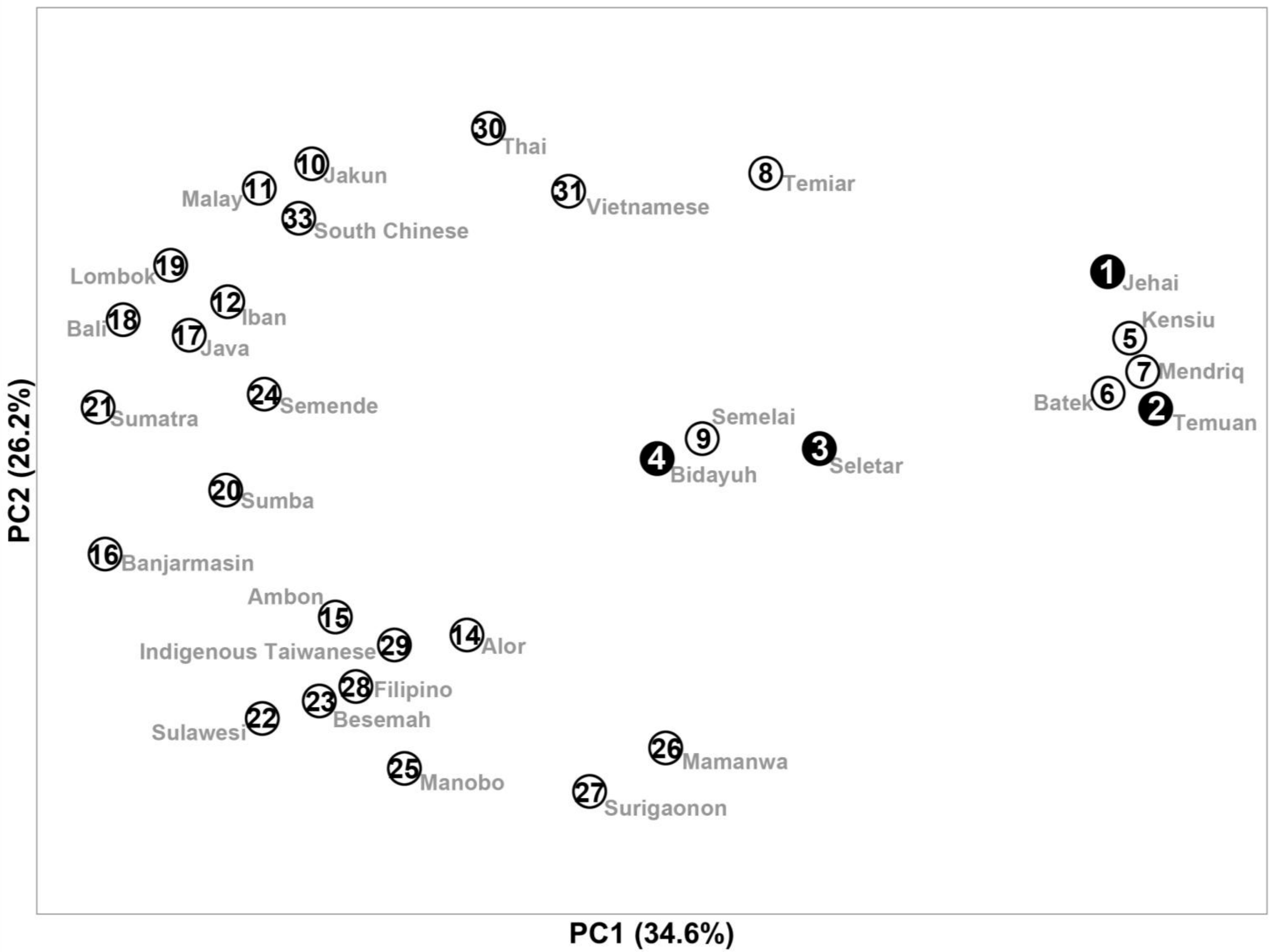


Figure 3.5: Principal Component Analysis (PCA) plot based on mtDNA haplogroup frequencies. Population numbers correspond to Table 3.1 and Figure 3.2. Black circles are from current study, white circles are from literature.

The M haplogroups observed in this study included those which were considered indigenous to the *Orang Asli*, namely M21a and M22 (Macaulay et al. 2005; Hill et al. 2006). Not found anywhere else outside of Southeast Asia, M21a was most frequent in the Temuan and Jehai as well as other Negrito subgroups in West Malaysia (Hill et al. 2006). Outside of West Malaysia, M21a was also present in appreciable frequencies in the Sakai (also a Negrito group) and in the Chiang Mai population from Thailand (Fucharoen et al. 2001) and very rarely in some Philippine populations (Tabbada et al. 2009; Gunnarsdottir et al. 2010). The other M21 subtypes, M21b and M21c which were reported at low frequencies in the *Orang Asli* (Hill et al. 2006) but frequent in the Moken of Myanmar (Dancause et al. 2009), were not observed in any of our current samples. The coalescent time of 49,500±5,400 YBP for all M21 lineages suggests deep ancestry in mainland Southeast Asia and at some point in time it may have spread to as far as the Philippines. M22 was earlier reported in the Proto Malays (Macaulay et al. 2005) and recent reports also showed that it was present in the Vietnamese (Peng et al. 2010) and Southern Chinese (Kong et al. 2011) but has so far not been reported in any island Southeast Asians. These M22 lineages have a coalescent date of around 29,100±13,200 YBP.

Haplogroup E, which was proposed to be a marker for post-glacial expansion centering in Island Southeast Asia (Soares et al. 2008), was found in the form of E1b in the Seletar and Bidayuh with a coalescent time of 6,700±5,200 YBP. Haplogroup M7 lineages that are present in the Malaysian samples included M7c3c in the Bidayuh. This haplogroup appears to be restricted to Southeast Asia and was suggested to be a marker for the Austronesian expansion during mid-Holocene (Hill et al. 2007). Our age estimate of 5,100±12 YBP for this haplogroup seems to agree with this idea. Other M7 lineages found in the Temuan include M7b1 and M7c2 and they coalesce with lineages from the mainland (Kong et al. 2003) around 11,700±7,000 YBP and 22,500±9,800 YBP, respectively. In addition to the above haplogroups, there are several haplogroups in the Malaysian samples which have not been reported in any Southeast Asian population to date. These included G1c in the Seletar and M74b and M20 in the Bidayuh. G1c was earlier reported in Northern Asians including Koreans (Derenko et al. 2007) and Han Chinese (Kong et al. 2003). The ancestral M74a haplotype was confined to southern Chinese populations (Kong et al. 2011) whereas the derived type M74b was found in the Bidayuh in Borneo and Hani of south China. The M74b1 subtype has been found in Surigaonon and Mamanwa in the Philippines (Gunnarsdottir et al. 2010) and also in the Besemah in Sumatra (Gunnarsdottir et al. 2011). The coalescent time of 27,500±7,500 YBP for M74b suggests a dispersal originating from southern China and into island Southeast Asia. Haplogroup M20 found in one Bidayuh individual coalesces with the M20 haplotype in a southern Chinese group (Kong et al. 2011) around 4,700±32 YBP. These two M20 lineages clustered with haplogroup M51 which was found in the Cham of Vietnam (Peng et al. 2010) and the Besemah in Sumatra (Gunnarsdottir et al. 2011).

As with haplogroup M, there exist rare N lineages which were previously reported in the *Orang Asli*, namely N21, N22 and R21 (Hill et al. 2006; Hill et al. 2007). N21 and N22 make up almost 40% of the mtDNA diversity in the Temuan whereas R21 was mostly restricted to the Negrito groups, Jehai and Kensiu. N21 branches directly from the N founder node and was earlier thought to have originated in island Southeast Asia and subsequently spread to West Malaysia based on control region sequences (Hill et al. 2007). However based on the NJ tree in Figure 3, the N21 lineages in the Temuan appeared to be derived from an ancestral type found in the Cham of Vietnam (Peng et al. 2010), implying an origin in Indochina during Pleistocene (29,300±15,500 YBP) and it’s dispersal appears to be limited to Sundaland which encompassed West Malaysia, Sumatra even up to the Alor islands (Hill et al. 2007). N22 appears to be limited to the Temuan, as observed in this study and by Hill et al. (2006) although it appears very low frequencies in the Philippines (Tabbada et al. 2009), Sumatra (Gunnarsdottir et al. 2011) and

Sumba islands (Hill et al. 2007). Like N21, the coalescent time of 24,800±13,800 YBP suggests deep ancestry in Sundaland.

Haplogroup R21 appears to be limited to Negrito populations in West Malaysia, although it was also found at appreciable frequencies in the Senoi (Hill et al. 2006), suggesting gene flow between the indigenous Negritos with the incoming Senoi from Indochina (Glover and Bellwood 2004). All R21 lineages coalesce with haplogroup P4, found mostly in Australian Aborigines and Papuans (Fig. 4), at approximately 47,100±5,100 YBP possibly linking the Negritos with the first migration to the Southeast Asian region. Haplogroup N9a is widespread in East Asia, but the subclade N9a6 appears to be restricted to Southeast Asian populations where it is found at low frequencies in Sumatra and Java, Indonesia, but not in the Philippines or Taiwan (Hill et al. 2007). However we found N9a6 and its daughter clade N9a6a to be quite frequent in the Malaysian groups, particularly in the Bidayuh and Seletar where frequencies reached 35% and 71%, respectively. N9a6 diverged from mainland N9a approximately during the late Pleistocene and dispersed south where it further diversified and spread to Borneo and Java.

Haplogroup B which is characterized by a 9-bp deletion at position 8,272 is fairly common in island Southeast Asia, particularly in the Polynesian islands. The distribution of this haplogroup is varied amongst the Malaysian populations, with B4a and B5b found in the Bidayuh, B4b and B6 in the Temuan and B4c in the Seletar. The two B4a lineages in the Bidayuh included B4a1a1a, also known as the Polynesian motif because it reaches near fixation in the Polynesian islands. The other is an undefined B4a haplogroup, which shares the same basal mutations as B4a but could not be further designated to any of its daughter clades. B4a and its sub-lineages are particularly common in the Philippines (Tabbada et al. 2009) and indigenous Taiwanese (Trejaut et al. 2005) and may have arisen around the vicinity of South China during the Pleistocene period and eventually dispersed to island Southeast Asia via Taiwan and Philippines. The presence of B4a1a1a in the Bidayuh may reflect recent gene flow from the

Pacific during the mid-Holocene period (Soares et al. 2011). The branching patterns of the NJ tree (Fig. 3) show that the ancestral types of haplogroups B4b, B4c and B5b were found in South Chinese populations, suggesting an origin in the mainland and dispersal to island Southeast Asia. Interestingly, the B4c2 haplogroup found in the Seletar was also found extracted from ancient Negrito hair samples (Ricaut et al. 2006), indicating a diffusion from the mainland during the late-Pleistocene. Haplogroup F is another common clade in Southeast Asia, with F1a1a present at 12.5% in the Jehai and was previously reported to also be frequent in the Temiar, a Senoi group (Hill et al. 2006). Haplogroup F1a'c shares the same basal mutations as F1a except at nucleotide position 4,086. It is present at 30% in the Bidayuh and coalesces with other F1a1 lineages at 18,200±6,900 YBP. Haplogroups F1b and F1c are mostly restricted to South China and this suggests that F1a'c and its derived types might also have originated from there and later spread to island Southeast Asia.

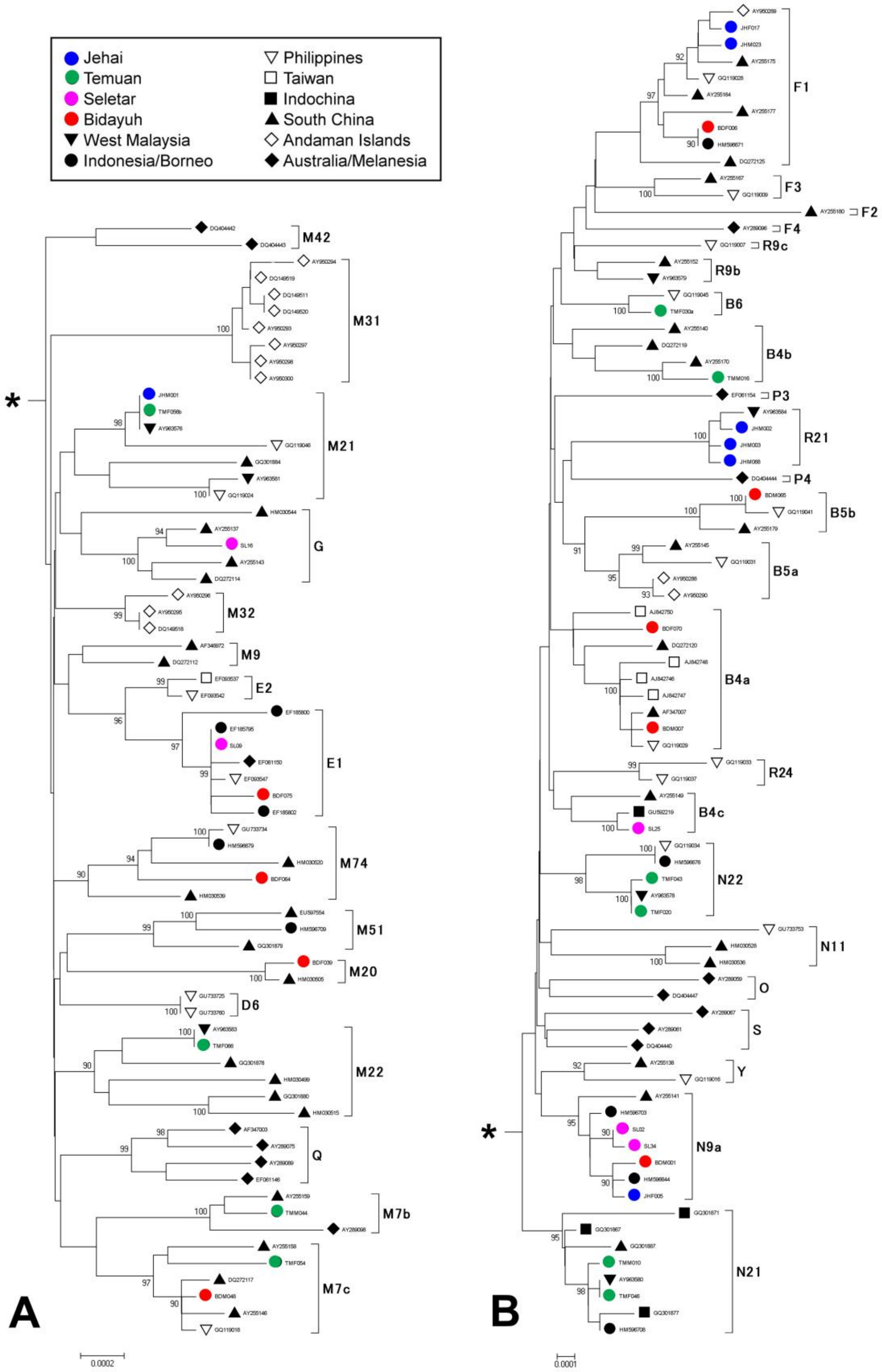


Figure 3.6: Neighbor-joining tree of mtDNA lineages constructed using coding-region sequences with 500 bootstrap replicates. Bootstrap values above 90% are shown. Asterisks indicate the branch connecting to the tree root. A) Subtree of haplogroup M lineages; B) Subtree of hapologroup N lineages.

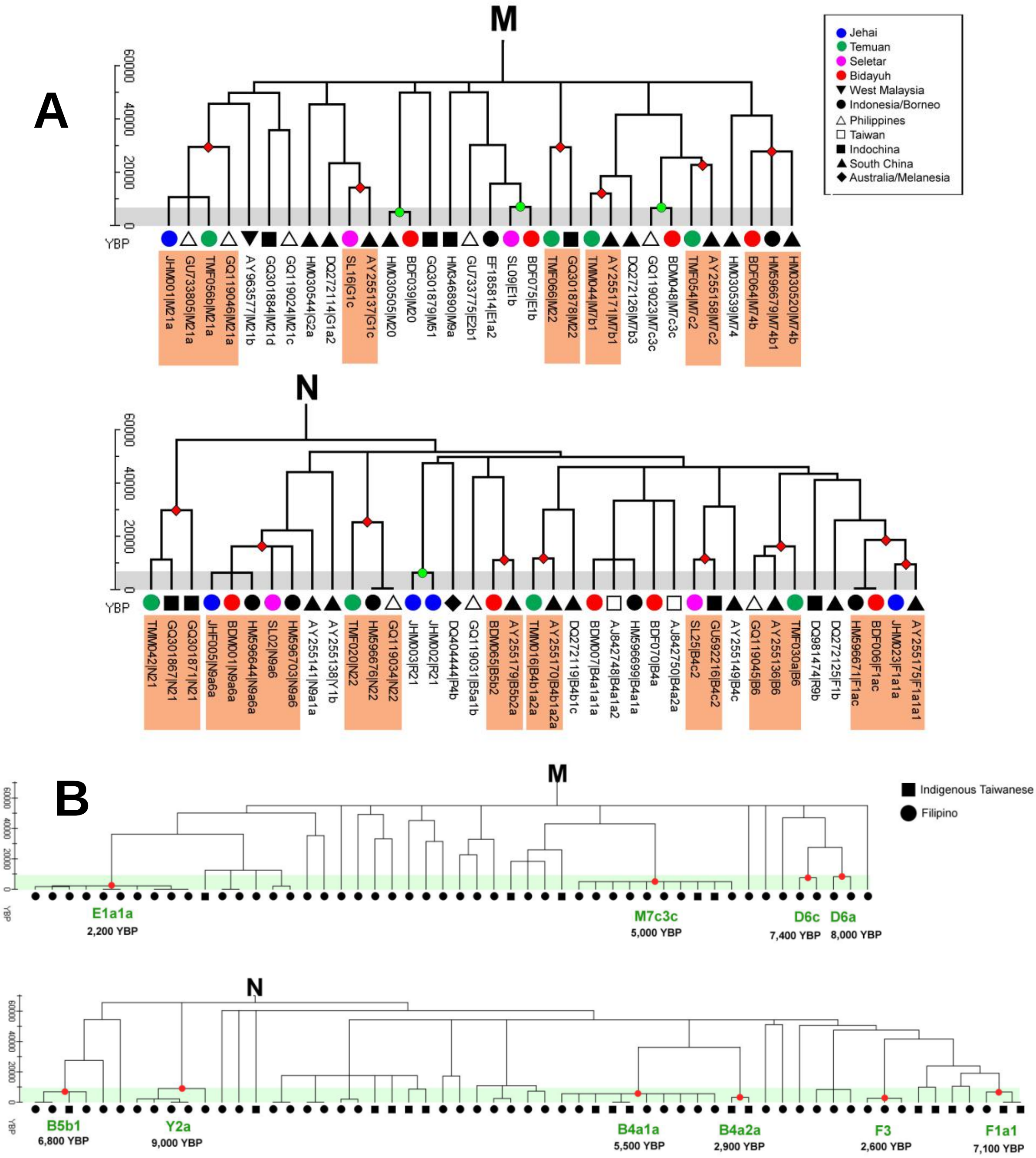


Figure 3.7: A) Maximum-likelihood tree constructed using mtDNA coding-region sequences. The molecular clock was calibrated with a mutation rate of 1.37 x 10^{-8} substitutions per site per year. Gray horizontal bar represents time frame of the Austronesian expansion from 7,000 years ago and haplogroups with coalescent times within that period are indicated with green dots. Haplogroups which support the ‘early train’ hypothesis are indicated with red diamonds and orange boxes. B) Maximum-likelihood tree of coding region mtDNA sequences from Filipino and indigenous Taiwanese (Tabadda et al. 2009; Gunnarsdottir et al. 2011; Loo et al. 2011). Horizontal green bar indicates time frame for the Austronesian expansion (earlier than 10,000 YBP). Haplogroups associated with the Austronesian expansion are shown together with their coalescence time and indicated by red dots on the nodes.

Table 3.4: Age estimates of selected haplogroups based on mtDNA coding-region sequences using Maximum-Likelihood (ML) and Bayesian MCMC methods

Haplogroup	**Maximum-Likelihood (SE)**	**Bayesian (95% HPD)**[a]
M	53,600 (10,700)	62,100 (35,800-90,900)
M20	4,700 (32)	5,900 (780-12,300)
M21a	29,200 (13,500)	22,600 (900-38,900)
M22	29,100 (13,200)	30,300 (21,600-62,400)
M74	41,100 (9,900)	39,700 (20,600-61,200)
M74b	27,500 (7,500)	26,100 (12,800-41,800)
M7	41,300 (6,700)	43,400 (22,900-66,100)
M7c2	22,500 (9,800)	22,300 (9,700-37,800)
M7c3c	6,400 (12)	5,900 (500-13,100)
M7b1	11,700 (7,000)	12,100 (4,100-21,800)
E	29,900 (9,500)	26,400 (12,300-44,800)
E1b	6,700 (5,200)	5,500 (970-11,200)
G1c	13,900 (7,000)	10,200 (3,200-18,600)
N	55,800 (13,600)	64,500 (36,300-94,700)
N21	29,300 (15,500)	26,400 (8,600-45,300)
N9a6	15,300 (7,700)	11,500 (4,600-19,700)
N9a6a	5,900 (11,000)	5,800 (1,500-10,300)
N22	24,800 (13,800)	19,400 (7,600-33,000)
R	49,400 (7,700)	48,500 (34,600-86,000)
R21	5,900 (10)	6,100 (960-12,700)
B5	41,800 (3,700)	43,200 (23,400-66,900)
B5b2	10,700 (6,600)	11,800 (3,800-21,700)
B4a	32,900 (10,900)	27,900 (14,200-44,100)
B4a1a1a	10,900 (504)	10,100 (3,100-18,400)
B4b	29,600 (8,700)	25,300 (11,800-41,100)
B4b1a2	11,200 (94)	9,100 (2,700-16,800)
B4c	30,700 (10,700)	26,100 (10,900-41,400)
B4c2	10,900 (9,100)	9,100 (1,700-17,600)
B6	15,800 (5,200)	14,400 (5,300-25,700)
F1	25,600 (6,100)	21,800 (10,700-34,800)
F1a'c	1,000 (13,400)	1,300 (330-7,900)
F1a1a	9,000 (5,100)	7,200 (1,900-13,600)

[a] 95% highest posterior density interval

3.3.5 Bayesian Skyline Plot analysis

The Bayesian Skyline Plots (BSP) in the Jehai, Temuan, Bidayuh and Seletar which were generated using coding-region sequences are shown in Figure 3.8. The BSP plots in all four populations gave a similar pattern which involved a constant population size from around 30,000 YBP followed by a decrease at around 7,000 YBP. Assuming a generation time of 25 years, the effective population size of females ranged from 2,400 to 4,000 in the Seletar and Jehai and 10,000 to 18,000 in the Temuan and Bidayuh during the time period between 30,000 to 7,000 YBP. The observed patterns did not indicate any signals of population expansion as suggested in other worldwide populations (Atkinson, Gray, and Drummond 2008; Fagundes et al. 2008) but did agree with a trend of decreasing population size starting around 6,000 to 8,000 YBP as observed in some Philippine populations (Gunnarsdottir et al. 2010). Although the exact reasons for this population decrease are unknown, one can speculate that it may be due to a disease outbreak or some natural disaster affecting Southeast Asia.

The BSP plots also showed a trend of increasing population size in all four groups around 1,000 YBP. Increases in population size tend to be associated with agriculture (Bellwood 2005) and this suggests that even in the ‘supposedly’ agricultural Austronesian groups, the shift from their traditional hunter-gathering tradition to agriculture occurred quite recently.

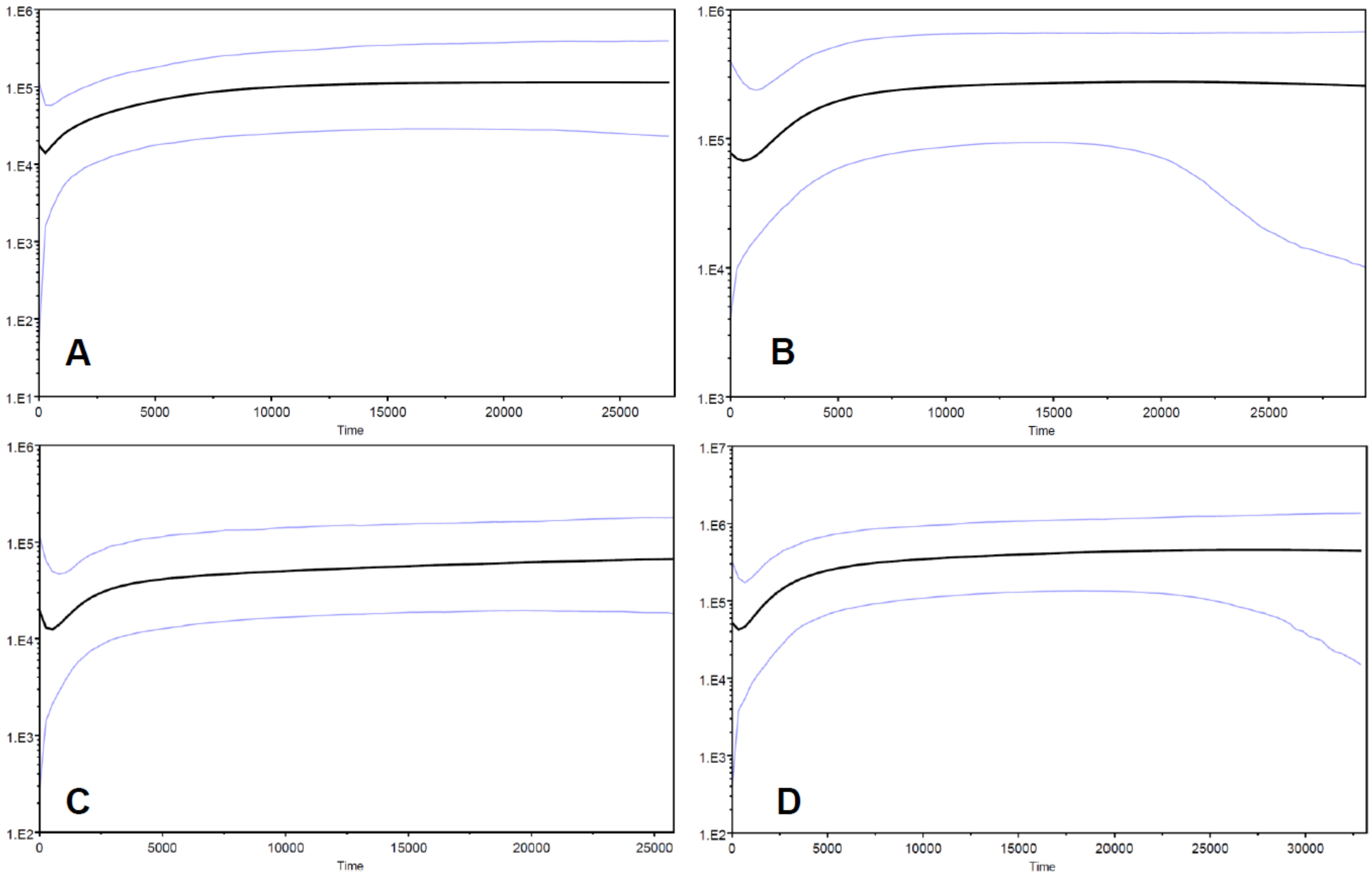


Figure 3.8: Bayesian Skyline Plots in the A) Jehai; B) Temuan; C) Seletar and D) Bidayuh. The y-axis is the product of effective female population size and generation time; x-axis is time in years. Black lines indicate mean values while blue lines denote the 95% HPD bounds.

3.4 Discussion

In this chapter, I report for the first time mtDNA diversity in four indigenous Malaysian populations using complete sequence data from all individuals sampled. This is in contrast to most studies in which complete sequencing was selectively performed on limited haplotypes based on control region diversity. Such biased sampling can lead to exaggerated results in some analyses as demonstrated in Gunnarsdottir et al. (2010). A striking feature that we observed from out data was the limited mtDNA diversity in the Seletar. With only four distinct haplogroups detected, it is reflected in the very low haplotype diversity statistic of 0.54, although not as extreme as the value of 0.167 observed in the Moken Sea Gypsies from Myanmar (Dancause et al. 2009). This limited mtDNA diversity in the Seletar may be the result of genetic drift, exacerbated by their small population size which numbers approximately 800 individuals (Nicholas 2000). This may partly explain how the haplogroup N9a6 which was reported at low frequencies in island Southeast Asia (Hill et al. 2007) rose to such high frequencies in the Selatar.

The data generated in this study provide insights into the migratory and demographic histories of Southeast Asian populations. As it stands, current data does not appear to find any similarities in extant mtDNA lineages of the Negrito groups from the Andaman, West Malaysia and the Philippines. The mtDNA diversity in each of those groups are marked by their distinctly indigenous and relict markers, namely M31 and M32 in the Andamanese (Thangaraj et al. 2003; Thangaraj et al. 2005; Barik et al. 2008), N11b (labeled as N* in the article) in the Mamanwa of the Philippines (Gunnarsdottir et al. 2010) and R21 in the Jehai and Kensiu from West Malaysia. Those mtDNA lineages have a time depth ranging from 30,000 to 50,000 YBP (Macaulay et al. 2005; Thangaraj et al. 2005; Gunnarsdottir et al. 2010), indicating their long-term presence in the Southeast Asian region, probably dating back to the original inhabitants of the region. It also appears that these Negrito groups have experienced substantial gene flow from their neighboring populations, introducing haplogroups such as M21a, F1a1a and N9a6a in the case of Malaysian Negritos. The influence of this admixture is evident from genome-wide SNP data in the Philippine and Malaysian Negritos (Abdulla et al. 2009; Jinam TA, unpublished data). Although the relatedness between these geographically distinct Negrito groups is not apparent from the point of view of the maternally inherited markers, a more exhaustive survey involving more populations and a denser set of autosomal SNP markers may yet possibly paint a clearer picture regarding their interesting past.

Regarding the Austronesian groups, our data clearly points to a more substantial influence from the mainland in shaping the mtDNA diversity of the three Austronesian groups in this study, namely the Temuan, Seletar and Bidayuh. The putative markers for the 'Out of Taiwan' expansion, B4a1a and M7c3c account for less than 10% of the mtDNA lineages in all three Austronesian groups combined. Furthermore, other markers such as Y2, D5, M7b3, F3b and F4 which were proposed to have followed the movement out of Taiwan (Hill et al. 2007; Tabbada et al. 2009), were not observed in any of the Austronesian groups in our study. An alternative explanation for the lack of 'Out of Taiwan' haplogroups in the Austronesian groups we studied may be that the Austronesian expansion involved incorporation of females from existing populations, rather than replacing them. However, a study by Jordan et al. (2009) suggested that the ancestral Austronesian populations mostly practiced matrilocal post-marital residence which would mean that the expanding Austronesian populations were more likely to retain their own mtDNA lineages. This puts some doubt to the idea of incorporation of existing females by the expanding Austronesians which would mask the Out of Taiwan signal.

Instead we found a sizeable proportion of haplogroups with links to the mainland around the vicinity of Indochina or South China with ages predating the Austronesian expansion. This is characterized by haplogroups M21a, N9a6, N21, N22 and F1a'c which account for more than 60% of the mtDNA lineages in the three Austronesian groups. This is in addition to the rare haplogroups M74b, M22, G1c, M7b1, B5b2, M7c2 and B4c2 which also has roots in the mainland. The age of those haplogroups range from 30,000 to 10,000 YBP, corresponding to the late Pleistocene to early Holocene period (Fig. 3.7A).

We therefore propose an 'early train' hypothesis for explaining the mtDNA diversity observed in the indigenous Malaysian samples, as pictured in Figure 6a. It essentially involved migration(s) originating from Indochina or South China which spread south to West Malaysia, Sumatra, Java and Borneo which were still connected as Sundaland. The timing of this migration may have ranged from 30,000 to 10,000 YBP based on the age estimates of haplogroups indicated in Figure 4. There is possibility that the migration could be even more recent, thus the haplogroup age estimates of 30,000 YBP may represent the upper limit for the possible time window of migration. However, even if we take age estimates of mtDNA lineages as the upper limit for the time of human migration events, we found that age estimates of mtDNA haplogroups associated with the Austronesian expansion (Figure 3.7B) are not too far off from the time estimated by archaeological data, which is 5,000 YBP. The dating of ~11,000 YBP for the Hoabinhian stone artifacts found in West Malaysia (Glover and Belwood 2004) corresponds to the lower limit for the migration window according to our mtDNA age estimates. In any case, our data points to possible migration event(s) which precede the proposed Austronesian expansion, hence the term 'early' train. The dichotomy in Austronesian populations observed along the second principal component in the PCA plot (Fig.3.5) lends further support to this model. However, it should be noted that interpretation of PCA analysis which was based on mtDNA haplogroup frequencies should be done with caution due to the relatively small number of individuals per population sampled in this study.

Our proposed 'early train' movement from the mainland does not exclude the episode of a Neolithic expansion from Taiwan involving Austronesian-speaking agriculturalists (Bellwood 2005). The presence of haplogroups B4a1a and M7c3c in the Bidayuh from Borneo can be taken as an indication for the 'Out of Taiwan' expansion or may reflect back migration from near Oceania to island Southeast Asia (Soares et al. 2011). It appears that Borneo may be an intersection between the 'early train' movement(s) via Sundaland and the 'express train' from Taiwan via the Philippine islands. We did find that the Bornean group that we sampled, the Bidayuh, had closer affinities to the Proto Malay from West Malaysia whereas the other two Borneo populations, Banjarmasin (Hill et al. 2007) and Iban (Simonson et al. 2011) have closer affinities to island and continental groups, respectively. This highlights the heterogeneity of the origins of the populations in Borneo who are composed of various tribes with diverse languages and cultures.

The proposed 'early train' dispersal of maternally inherited markers from the mainland prior to the Neolithic Austronesian expansion was earlier hinted by Hill et al. (2006; 2007) from extensive sampling of populations in island Southeast Asia but based mainly off control-region mtDNA sequences. Archaeological data points to the existence of the Hoabinhian tradition, characterized by flaky pebble tools (Glover and Bellwood 2004). The Hoabinhian are thought to have emerged from Indochina during the early Holocene period (~11,000 YBP) and spread southwards via West Malaysia, possibly bringing with them Austro-Asiatic languages (Bellwood 2007) and their mtDNA lineages. Other lines of evidence which corroborate our data include Y-chromosomal markers, which suggests a Paleolithic (30,000 to 15,000 YBP) contribution from mainland Asia (Karafet et al. 2010) and also data from pig domestication which proposes a migration from East Asia via Sundaland and into the Pacific region (Larson et al. 2007). However, some caveats are in order when interpreting these results. It should be noted that the age of an mtDNA haplotype does not necessarily equate to its age in a population. It may well be possible that an 'old' haplotype was introduced into a population by recent migrations. It would therefore be desirable in future studies to vigorously test our 'early train' model against other competing and plausible scenarios using demographic modeling, such as those reported by Batini et al. (2011).

Taken together, our results suggest an ‘early train’ wave(s) of migration originating from South China or Indochina during late Pleistocene to early Holocene (30,000 to 10,000 YBP), predating the Neolithic expansion from Taiwan (Glover and Bellwood 2004; Bellwood 2005; Bellwood 2007). We do not refute that the Out of Taiwan migration had taken place, but it looks improbable that it contributed significantly to the mtDNA diversity in Austronesian groups, particularly those west of the Wallace line. In conclusion, our data suggests a more intricate migration history than the generally accepted, if not oversimplified, two-wave hypothesis regarding the peopling of Southeast Asia.

CHAPTER 4

Population history and genetic affinities of Japanese populations based on genome-wide SNP data

4.1 Introduction

The origin of the Japanese population is another topic that has long been debated by archaeologists, linguists, anthropologists and now most recently by geneticists. Archaeological records point to the arrival of modern humans to the Japanese Archipelago to about 40,000 YBP (Imamura 1996). During this Paleolithic period up until the last glacial maximum (~20,000 YBP), the current Japanese islands are connected to the mainland via the current Korean peninsula on the south and via the current Sakhalin island to the north. This allows for episodes of human movements between the mainland and the Japanese Archipelago. The next division in the prehistory of Japan is called the Jomon period beginning from approximately 15,000 YBP. The Jomon period is characterized by distinct cord marks found on pottery and is further subdivided into incipient, initial, early, mid-, late, final and epi-Jomon stages. Following the Jomon period, the subsequent stage is known as the Yayoi period, beginning from ~3,000 YBP. The defining feature of Yayoi period is the introduction of agriculture, as opposed to the hunter-gatherer lifestyle practiced by the Jomon people. Agriculture is thought to be introduced by the human migrants from the Asian mainland via the Korean peninsula (Hudson 2006).

With these two major timelines in mind, several theories have been proposed regarding the origins of the current Japanese populations (Suzuki 1983). One of them is the replacement theory which posits that the incoming Yayoi migrants totally displaced the existing Jomon populations in the Japanese Archipelago but this theory failed to gain much support and has fallen out of favor. The other two models are known as the transformation theory and the admixture theory.

According to the transformation theory, the current Japanese populations originated from an ancient migration from the mainland during the Pleistocene (~30,000 YBP) and eventually formed the Jomon people. It supposes that these Jomon people gradually transformed or evolved into the current Japanese population and that the Yayoi migration did not contribute significantly to the genetic makeup of the modern Japanese (Suzuki 1963; Mizoguchi 1968). The admixture theory is also known as the dual-structure hypothesis (Hanihara 1991). In this theory, the ancestors of the Jomon people are believed to have originated in Southeast Asia and migrated to the Japanese Archipelago during the Pleistocene period. During the Yayoi period, the incoming migrants from the Korean peninsula gradually pushed the indigenous Jomon to the southern and northern islands of Ryukyu and Hokkaido, respectively. In the process, there were some admixture events between the Yayoi and Jomon which led to origins of the current Japanese populations in the main island of Hondo and Kyushu. The indigenous groups known as the Ainu in Hokkaido and the Ryukyuans in the southern Ryukyu Islands are believed to be direct descendants of the Jomon people and that they experienced less admixture with the Yayoi than their counterparts on the Hondo and Kyushu islands.

It should be noted that all the above hypotheses were mainly based on morphological features such as cranial and dental measurements (Suzuki 1983). A spate of studies employing genetic markers such as mitochondrial, Y-chromosomal and autosomal DNA also tried to tackle the issue surrounding the origins of the Japanese. Support for the dual-structure hypothesis included studies based on mitochondrial DNA sequences (Horai et al. 1996; Tanaka et al. 2004) and Y-chromosomal SNP and Short Tandem Repeat (STR) polymorphisms (Hammer et al. 2005). Analyses involving 'classical' protein markers led to a partial support of the dual-structure model, supporting the shared ancestry of the Ainu and Ryukyuans dating back to the Jomon people, but found no close relationship between them and Southeast Asians (Omoto and Saitou 1997).

Another study by Nei (1995) showed the Japanese populations, including the Ainu and Ryukyuans, have a closer affinity to northern East Asians than to Southeast Asians, leading him to propose an out-of-Northeast-Asia theory which was not too dissimilar from the earlier transformation theory (Suzuki 1963; Mizoguchi 1968).

Previous genetic studies tended to rely on uniparentally-inherited markers or on limited numbers of autosomal markers. A recent study using genome-wide SNP data showed clear differences between the Ryukyuans and other Japanese from the main islands and also some genetic substructure within the Hondo Japanese (Yamaguchi-Kabata et al. 2008). Another study involving 73 Asian populations which included Ryukyuans and Hondo Japanese suggests a mainly south to northern migration into the origins of East Asians (Abdulla et al. 2009). However, both of those genome-wide studies did not include data from the Ainu. The availability of archival DNA in the Ainu as well as the advancement of high throughput SNP genotyping for this study allows us to examine in closer detail the genetic substructure within Japanese populations and ultimately try to answer questions regarding which model of Japanese origins (Suzuki 1983) would be best supported by the genome-wide SNP data.

4.2 Materials and methods

4.2.1 Sample data, ethical approval & SNP genotyping

Blood samples of the Ainu people were collected by Keiichi Omoto and his colleagues from late 1970s to early 1980s in Hidaka District of Hokkaido. Extracted DNA samples were used for mitochondrial DNA and Y chromosome studies (Harihara et al. 1988; Horai et al, 1996; Tajima et al. 2002, 2004; Hammer et al. 2006) and have since been archived at the University of Tokyo Medical School. Following the advancement of SNP genotyping technology, the archival samples were used for high throughput genotyping. In addition, individuals from the Ryukyu Islands were recruited at the Department of Medical Genetics, University of the Ryukyus Graduate School of Medicine from April 2004 to 2008.

A total of 36 Ainu and 38 Ryukyuan samples were genotyped using the Affymetrix Genome-Wide SNP 6.0 microarray platform. All genotyping experiments were conducted by technicians in Professor Katsushi Tokunaga's lab at the Department of Human Genetics, University of Tokyo. The workflow of the genotyping experiments was essentially similar to the Affymetrix Xba I assay (Chapter 2) except that in the Genome-Wide SNP 6.0 assay, two separate reaction enzymes (*Nsp I* and *Sty I)* were used. At the end of the workflow, the raw intensity files were converted to genotype calls using Affymetrix's Birdseed Ver2 algorithm, resulting in 906,600 SNP genotypes per sample.

In addition to the Ainu and Ryukyu populations, SNP genotype data from 200 mainland Japanese mostly from the Kanto area (Nishida et al. 2008) which were generated using the same method, were obtained. These three groups (Ainu, Ryukyuan and Kanto Japanese) form the Japanese population dataset. This study was approved by the Research Ethics Committee of the Faculty of Medicine, The University of Tokyo. This Japanese population dataset was further augmented with genotype data from four HapMap populations, namely Yorubans from Africa (YRI), Europeans (CEU), Han Chinese from Beijing (CHB) and Japanese from Tokyo (JPT) which were also generated using the same genotyping method.

4.2.2 Data filtering & quality checks

From an initial number of 906,600 SNPs, filtering was done to exclude SNPs from the mitochondrial, X, and Y chromosomes. Duplicate SNPs and SNPs without a dnSNP ID were also filtered out, resulting in a total of 868,257 remaining SNPs. Individual samples with poor genotyping performances were further filtered out based on the Affymetrix contrast quality control (cQC) threshold of 0.04, as recommended by the manufacturer. Three Ryukyuan and two Kanto Japanese samples were omitted based on this criterion. However, in the Ainu population, only 13 out of 36 individuals passed the cQC threshold. This was probably due to the degradation of DNA quality in the archival samples. In order to maximize the number of Ainu individuals to be used for downstream analysis, further SNP filtering was done based on confidence scores for each SNP generated during genotype calling using the Affymetrix Birdseed Ver2 algorithm. In general, SNPs with a confidence score more than 0.1 are more likely to have failed genotype calling (i.e. 'no calls'). By visually inspecting genotype cluster graphs of random SNPs with confidence scores ranging from 0.1 to 0.004, a more stringent cutoff of 0.008 was used to exclude under-performing SNPs while retaining the maximum number of individuals.

Thus, based on this criterion, 212,448 SNPs were omitted from the set of 36 Ainu individuals, resulting in 656,237 remaining SNPs. SNP data in the Ainu and all other populations which were generated using the Affymetrix Genome-Wide 6.0 Assay were further filtered to remove SNPs with call rate less than 95% and which deviate from Hardy-Weinburg equilibrium ($p<0.001$). The filtering steps were done on each population separately and the number of SNPs filtered out are shown in Table 4.1. After merging SNP data from all populations, the final number of SNPs in the Japanese and HapMap datasets were 644,149.

Table 4.1: SNP filtering applied to the Japanese and HapMap dataset

Population	n	Number of SNPs omitted		Remaining SNP
		Genotyping call rate (<95%)	HWE (p<0.001)	
Ainu[a]	36	0	449	655,788
Ryukyu	35	29,874	538	837,845
Kanto Japanese	198	17,169	1888	849,200
CHB	42	3,069	336	864,852
JPT	45	5,004	446	862,807
CEU	89	4,887	514	862,856
YRI	89	4,780	706	862,771

[a] Starting from 656,237 SNPs

4.2.3 Merging with other population data

In addition to the Japanese and HapMap population datasets, we also included SNP data from other populations available from public databases. These included the Human Genome Diversity Project (HGDP-CEPH) dataset which consists of 650,000 SNPs from 51 worldwide populations (Li et al. 2008) and the Pan-Asian SNP (PASNP) dataset which consists of 54,794 SNPs from 73 populations from Asia (Abdulla et al. 2009). The number of SNPs that overlap between the Japanese-HapMap datasets with the HGDP-CEPH panel was 114,001. After applying filters (excluding SNP with less than 95% genotype call rate and minor allele frequency less than 1%) in the merged dataset, there were 101,562 SNP remaining. For merging the Japanese-HapMap population datasets with the PASNP data, 15,526 overlapping SNPs from both datasets were extracted and merged. After applying the same filtering criteria as above, the number of remaining SNP was 14,997. The combination of Japanese, HapMap, HGDP-CEPH and PASNP datasets yielded only 4,237 overlapping SNPs. All filtering and merging steps were done using PLINK software (Purcell et al. 2007).

4.2.4 Data analysis

Subsequent analysis was done using different combinations of the above datasets. For the merged data from all datasets, only populations from East Asia (Table 4.2) were used for analysis. Principal Component Analysis was done using the *smartpca* program from the EIGENSOFT software package (Patterson, Price, and Reich 2006). To examine population structure and admixture, a more computationally efficient STRUCTURE-like program called *frappe* which is based on maximum-likelihood methods (Tang et al. 2005), was used. Population-based phylogenetic trees were constructed based on SNP allele frequencies using CONTML program from the PHYLIP package (Felsenstein 2005) with 100 bootstrap replicates.

Table 4.2: East Asian populations from HDGP-CEPH and PASNP datasets which were merged with the Japanese-HapMap datasets

Population ID	Ethnicity	n	Dataset	Geographical origin
Dai	Dai	10	HGDP-CEPH	China
Daur	Daur	9	HGDP-CEPH	China
Han	Han	44	HGDP-CEPH	China
Hezhen	Hezhen	9	HGDP-CEPH	China
Japanese	Japanese	28	HGDP-CEPH	Japan
Lahu	Lahu	8	HGDP-CEPH	China
Miaozu	Miaozu	10	HGDP-CEPH	China
Mongolia	Mongolia	10	HGDP-CEPH	China
Naxi	Naxi	8	HGDP-CEPH	China
Oroqen	Oroqen	9	HGDP-CEPH	China
She	She	10	HGDP-CEPH	China
Tu	Tu	10	HGDP-CEPH	China
Tujia	Tujia	10	HGDP-CEPH	China
Xibo	Xibo	9	HGDP-CEPH	China
Yakut	Yakut	25	HGDP-CEPH	Siberia
Yizu	Yizu	10	HGDP-CEPH	China
AX-AM	Ami	10	PASNP	Taiwan
AX-AT	Atayal	10	PASNP	Taiwan
CN-CC	Zhuang	26	PASNP	China
CN-HM	Hmong	26	PASNP	China
CN-UG	Ugyur	26	PASNP	China
CN-SH	Han	21	PASNP	China
CN-JN	Jinuo	29	PASNP	China
JP-ML	Mainland Japanese	71	PASNP	Japan
JP-RK	Ryukyu	49	PASNP	Japan
CN-WA	Wa	56	PASNP	China
KR-KR	Korean	90	PASNP	Korea
TW-HA	Han	80	PASNP	Taiwan
CN-JI	Jiamao	31	PASNP	China
CN-GA	Han	30	PASNP	China

4.3 Results

4.3.1 Genetic substructure within Japanese populations

To have an initial idea on how the three Japanese groups relate to other global populations, Principal Component Analysis (PCA) was done using approximately 600,000 SNP data from the Japanese and HapMap datasets. Figure 4.2(A) shows that all individuals fall into three clusters corresponding to Africans, Europeans and East Asians. The Ainu and Ryukyuans are clustered with other East Asian populations and showed no clear relationships with current African or European populations. It was the same case when African populations were omitted from the PCA analysis, shown in Figure 4.2(B). At this point however, we start to see some extensive inter-individual variation among the Ainu compared to other populations. To have a closer look at the relationships between East Asian populations, PCA was performed only on the three Japanese populations and the HapMap Han Chinese (CHB). The resulting PCA plots are shown in Figure 4.3. The first principal component (PC) explains 1.8% of the variation between individuals and separates the Ainu from the rest of the Japanese and Chinese populations. The second PC distinguished between the Ryukyuans, Kanto Japanese and Han Chinese reminiscent of a South (Ryukyuan) to North (Han Chinese) cline. The third PC separates several Kanto Japanese from the others and interestingly these individuals appeared closer to some Ainu individuals along the third PC plane.

An interesting pattern observed in Figure 4.3 is the substantial inter-individual variation amongst the Ainu compared to other Japanese and Han Chinese populations. To further examine this interesting pattern, PCA was performed on the three Japanese populations separately and the results are shown in Figure 4.4. For each of the three Japanese populations, several outlier individuals were observed. In the case of the Ryukyuans and Kanto Japanese, these outlier individuals do not seem to be related to any particular population based on their coordinates in Figure 4.3. In the Ainu however, we observed three individuals who cluster with the Kanto

Japanese and five other individual outliers who do not seem to be related to any other population. The presence of these outliers in the Ainu resulted in a triangular-like pattern in the PCA plot in Figure 4.4(A) and may be the result of recent admixture involving two different source populations.

To see if this was the case, we calculated the allele sharing distance between the Ainu and Kanto Japanese individuals. As shown in Figure 4.5, there is clear trend involving the allele sharing distances and the PC1 coordinates between the Ainu and Kanto Japanese. Ainu individuals within the Kanto Japanese cluster had the least genetic distance with the Kanto Japanese and conversely, Ainu individuals farthest from the Kanto Japanese on the PC1 axis tend to have greater genetic distance from the Kanto Japanese. These observations indicate that the allele sharing between the Ainu and the Kanto Japanese was the result of relatively recent and continuous episodes of gene flow with the mainland Japanese.

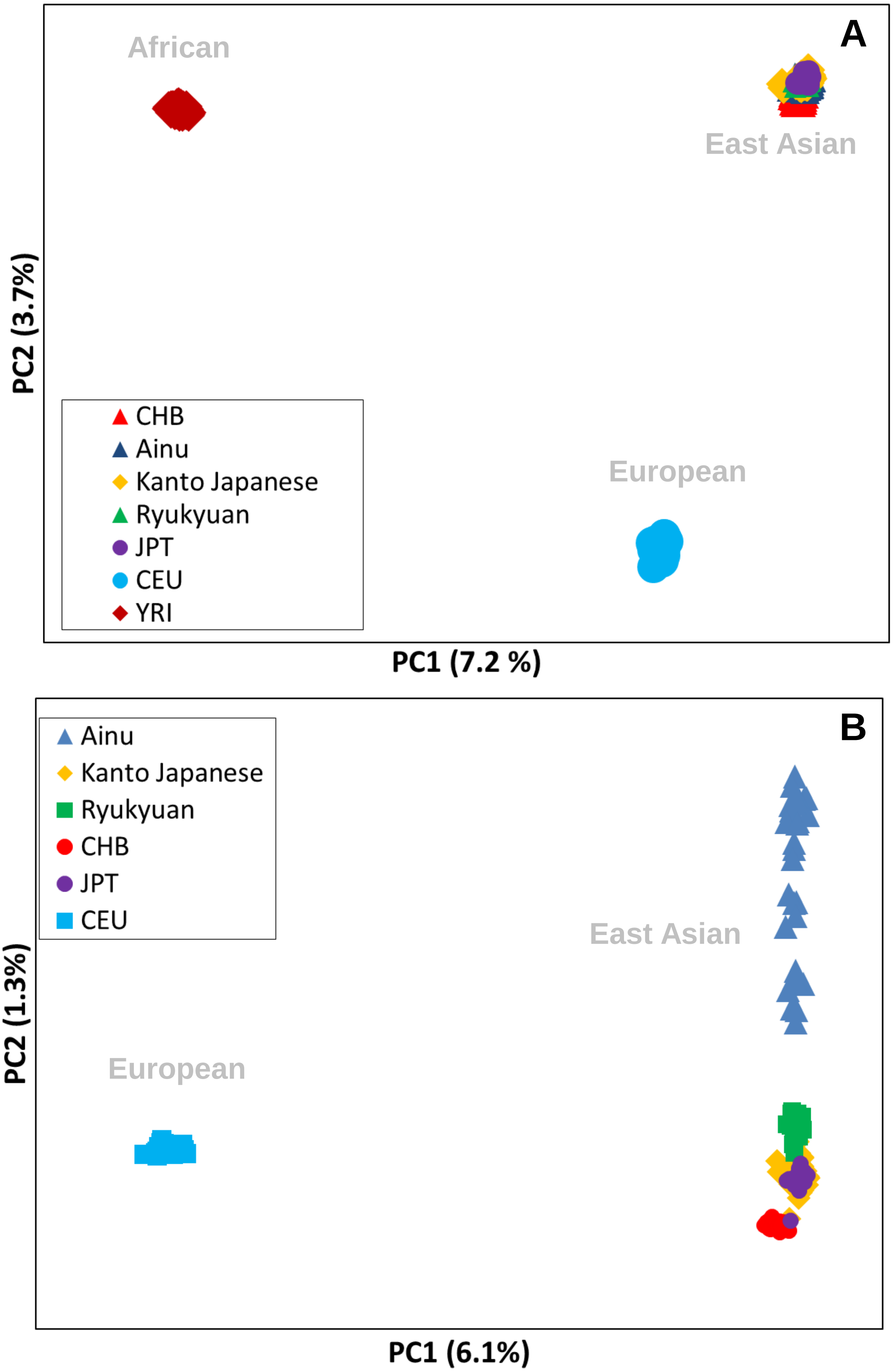


Figure 4.1: PCA plots of Japanese and HapMap individuals. **A)** All individuals included. **B)** Without Africans (YRI).

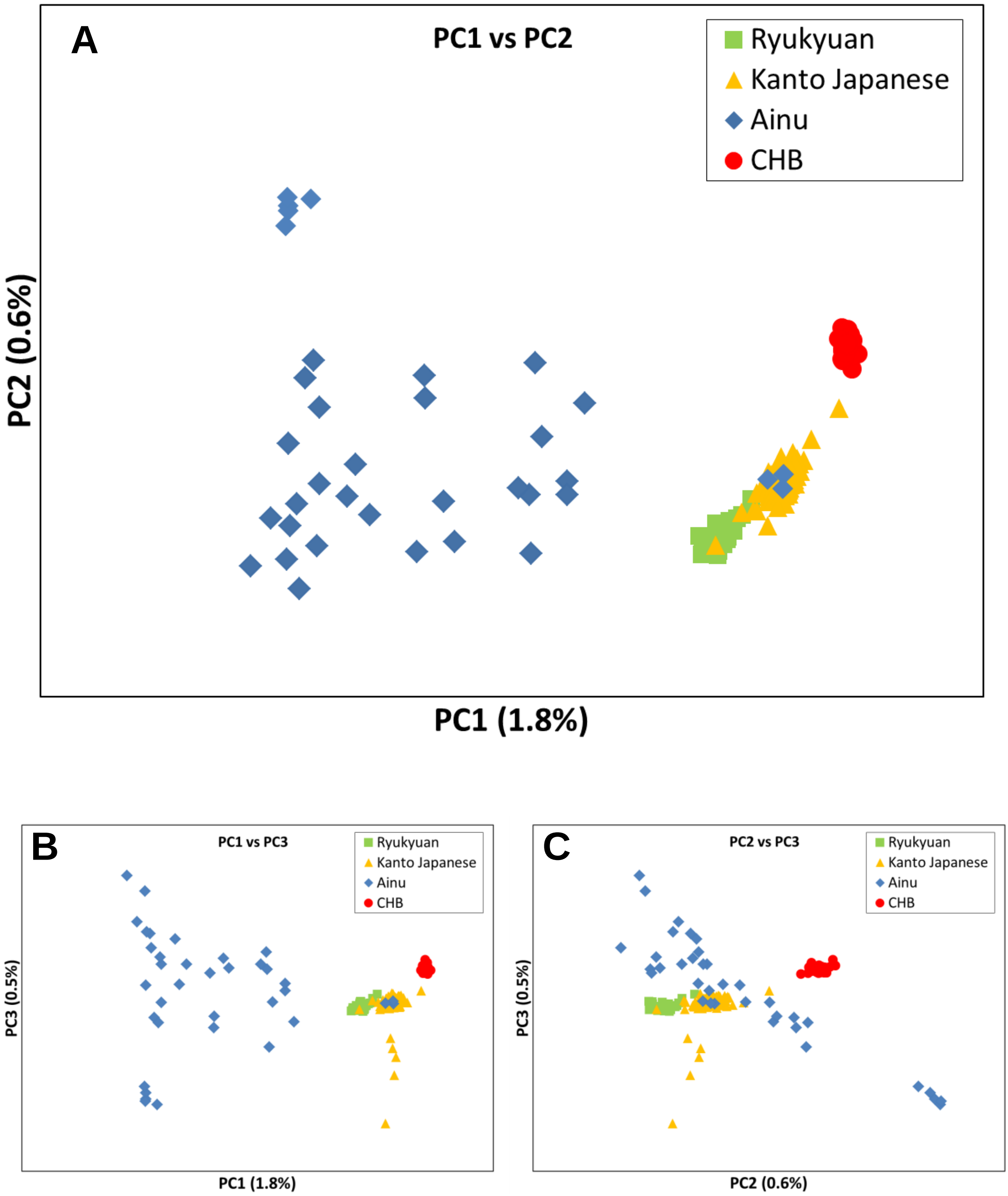


Figure 4.2: PCA plots of only Japanese and HapMap Han Chinese (CHB) individuals. **A)** PC1 vs. PC2; **B)** PC1 vs PC3; **C)** PC2 vs PC3.

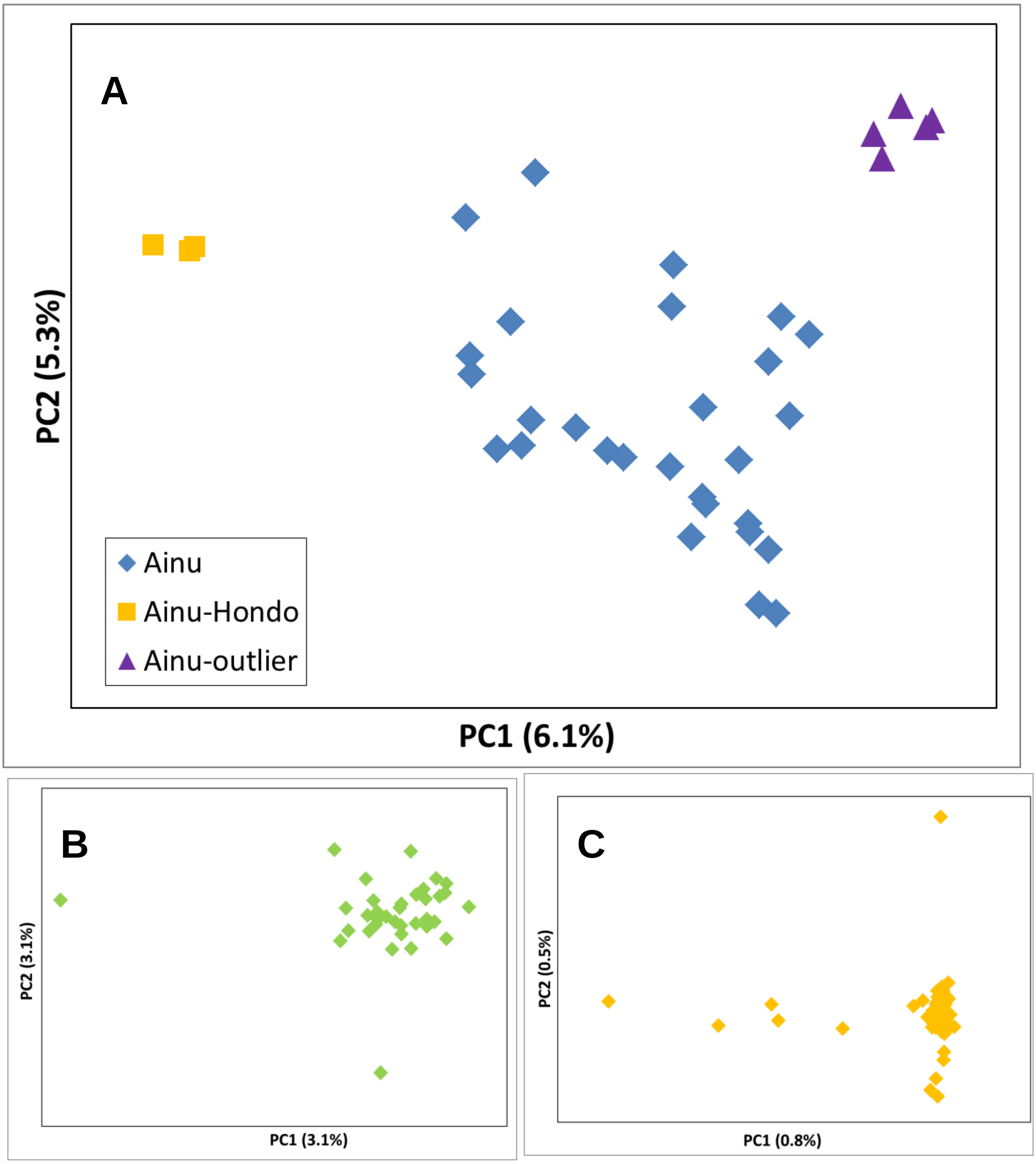


Figure 4.3: Separate PCA plots of Japanese groups. **A)** Ainu; **B)** Ryukyuan; **C)** Kanto Japanese.

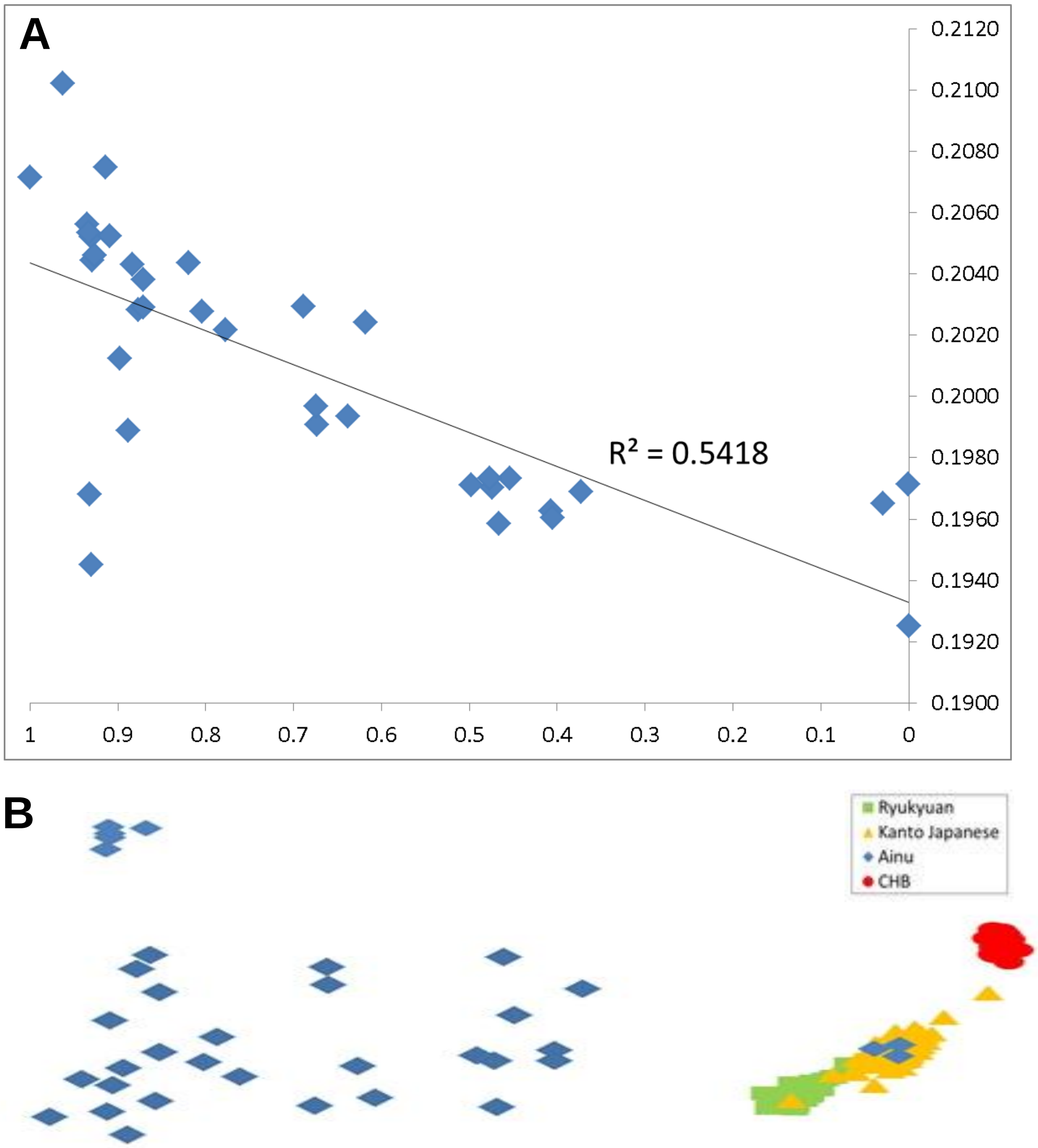


Figure 4.4: Correlation between allele sharing distance with Principal Component 1 coordinates between the Ainu and Kanto Japanese. **A)** Each dot represents an Ainu individual and the y-axis values are the average allele sharing distances between Ainu and Kanto Japanese individuals. The x-axis values are the PC1 coordinates, normalized to range from 0 to 1 so that the value of 0 represents individuals within the Kanto Japanese cluster. **B)** PCA plot depicting the positions of the Ainu individuals along the first principal component (x-axis) relative to the Kanto Japanese.

The Neighbor-Joining tree which was constructed from the allele sharing distance matrix between the Japanese and Han Chinese individuals (Figure 4.6) shows the positions of the admixed Ainu within the cluster of Kanto Japanese individuals, otherwise all other Ainu are clustered in one branch. When looking at the allele sharing distances among individuals within each group separately, we found that the variance in the mean allele sharing distance between Ainu individuals was the highest at 6.89 x 10^{-5} compared to 5.63 x 10^{-7} and 7.15 x 10^{-7} in the Ryukyuan and Kanto Japanese, respectively.

The results of *frappe* analysis are shown in Figure 4.7. When k=2, the ancestry components corresponds to Ainu (dark blue) and other East Asians (orange). It shows that the Ainu already had varying degrees of East Asian components as implied in the PCA analysis. Conversely, all other populations had different amounts of Ainu ancestry, with the highest in the Ryukyuans, followed by the Kanto Japanese and the lowest in the Han Chinese. As k is increased to 3, the new component corresponds to the Han Chinese (red) and is found at approximately 30% in the Kanto Japanese. At k=4, the outlier individuals initially observed in the PCA plot (Figure 4.4) were differentiated from the rest of the Ainu and are indicated in purple. Finally the ancestry component of the Ryukyuans was observed when k=5. In general, the *frappe* results appear to be consistent with the PCA analysis, which shows varying amounts of admixture in the Ainu with the Kanto Japanese and also confirms the presence of another source population which may contribute to the genetic structure in the Ainu.

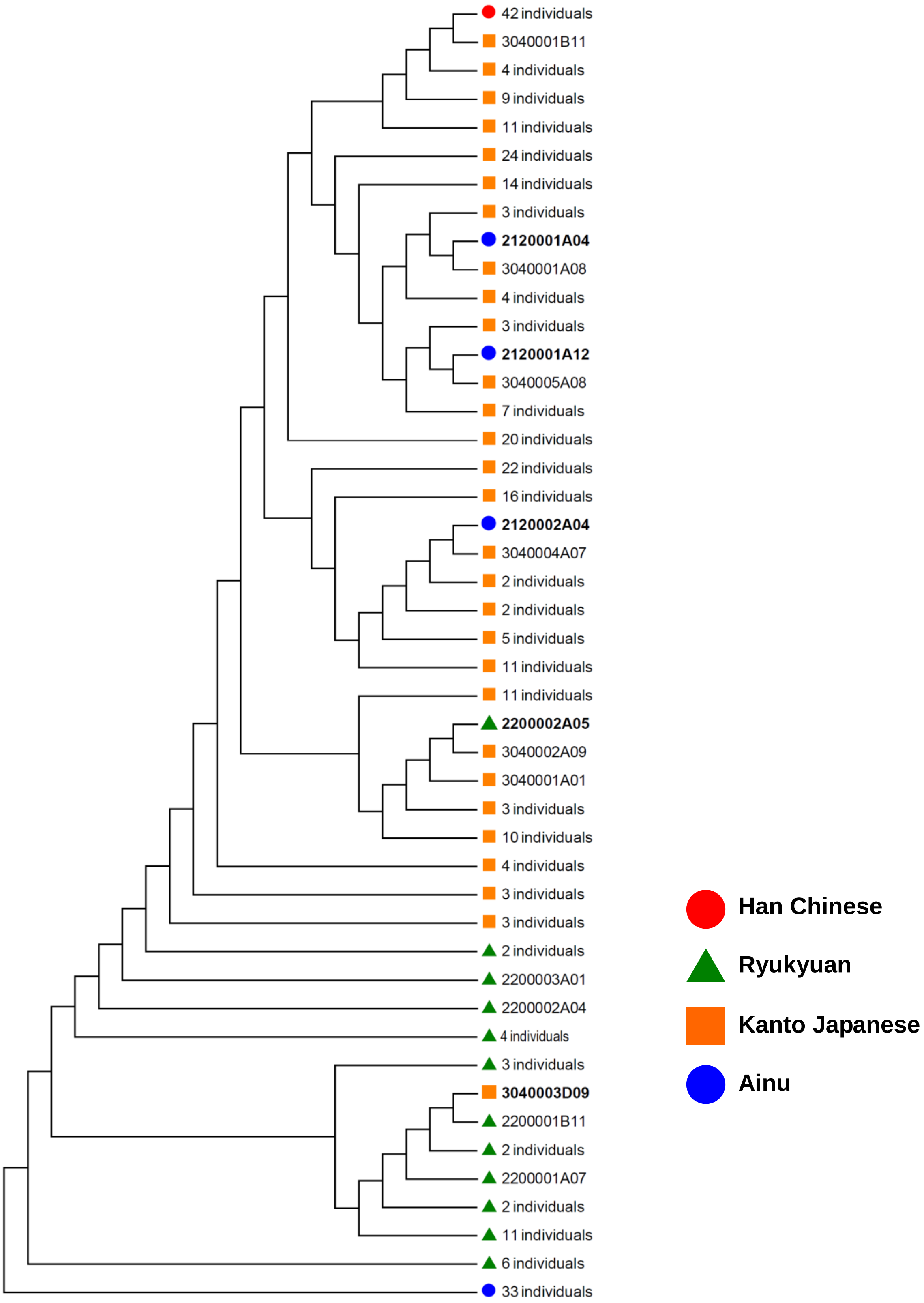


Figure 4.5: Neighbor-Joining tree of Japanese and Han Chinese based on allele sharing distances between individuals. Clades with two or more individuals are compressed. Individuals are indicated by numerical IDs and their respective population symbols.

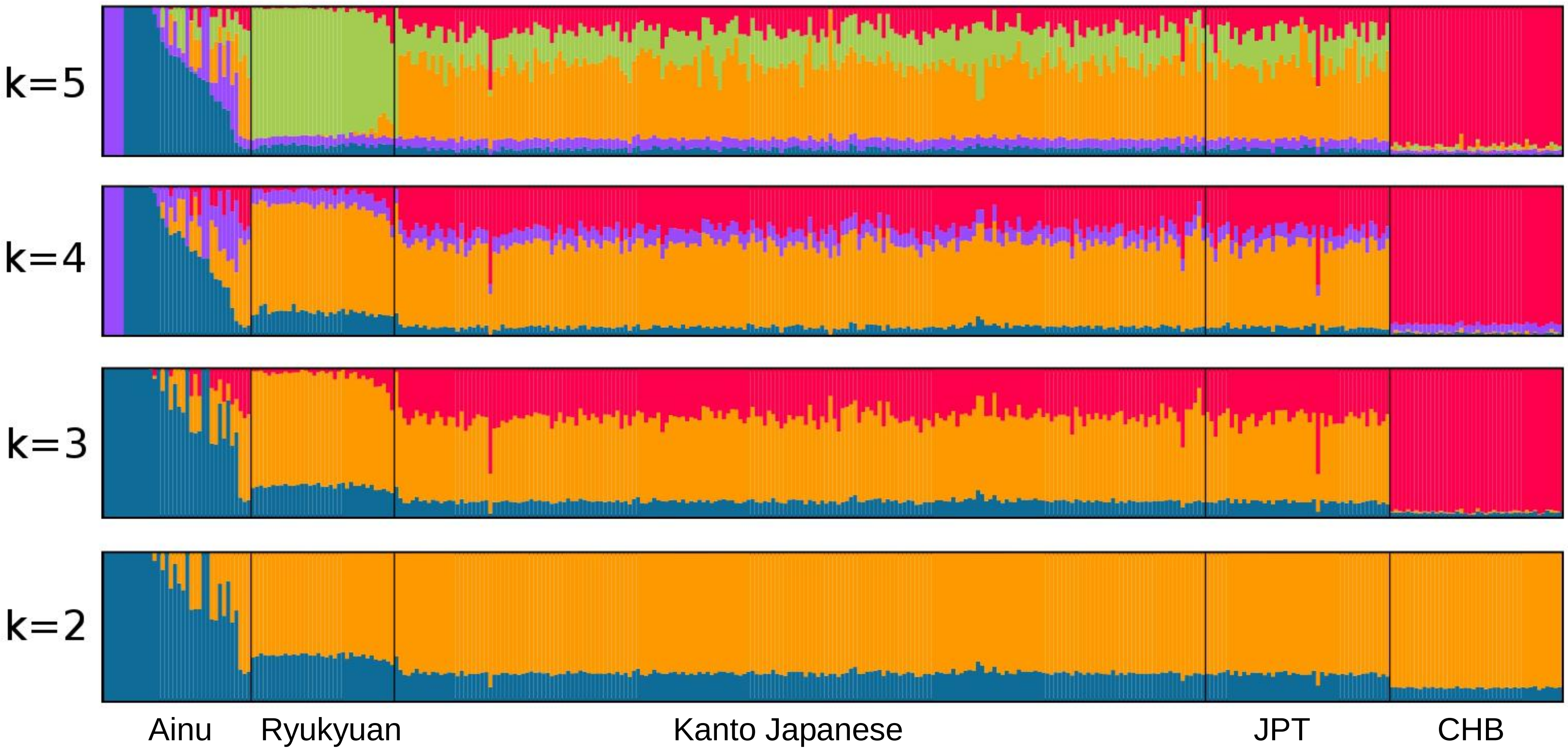


Figure 4.6: Results of *frappe* analysis from k=2 to k=5 in the three Japanese groups and HapMap Japanese (JPT) and Han Chinese (CHB). Each individual is represented by a vertical line and different colors represent different ancestry components at each k.

4.3.2 Phylogenetic analysis and relationships between populations

To see the relationships between the three Japanese groups with other worldwide populations, phylogenetic trees were constructed using merged SNP data with the HGDP-CEPH and PASNP datasets. The Neighbor-Joining (NJ) tree of the Japanese and HGDP-CEPH populations based on allele frequencies of ~100,000 SNPs is shown in Figure 4.8 whereas the NJ tree using only Japanese and PASNP populations using ~14,000 SNPs is shown in Figure 4.9. In both NJ trees, the three Japanese groups clustered with other East Asian populations. In figure 4.8, the populations that are geographically closer to the Japanese are the Hezhen, Daur, Oroqen and Mongolians and they occupy the basal locations in the East Asian clade. In Figure 4.9 the Japanese are closest to the Koreans (KR-KR) and the Han Chinese from Shanghai (CH-SH) and Taiwan (TW-HA, TW-HB). Comparisons with only East Asian populations confirm the three Japanese populations tend to group together, with the Koreans the only other population in this ‘Japanese’ clade (Figure 4.10A). In all NJ trees the Ainu and Ryukyuans cluster together despite the fact that their geographical locations are the two opposing ends of the Japanese Archipelago. An unrooted NJ tree using only the Japanese, Koreans and Han Chinese data in Figure 4.10B show the close relationship between the Ainu and Ryukyuans is supported by 100% bootstrap probability. The very short branch leading to the Kanto Japanese as well as its intermediate position between the Ainu-Ryukyu and Chinese-Korean branches suggest that the current Kanto Japanese may be the result of admixture between these two ancestral population sources.

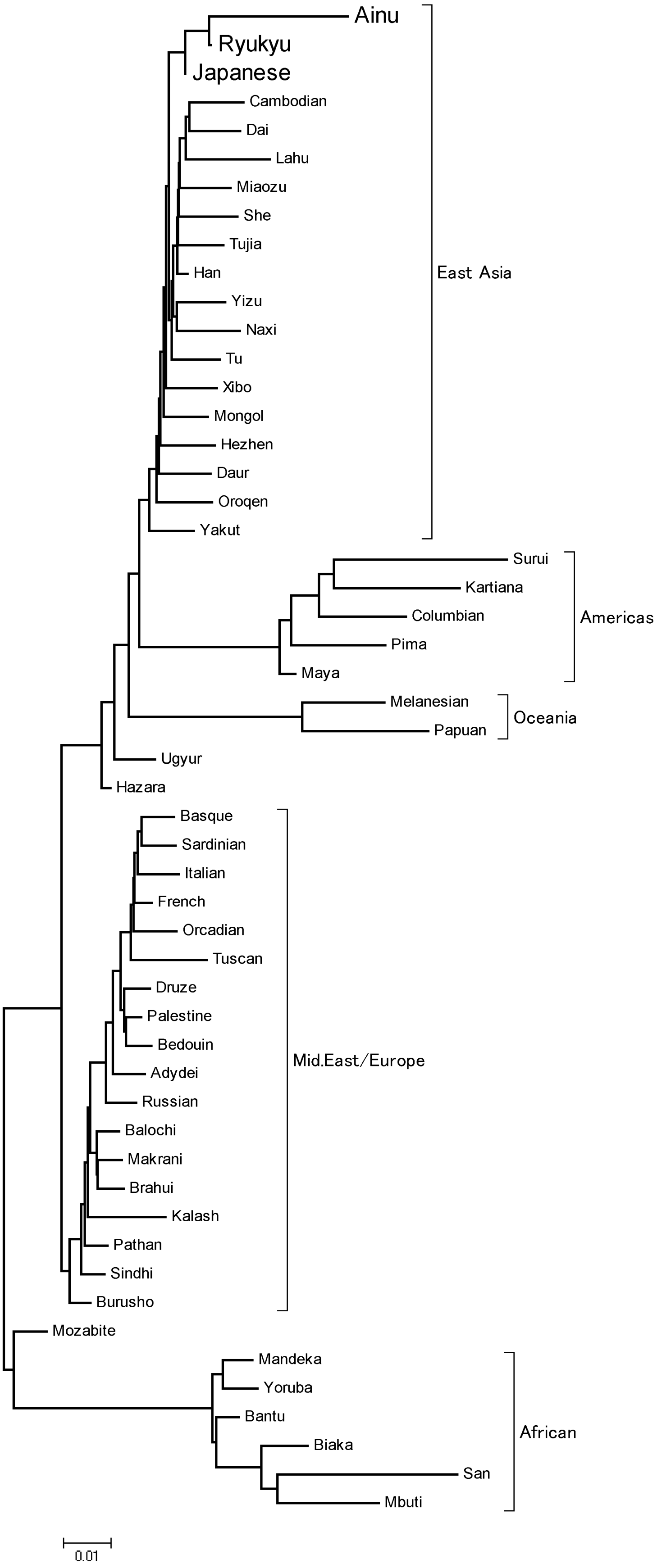


Figure 4.7: Neighbor-Joining tree of Japanese and HGDP-CEPH populations

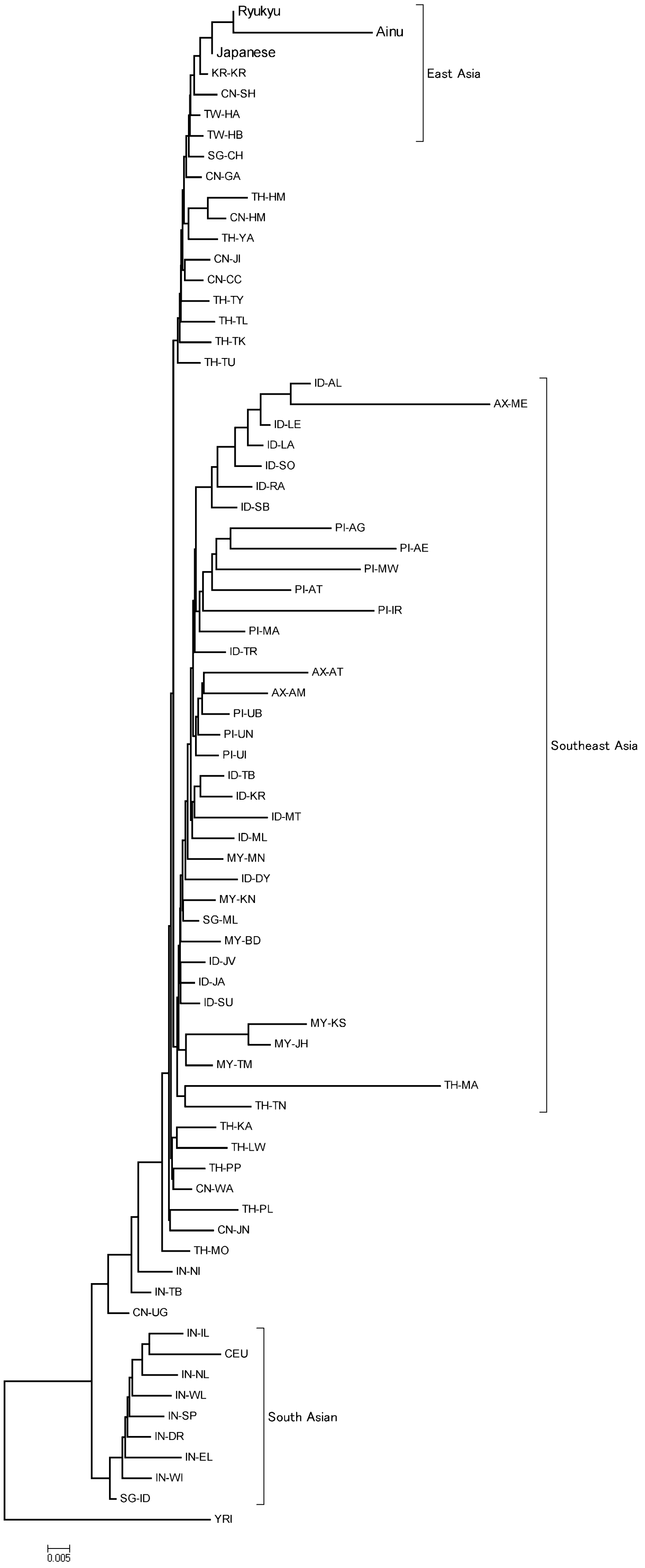


Figure 4.8: Neighbor-Joining tree showing topology of Japanese and PASNP populations

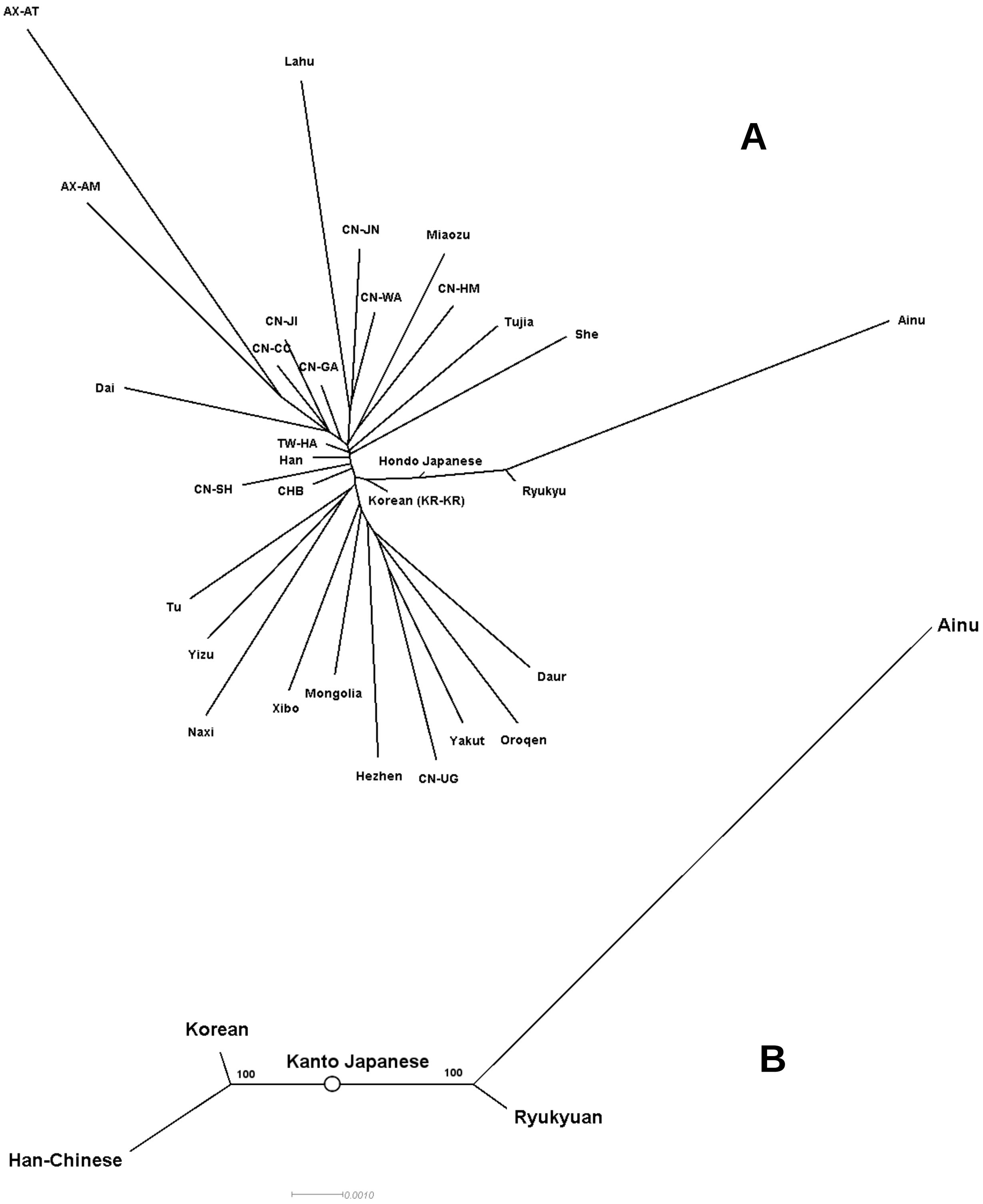


Figure 4.9: Unrooted Neighbor-Joining tree depicting the relationships between A) the Japanese and other East Asian populations and B) the Japanese, Koreans and Chinese. Numbers indicate bootstrap support for the branches (in percentage).

4.4 Discussion

This chapter reports the genetic variation in the Japanese, including two indigenous minority groups, using genome-wide SNP data. While genome-wide analyses have been reported previously in the Ryukyuans and mainland Japanese (Yamaguchi-Kabata et al. 2008; Abdulla et al. 2009), this is the first report of genetic variation within the Ainu using dense genome-wide SNP data. Previous genetic studies in the Ainu were mostly on population level, assuming that the individuals are relatively homogenous. However an interesting pattern emerged when we analyzed them at the individual level using PCA. It appears that the Ainu individuals displayed greater variation between individuals compared to other Japanese populations, as shown by their scattered positions in the PCA plots (Figures 4.2 to 4.4) and the large amount of variance in their allele sharing distances. Such a scattered pattern on the individual-based PCA plots were most often observed in other indigenous populations such as Australian Aboriginals (McEvoy et al. 2010), Latinos (Bryc et al. 2010), Negritos (Chapter 2) or Indians (Reich et al. 2009). In those studies, a typically linear pattern was observed in the PCA plots, indicating an admixture gradient with a neighboring source population. Such was also the case in the Ainu, whereby the source of the admixture was the mainland Japanese, as indicated in Figure 4.5. However admixture with the mainland Japanese can only partially explain the PCA pattern observed within the Ainu, as several other outlier individuals were observed in the PCA plots as well as in the *frappe* analysis.

Unlike admixture with the mainland Japanese, it is difficult to ascertain the other potential source of admixture in the Ainu without a proper source population. It can only be assumed at this point, that the source population may be somewhere from Northeast Asia, judging from the position of the outlier Ainu individuals who are above the Han Chinese in Beijing along PC2 (y-axis) in Figure 4.3(A). Although it is merely a conjecture at this point, previous studies do support the idea of contact with Northern populations which may have contributed to the genetic diversity in the Ainu. Archaeological data points to an introduction of Satsumon cultures by the Okhotsk peoples from Sakhalin into Hokkaido during the 7th to 10th centuries (Imamura 1996; Hudson 2006). Genetic studies using mitochondrial DNA (Tajima et al. 2004) and Human Leukocyte Antigen (HLA) loci (Tokunaga et al. 2001) further supplements this idea by showing close affinities between the Ainu and the Nivkhi from Sakhalin Island. It would therefore be interesting to collect samples from populations from that area in future studies to have a clearer view of the relationships between the Ainu and Northeast Asian populations.

Regarding the origins of the Japanese population, phylogenetic trees in Figures 4.8 to 4.10 shows that the three Japanese populations (Kanto, Ryukyuan and Ainu) tended to cluster together and that they are placed within the clade consisting of other East Asian populations. Our study failed to show any evidence of close relationships between the Ainu and Ryukyuans with current Southeast Asian populations, as proposed by the dual-structure model. The link with Southeast Asia was based mostly on cranial and dental measurements on living populations and archaeological specimens (Hanihara 1991). Indeed, most genetic studies so far failed to show any conclusive links between the Ainu and Ryukyuans with Southeast Asian populations. Instead, most of the genetic studies pointed to closer affinities with Northeast Asian populations (Nei 1995; Omoto 1995; Omoto and Saitou 1997) which corroborates the results that we obtained in this study. This puts further doubt into one aspect of the dual-structure model, which is the origin of Jomon ancestors from Southeast Asia.

However, we did demonstrate that the Ainu and Ryukyuans have much closer affinities to each other than they are to the mainland Japanese. This is shown in the phylogenetic trees in Figures 4.8 to 4.10 which show that these two groups always formed a clade. The 100% bootstrap support for the branching pattern lends further confidence to this idea. The close association between the Ainu and Ryukyuans despite their current geographical locations which is at the two opposing ends of the Japanese Archipelago can be interpreted as them having a shared common ancestry probably dating back to the Jomon period. The idea of the Ainu and Ryukyuans having Jomon roots was earlier established based on the similarities of bone and tooth measurements between Jomon archaeological samples and Ainu/Ryukyu individuals (Hanihara 1991). Analyses of autosomal genetic markers also showed close affinities between the Ryukyuans and Ainu (Nei 1995; Omoto and Saitou 1997; Tokunaga et al. 2001).

Regarding the origins of the mainland Japanese, the unrooted tree in Figure 4.10(B) places the Kanto Japanese in an intermediate position between the Chinese/Korean and Ainu/Ryukyuan branches. This observation coupled with the very short external branch in the Kanto Japanese implies that they may have been the result of admixture between the ancestors of the Chinese/Koreans and the Ainu/Ryukyuans. The results of *frappe* analysis at k=3 onwards (Figure 4.7) also shows that the Kanto Japanese received substantial genetic contributions from the mainland (Chinese) but also shows some degree of Ainu/Ryukyuan components which are unsubstantial in the Han Chinese. These observations appear to support some partial aspects of the dual-structure model, which supposes that the Ainu and Ryukyuans shared common ancestry with the Jomon people and that the mainland Japanese is a result of admixture between the Jomon and Yayoi ancestral populations.

Taken together, our analysis of a dense set of genome-wide SNP in the Japanese populations reveal greater genetic variation within individuals of the Ainu group, brought about by admixture with the mainland Japanese and possibly another population from Northeast Asia. Regarding the origins of the Japanese population in general, our data supports some aspects of the dual-structure model in that the Ainu and Ryukyuans have shared genetic ancestry and that the mainland Japanese are the result of admixture between ancestral Yayoi and Jomon peoples. However, phylogenetic tree analysis did not seem to support the idea of a Southeast Asian origin of Jomon peoples but shows a closer affinity to Northeast Asian populations.

CHAPTER 5

General discussion and conclusions

The main themes covered in this thesis relate to the genetic diversity and migration histories of indigenous groups in Southeast Asia and East Asia. Using genome-wide SNP data, I explored the genetic substructure within these indigenous groups and distinct patterns were observed on the PCA plots involving the Negritos from West Malaysia (Chapter 2) and the Ainu from Japan (Chapter 4). Measures of genetic distance between individuals as well as STRUCTURE and *frappe* analysis indicate that those patterns were most likely the result of sustained admixture with surrounding populations. Admixture was not restricted to those two indigenous groups who had a long history with their current geographical location, but was also evident in the relatively recent migrant populations.

These results demonstrate the influence of surrounding populations to the genetic diversity in indigenous Malaysian populations which also contributes to the genetic substructure in these indigenous groups. The presence of admixed individuals may have a bearing on the design and sampling strategy of future population genetic studies. Population-based phylogenetic analyses tend to assume populations are a static and panmictic unit but our results indicate admixed individuals may affect measures such as genetic distance. This is even more critical when sampling small number of individuals from a highly admixed population. Thus future sampling particularly involving indigenous populations should consider having a larger sample size to accommodate the possibility of sampling too many admixed individuals. Population substructure as a result of admixture should also be taken into consideration especially when conducting association studies as the presence of population stratification may lead to increased false positive associations (Tian, Gregersen, and Seldin 2008; Yamaguchi-Kabata et al. 2008).

Analysis of genome-wide SNP and uniparentally inherited markers such as mtDNA has most often been applied in studying past human movements and ultimately the origins of specific populations. Comparative analysis of complete mtDNA sequences in the Negrito and Austronesian groups in Malaysia with various other populations from Southeast Asia suggests a more complex history regarding the peopling of Southeast Asia than the more simplified, two-wave model involving an early southern, coastal dispersal from Africa and an Austronesian expansion from Taiwan, would suggest. Using a combination of maximum-likelihood phylogenetic analysis, TMRCA estimates of mtDNA lineages and PCA, our results indicate a long term presence followed by isolation in the Southeast Asian region by the Negritos in West Malaysia, consistent with the initial wave via the coastal route. We also found evidence of an earlier population movement originating from South China or Indochina during late Pleistocene to early Holocene (~30,000 to 10,000 YBP) with regards to the history of the Austronesians. This earlier movement predates the Neolithic expansion from Taiwan around 5,000 to 7,000 YBP but our results do not refute the Out of Taiwan migration event took place. However it looks improbable that it was responsible for the origins of all Austronesian speakers, particularly those west of the Wallace line. A more likely scenario to explain our observations would be an adoption of Austronesian languages by extant Southeast Asian populations, while allowing for some degree of gene flow of mtDNA lineages from Taiwan.

The plausible scenarios involving human movement in Southeast Asia are summarized in Figure 5.1. In panel A) the first arrival of humans to the region was via a southern coastal route around 50,000 YBP. This is followed by a south to north migration into East Asia as proposed in the PASNP paper (Abdulla et al. 2009). Although no specific dates were mentioned, we speculate that it could have happened after the initial settlement event. Panel B) shows our proposed ‘early train’ dispersal from Indochina or South China around 30,000 to 10,000 YBP based on the haplogroup frequencies, tree phylogenies and TMRCA estimates of mtDNA lineages. This is followed by the Out of Taiwan migration shown in panel (C) which contributes to the genetic make-up of some populations in island Southeast Asia and the large part of the Pacific islands. A fairly recent event is shown in panel (D) which involves gene flow from India and is largely restricted to the populations in West Malaysia and Sumatra. This is indicated by the Indian ancestry in those populations in the STRUCTURE analysis (Chapter 2). The presence of Indian-specific Y-haplogroups in those populations (Karafet et al. 2010) as well as archaeological evidence pointing to contact as early as 4th century BC (2,500 YBP) (Bellina and Glover 2004) seems to support this idea.

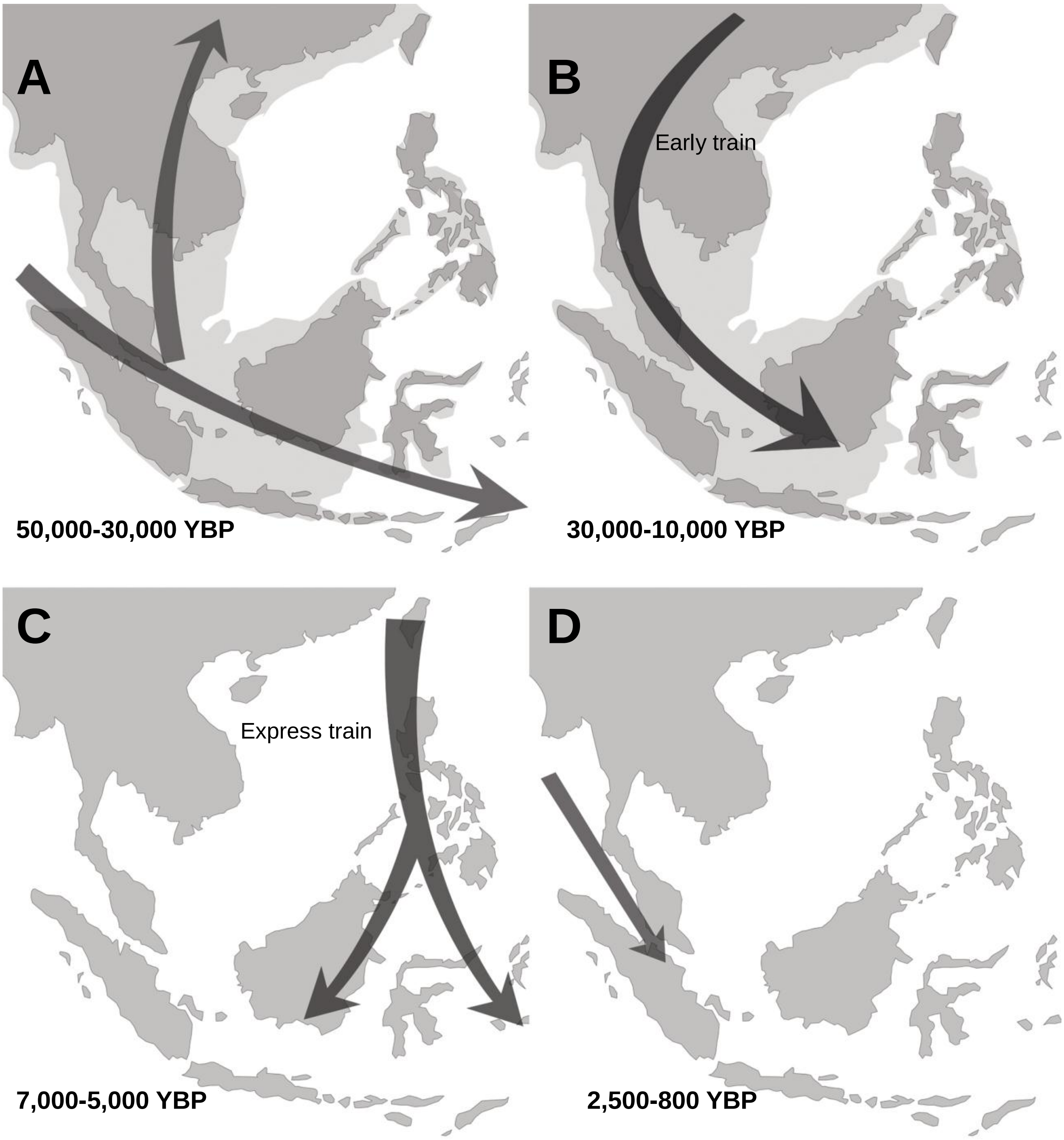


Figure 5.1: Plausible migration routes to Southeast Asia based on genetic data. A) First arrival of modern humans via the southern coastal route, followed by a northward migration to East Asia. B) Southward migration from Indochina/South China towards Sundaland C) Out-of-Taiwan migration in to island Southeast Asia and the Pacific. D) Influence from India in Malay & Sumatran populations

Regarding the history and origins of Japanese populations, our analysis of a dense set of genome-wide SNP supports some aspects of the dual-structure model which proposes admixture between the ancient Jomon with fairly recent Yayoi migrants. We found that the Ainu and Ryukyuans have shared genetic ancestry which probably dates back to the Jomon period and that the mainland Japanese are the result of admixture between ancestral Yayoi and Jomon peoples. On a lighter note, one of the first reports of Ainu-Ryukyuan relatedness based on bone structure was by von Baelz (1911), later cited by Hanihara (1991). It seems that has taken 100 years (at the time of writing) for this theory to be shown true by genome-wide SNP data. As for the other aspect of the dual-structure model which posits a Southeast Asian origin of the Jomon people, our data shows a closer affinity of the Ainu/Ryukyu to Northeast Asian populations. Although this initial result seems to put that aspect of the model in doubt, a much more detailed and expansive future study would be required to address the issue.

In the genome-wide SNP analysis in chapters 2 and 4, PCA was used to infer the relationships between individuals and in general, individuals tend to cluster according to their respective populations. This was expected because individuals from the same population tend to share the same alleles due to interbreeding of individuals from the same population. However the PCA analysis did show some peculiar patterns such as the ‘comet-like’ pattern in the Negritos, Temuan and Bidayuh (Chapter 2) and the triangular-like scattering of Ainu individuals (Chapter 4). A logical explanation for these observations would be recent admixture. An admixed individual receives equal genetic contribution from both parental populations, thus we would expect the individual to be intermediate between the two parental populations if plotted along a linear vector as in the PCA analysis. If this admixture process was continuous and involved admixed individuals interbreeding with either of the parental populations, we would expect to see an admixture gradient akin to a ‘comet-like’ pattern on PCA plots. This has been confirmed by Patterson et al. (2006) who simulated admixed individuals and plotted them on PCA. The same ‘comet-like’ pattern was observed in the PCA plot using the simulated data. In the case of the Ainu, the ‘comet-like’ pattern was not observed but more of a triangular-like pattern. A possible explanation for this is that admixture involved three source populations, namely ‘pure’ Ainu, Hondo Japanese and another population presumably from Sakhalin based on evidence from other genetic markers. Another concern was the quality of the Ainu DNA which was kept in cold-storage for several years prior to genotyping. Data filtering was performed meticulously in order to keep only good quality SNP while keeping the maximum number of individuals. To show that genotyping error did not contribute to the observed PCA plot in the Ainu, I performed PCA using SNP obtained from different steps of filtering based on genotyping call rate (90%, 95% and 100% call rates). All three resulting PCA plots did not show any difference, particularly regarding the Ainu (Appendix Figure A4), so we can conclude that the PCA pattern was not due to genotyping error.

In summary, my results demonstrate the influence of surrounding populations to the genetic diversity in indigenous Malaysian and Japanese populations which also contributes to the genetic substructure in these indigenous groups. The presence of admixed individuals has to be considered when designing sampling strategies for future population genetic studies as well as when conducting and interpreting results of association studies. Regarding the history and origins of Austronesians in Southeast Asia, results suggest an earlier movement originating from Indochina around 30,000 to 10,000 YBP which has more impact on the mtDNA diversity of indigenous Austronesians in West Malaysia and Borneo than the proposed Out of Taiwan expansion around 7,000 YBP. As for the origins of the Japanese population, my data support some aspects of the dual-structure model in that the Ainu and Ryukyuans have shared genetic ancestry and that the mainland Japanese are the result of admixture between ancestral Yayoi and Jomon peoples. However, our data does not indicate a Southeast Asian origin of Jomon peoples but shows a closer affinity to Northeast Asian populations.

REFERENCES

Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen C-H, Chen J, Chen Y-T, et al. 2009. Mapping human genetic diversity in Asia. Science 326:1541-1545.

Adelaar KA, Himmelmann N. 2005. The Austronesian languages of Asia and Madagascar. Routledge

Andaya BW, Andaya LY. 1984. A History of Malaysia. Palgrave Macmillan

Andaya LY. 2001. The Search for the Origins of Melayu. Journal of Southeast Asian Studies 32:315-330.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. 1981. Sequence and organization of the human mitochondrial genome. Nature 290(5806):457-65.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23(2):147.

Atkinson QD, Gray R. D, Drummond A. J. 2008. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. Molecular biology and evolution 25:468.

Barik SS, Sahani R, Prasad BVR, Endicott P, Metspalu M, Sarkar BN, Bhattacharya S, Annapoorna PCH, Sreenath J, Sun D, et al. 2008. Detailed mtDNA genotypes permit a reassessment of the settlement and population structure of the Andaman Islands. American Journal of Physical Anthropology 136:19-27.

Barker G, Barton H, Bird M, Daly P, Datan I, Dykes A, Farr L, Gilbertson D, Harrisson B, Hunt C, et al. 2007. The `human revolution' in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). Journal of Human Evolution 52:243-261.

Bellina B, Glover I. 2004. The archaeology of early contact with Indian and the Mediterranean world, from the fourth century BC to the fourth century AD. In: Glover I, Bellwood P, editors. Southeast Asia: from prehistory to history. London and NewYork: Routledge.

Bellwood PS. 2005. The first farmers: the origins of agricultural societies. Wiley-Blackwell

Bellwood P. 2007. Prehistory of the Indo-Malaysian Archipelago. ANU E Press

Blust R. 1995. The prehistory of the Austronesian-speaking peoples: A view from language. J World Prehist 9:453-510.

Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, Hammer M, Bustamante CD, Ostrer H. 2010. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proceedings of the National Academy of Sciences 107:8954-8961.

Carey I. 1976. Orang Asli: Aboriginal Tribes of Peninsular Malaysia. Kuala Lumpur and New York: Oxford University Press

Cavalli-Sforza L, Feldman MW. 2003. The application of molecular genetic approaches to the study of human evolution. Nat. Genet 33 Suppl:266-275.

Chang YM, Swaran Y, Phoon YK, Sothirasan K, Sim HT, Lim KB, Kuehn D. 2009. Haplotype diversity of 17 Y-chromosomal STRs in three native Sarawak populations (Iban, Bidayuh and Melanau) in East Malaysia. Forensic Science International: Genetics 3:e77-e80.

Dancause KN, Chan CW, Arunotai NH, Lum JK. 2009. Origins of the Moken Sea Gypsies inferred from mitochondrial hypervariable region and whole genome sequences. J Hum Genet 54:86-93.

Davison A, Chiba S, Barton NH, Clarke B. 2005. Speciation and Gene Flow between Snails of Opposite Chirality. PLoS Biol 3:e282.

Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, Perkova M, Dorzhu C, Luzina F, Lee H, Vanecek T. 2007. Phylogeographic Analysis of Mitochondrial DNA in Northern Asian Populations. The American Journal of Human Genetics 81:1025-1041.

Dhaliwal JS, Shahnaz M, Azrena A, Irda YA, Salawati M, Too CL, Lee YY. 2010. HLA polymorphism in three indigenous populations of Sabah and Sarawak. Tissue Antigens 75:166-169.

Diamond JM. 1988. Express train to Polynesia. Nature 336:307-308.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol 7:214.

Fagundes NJR, Kanitz R, Eckert R, Valls ACS, Bogo MR, Salzano FM, Smith DG, Silva Jr. WA, Zago MA, Ribeiro-dos-Santos AK. 2008. Mitochondrial Population Genomics Supports a Single Pre-Clovis Origin with a Coastal Route for the Peopling of the Americas. The American Journal of Human Genetics 82:583-592.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package). Department of Genome Sciences, University of Washington, Seattle

Glover I, Bellwood PS. 2004. Southeast Asia: from prehistory to history. London and NewYork: Routledge

Gray RD, Drummond AJ, Greenhill SJ. 2009. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. Science 323:479 -483.

Gray RD, Jordan FM. 2000. Language trees support the express-train sequence of Austronesian expansion. Nature 405:1052-1055.

Gunnarsdottir ED, Li M, Bauchet M, Finstermeier K, Stoneking M. 2010. High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Research 21:1-11.

Gunnarsdottir ED, Nandineni MR, Li M, Myles M, Gil D, Pakendorf B, Stoneking M. 2011.

Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. Nat. Commun. 2: 228.

Hammer MF, Karafet TM, Park H, Omoto Keiichi, Harihara Shinji, Stoneking Mark, Horai Satoshi. 2005. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet 51:47-58.

Hanihara K. 1991. Dual structure model for the population history of the Japanese. Japan Review 2:1-33.

Harunari H, Imamura M. 2004. The real age of the Yayoi period. Tokyo: Gakusei-sha.

Hatin WI, Nur-Shafawati AR, Zahri M-K, Xu S, Jin L, Tan S-G, Rizman-Idid M, Zilfalil BA. 2011. Population Genetic Structure of Peninsular Malaysia Malay Sub-Ethnic Groups. PLoS One 6.

Hill C, Soares P, Mormina M, Macaulay V, Meehan W, Blackburn J, Clarke D, Raja J, Ismail P, Bulbeck D, et al. 2006. Phylogeography and Ethnogenesis of Aboriginal Southeast Asians. Molecular Biology and Evolution 23:2480 -2491.

Hill C, Soares P., Mormina M., Macaulay V., Clarke D., Blumbach PB, Vizuete-Forster M, Forster P., Bulbeck D., Oppenheimer S., et al. 2007. A mitochondrial stratigraphy for island southeast Asia. The American Journal of Human Genetics 80:29–43.

Horai S., Murayama K, Hayasaka K, Matsubayashi S, Hattori Y, Fucharoen G, Harihara S., Park KS, Omoto K., Pan IH. 1996. mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. Am J Hum Genet 59:579-590.

Hudson MJ. 2006. Japanese Beginnings. In: Tsutsui WM, editor. A Companion To Japanese History. United Kingdom: Blackwell Publishing. pp. 11-29.

Imamura K. 1996. Prehistoric Japan: New Perspectives on Insular East Asia. First Am edition. Honolulu: University of Hawaii Press

Ingman M, Kaessmann H, Paabo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708-713.

Jinam T, Saitou N, Edo J, Mahmood A, Phipps M. 2010. Molecular analysis of HLA Class I and Class II genes in four indigenous Malaysian populations. Tissue Antigens 75:151-158.

Jinam T. 2007. Health, Human Leukocyte Antigens and genomic diversity in indigenous populations of Malaysia. M.Sc Thesis. University of Malaya, Malaysia.

Karafet TM, Hallmark B, Cox MP, Sudoyo H, Downey S, Lansing JS, Hammer MF. 2010. Major East–West Division Underlies Y Chromosome Stratification across Indonesia. Molecular Biology and Evolution 27:1833 -1844.

Kong Qing-Peng, Sun C, Wang Hua-Wei, Zhao M, Wang W-Z, Zhong L, Hao X-D, Pan H, Wang S-Y, Cheng Y-T, et al. 2011. Large-Scale mtDNA Screening Reveals a Surprising Matrilineal Complexity in East Asia and Its Implications to the Peopling of the Region. Molecular Biology and Evolution 28:513 -522.

Kong Qing-Peng, Yao Yong-Gang, Sun C, Bandelt H-J, Zhu C-L, Zhang Ya-Ping. 2003. Phylogeny of East Asian Mitochondrial DNA Lineages Inferred from Complete Sequences. The American Journal of Human Genetics 73:671-676.

Larson G, Cucchi T, Fujita M, Matisoo-Smith E, Robins J, Anderson A, Rolett B, Spriggs M, Dolman G, Kim T-H, et al. 2007. Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in Island Southeast Asia and Oceania. Proceedings of the National Academy of Sciences 104:4834 -4839.

Leavesley M, Chappell J. 2004. Buang Merabak: additional early radiocarbon evidence of the colonisation of the Bismarck Archipelago, Papua New Guinea. Antiquity 78.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza Luigi L, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. 319.

Macaulay V. 2005. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. Science 308:1034-1036.

Macaulay Vincent, Hill Catherine, Achilli A, Rengo Chiara, Clarke Douglas, Meehan William, Blackburn James, Semino O, Scozzari R, Cruciani Fulvio, et al. 2005. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. Science 308:1034 -1036.

Majumder PP. 2010. The Human Genetic History of South Asia. Current Biology 20:R184-R187.

McEvoy BP, Lind JM, Wang ET, Moyzis RK, Visscher PM, van Holst Pellekaan SM, Wilton AN. 2010. Whole-Genome Genetic Diversity in a Sample of Australians with Deep Aboriginal Ancestry. The American Journal of Human Genetics 87:297-305.

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay Vincent, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, et al. 2003. Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100:171-176.

Mizoguchi Y. 1968. Contributions of prehistoric Far East Populations to the population of modern Japan: A Q-mode path analysis based on cranial measurements. In: Akazawa T, Aikens C, editors. Prehistoric Hunter-Gatherers in Japan. Tokyo: University of Tokyo Press. pp. 107-136.

Nei M. 1995. The origins of human populations: genetic, linguistic, and archeological data. In: Brenner S, Hanihara K, editors. The Origin and Past of Modern Humans as Viewed from DNA. Singapore and London: World Scientific. pp. 71-91.

Nicholas C. 2000. The Orang Asli and the Contest for Resources. Indigenous Politics, Development and Identity in Peninsular Malaysia. Denmark: IWGIA & Center for Orang Asli Concerns

O’Connell JF, Allen J. 2004. Dating the colonization of Sahul (Pleistocene Australia-New Guinea): a review of recent research. Journal of Archaeological Science 31:835-853.

Omoto K., Saitou N. 1997. Genetic origins of the Japanese: a partial support for the dual structure hypothesis. American journal of physical anthropology 102:437–446.

Omoto K. 1983. Genetic polymorphism of Japanese. In: Ikeda J, editor. Anthropology Lecture Series 6: The Japanese II. Tokyo: Yuzankaku. pp. 217-263.

Omoto K. 1995. Genetic diversity and the origins of the “Mongoloids.” In: Brenner S, Hanihara K, editors. The Origin and Past of Modern Humans as Viewed from DNA. Singapore and London: World Scientific. pp. 92-109.

Oota H, Pakendorf B, Weiss G, von Haeseler A, Pookajorn S, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking Mark. 2005. Recent origin and cultural reversion of a hunter-gatherer group. PLoS Biol 3:e71.

Oppenheimer SJ, Richards Michael. 2001. Slow boat to Melanesia? Nature 410:166.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat 30:E386-394.

Patterson N, Price AL, Reich D. 2006. Population Structure and Eigenanalysis. PLoS Genet 2:e190.

Peng M-S, Quang HH, Dang KP, Trieu AV, Wang H.-W., Yao Y.-G., Kong Q.-P., Zhang Y.-P. 2010. Tracing the Austronesian Footprint in Mainland Southeast Asia: A Perspective from Mitochondrial DNA. Molecular Biology and Evolution 27:2417-2430.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet 81:559-575.

Reich D, Thangaraj Kumarasamy, Patterson N, Price AL, Singh Lalji. 2009. Reconstructing Indian population history. Nature 461:489-494.

Ricaut FX, Bellatti M, Lahr MM. 2006. Ancient mitochondrial DNA from Malaysian hair samples: Some indications of Southeast Asian population movements. American Journal of Human Biology 18:654–667.

Ruiz-Linarez A, Minch E, Meyer D, Cavalli-Sforza L. 1995. Analysis of classical and DNA markers for reconstructing human population history. In: Brenner S, Hanihara K, editors. The Origin and Past of Modern Humans as Viewed from DNA. Singapore and London: World Scientific. pp. 71-91.

Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, Braid M, Clarke DJ, Loo Jun-Hun, Thomson N, et al. 2011. Ancient Voyaging and Polynesian Origins. The American Journal of Human Genetics 88:239-247.

Soares P, Trejaut J, Loo Jun-Hun, Hill C, Mormina M, Lee C-L, Chen Yao-Ming, Hudjashov G, Forster P, Macaulay V, et al. 2008. Climate change and postglacial human dispersals in southeast Asia. Mol. Biol. Evol 25:1209-1218.

Suzuki H. 1963. Skeletal bones of the Japanese people. Tokyo: Iwanami Shoten

Suzuki H. 1983. The root of Japanese viewed from bones. Tokyo: Iwanami Shoten

Tabbada KA, Trejaut J., Loo J.-H., Chen Y.-M., Lin M., Mirazon-Lahr M, Kivisild T, De Ungria MCA. 2009. Philippine Mitochondrial DNA Diversity: A Populated Viaduct between Taiwan and Indonesia? Molecular Biology and Evolution 27:21-31.

Tajima A, Pan IH, Fucharoen G, Fucharoen S, Matsuo M, Tokunaga K, Juji T, Hayami M, Omoto K, Horai S. 2002. Three major lineages of Asian Y chromosomes: implications for the peopling of east and southeast Asia. Hum Genet 110(1):80-8.

Tajima A, Hayami M, Tokunaga Katsushi, Juji Takeo, Matsuo M, Marzuki S, Omoto Keiichi, Horai Satoshi. 2004. Genetic origins of the Ainu inferred from combined DNA analyses of maternal and paternal lineages. Journal of Human Genetics 49:187-193.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution 10:512 -526.

Tamura K, Peterson D, Peterson N, Stecher G, Nei Masatoshi, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution [Internet]. Available from: http://mbe.oxfordjournals.org/content/early/2011/05/04/molbev.msr121.abstract

Tanaka M, Cabrera VM, González AM, Larruga JM, Takeyasu T, Fuku N, Guo LJ, Hirose R, Fujita Y, Kurata M, et al. 2004. Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome research 14:1832.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol 28:289-301.

Teng YS, Tan SG. 1979. Genetic evidence of gene flow from Indians to Malays. Journal of Human Genetics 24:1–8.

Thangaraj K., Chaubey G, Kivisild T., Reddy A. G, Singh VK, Rasalkar AA, Singh L. 2005. Reconstructing the origin of Andaman Islanders. Science 308:996.

Thangaraj K, Singh L, Reddy AG, Rao VR, Sehgal SC, Underhill PA, Pierson M, Frame IG, Hagelberg E. 2003. Genetic Affinities of the Andaman Islanders, a Vanishing Human Population. Current Biology 13:86-93.

Tian C, Gregersen PK, Seldin MF. 2008. Accounting for ancestry: population substructure and genome-wide association studies. Human Molecular Genetics 17:R143-R150.

Tokunaga K., Ohashi J, Bannai M, Juji T. 2001. Genetic link between Asians and native Americans: evidence from HLA genes and haplotypes. Human Immunology 62:1001–1008.

Torroni A., Rengo C., Guida V, Cruciani F., Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M., et al. 2001. Do the four clades of the mtDNA haplogroup L2 evolve at different rates. The American Journal of Human Genetics 69:1348–1356.

Trejaut Jean A., Kivisild Toomas, Loo JH, Lee CL, He CL, Hsu CJ, Li ZY, Lin Marie. 2005. Traces of Archaic Mitochondrial Lineages Persist in Austronesian-Speaking Formosan Populations. Plos Biol 3:e247.

Vlieland CA. 1934. The population of the Malay Peninsula: A study in human migration. Geographical Review 24:61–78.

von-Baelz E. 1911. Die Riu-Kiu-Insulaner, die Aino und andere kaukaiser-ahnliche Reste in Ostasien. Korres Blatt Dtsch Ges Anthrop Ethnol Urgesch 42:187-191.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38:1358-1370.

Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, Nakamura Y, Kamatani N. 2008. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. Am. J. Hum. Genet 83:445-456.

APPENDICES

Table A1: List of 11 pairs of PCR primers for complete mtDNA sequencing (from Torroni et al. 2001)

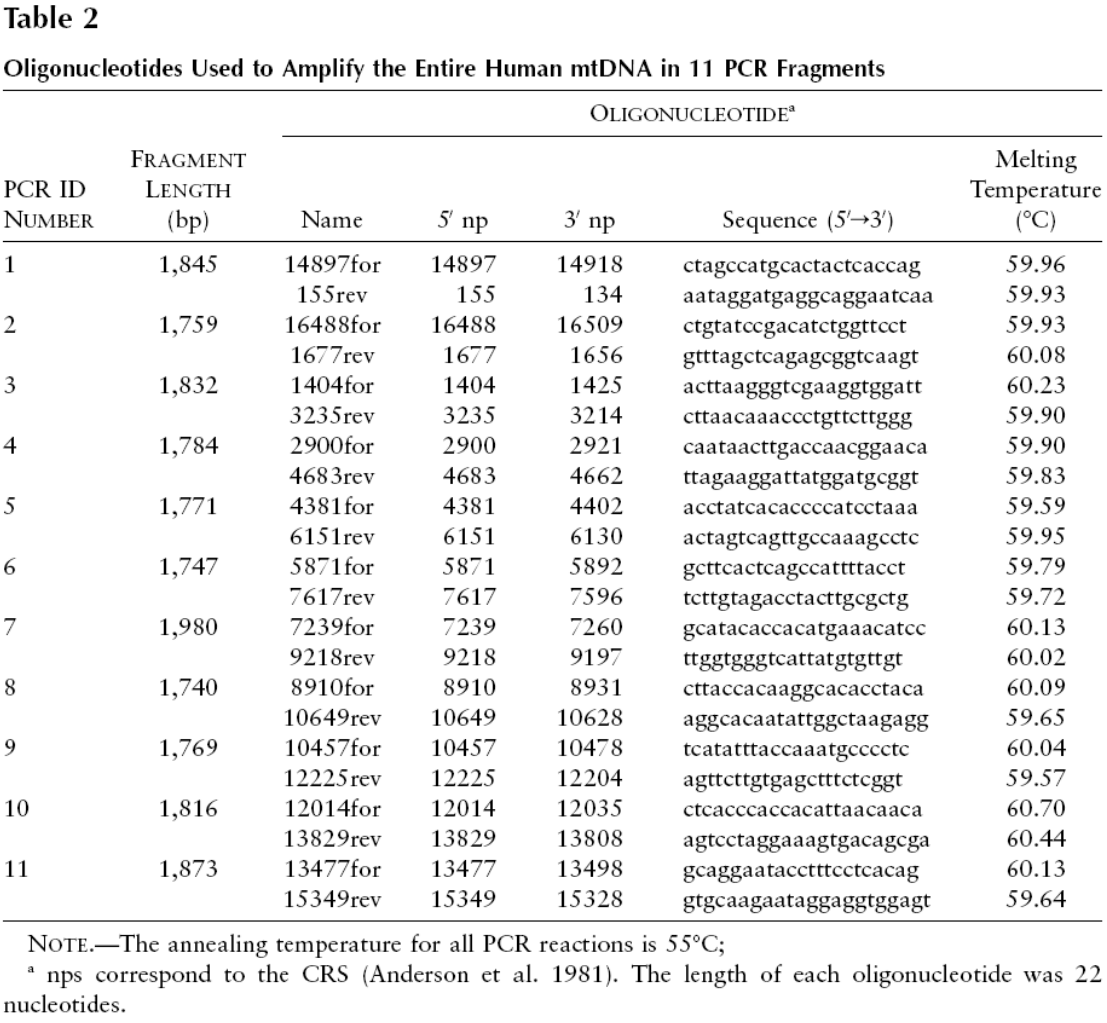


Table 2

Oligonucleotides Used to Amplify the Entire Human mtDNA in 11 PCR Fragments

PCR ID Number	Fragment Length (bp)	Oligonucleotide[a] Name	5′ np	3′ np	Sequence (5′→3′)	Melting Temperature (°C)
1	1,845	14897for	14897	14918	ctagccatgcactactcaccag	59.96
		155rev	155	134	aataggatgaggcaggaatcaa	59.93
2	1,759	16488for	16488	16509	ctgtatccgacatctggttcct	59.93
		1677rev	1677	1656	gtttagctcagagcggtcaagt	60.08
3	1,832	1404for	1404	1425	acttaagggtcgaaggtggatt	60.23
		3235rev	3235	3214	cttaacaaaccctgttcttggg	59.90
4	1,784	2900for	2900	2921	caataacttgaccaacggaaca	59.90
		4683rev	4683	4662	ttagaaggattatggatgcggt	59.83
5	1,771	4381for	4381	4402	acctatcacaccccatcctaaa	59.59
		6151rev	6151	6130	actagtcagttgccaaagcctc	59.95
6	1,747	5871for	5871	5892	gcttcactcagccattttacct	59.79
		7617rev	7617	7596	tcttgtagacctacttgcgctg	59.72
7	1,980	7239for	7239	7260	gcatacaccacatgaaacatcc	60.13
		9218rev	9218	9197	ttggtgggtcattatgtgttgt	60.02
8	1,740	8910for	8910	8931	cttaccacaaggcacacctaca	60.09
		10649rev	10649	10628	aggcacaatattggctaagagg	59.65
9	1,769	10457for	10457	10478	tcatatttaccaaatgcccctc	60.04
		12225rev	12225	12204	agttcttgtgagctttctcggt	59.57
10	1,816	12014for	12014	12035	ctcacccaccacattaacaaca	60.70
		13829rev	13829	13808	agtcctaggaaagtgacagcga	60.44
11	1,873	13477for	13477	13498	gcaggaatacctttcctcacag	60.13
		15349rev	15349	15328	gtgcaagaataggaggtggagt	59.64

NOTE.—The annealing temperature for all PCR reactions is 55°C;

[a] nps correspond to the CRS (Anderson et al. 1981). The length of each oligonucleotide was 22 nucleotides.

Table A2: List of 32 sequencing primers for complete mtDNA sequencing (from Torroni et al. 2001)

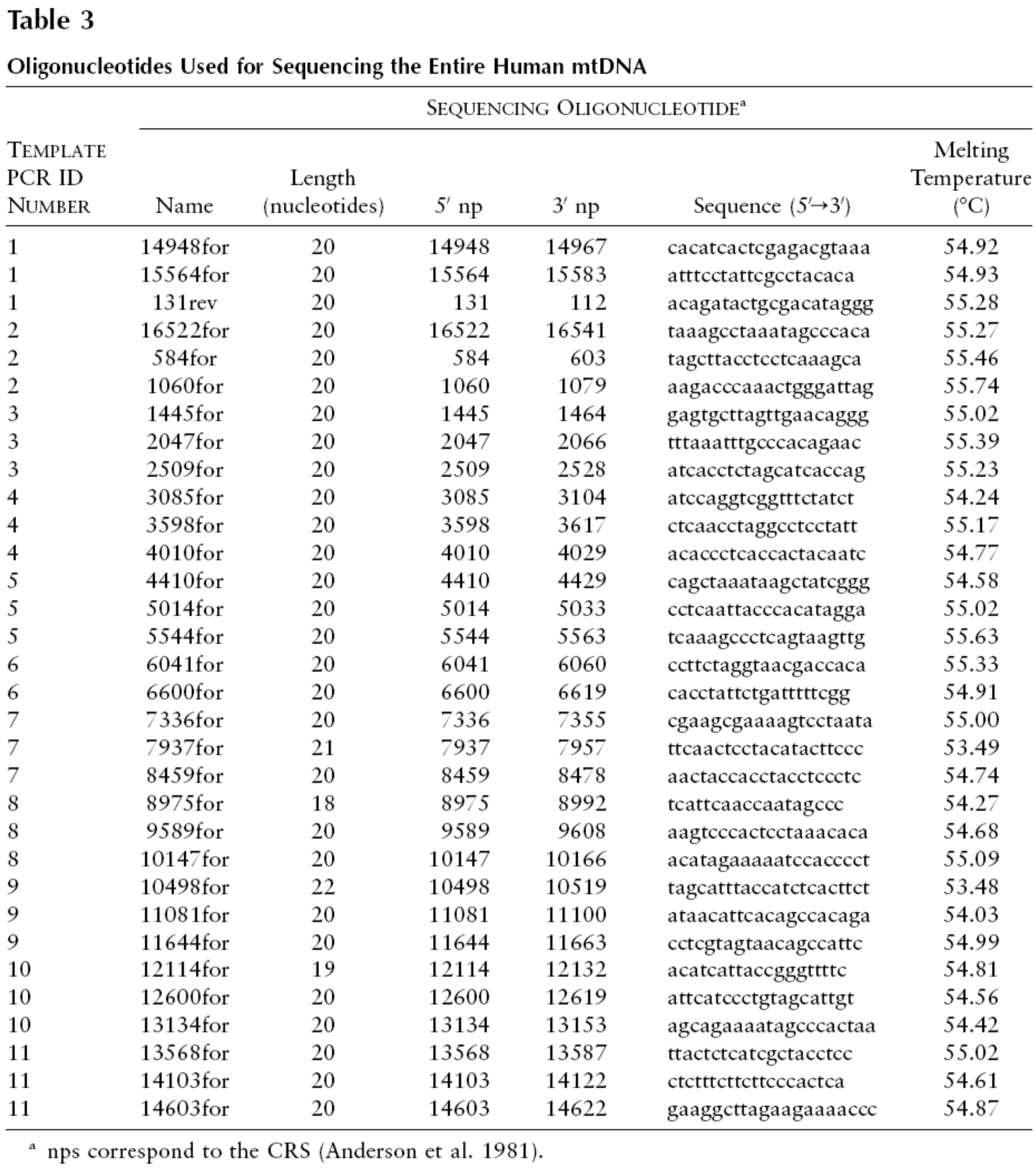


Table 3

Oligonucleotides Used for Sequencing the Entire Human mtDNA

	Sequencing Oligonucleotide[a]					
Template PCR ID Number	Name	Length (nucleotides)	5′ np	3′ np	Sequence (5′→3′)	Melting Temperature (°C)
1	14948for	20	14948	14967	cacatcactcgagacgtaaa	54.92
1	15564for	20	15564	15583	atttcctattcgcctacaca	54.93
1	131rev	20	131	112	acagatactgcgacataggg	55.28
2	16522for	20	16522	16541	taaagcctaaatagcccaca	55.27
2	584for	20	584	603	tagcttacctcctcaaagca	55.46
2	1060for	20	1060	1079	aagacccaaactgggattag	55.74
3	1445for	20	1445	1464	gagtgcttagttgaacaggg	55.02
3	2047for	20	2047	2066	tttaaatttgcccacagaac	55.39
3	2509for	20	2509	2528	atcacctctagcatcaccag	55.23
4	3085for	20	3085	3104	atccaggtcggtttctatct	54.24
4	3598for	20	3598	3617	ctcaacctaggcctcctatt	55.17
4	4010for	20	4010	4029	acaccctcaccactacaatc	54.77
5	4410for	20	4410	4429	cagctaaataagctatcggg	54.58
5	5014for	20	5014	5033	cctcaattacccacatagga	55.02
5	5544for	20	5544	5563	tcaaagccctcagtaagttg	55.63
6	6041for	20	6041	6060	ccttctaggtaacgaccaca	55.33
6	6600for	20	6600	6619	cacctattctgatttttcgg	54.91
7	7336for	20	7336	7355	cgaagcgaaaagtcctaata	55.00
7	7937for	21	7937	7957	ttcaactcctacatacttccc	53.49
7	8459for	20	8459	8478	aactaccacctacctccctc	54.74
8	8975for	18	8975	8992	tcattcaaccaatagccc	54.27
8	9589for	20	9589	9608	aagtcccactcctaaacaca	54.68
8	10147for	20	10147	10166	acatagaaaaatccacccct	55.09
9	10498for	22	10498	10519	tagcatttaccatctcacttct	53.48
9	11081for	20	11081	11100	ataacattcacagccacaga	54.03
9	11644for	20	11644	11663	cctcgtagtaacagccattc	54.99
10	12114for	19	12114	12132	acatcattaccgggttttc	54.81
10	12600for	20	12600	12619	attcatccctgtagcattgt	54.56
10	13134for	20	13134	13153	agcagaaaatagcccactaa	54.42
11	13568for	20	13568	13587	ttactctcatcgctacctcc	55.02
11	14103for	20	14103	14122	ctctttcttcttcccactca	54.61
11	14603for	20	14603	14622	gaaggcttagaagaaaaccc	54.87

[a] nps correspond to the CRS (Anderson et al. 1981).

Figure A1: Gel electrophoresis of 11 PCR products prior to annealing temperature optimization. Correct PCR amplicons are around 1.8kb in length. Presence of shorter fragments indicates unspecific amplification.

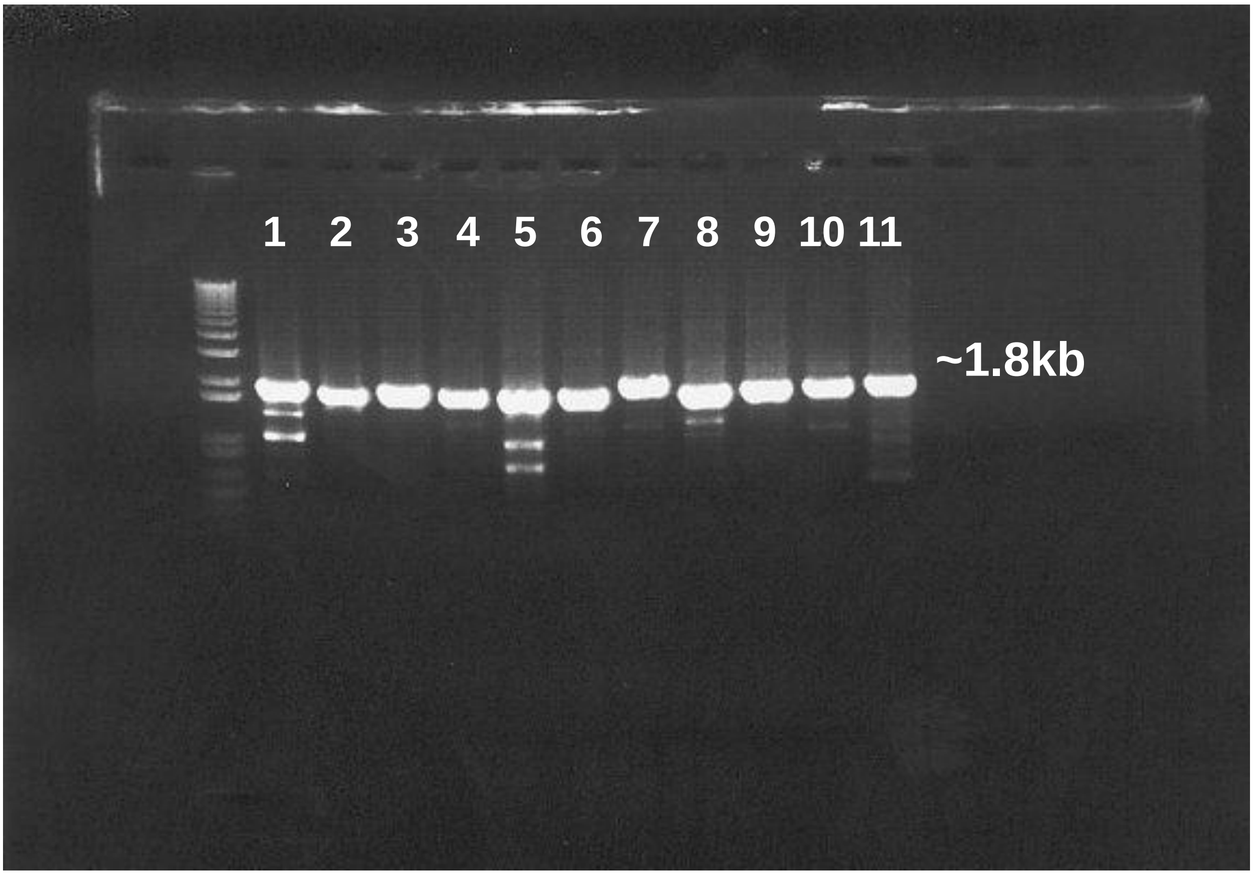
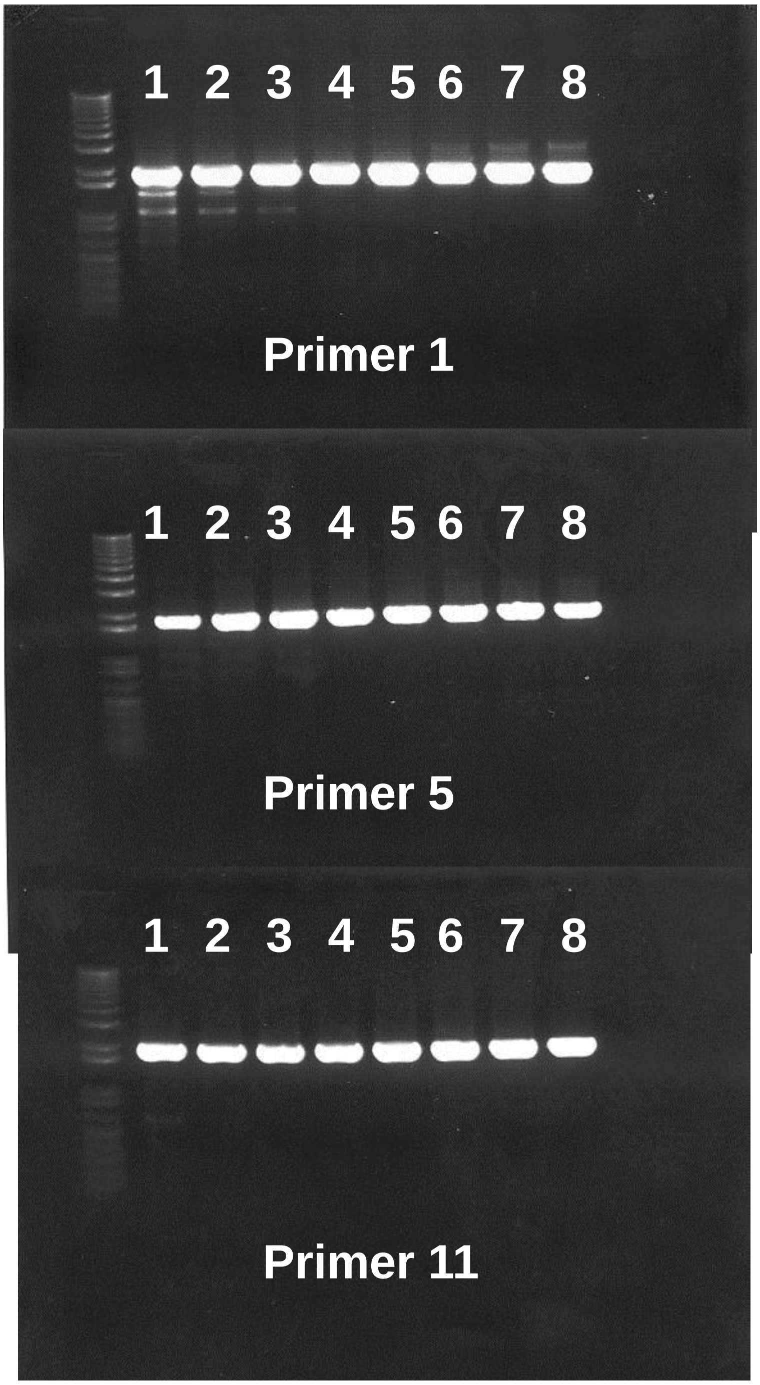


Figure A2: Gel electrophoresis after gradient PCR using annealing temperatures from 51°C to 65°C.



Lane labels
1: 51.4ºC
2: 53.8ºC
3: 56.2ºC
4: 58.6ºC
5: 60.9ºC
6: 63.0ºC
7: 64.5ºC
8: 65.0ºC

Figure A3: Gel electrophoresis after optimizing annealing temperatures to 60°C for all 11 primer pairs

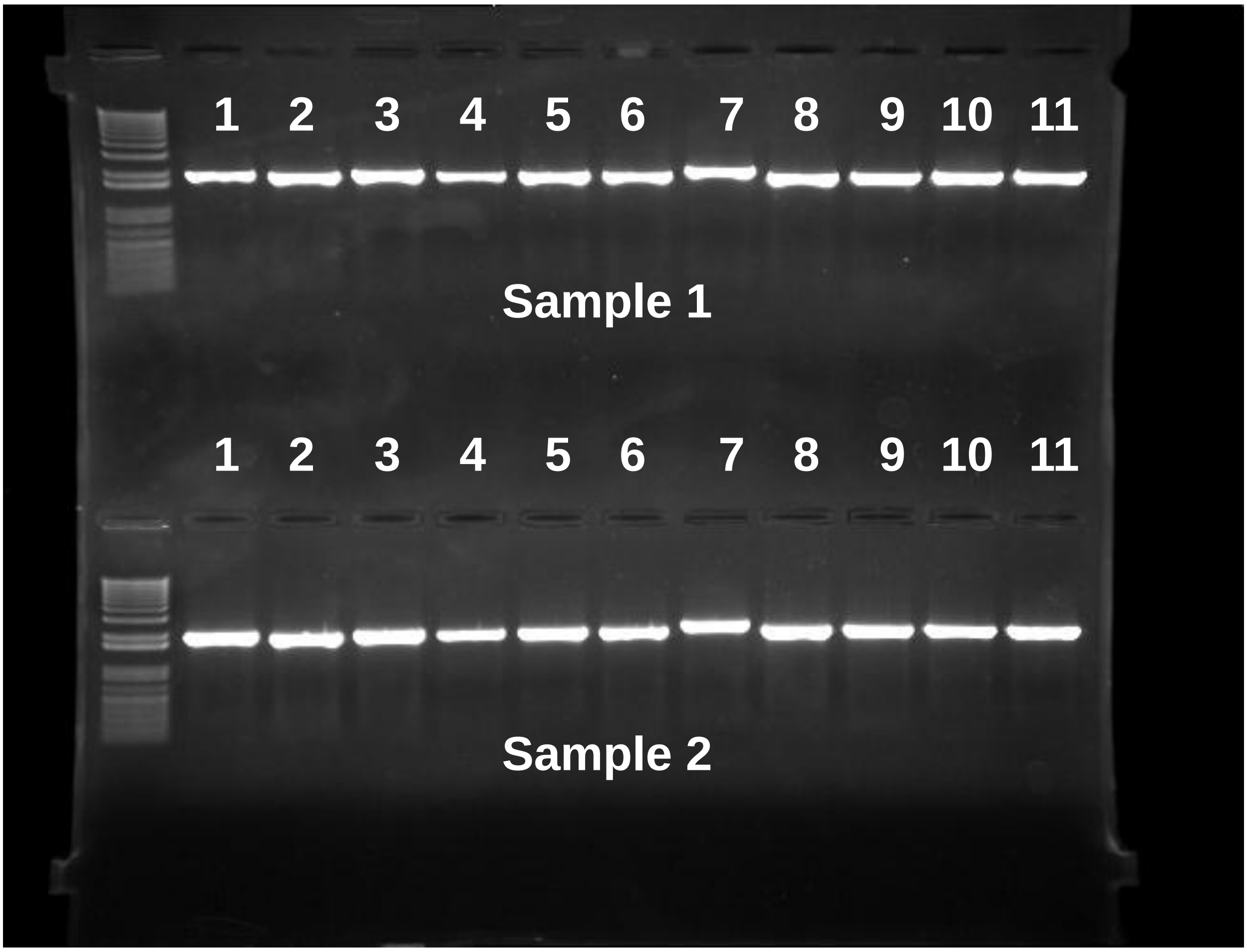
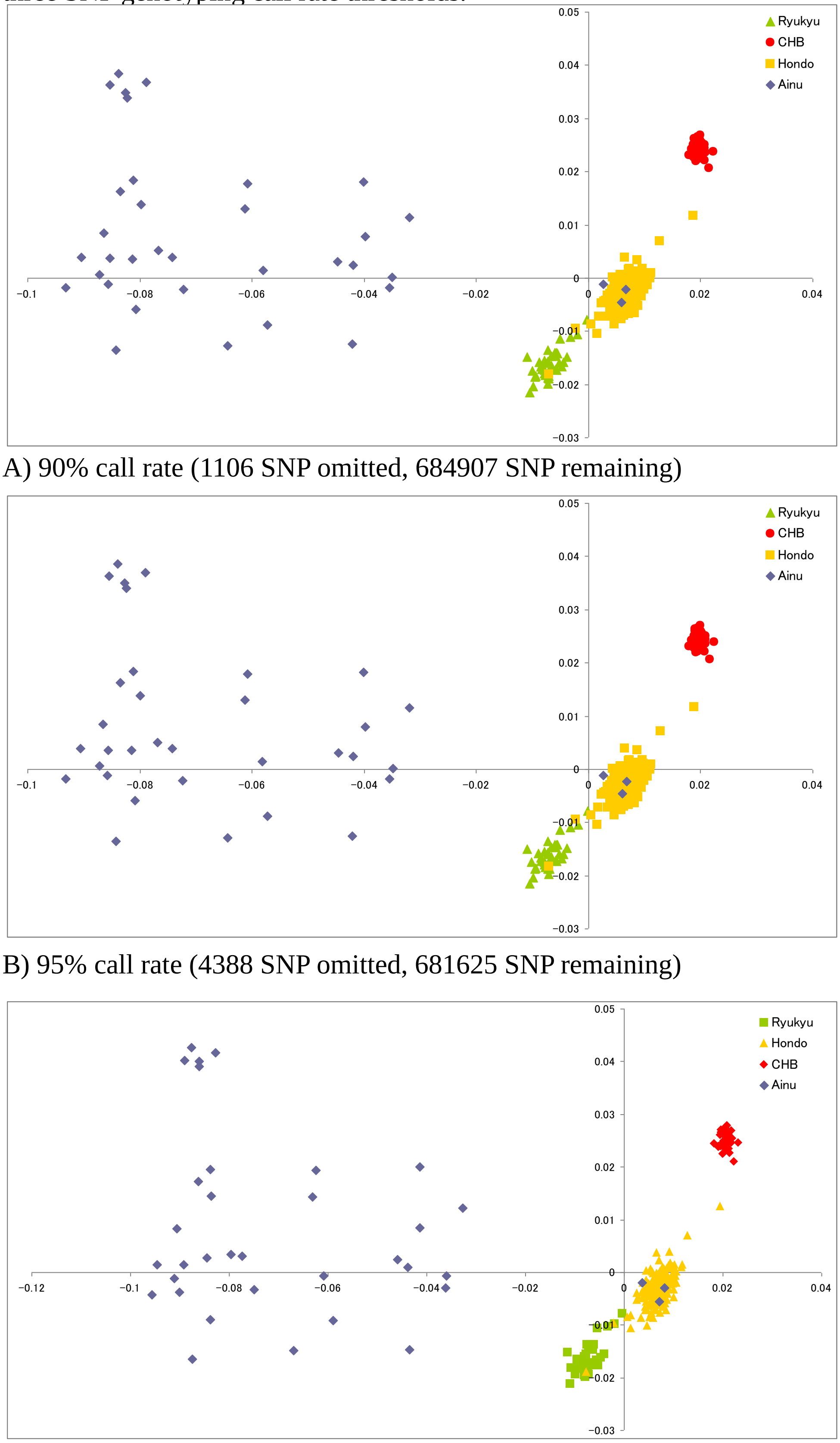


Figure A4: PCA plots of Japanese and Han-Chinese populations using SNP sets obtained from three SNP genotyping call rate thresholds.



A) 90% call rate (1106 SNP omitted, 684907 SNP remaining)

B) 95% call rate (4388 SNP omitted, 681625 SNP remaining)

C) 100% call rate (215061 SNP omitted, 470952 SNP remaining)