

**ASSESSMENT AND ANALYSIS OF
GENOMIC DIVERSITY AND BIOMARKERS IN
SABAHAN INDIGENOUS POPULATIONS**

KEE BOON PIN

**THESIS SUBMITTED IN FULFILMENT
OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

**FACULTY OF MEDICINE
UNIVERSITY OF MALAYA
KUALA LUMPUR
MALAYSIA**

2014

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: KEE BOON PIN

(I.C/Passport No: 840906-06-5305)

Registration/Matric No: MHA100002

Name of Degree: DOCTOR OF PHILOSOPHY

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

ASSESSMENT AND ANALYSIS OF GENOMIC DIVERSITY AND BIOMARKERS IN SABAHAN
INDIGENOUS POPULATIONS

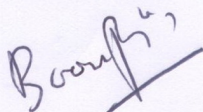
Field of Study: MOLECULAR BIOLOGY (GENETICS)

I do solemnly and sincerely declare that:

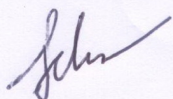
- (1) I am the sole author/writer of this Work;
- (2) This Work is original;
- (3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- (4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
- (5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
- (6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature

Date 15 JAN 2014



Subscribed and solemnly declared before,



Witness's Signature

Date 15 JAN 2014

Name: Prof. Dr. Chua Kek Heng

Designation: Professor

ABSTRACT

Ever since the proposal of the recent African origin of modern humans and its various opposing concepts, evolutionary studies have been focusing on the discovery of timelines for such events based on traceable records in archaic remains and contemporary human populations. Although the migration of anatomically modern humans between the 7 major continents in the near 120,000 years has been explained and fit well in the “out of Africa” hypothesis, human movement within these continents remains elusive. Human demic events in Southeast Asia region have also been the heat of debate among researchers. It has been proposed that this region was populated by a recent migration wave from Taiwan about 5,000 years ago, accompanying by the expansion of Austronesian languages into various parts of this region and further into Oceania – known as the “out of Taiwan” model. However, genetic studies in the Southeast Asia have been centred on populations lying along the proposed migratory paths, i.e., southern China, Taiwan, Philippines, Eastern Indonesia, and Near Oceania. The study on population in other parts of this region could shed important information, to complement the proposed model, on the understanding of the historical migration within Southeast Asia and also between other neighboring regions. Sabah, at the northern tip of Borneo Island, is strategically located in the centre of Southeast Asia. In the present study, the aim was to characterize the genetic structure of the 3 major indigenous populations in Sabah, i.e., Kadazan-Dusun, Bajau, and Rungus, and correlate them to the migration patterns in this region. A total of 639 indigenous individuals were recruited and genomic DNA was extracted from the blood/buccal samples. Polymorphisms on the nuclear DNA (VNTRs and InDels) were accessed by direct PCR method. In addition, typing of 15 STR markers on each sample was completed via fragment analysis study. Furthermore, the mitochondrial DNA was examined by the screening of the 9-bp deletion in the region V and the nucleotide

sequence of the 3 hypervariable regions in the D-loop was determined via sequencing reactions. The genetic data generated was subsequently subjected to statistical and comparative analysis. In an overview, these indigenous populations were shown to have high genetic similarity (AMOVA < 5 %). The Kadazan-Dusun and Rungus populations exhibited a closer relationship compared to the Bajaus. Based on the mitochondrial lineages, different waves/directions of dispersal into the Borneo Island that perhaps shaped the genetic discrepancies of the Bajau with the Kadazan-Dusun and Rungus groups were proposed. The Sabahan Bajau population could have persisted and originated from South Philippines since the earliest entry about 50,000 years ago. There was more interaction found in the Bajau with the surrounding lineages, such as East Asia, Mainland SEA, South Asia, and Oceania, which contributed to their high diversity. The Kadazan-Dusun and Rungus on the other hand, may have arrived in nearer timeframes, possibly following a western route through the Palawan Islands after their exodus from Taiwan some 5,000 to 10,000 years ago.

ABSTRAK

Sejak cadangan dikemukakan bahawa manusia moden berasal dari Afrika, dan juga pelbagai konsep yang bertentangan, kajian evolusi telah memberi tumpuan kepada penemuan garis masa untuk acara sedemikian berdasarkan rekod tinggalan kuno dan kontemporari populasi manusia. Walaupun penghijrahan manusia moden antara 7 benua utama dalam 120,000 tahun yang berhampiran ini telah dijelaskan dalam "out of Africa" hipotesis, pergerakan manusia dalam benua masing-masing masih sukar difahami. Peristiwa demik manusia di rantau Asia Tenggara juga telah menimbulkan perdebatan kuat di kalangan penyelidik. Ianya telah dicadangkan bahawa rantau ini telah didiami oleh manusia dari gelombang penghijrahan dari Taiwan kira-kira 5,000 tahun lalu, di mana ia disertakan dengan perkembangan bahasa Austronesia ke pelbagai bahagian di rantau Asia Tenggara dan seterusnya ke Oceania. Model tersebut dikenali sebagai "out of Taiwan". Walau bagaimanapun, kajian genetik dalam Asia Tenggara hanya tertumpu kepada penduduk dalam kawasan yang terletak di sepanjang laluan migrasi seperti yang dicadangkan dalam model "out of Taiwan", iaitu, China Selatan, Taiwan, Filipina, Indonesia Timur, dan "Near Oceania". Kajian ke atas penduduk di bahagian lain di rantau ini boleh menyumbangkan maklumat penting dan melengkapkan model tersebut, terutamanya pada pemahaman penghijrahan di Asia Tenggara dan juga antara wilayah yang perjiranan. Sabah terletak di hujung utara Pulau Borneo dan mempunyai lokasi yang strategik di tengah Asia Tenggara. Dalam kajian ini, objektifnya adalah untuk mencirikan struktur genetik 3 kumpulan penduduk bumiputra yang utama di Sabah, iaitu kaum Kadazan-Dusun, Bajau, dan Rungus, dan mengaitkan mereka kepada corak penghijrahan di rantau ini. Seramai 639 individu yang berasal daripada bumiputra Sabah disertakan dan DNA genomik telah dikeluarkan dari sampel darah/buccal. Polimorfisme pada DNA nuklear ("VNTR" dan "InDel") telah diuji dengan kaedah "PCR". Di samping itu, 15 penanda "STR" telah ditaipkan untuk setiap

sampel melalui kajian analisis fragmen. Selain daripada itu, DNA mitokondria telah diperiksa untuk mengesan delisi 9-bp dan turutan nukleotida pada 3 kawasan “hypervariable” dalam lingkungan-D telah ditentukan melalui reaksi jujukan. Kemudiannya, analisis statistik dan perbandingan dijalankan atas data genetik yang dihasilkan daripada kajian terdahulu. Secara umumnya, penduduk bumiputra Sabah telah dibuktikan mempunyai persamaan genetik yang tinggi (AMOVA < 5 %). Antaranya, kaum Kadazan-Dusun dan Rungus menunjukkan hubungan genetik yang lebih rapat berbanding dengan kaum Bajau. Berdasarkan ujian keturunan mitokondria, kewujudan gelombang yang berbeza dalam arahan penyebaran ke Pulau Borneo yang akhirnya membentuk perbezaan genetik di antara kaum Bajau dengan kaum Kadazan-Dusun dan Rungus telah dicadangkan. Kaum Bajau dicadangkan sampai di Filipina Selatan dalam Asia Tenggara menuruti migrasi dari Afrika seawal-awalnya 50,000 tahun lalu. Kaum Bajau juga mempunyai lebih interaksi dengan penduduk di kawasan sekitar, seperti Asia Timur, Tanah Besar Asia Tenggara, Asia Selatan, dan Oceania. Kaum Kadazan-Dusun dan Rungus bagaimanapun, mungkin tiba di Pulau Borneo dalam jangkamasa yang dekat, mungkin berikutan laluan barat melalui Kepulauan Palawan mengikuti gelombang yang keluar dari Taiwan dalam masa 5,000 hingga 10,000 tahun yang lalu.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my greatest gratitude to my supervisors, Professor Dr. Chua Kek Heng and Dr. Lian Lay Hoong, for their unequivocal support and untiring guidance throughout the entire period of my candidature. I have learnt invaluable values from them that have certainly strengthened my faith and interest in scientific research. I also appreciate their enlightening advice and deep insights that have been my continuous inspiration.

My sincere thanks also go to the 639 indigenous volunteers from the Sabah state for their kind willingness to contribute their DNA samples for our study. I would also take the opportunity to thank our collaborator, Associate Professor Dr. Lee Ping Chin from Universiti Malaysia Sabah (UMS), for conducting the field trip for sample collection. I am also thankful to all personnels who have involved and helped in the process.

A very huge thank to all lecturers and staff from the departments of Biomedical Science and Molecular Medicine. Not forgetting the administrative staff from the Dean's office of Faculty of Medicine and University of Malaya. My heartfelt gratitude for their help in all means, directly and indirectly, towards the successful submission of my doctoral thesis.

Special thanks to all members of Clinical Chemistry Laboratory – Ms. Chew Ching Hoong, Ms. Eva Puah Suat Moi, Ms. Chai Hwa Chia, and Ms. Lau Tze Pheng, and also to all graduate students from the department. Thank you for the good company and the fun sharing.

Last but not least, I would like to express my gratitude to my parents and family for their patience, understanding, and believe in me. I am also thankful for their support and encouragement. Thanks to Ms. Tan Kim Kee for her constant care and support through all the difficulties.

Finally, I would like to acknowledge the financial support received for the study from University of Malaya (PPP #PS145/2008C, #PS234/2010A), Ministry of Science, Technology and Innovation (eScience Fund #12-02-03-2045), and Ministry of Higher Education (FRGS #FP024/2009).

LIST OF CONTENTS

	PAGE
TITLE PAGE	I
ORIGINAL LITERARY WORK DECLARATION	II
ABSTRACT	III
ABSTRAK	V
ACKNOWLEDGEMENTS	VII
LIST OF CONTENTS	IX
LIST OF FIGURES	XIV
LIST OF TABLES	XVIII
ABBREVIATIONS	XX
LIST OF APPENDICES	XXIII
CHAPTER 1 INTRODUCTION	1
CHAPTER 2 LITERATURE REVIEW	4
2.1 The emergence of modern humans	4
2.2 Migration of modern humans	10
2.2.1 Single origin hypothesis	10
2.2.2 Multiregional model	13
2.3 The land below the wind - Sabah	15
2.3.1 Kadazan-Dusun	16
2.3.2 Rungus	17
2.3.3 Bajau	17
2.4 Migration patterns within Southeast Asia	18
2.5 The human genome	23
2.5.1 Nuclear DNA	23
2.5.2 Mitochondrial DNA	24
2.5.3 Genetic polymorphisms	27
2.5.3.1 Variable number of tandem repeats	27
2.5.3.1(a) Dopamine transporter gene	28
2.5.3.2 Microsatellite	31
2.5.3.3 InDel markers	34
2.5.3.3(a) <i>Alu</i> insertions	35
2.5.3.3(b) Mitochondrial 9-bp deletion	38
2.5.3.4 Single nucleotide polymorphisms	39

	PAGE
2.5.4 Polymorphisms in mtDNA	41
2.5.4.1 Mitochondrial haplogroups and migration	41
2.5.4.2 Nomenclature	44
CHAPTER 3 MATERIALS AND METHODS	45
3.1 Materials	45
3.1.1 Solution, reagents, and buffer preparation	45
3.1.1.1 2 M NaOAc, pH 5.6	45
3.1.1.2 NaCl, 5 M and 6 M	45
3.1.1.3 Ethanol, 70 % (v/v)	45
3.1.1.4 Proteinase K buffer, 5 X strength	46
3.1.1.5 Proteinase K solution, 10 mg/ml	46
3.1.1.6 Red Cell Lysis buffer, 10 X strength	46
3.1.1.7 SDS solution, 20 % (w/v)	47
3.1.1.8 MgCl ₂ ·6H ₂ O, 1 M	47
3.1.1.9 TRIS-HCl, 1 M, pH 7.5	47
3.1.1.10 EDTA, 0.5 M, pH 8.0	47
3.1.1.11 Double distilled water	48
3.1.2 Commercially available reagents	49
3.1.3 Commercially available kits	49
3.1.4 Analysis software	50
3.1.5 Essential research services	51
3.2 Research methodologies	52
3.2.1 Sterilization techniques	52
3.2.2 Volunteer recruitment and sample collection	53
3.2.3 Genomic DNA extraction	54
3.2.4 Quantity and quality assessment of DNA samples	56
3.2.5 Genotyping of samples	57
3.2.5.1 Nuclear DNA markers: VNTRs	59
3.2.5.2 Nuclear DNA markers: InDels	61
3.2.5.3 Nuclear DNA markers: STRs	65
3.2.5.3(a) Amplification stage	66
3.2.5.3(b) Fragment analysis stage	68
3.2.5.3(c) Raw data collection and allele calling	69
3.2.5.3(d) Non-allelic variant	70

	PAGE
3.2.5.3(d)(i) PCR and ligation reactions	70
3.2.5.3(d)(ii) Colony selection and screening	72
3.2.5.3(d)(iii) Plasmid preparation	73
3.2.5.3(d)(iv) Plasmid sequencing	74
3.2.5.3(e) Statistical analysis	75
3.2.5.4 MtDNA markers: 9-bp deletion	78
3.2.5.5 MtDNA markers: SNPs in control region	80
3.2.5.5(a) PCR amplification and product purification	80
3.2.5.5(b) Sequencing reactions	82
3.2.5.5(c) Use of additional primers	82
3.2.5.5(d) Cloning of problematic samples	85
3.2.5.5(e) MtDNA sequence analysis	86
3.2.5.5(e)(i) Haplotype and polymorphism frequencies	86
3.2.5.5(e)(ii) Haplogroup assignment	87
3.2.5.5(e)(iii) Quality assurance and data deposition	87
3.2.5.5(f) MtDNA: Principal component analysis	90
3.2.5.5(g) MtDNA: Phylogenetic analysis	91
CHAPTER 4 RESULTS	92
4.1 DNA extraction	92
4.2 Nuclear DNA markers: VNTRs	92
4.2.1 DAT-1 3'UTR 40-bp VNTR	93
4.2.2 DAT-1 Intron-8 30-bp VNTR	96
4.2.3 Forensic and population parameter evaluation	99
4.3 Nuclear DNA markers: InDels	102
4.3.1 InDels: Distribution among indigenous populations	105
4.3.2 InDels: Marker evaluation	107
4.3.3 InDels: Population differentiation analysis	109
4.3.4 InDels: Principal component analysis	111
4.3.4.1 PCA of world populations	111
4.3.4.2 PCA of populations in the neighboring regions	114
4.3.5 InDels: Phylogenetic assessment	116
4.3.5.1 Phylogenetic assessment with world populations	116
4.3.5.2 Phylogenetic assessment in neighboring regions	118

	PAGE
4.4 Examination of autosomal STR markers	120
4.4.1 STR distribution in the Sabahan indigenous populations	120
4.4.1.1 Kadazan-Dusun	124
4.4.1.2 Bajau	127
4.4.1.3 Rungus	130
4.4.2 Identification of non-allelic variant	133
4.4.3 STRs: Population differentiation analysis	135
4.4.4 STRs: Population structure analysis	137
4.4.5 STRs: Cluster analysis – Principal coordinate analysis	141
4.4.5.1 PCoA of world populations	141
4.4.5.2 PCoA of neighboring populations	144
4.4.6 STRs: Cluster analysis – Phylogenetic analysis	147
4.4.6.1 Phylogenetic analysis of world populations	147
4.4.6.2 Phylogenetic analysis of neighboring populations	150
4.5 Assessment of mtDNA	152
4.5.1 Intergenic 9-bp deletion marker	152
4.5.2 SNPs in the control region	154
4.5.2.1 Sequence quality check	154
4.5.2.2 Distribution of genetic polymorphisms	156
4.5.2.3 Sequencing of problematic samples	161
4.5.2.4 Haplogroup determination	163
4.5.2.5 Haplotype analysis - individual population	169
4.5.2.6 Haplotype analysis - combined study	171
4.5.2.7 Principal component analysis	174
4.5.2.8 Phylogenetic analysis	176
CHAPTER 5 DISCUSSION	179
5.1 VNTRs in the Sabahan indigenous populations	179
5.1.1 Distribution in world and neighboring populations	180
5.2 <i>Alu</i> elements	187
5.2.1 <i>Alu</i> insertions in the Sabahan indigenous populations	189
5.2.2 World distribution of <i>Alu</i> insertions	191
5.2.3 PCA plots and NJ trees	194
5.3 STRs within the Sabahan indigenous populations	198

	PAGE
5.3.1 STRs: Non-allelic variant	201
5.3.2 STRUCTURE of the Sabahan indigenous populations	204
5.3.3 STRs: Clustering of world populations	205
5.3.4 STRs: Clustering of SEA neighboring populations	208
5.4 Examination of mtDNA	211
5.4.1 Mitochondrial intergenic 9-bp deletions	211
5.4.2 Mitochondrial haplogroups	216
5.4.3 Haplogroups in the Sabahan indigenous populations	219
5.4.3.1 Haplogroup B4a	221
5.4.3.2 Haplogroup R9	223
5.4.3.3 Haplogroup E	225
5.4.3.4 Haplogroup M7	227
5.4.3.5 Haplogroup F	232
5.4.3.6 Haplogroup D	234
5.4.3.7 Haplogroups N and Y	236
5.4.4 Mitochondrial haplogroups: PCA	238
5.4.5 Phylogenetic assessment of mtDNAs	241
5.5 Summary and conclusion	256
5.5.1 Variable number of tandem repeat	256
5.5.2 <i>Alu</i> insertion	258
5.5.3 Short tandem repeat	258
5.5.4 Mitochondrial DNA	259
LIMITATIONS AND FUTURE STUDIES	261
REFERENCES	264
APPENDICES	290

LIST OF FIGURES

	PAGE
Figure 2.1 : Distribution of hominin species after divergence from the chimpanzee lineage, based on fossil evidence unearthed from various regions of the world	7
Figure 2.2 : Speciation of modern humans and geographic distribution of <i>H. sapiens</i> along the evolution	8
Figure 2.3 : Different models describing the origin of modern humans	11
Figure 2.4 : Geographical location of Sabah state in East Malaysia, situated at the northern tip of the island of Borneo	15
Figure 2.5 : Diagrammatic representation of the proposed migration patterns as suggested by their respective proponents	20
Figure 2.6 : Structure of mtDNA	25
Figure 2.7 : Diagrammatic illustration of polymorphic variants of VNTRs in the 3'UTR and Intron-8 of DAT-1 gene	30
Figure 2.8 : Schematic representation of the formation of +1 and -1 repeat alleles by replication slippage	32
Figure 2.9 : Diagram showing the insertion of an <i>Alu</i> element into a genome sequence resulting in the elongation of the particular stretch of sequence (“+” allele) than the original sequence without insertion (“-” allele)	36
Figure 2.10 : Illustration of the position of region V and the 9-bp deletion in mtDNA	38
Figure 2.11 : Migration route of humans out of Africa as suggested by genetic data from mtDNA	43
Figure 3.1 : Outline of the genetic studies carried out on the Sabahan indigenous populations, via the examination of multiple loci in different genome subsets	58
Figure 3.2 : PCR cycling parameters employed in the amplification of Powerplex 16 system	67
Figure 3.3 : PCR cycling parameters and primer sequences for mt 9-bp deletion examination	79
Figure 3.4 : PCR cycling parameters and primer sequences for amplification of mt HV regions	81

	PAGE
Figure 3.5 : Sequencing strategy of mt HV regions in the control region, 1,121 bp spanning from positions 16,024 to 576	84
Figure 4.1 : Visualization of PCR amplicon of the DAT-1 3'UTR 40-bp VNTR on 3 % (w/v) native agarose gel stained with EtBr	93
Figure 4.2 : Separation patterns of different genotypes of DAT-1 Intron-8 30-bp VNTR on 3 % (w/v) EtBr-stained native agarose gel	96
Figure 4.3 : Band patterns of various genotypes for HS3.23, TPA25, and HS4.32 on 2 % (w/v) native agarose gel, pre-stained with EtBr	103
Figure 4.4 : Band patterns of various genotypes for APO, PV92, and B65 on 2 % (w/v) native agarose gel, pre-stained with EtBr	104
Figure 4.5 : PC plot based on <i>Alu</i> insertion frequencies of the 3 Sabahan indigenous and world populations	112
Figure 4.6 : Diagram showing PCA and biplot constructed using insertion frequencies of 4 <i>Alu</i> markers of the Sabahan indigenous groups and populations in the neighboring regions	115
Figure 4.7 : Radiated and unrooted NJ tree constructed based on insertion frequencies of <i>Alu</i> markers from the world populations	117
Figure 4.8 : Unrooted phylogenetic tree depicts relationships of Kadazan-Dusun, Bajau, and Rungus with their neighboring populations	119
Figure 4.9 : Diagram showing the electropherogram of fluorescein-labeled STR loci (D3S1358, TH01, D21S11, D18S51, and Penta E)	121
Figure 4.10 : Diagram showing the electropherogram of JOE-labeled STR loci (D5S818, D13S317, D7S820, D16S539, and Penta D)	122

	PAGE
Figure 4.11 : Diagram showing the electropherogram of TMR-labeled STR loci (sex-determining Amelogenin, vWA, D8S1179, TPOX, and FGA)	123
Figure 4.12 : Identification of allele 15 for FGA locus	134
Figure 4.13 : Two approaches for the inference of best-fit <i>K</i>	138
Figure 4.14 : Bar plot displays individual ancestry estimates of <i>K</i> structure of 2 to 6 of the studied populations	140
Figure 4.15 : Distribution of populations on the PCoA plot built from allelic frequencies of 13 STR loci	143
Figure 4.16 : PCoA plot of 40 populations in SEA and the neighboring regions (South Asia, East Asia, and Oceania)	146
Figure 4.17 : NJ tree showing the clustering pattern of the world populations	148
Figure 4.18 : NJ tree constructed based on the genetic distance derived from 12 STR loci for the SEA-neighboring populations	151
Figure 4.19 : A 3.5 % (w/v) EtBr-stained native agarose gel showing PCR amplicon for the screening of mt 9-bp deletion in region V	153
Figure 4.20 : Drawings of mt lineages among 150 Kadazan-Dusun individuals as represented by network and torso	155
Figure 4.21 : Electropherograms of DNA sequencing via forward primer for purified amplicon and cloned vector, in comparison to rCRS	162
Figure 4.22 : Bar chart showing the frequencies of mt haplogroups in Kadazan-Dusun, Bajau, and Rungus populations	167
Figure 4.23 : Diagram illustrating the division of haplotypes for the Kadazan-Dusun, Bajau, and Rungus populations	171
Figure 4.24 : PCA plot illustrating clusters of populations based on mt haplogroup frequencies in East Asia and SEA regions	175
Figure 4.25 : Condensed NJ tree constructed via 1,036 mtDNAs	177
Figure 5.1 : Paths of evolution of <i>Alu</i> elements	188
Figure 5.2 : Line chart shows insertion frequencies of <i>Alu</i> markers (APO, B65, HS4.32, PV92, and TPA25) in modern human populations from different continents	192

	PAGE
Figure 5.3 : Diagrammatic representation of mt 9-bp deletion frequencies as seen in various modern human populations around the world	214
Figure 5.4 : Clustering of haplogroups R14, R22, and N8 in the NJ tree	242
Figure 5.5 : Segregation of Bajau individuals in the clusters of haplogroups B4a2b, B4a1a1a (Polynesian motif), and Y2	243
Figure 5.6 : B5b individuals of the Bajau group	244
Figure 5.7 : The division of Kadazan-Dusun and Bajau individuals in the B4b1 cluster	245
Figure 5.8 : Sub-clades of haplogroups F3b and F3b1 for Bajau and Rungus individuals in the NJ tree	246
Figure 5.9 : D5b1c1 clusters for Kadazan-Dusun and Rungus individuals	247
Figure 5.10 : Haplogroups B4a1a and E1b	248
Figure 5.11 : Haplogroup E1a1a	251
Figure 5.12 : Clustering of M7c3c individuals in the Sabahan indigenous populations	253
Figure 5.13 : Clade of haplogroup M7b1	255
Figure 5.14 : “Out of Taiwan” expansion and the routes of migration taken by the ancestors of Kadazan-Dusun, Rungus, and Bajau populations into the Borneo Island	260

LIST OF TABLES

	PAGE
Table 2.1 : Timeline and summary of the evolution of hominin	5
Table 3.1 : PCR components for the amplification of VNTR markers	60
Table 3.2 : Sequences of primers used to amplify the targeted markers in the VNTR study	60
Table 3.3 : Chromosomal locations and primer sequences of the 6 <i>Alu</i> insertions investigated in the present study	62
Table 3.4 : Reaction components for amplification of Powerplex 16 system	66
Table 3.5 : Reagents added in the ligation mixture	71
Table 3.6 : Sequences of internal primers used for re-sequencing of samples with homopolymeric C-stretches	83
Table 4.1 : Distribution of alleles and genotypes of DAT-1 3'UTR 40-bp VNTR polymorphisms in Kadazan-Dusun, Bajau, and Rungus	95
Table 4.2 : Allelic and genotypic frequencies of DAT-1 Intron-8 30-bp VNTR polymorphisms in Kadazan-Dusun, Bajau, and Rungus	98
Table 4.3 : Forensic parameters and population differentiation study of 2 VNTRs in the DAT-1 gene	99
Table 4.4 : Allelic and genotypic distributions of 6 <i>Alu</i> insertions in genotyped samples from Kadazan-Dusun, Bajau, and Rungus, Sabah	106
Table 4.5 : PD, PE, PIC, TPI, and HWE of <i>Alu</i> markers in the Sabahan indigenous populations	107
Table 4.6 : Insertion frequency, population differentiation analysis, and AMOVA of 6 <i>Alu</i> markers in Kadazan-Dusun, Bajau, and Rungus populations	109
Table 4.7 : Allelic distribution of 15 autosomal STR markers in the Kadazan-Dusun population	125
Table 4.8 : Allelic distribution of 15 autosomal STR markers in the Bajau population	128

	PAGE
Table 4.9 : Allelic distribution of 15 autosomal STR markers in the Rungus population	131
Table 4.10 : Summary of the genetic differentiation and AMOVA analyses performed on 15 autosomal STR markers in Kadazan-Dusun, Bajau, and Rungus individuals	136
Table 4.11 : MtDNA sites with more than 1 observed variant in the Sabahan indigenous individuals	156
Table 4.12 : Summary of the distribution of genetic polymorphisms present in all 3 mt HV regions in 150 Kadazan-Dusun individuals	158
Table 4.13 : Summary of the distribution of genetic polymorphisms present in all 3 mt HV regions in 150 Bajau individuals	159
Table 4.14 : Summary of the distribution of genetic polymorphisms present in all 3 mt HV regions in 150 Rungus individuals	160
Table 4.15 : Distribution of mt haplogroups determined in 450 Sabahan indigenous individuals	165
Table 4.16 : Distribution of haplotypes observed within respective examined indigenous populations	170
Table 4.17 : Numbers of haplotypes that are shared and most frequently found in the 3 indigenous populations in Sabah	173
Table 5.1 : Global allelic distribution of VNTR in the DAT-1 3'UTR	181
Table 5.2 : Global distribution of alleles for DAT-1 Intron-8 VNTR in populations from different regions	185
Table 5.3 : Comparison of the efficiency of the 15 STR markers in different populations	199
Table 5.4 : Frequencies of FGA allele 15	202
Table 5.5 : Frequency review of the intergenic 9-bp deletion within the mt coding region at nps 8,271 to 8,279 in world populations	211
Table 5.6 : Tree of global mt haplogroups	217
Table 5.7 : Advantages and limitations of the 4 subsets of genetic markers used in the study	257

ABBREVIATIONS

A	...	Adenine
AD	...	<i>Anno Domini</i>
AGE	...	Agarose Gel Electrophoresis
ALFRED	...	The Allele Frequency Database
AMEL	...	Amelogenin
AMH	...	Anatomically modern human
AMOVA	...	Analysis of molecular variant
BP	...	Before present
bp	...	Basepair
C	...	Cytosine
CO II	...	Cytochrome Oxidase subunit II
CODIS	...	Combined DNA Index System
D1	...	First principal component
D2	...	Second principal component
D _A	...	Genetic distance
D-loop	...	Displacement loop (mitochondrial)
DAT-1	...	Dopamine transporter (gene)
DNA	...	Deoxyribonucleic acid
dNTP	...	Deoxynucleotide mix
D _{ST}	...	Inter-population gene diversity
ds	...	Double stranded
e.g.	...	For example
EDTA	...	Ethylenediaminetetraacetic acid
EtBr	...	Ethidium Bromide
Etc	...	et cetera
FL	...	Fluorescein
G	...	Guanine
g	...	Gram
G _{IS}	...	Inbreeding coefficient
G _{ST}	...	Coefficient of gene differentiation
<i>H.</i>	...	<i>Homo</i> (genus)
H-strand	...	Heavy strand (mitochondrial)
H _{Obs}	...	Observed heterozygosity

H _{Exp}	...	Expected heterozygosity
HV	...	Hypervariable region (mitochondrial)
H _S	...	Intra-population gene diversity
H _T	...	Total gene diversity
HWE	...	Hardy-Weinberg equilibrium
i.e.	...	id. Est (that is)
ILS	...	Internal lane standard
InDel	...	Insertion-deletion
ISEA	...	Island Southeast Asia
JOE	...	6-carboxy-4',5'-dichloro-2',7'-dimethoxyfluorescein
Kb	...	Kilo basepair
KCA	...	Kadazan Cultural Association
Kv	...	Kilo volt
L	...	Litre
LINE	...	Long interspersed element
L-strand	...	Light strand (mitochondrial)
M	...	Molar
Ma	...	Megaannum
mya	...	Million years ago
MCRA	...	Most Common Recent Ancestor
mg	...	Milligram
ml	...	Millilitre
mt	...	Mitochondrial
MW	...	Molecular weight
NCBI	...	National Center for Biotechnology Information
NJ	...	Neighbor Joining
nm	...	Nanometer
NTC	...	Non-template control
OD	...	Optical density
PCA	...	Principal component analysis
PCoA	...	Principal coordinate analysis
PD	...	Power of discrimination
PE	...	Power of exclusion
PCR	...	Polymerase chain reaction
PIC	...	Polymorphism information content

p.s.i.	...	Pound-force per square inch
QM	...	Quasi median
rCRS	...	Revised Cambridge reference sequence
RFLP	...	Restricted fragment length polymorphism
RNA	...	Ribonucleic acid
rRNA	...	Ribosomal RNA
rpm	...	Round per minute
RSRS	...	Reconstructed sapiens reference sequence
sdH ₂ O	...	Double distilled water
SEA	...	Southeast Asia
SINE	...	Short interspersed element
SNP	...	Single nucleotide polymorphism
STR	...	Short tandem repeat
T	...	Thiamine
TA	...	Annealing temperature
TBE	...	Tris-borate-EDTA
TMR	...	Carboxy-tetramethylrhodamine
tRNA	...	Transfer RNA
TPI	...	Typical paternity index
UV	...	Ultraviolet
U	...	Unit (enzyme)
VNTR	...	Variable number of tandem repeat
xg	...	Gravity force
3'UTR	...	3' untranslated region
°C	...	Degree Celsius
µg	...	Microgram
µl	...	Microlitre
µM	...	Micromolar
>	...	More than
<	...	Less than
~	...	About/approximately

LIST OF APPENDICES

	PAGE
Appendix 1 : Sample of informed consent form for volunteers	290
Appendix 2 : Approval letter for ethical clearance (reference number: 770.21)	291
Appendix 3 : Approval letter for ethical clearance (reference number: 612.16)	292
Appendix 4 : Frequencies of <i>Alu</i> insertions in various populations used for the construction of PCA and NJ phylogenetic trees	293
Appendix 5 : Genetic distance (D_A) of 48 populations generated from allelic frequencies of 13 STR loci	294
Appendix 6 : Genetic distance (D_A) of 40 populations generated from allelic frequencies of 12 STR loci	295
Appendix 7 : Mt polymorphisms of 150 Kadazan-Dusun individuals and their determined haplogroups, presented in the EMPOP database format	296
Appendix 8 : Mt polymorphisms of 150 Bajau individuals and their determined haplogroups, presented in the EMPOP database format	298
Appendix 9 : Mt polymorphisms of 150 Rungus individuals and their determined haplogroups, presented in the EMPOP database format	300
Appendix 10 : Frequencies of mt haplogroup of various populations used in the construction of PCA	302
Appendix 11 : Publications	303
Appendix 12 : Presentation in conference	305

CHAPTER 1

INTRODUCTION

INTRODUCTION

The population movements within Southeast Asia have been studied and explained by researchers from different principal disciplines, i.e., archeology, anthropology, evolution, and linguistic, and a number of proposals/suggestions have been raised. Among them, the “out of Taiwan” model by Bellwood has gained general acceptance. According to the model, all Austronesian speakers in the contemporary Southeast Asia descended from Taiwanese migrants (originated from southern China), who travelled into Indonesia and Near Oceania through Philippines about 5,000 years ago. Hence, genetic studies in this region have focused mainly on the populations along the proposed migratory path, namely the southern Chinese, Taiwanese, Filipino, Indonesian, and Oceania. Populations in other parts of the Southeast Asia region received relatively less attention, including those residing in the Borneo Island. Southeast Asia is the homeland to a variety of ethnic groups from different language families and is well-known for its high degree of diversity. It remains speculative whether or not the rich diversity in Southeast Asia can be explained solely by the Neolithic expansion from Taiwan, which dates back few thousand years ago. Hence, it would be interesting to examine the genetic structure of other Austronesian populations in this region, which may in turn shed important information on the local movement events. The Borneo Island is located in the centre of Island Southeast Asia, surrounded by the Mainland, Indonesia, and Philippines. Its strategic location would serve as a vital transfer station for ancient migrants who moved between the various regions in Southeast Asia. The Borneo Island houses a spectrum of Austronesian speakers that were believed to have originated from Taiwan. Often, these populations are not treated as separated groups but are included as a single population in some large scale genetic studies. On the other hand, previous studies conducted on some of these populations (Iban, Bidayuh, and Melanau) focused on the population and forensic evaluation, but did not relate them to

the movement events in this region. Although there have been a limited number of publications on the indigenous populations in the Sabah, the markers screened in these studies were less informative and uncommonly used for comparison analysis.

Thus, the objective of our study was to access the genomic diversity of the Sabahan indigenous populations and characterize their genetic structure. The specific aims were as follow:

- (a) To study and analyze the genomic diversity in the Sabahan indigenous population, i.e. looking specifically into different sets of polymorphisms.
- (b) To generate genetic data and variant frequencies of the respective loci, and later examine if there is any pattern linking these subsets of genetic polymorphisms (unique haplotypes).
- (c) Using the data generated, calculates various basic statistical population parameters, as this would help to characterize the structure of the population studied.
- (d) To construct phylogenetic trees that correlate all the populations in the study, in order to access the genetic relevance among these populations.
- (e) To develop markers that can be used effectively in forensic investigation, anthropology, and evolutionary studies for the local population.

In the present study, the genomic diversity of 3 of the largest indigenous populations residing in the Sabah state, i.e., Kadazan-Dusun, Bajau, and Rungus, was accessed. Although they live in close proximity and share the same branch of the Austronesian language, they each have a distinctive characteristic, in terms of culture, custom, and living traditions. By examination and characterization of the genetic structure of these indigenous populations, they were related to the migration patterns within the region. In order to generate a comprehensive representation of their genomes, different types of

genetic polymorphisms were examined for the 2 fractions of genomes (chromosomal and mitochondrial DNAs). Statistical analyses were conducted for intra- and inter-population differences. In addition, the mitochondrial lineages (one of the most extensively studied uniparental markers) were interpreted and intergrated to suggest some possible migration patterns and origins of the Kadazan-Dusun, Bajau, and Rungus populations.

CHAPTER 2

LITERATURE REVIEW

2 LITERATURE REVIEW

2.1 The emergence of modern humans

The evolution of the modern humans has been a long process since the day it branched off from the common ancestor with its cousin species, the gorillas and chimpanzees, about 4 to 8 million years ago (Ma) (Chen & Li, 2001). The speciation continued along the lineage, accompanied by the existence of species from few genera before the occurrence of anatomically modern human (AMH) (Table 2.1 and Figure 2.1). The immediate species after divergence from the chimpanzee lineage came from the genus *Sahelanthropus*, followed by the genera *Orrorin* and *Ardipithecus* (Brunet, et al., 2002; Haile-Selassie, 2001; Senut, et al., 2001). All species in these genera had gradually adapted to walk upright on 2 legs – bipedalism.

Later, about 4 to 1.8 Ma, the genus *Australopithecus* emerged in eastern African and spread throughout the region (Ward, Leakey, & Walker, 1999). This genus were true bipedals, making their hands free for grasping or carrying food, unlike their ancestral genera who were fugitative bipedals – unable to walk or run on legs for too long. Before their extinction, the species in genus *Australopithecus* gave rise to 2 unique genera, i.e., *Paranthropus* and *Homo*, which co-existed at the same time (Wood & Strait, 2004). However, the genus *Paranthropus*, which had poorer adaptation to the environment, died out in 2 million years, leaving no descendants in its lineage. On the other hand, the genus *Homo* had adapted well and continued to evolve and eventually lead to the emergence of modern humans.

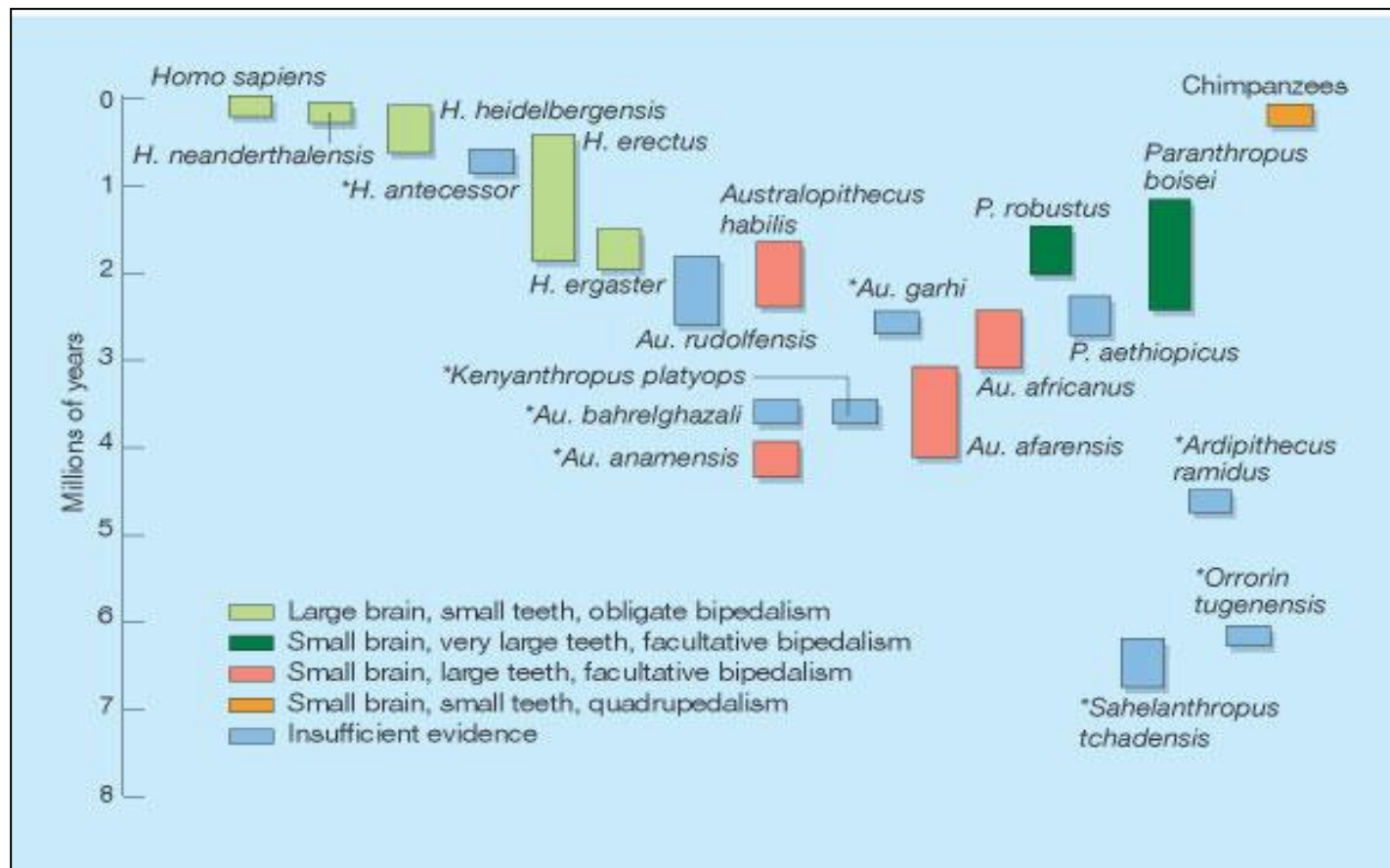
Homo sapiens (*H. sapiens*) have been living from 250,000 years ago until present and it is the only surviving species in the genus (Figure 2.2). *H. sapiens* is believed to have descended from *H. erectus*, with evidence from the expansion of the skull cavity and the ability to develop and elaborate stone tools (Plummer, 2004). Paleoanthropologists

Table 2.1 : Timeline and summary of the evolution of hominin.

Period of occurrence (Ma)	Genus	Species	Note
7	<i>Sahelanthropus</i>	<ul style="list-style-type: none"> • <i>tchadensis</i> 	Strong arguments on whether it should be placed as the common ancestor of humans and chimpanzees or only humans
6	<i>Orrorin</i>	<ul style="list-style-type: none"> • <i>tugenensis</i> 	Second oldest known ancestor of hominin; bipedal
5.5 - 4.4	<i>Ardipithecus</i>	<ul style="list-style-type: none"> • <i>kadabba</i> • <i>ramidus</i> 	One of the very early hominins; facultative biped that could not run or walk for too long
4 - 1.8	<i>Australopithecus</i>	<ul style="list-style-type: none"> • <i>anamensis</i> • <i>afarensis</i> • <i>bahrelghazali</i> • <i>africanus</i> • <i>garhi</i> • <i>sediba</i> 	A significant genus that evolved in and spread throughout Africa; give rise to 2 distinctive genera, i.e., <i>Homo</i> and <i>Paranthropus</i>
3 - 2.7	<i>Kenyanthropus</i>	<ul style="list-style-type: none"> • <i>platyops</i> 	Proposed to represent a new hominine genus – the <i>Kenyanthropus</i> , yet others regarded it as a member of the <i>A. afarensis</i>
3 - 1.2	<i>Paranthropus</i>	<ul style="list-style-type: none"> • <i>robustus</i> • <i>boises</i> • <i>aethiopicus</i> 	Co-existed with species of the genus <i>Homo</i> ; extinct as a result of poor adaptation
2 - present	<i>Homo</i>	<ul style="list-style-type: none"> • <i>habilis</i> • <i>ergaster</i> • <i>erectus</i> • <i>heidelbergensis</i> • <i>neanderthalensis</i> • <i>sapiens</i> 	The genus that includes modern humans and other related species

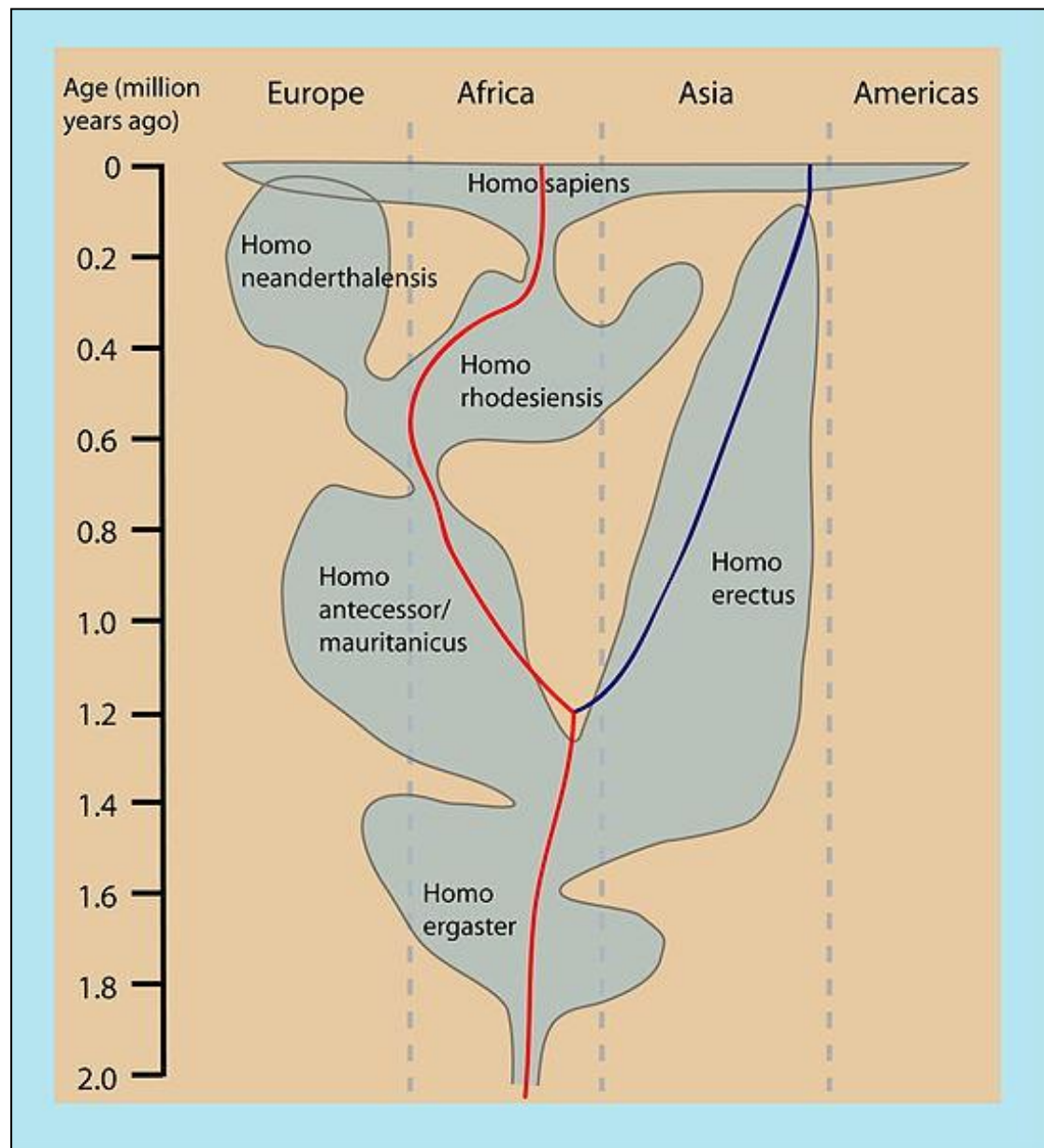
Table 2.1 : continuation.

Period of occurrence (Ma)	Species	Note
2.3 – 1.4	<i>Homo habilis</i>	The first member of genus <i>Homo</i> ; also known as “handy man” as the first body remains were discovered beside many tools; had less-specialized diet; give rise to <i>H. ergaster</i>
1.9 – 1.4	<i>Homo ergaster</i>	Used improved and relatively more complex stone tools than <i>H. habilis</i> ; suggested to give rise to <i>H. erectus</i> in Asia; alternatively, was also regarded as a form of <i>H. erectus</i> living in Africa
1.8 – 0.2	<i>Homo erectus</i>	First to live in coordinated community as gatherer-hunter; utilized voice to communicate among community members; ability to use and control fire
0.6 - .035	<i>Homo heidelbergensis</i>	Give rise to Neanderthals and modern humans; first to bury the dead; may have acquired languages, in primitive form though; have taller and larger figure than its ancestors
0.35 – 0.03	<i>Homo neanderthalensis</i>	Close relationship to modern humans; co-existed and interbred with modern humans; no fossil records found in Africa, thought to migrate out of Africa and evolve in Europe and central Asia; used sophisticated tools and built dwellings with animal bones/skins
0.2 - present	<i>Homo sapiens</i>	The only surviving member of genus <i>Homo</i> ; emerged and evolved in Africa; left Africa and colonized all parts of the world



(Wood, 2002)

Figure 2.1 : Distribution of hominin species after divergence from the chimpanzee lineage, based on fossil evidence unearthed from various regions of the world.



(Reed, Smith, Hammond, Rogers, & Clayton, 2004)

Figure 2.2 : Speciation of modern humans and geographical distribution of *H. sapiens* along the evolution.

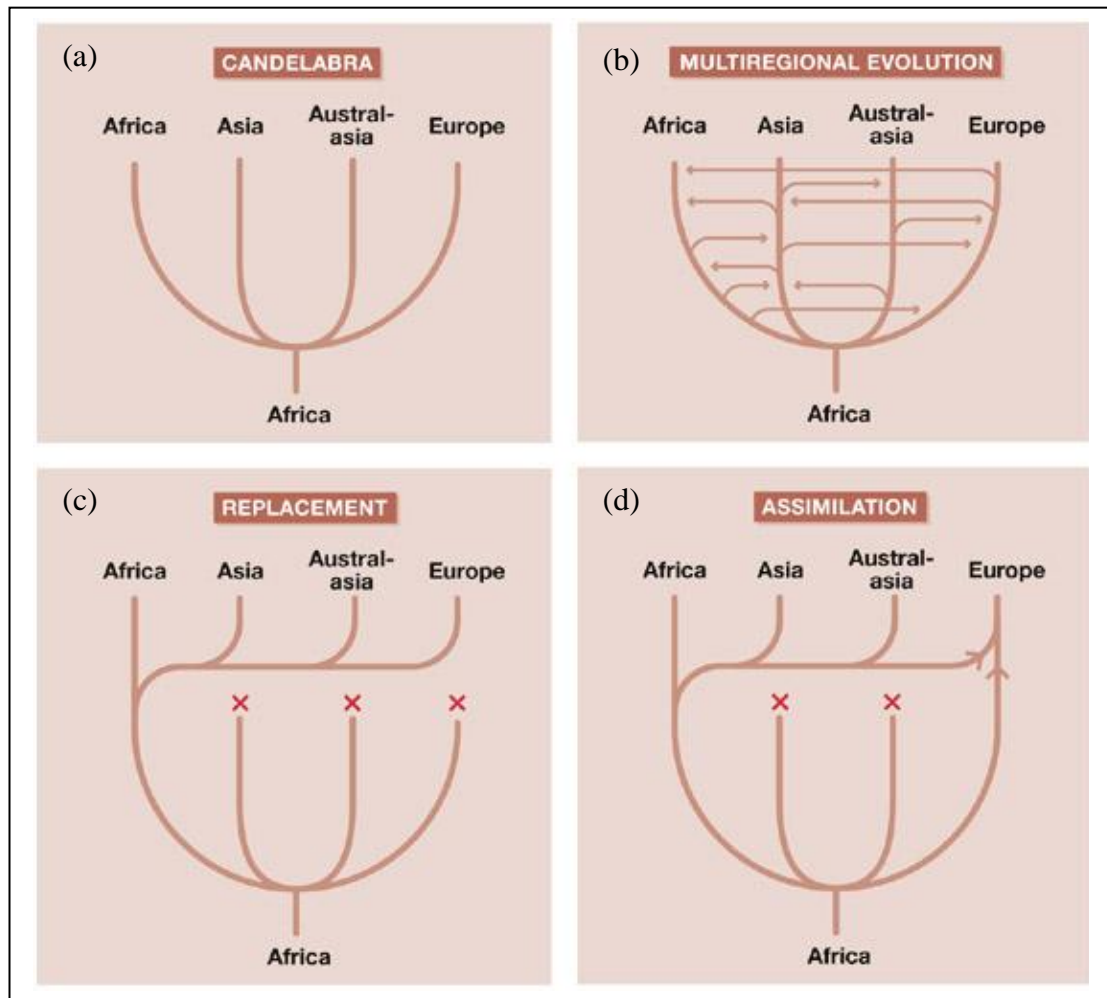
proposed that *H. sapiens* may have originated and evolved from a population of *H. erectus* in Africa, while other groups of *H. erectus* colonized other parts of the world (Anton, 2003). Later, the *H. sapiens* migrated out and gradually replaced the existing *H. erectus* that was dispersed much earlier in time (Aiello & Wells, 2002).

2.2 Migration of modern humans

In the study of human evolution, the migration of modern humans, as in how and where, has been greatly debated by researchers from all over the world (Armitage, et al., 2011; Oppenheimer, 2009; Rasmussen, et al., 2011). Despite the extensive studies, there is yet to be a conclusive and mutual agreement among scientists. There are a number of theories and hypotheses being proposed to describe the patterns and origin of modern humans (Figure 2.3). These theories can be grouped mainly into 2 categories based on one aspect, i.e., whether modern humans arose from a single origin or multiple origins. As the name infers, single origin theories, or monogenesis, suggest that present humans are descendants of a group of ancestral humans who lived and evolved in a region before they spread to other parts of the world (Stringer & Andrews, 1988). Conversely, scientists who believe in the multiregional model, suggest that modern humans originated from ancestors that arose and evolved individually in different parts of the world – polygenesis (Fruyer, Wolpoff, Thorne, Smith, & Pope, 1993).

2.2.1 Single origin hypothesis

The single origin hypothesis is the prevailing model opposed to the multiregional origin hypotheses. This hypothesis proposes that modern humans originated and evolved in Africa before their dispersal to the rest of the world (Stringer & Andrews, 1988). This speculation was first suggested by a very renowned natural geneticist - Charles Darwin, in the year 1871. He thought that the common ancestor of all humans once lived in Africa based on his observation (Darwin, 1871). He observed that all living mammals in every parts of the world are in close relation to the extinct species in the same region. Therefore, he proposed that human ancestor may have probably co-existed with their closest allied species, the chimpanzee and gorilla - in Africa (Darwin, 1871). His view



(Stoneking, 2008)

Figure 2.3 : Different models describing the origin of modern humans. (a) Candelabra: refers to parallel evolution where humans in different regions evolved independently. (b) Multiregional: arrows between the 4 main streams indicate gene flow between humans in different regions. (c) Replacement: suggest that modern humans evolved in Africa and migrated out to replace archaic humans, without interbreeding with them. (d) Assimilation: a modification to replacement model that proposed certain degree of assimilation between modern and archaic humans, based on findings in genetic studies.

on the origin of humans remained speculative until 1980s, where scientific research and technologies have advanced greatly and allowed more genetic studies to be carried out in greater detail and depth.

Darwin's concept has gained support through the discoveries of various fossils of hominids within and outside of Africa. In Africa, anthropologists unearthed numerous pieces of fossils that belonged to hominids of different ages ranging from 7,000,000 to 12,000 years ago (Leakey & Walker, 1997). The oldest fossils belonged to the common ancestor of all hominans – *Sahelanthropus tchadensis*, while the youngest fossil was identified as a member of *H. sapiens* (Brunet, et al., 2002; Wendorf, et al., 1984). In addition, all fossils outside of Africa, except for the one that was found in the Qafzeh cave in Israel, have been dated to more recent times of less than 60,000 years ago (Bowler, et al., 2003; Olley, Roberts, Yoshida, & Bowler, 2006). These findings have favoured the concept of Africa as the single point of origin for all modern humans. The discovery of fossils of *H. sapiens* in the Qafzeh cave was dated 80,000 to 100,000 years. Scientists explained that these humans may have come from one of the waves of the earlier unsuccessful migrations, ended up extinct and were replaced by Neanderthals or moved back to Africa about 70,000 to 80,000 years ago (Olivieri, et al., 2006). Based on the remnants of tools discovered, archaeologists postulate that the exodus from Africa may have taken place as early as 125,000 years ago (Armitage, et al., 2011). These findings elaborate that the great migration took few failed attempts before the successful one about 60,000 years ago.

Other than fossils, genetic data generated from archaic human samples and present living individuals from different continents has shed more light into the understanding of human evolution and their migration patterns. Among the genetic materials, mitochondrial (mt) and Y-chromosomal genomes are recognized as the best candidates to be used for evolutionary studies. This is because they are passed on entirely from the

parents without being subjected to recombination that creates diversity among generations (Underhill & Kivisild, 2007). The Y-chromosome is inherited and passed down from fathers to sons along the lineage, whereas mt genome is only inherited by children from mothers (uniparental genomes). Thus, it is possible to trace and estimate the coalescent ages of all individuals in a lineage by analyzing their genomes with the help of statistical tools (Mitchell & Hammer, 1996).

2.2.2 Multiregional model

The multiregional hypothesis is the major competing model against the single origin hypothesis. This hypothesis puts forward the thought that modern humans appeared about 2 Ma and the subsequent evolution events happened continuously within a single species that includes earlier archaic humans and later modern humans (Wolpoff, et al., 1988). The hypothesis proposes that the evolution took place in all human populations in every region worldwide (Wolpoff, Hawks, & Caspari, 2000).

The idea of multiregional evolution of modern humans was first suggested by a German anatomist cum anthropologist – Franz Weidenreich, who tried to answer the similarities that he observed in archaic human fossils and modern humans. He also suggested that genes that are generally adaptive among a species (such as intelligence and communication) flow across regions quickly and are shared by all species (Washburn, 1964). Conversely, genes that emerged from local adaptation would not flow across regions and only shared by species within that particular region.

The term “multiregional” was coined by Milford Wolpoff, replacing the word “polycentric” initially used by Franz Weidenreich (Wolpoff, et al., 2000). Milford Wolpoff emphasizes that the dissimilarities in modern humans in different regions, such as Africa, Asia, Europe, and Australia, are resulted from a phenomenon known as

regional continuity – the appearance of common traits within a geographical region for a long period of time (Wolpoff, et al., 2000). He also refused to adopt the concept of parallel evolution into the model and alternatively suggested that clinal variation plays an essential role in shaping the evolution of modern humans. Wolpoff postulated that the evolution of *H. erectus* in all regions of the world did not occur independently (Wolpoff, et al., 2000). They adapted differently to local conditions and some isolated populations may have evolved in other directions that are so distinct to their counterparts in other regions. But through continuous interbreeding, replacement, genetic drift and selection, the gene flow among these populations maintains the direction of evolution as a whole, where all species in different regions evolve towards a general trend. Also with the gene flow, traits that are advantageous to all species would be spread and taken, while keeping regional adaptive genes to the local species. By that, it explains the presence of both similarities and dissimilarities among archaic and modern humans in all regions. The hypothesis has seen evidence in pieces of fossils in various regions, i.e., Southeast Asia (SEA), China, Europe, etc. (Duarte, et al., 1999; Shang, Tong, Zhang, Chen, & Trinkaus, 2007; Trinkaus, et al., 2003). Researchers have also found evidence in genetic materials that favors the multiregional hypothesis. The examination of European Neanderthal's DNA revealed that they shared more genetic variants, 1 % to 4 %, with living non-Africans than living Africans (Green, et al., 2010). In addition, Denisova hominin – a non-Neanderthal archaic human in southern Siberia, was also found to share 4 % to 6 % of its genome with living Melanesians, but not with any other living populations (Reich, et al., 2010). These findings illustrate the events of genetic flow among human species in different regions and interbreeding of archaic and modern humans, which is denied by the replacement model in the “out of Africa” hypothesis.

2.3 The land below the wind - Sabah

Sabah, one of the 13 states of Malaysia, is located at the northern part of the Borneo Island (Figure 2.4). It is often known as “the land under the wind” due to its strategic location situated right beneath the typhoon-region near Philippines. It is also the second largest state in Malaysia after Sarawak, and consists of 5 divisions that are further divided into 25 districts.



Figure 2.4 : Geographical location of Sabah state in East Malaysia, situated at the northern tip of the island of Borneo (Source: edited from Google map).

Sabah is also known for its ethnic pluralism. The population of Sabah comprises of 32 ethnic groups, in which 28 are indigenous. The people of Sabah speak more than 80 different languages and dialects. The indigenous people make up more than 60 % of the local population, besides the bumiputras, Chinese, Indians, etc. The largest indigenous group in Sabah is the Kadazan-Dusun (17.8 %), and followed by the Bajau (13.4 %) and the Murut (3.3 %). Other indigenous groups include the Kwijau, Illanun, Lotud, Rungus, Tambanuo, Dumpas, Mangka'ak, Suluk, Ilocano, Orang Sungai, etc. (Boutin & Boutin, 1985).

2.3.1 Kadazan-Dusun

The Kadazan-Dusun is the collective name given to the Kadazan and Dusun tribes, who are sharing both cultures and languages, in view to achieve unity between these tribes. The unification is resulted from the 5th Kadazan Cultural Association (KCA) Conference held in year 1989 (Puyok & Bagang, 2011). This move is to end the identity crisis between Kadazan and Dusun tribes that had started since early of 1960s that brought destructive impacts on local community. The main difference between Kadazan and Dusun is their traditional geographical distribution. Kadazans are dwellers of flat lands, cultivating paddy farming while Dusuns are inhabitants of hilly and mountainous regions (Mansur, Kogid, & Madais, 2010).

The word “Kadazan” means “the people of the land”. It was coined by the Dusun people who left their village to live in relatively more modernized area. “Kadazan” can be translated as “people who live in a more modern area and not in remote places like the Dusuns” and was then used to differentiate the modern Dusuns from those living in remote villages. The culture of Kadazans is so much influenced by activity they do for living – paddy farming (Mansur, et al., 2010).

The word “Dusun” is thought to have originated from a Malay term means “orchard” and the the Dusuns were regarded as the people living in the orchard (Ulluwishewa, Roskrige, Harmsworth, & Antaran, 2008). This may be because the Dusuns live in houses surrounded by various fruit trees. Dusun is a huge tribe with more than 30 sub-tribes, all speaking different dialects of Dusunic family language. The Dusuns usually settle at hilly area and valleys for agricultural activities. They also trade with coastal dwellers by carrying their harvest and forest products in exchange for salt, fish and other coastal products.

2.3.2 Rungus

The Rungus people can be found mainly in the northern region of Sabah, close to the Kudat district. Despite being a branch-off from the Kadazan-Dusun tribe, the Rungus have very distinctive features in their customs, cultures, dress, and language (Puyok & Bagang, 2011). Traditionally, rice cultivation and fruit planting have been the main income generator in the Rungus villages. However, due to the urge from urbanization, the younger generations have moved and are working in towns.

2.3.3 Bajau

The “Bajau” is a collective term for several indigenous tribes who are very much associated among each other. They are now the second largest indigenous group in Sabah, making up a total of 13.4 % of the local population. Bajau people, unlike the Kadazan-Dusun and Rungus, are a sea-oriented group, who live via a sea-faring manner (Schwerdtner Manez & Ferse, 2010). They fish on a small wooden vessels, called “perahu”, and live with their families in a larger boat, known as “lepa-lepa”. All these have allowed them to adapt well to life on the sea. They also engage in sea-trading activity, like trading their sea products to inland inhabitants. In 1950s, the Bajau people started to migrate to neighbouring regions, such as Sabah, Sarawak, Sulawesi, and Kalimantan, due to socio-economical pressure generated from conflicts and discrimination in southwestern Philippines (Nimmo, 1968). The origin of the Bajau, however, is not clearly known. It was believed that the ancestors of the Bajau people may be refugees who migrated from Johor to evade the attack of the Bugis during the Sultanate of Johor.

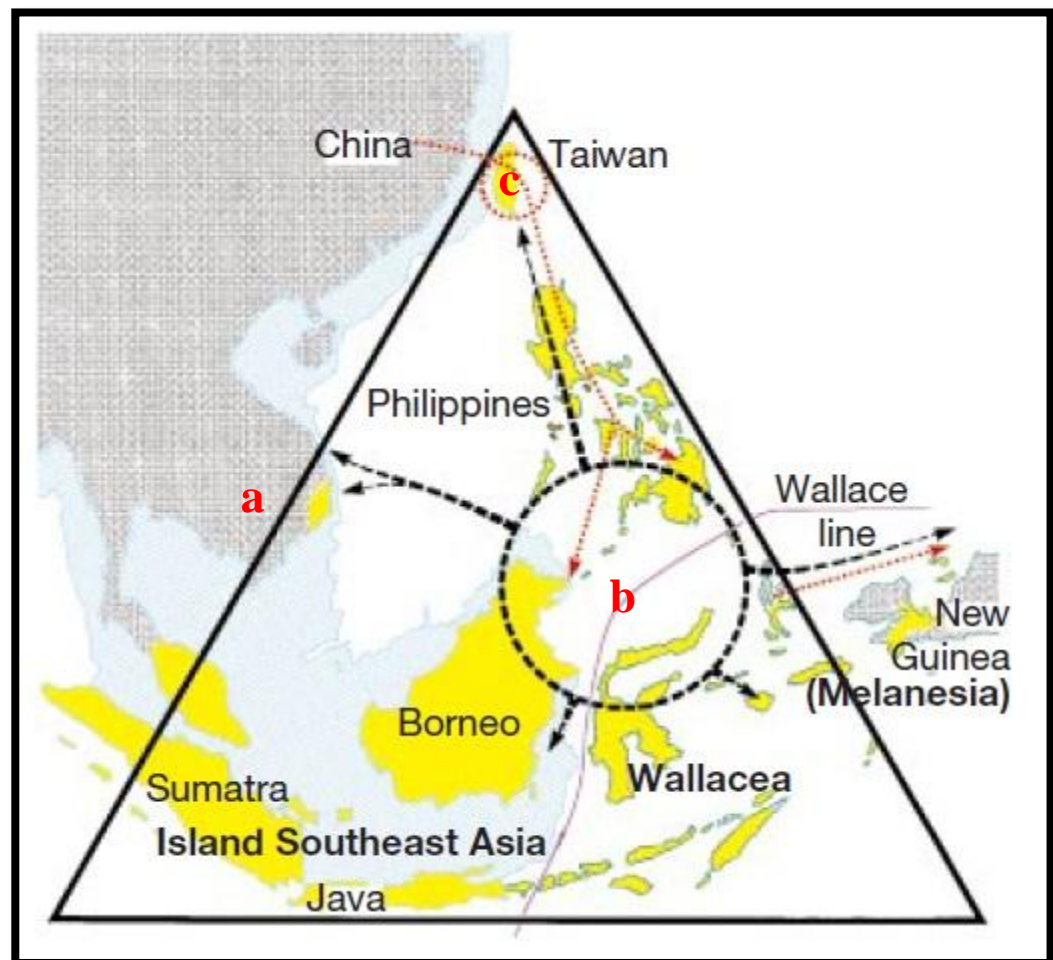
2.4 Migration patterns within Southeast Asia

Ever since the migration out of Africa, AMH started the colonization of various parts of the world with all means at different pace, primarily by walking. It has been generally accepted by researchers that Island SEA (ISEA) was first colonized by modern humans approximately 50,000 years ago (Detroit, et al., 2004). These people spread across ISEA and finally reached Australia and New Guinea. Thus, they are known as “Australo-Melanesian”. However, this initial wave of colonization was not able to explain the complexity of genetic diversity observed in the populations of present inhabitants in ISEA. The speculative events of prehistoric migration within ISEA in the near 50,000 years have been a heated topic among anthropologists (Demeter, et al., 2012; Reich, et al., 2011; Trejaut, Yen, Loo, & Lin, 2011) (Figure 2.5). And yet, there is no definite elucidation on the matter.

Through extensive studies, scholars have proposed a model regarding the settlement of ISEA – known as the “out of Taiwan” model (Bellwood, 2007). It was suggested that the settlement of ISEA occurred in 2 tiers of dispersal - the first dispersal was marked by the arrival of “Australo-Melanesian” people about 50,000 years ago. These people were then replaced by the Austronesians, who came to ISEA by sea from southern China through Taiwan, about 4,000 years back. The migration to ISEA was believed to be resulted from population expansion due to over-development of rice cultivation. This urged the rice farmers to continue seeking for new settlement elsewhere to accommodate the rapidly growing community, which in turn led to their dispersals into ISEA regions. The agricultural technology that was developed and forwarded within Austronesian community had been beneficial to them over the hunter-gathering Australoid. As a result, the Australoid foragers in the ISEA regions were slowly replaced or assimilated by the Austronesians. However, the spread did not get any further into Australia and New Guinea, most probably due to the geographical

constraints of ocean, leaving the Austroloid occupants in these regions undisturbed (Turner, 1987).

Along with the arrival of the Austronesians, they brought in the Austronesian languages - Melayo-Polynesian (Bellwood, 2007). They are now the mother-tongue of the inhabitants of these regions. The other 9 branches of the Austronesian languages are spoken by Taiwan Aboriginal populations (Blust, 1999). Hence, it was assumed that the Austronesian languages were developed in Taiwan. Then, the Melayo-Polynesian and its branches were being distributed during the second wave of colonization. Based on the “out of Taiwan” model, scientists also propose few variants that attempt to shape the model to fit various observations seen in the genetic pool of ISEA inhabitants. Amongst, the “express train to Polynesia” hypothesis suggests the movement of population from southern China into Taiwan about 6,000 years ago (Diamond, 1988). At 4,000 BP (years before present), these settlers expanded into Philippines and Indonesia, followed by Pacific region. This north to south migration was thought to occur quickly with minimum level of admixture with the existing populations along the way. In contrast, proponents of the “slow boat” hypothesis postulated that although the migration happened in the same direction, but with relatively slower speed which allows the migrants to have higher degree of assimilation with the local populations (Oppenheimer & Richards, 2001b).



a) Dark Triangle – Austronesia proposed by Meacham, where within lies the unidentified reconstructable homeland of ISEA origin.

b) Dark dashed circle – Solheim’s insular SEA model, the origin of ISEA arose in the centre of the circle near the Wallace’s line and spread in radical direction into neighboring regions, including Taiwan.

c) Red dashed line – Bellwood’s “out of Taiwan” model, colonization of ISEA began following the Neolithic expansion of rice farmers in north to south manner from southern China, across Taiwan, into Philippines, Borneo, and ISEA.

(Oppenheimer & Richards, 2001a)

Figure 2.5 : Diagrammatic representation of the proposed migration patterns as suggested by their respective proponents.

Another much debated issue regarding the migration within ISEA is the motive(s) that fuel the prehistoric migration. Many believed that the Neolithic migration was triggered by the expansion of farming activities by agriculturists from southern China (Bellwood, 2004). On the other hand, some researchers also suggest that the migration may not or not only due to the expansion of rice agriculture, but climate change in the region (Soares, et al., 2008). The continuous and rapid rising of sea levels had flooded almost half the Sunda landmass and may have caused disastrous effect on the population living along the coastlines. These populations may have then forced to refuge and seeking for better ground for living. These have then driven them to spread across Taiwan and subsequently ISEA. Similar dispersals have also been detected in western Europe during the last Ice Age (Gamble, Davies, Pettitt, & Richards, 2004). The increased glaciations forced the hunter-gathering groups in Europe to move to southern region.

The “out of Taiwan” hypothesis and its variants have been justified primarily based on linguistic approaches. Although the Austronesian languages are more diversified in the Taiwanese Aboriginal groups, 9 out of 10 subfamilies were spoken exclusively in Taiwan and only the Melayo-Polynesian branch was spread in ISEA and Polynesia. But it may not be correct to interpret that the Melayo-Polynesian branch must have derived from the Taiwanese Aboriginal groups. The root could be assigned in either Taiwan or ISEA. From there, researchers hypothesized that there might be an unidentified, ancient, and Asian-origin ancestry, in which the colonists emerged. This hypothesis is known as the “insular SEA” (Meacham, 1988). It suggests that the migration began in the centre of ISEA and spread to neighbouring regions, including Pacific, Philippines, Indonesia, Borneo Island, and Taiwan. The migration occurred in a radial form somewhere along the Wallace’s line – a line that separated Asia and Australia.

In 1980s, Meacham noticed that all branches of Austronesian languages were found encapsulated in a broad triangle, bordered by Taiwan, Sumatra, and Timor. He,

therefore, proposed to call these regions as “Austronesia”, where the Austronesian arose from. Meacham also mentioned that it would be strange to assume that Austronesian originated from southern China as there is no trace of Austronesian languages or its derivatives being spoken in regions other than Austronesia, including the proposed homeland. Meacham and colleagues also brought up the question about the migration direction; why would the migrants opted to cross the sea to ISEA, instead of travelling along the coastlines to Vietnam and Thailand that are accessible by walking (Meacham, 1988).

With the advancement of genetic testing technologies and prominent choices of markers available, these arguments can be tested and verified by the data obtained from present inhabitants and archaic human remains in these regions. Various studies have been conducted to answer and testify the models proposed regarding the peopling of ISEA, using both recombinant (autosomal loci) and non-recombinant markers (mitochondrial DNA and Y-chromosome) (Hurles, Sykes, Jobling, & Forster, 2005; Simonson, et al., 2011; Xu, Pugach, Stoneking, Kayser, & Jin, 2012). The outcomes of these studies remain inconclusive due to contradicting results and inappropriate data interpretation. Nevertheless, these limitations can be solved. First, more markers (especially non-recombinant markers) should be examined and included to construct phylogenetic networks with better resolution, thus providing more accurate representation of tested populations on each branches. Genetic data should be generated from individuals with pure lineage and the typing should cover as many groups as possible in the region.

2.5 The human genome

The entire human genome is approximately 3.3 billion bp in length and can be divided into 2 categories based on the locations where DNA resides in the cell, i.e., nuclear DNA and mitochondrial DNA (mtDNA) (Venter, et al., 2001).

2.5.1 Nuclear DNA

As the name implies, nuclear DNA is enclosed in the command centre of the cell, i.e., the nucleus. The nucleus, a membrane-bound organelle, coordinates all activities in the cell by mediating the gene expression levels. DNA presents as a long linear molecule and is tightly packed in the nucleus. It also interacts with protein structures, such as histone, in the form of chromosome. The entire nuclear genome of human consists of 22 pairs of autosomal chromosomes and a pair of sex chromosomes (Venter, et al., 2001). There are about 20,000 to 25,000 genes, which is roughly 1.5 % of the full genome size (Venter, et al., 2001). The remaining sequences consist of non-protein coding regions present as RNA genes, regulatory sequences, introns, transposons, and pseudogenes. These sequences do not give rise to functional protein production, but may play a crucial role in regulation of gene expression, repair and maintenance of DNA, markers of genetic diversity, and evolutionary artifacts (Wahls, Wallace, & Moore, 1990; Zuckerkandl & Cavalli, 2007). Nuclear DNA is inherited from both parents, where 23 segregated chromosomes are received from both father and mother individually.

2.5.2 Mitochondrial DNA

The mitochondrion has a distinctive genome from its counterpart – nuclear DNA. Mt genome is operating independently without restricted by the nuclear activities. It regulates its own gene expression, replication, and division. The entire mt genome is about 200,000 times smaller than the nuclear genome, i.e., only 16,569 bp (Anderson, et al., 1981). However, it has higher copy number in each cell, owing to the fact that some cells may have over 1,000 copies of mtDNA but only 1 copy of nuclear DNA. Mt genome presents as a circular and double stranded (ds) DNA, where both strands are notably different in the nucleotide composition (Anderson, et al., 1981). The strand with higher GC-residues is regarded as “Heavy (H) strand” and the opposite strand is known as “Light (L) strand” (Kasamatsu & Vinograd, 1974). Together, both strands code for a total of 37 gene products – 22 tRNA, 2 rRNA, and 13 protein structures (Figure 2.6). The H-strand codes for 28 genes (14 tRNA, 2 rRNA, and 12 proteins), whereas the L-strand transcripts for 9 (8 tRNA and 1 protein). The genetic information in the coding region of mtDNA is efficiently packed, where there is none or little non-coding sequence between each gene (Taanman, 1999). In the entire coding region of 15,447 bp, only 55 nucleotides (3.6 %) did not involve in gene transcription. In addition, gene overlapping is also observed in the coding region of mtDNA. There are 45 bases being shared for the production of ATPase synthase subunits 6 and 8 at position 8,857 to 8,572 (Anderson, et al., 1981).

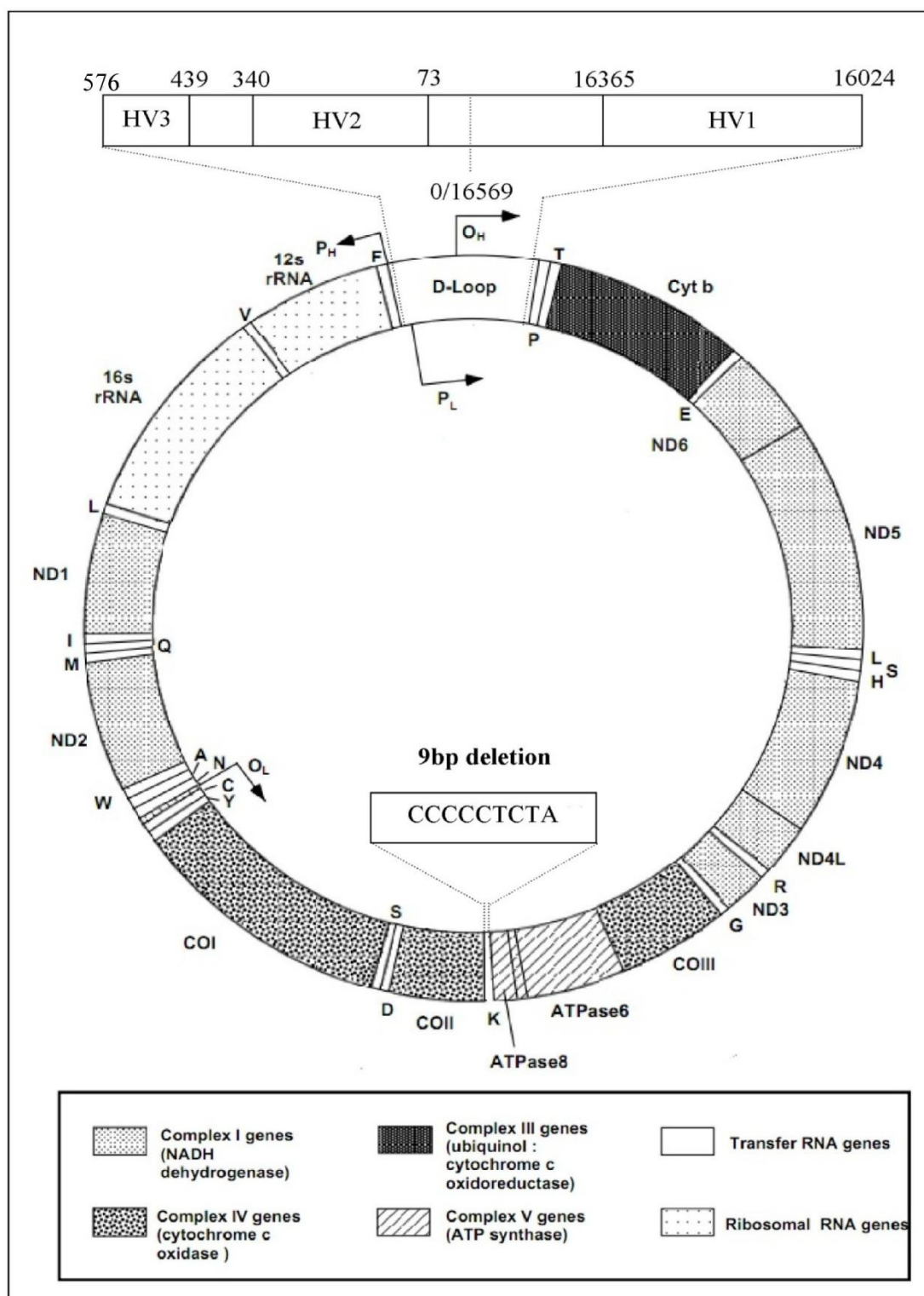


Figure 2.6 : Structure of mtDNA; including locations of 37 gene products, HV regions within the D-loop, and the 9-bp deletion (adapted and modified from MITOMAP).

Unlike the nuclear DNA, mtDNA is inherited maternally. The intact mt genome is passed on from mothers to sons and daughters without genetic recombination. Despite the fact that mitochondria present abundantly in the tail of sperm cells, the tail is restricted to enter the ovum by the selective membrane, allowing only the head to pass through for fertilization (Sutovsky, Navara, & Schatten, 1996). Furthermore, scientists have also found that the paternal mitochondria would be destroyed inside the egg (Nishimura, et al., 2006). Thus, the zygote only contains mitochondria exclusively from the ovum (mother). This phenomenon has made mtDNA extremely useful in studying maternal lineages in human populations.

2.5.3 Genetic polymorphisms

Scientific studies have revealed that all humans have nearly identical genetic composition, i.e., 99.9 % (Tchinda & Lee, 2006). The variations present between individuals are termed as genetic polymorphisms. These variations of genetic content give rise to the differences between every human being, be it physical or physiological. Some of these genetic variations contribute to the outlook of an individual, such as skin and pupil colors, wavy or straight hair, tall or short stature, etc. Meanwhile, others affect underlying biological activities and responses that cannot be visualized by naked eye. Studies have shown that the degree of disease susceptibility of an individual could be influenced by the patterns of genetic polymorphisms (Karwautz, et al., 2008; Settin, Abdel-Hady, El-Baz, & Saber, 2007). Examples of a few major types of genetic polymorphisms are, variable number of tandem repeats (VNTRs), tetranucleotide repeats (microsatellites), insertions/deletions (InDel), and single nucleotide polymorphisms (SNPs).

2.5.3.1 Variable number of tandem repeats

VNTR is another type of genetic polymorphism characterized by the repetition of a stretch of sequence, where the repeating blocks are located next to each other (Inglehearn & Cooke, 1990). The repeats are generated or reduced by errors that occur during recombination and replication. The repeating units can occur up to more than 100 times in certain loci. Therefore, it results in length variations among individuals. The alleles are also inheritable through simple Mendelian pattern. Thus, it can be used effectively in individual genetic identity matching, as well as paternity testing for inheritance examination.

VNTRs were the first markers being used in DNA profiling by an English geneticist, Dr. Alec Jeffreys, more than 20 years ago (Jeffreys, Wilson, & Thein, 1985). He developed a technique which is employed to examine these length variations present in the DNA sample, enabling him to perform the very first human identity test. These repeating regions were then coined as VNTR, and the technique employed by Dr. Jeffreys was named 'Restricted Fragment of Length Polymorphism' or RFLP. This early form of DNA profiling was helpful in solving English immigration and local homicide cases (Butler, 2005). Since then, the development of DNA profiling has been rapidly advancing and the technology is widespread (Foster & Laurin, 2012; Gill, et al., 1994; Liu, Scherer, Greenspoon, Chiesl, & Mathies, 2011).

According to size of the repeating unit, VNTRs are differentiated into 2 distinct categories, i.e., minisatellites and microsatellites. The minisatellite is a repeating sequence, ranging from 10 bp to 100 bp, and occurs in more than 1,000 locations in the entire genome. The microsatellite on the other hand, involves tandem repetition of core sequences that are relatively smaller than the minisatellites, which are generally 2 to 5 bases. In this study, 2 minisatellites that are present in the 3' untranslated region (3'UTR) and Intron-8 of the dopamine transporter (DAT-1) gene were studied.

2.5.3.1(a) Dopamine transporter gene

Dopamine is an important neurotransmitter in the brain and autonomic nervous system. The dopamine transporter removes dopamine that is released into the synaptic cleft, in order to prevent continuous stimulation of the post-synaptic neuron. The dopamine is then transported back to the synaptic knob of the neuron. Dopamine transporter is coded by DAT-1 gene and is mapped to the chromosome 5p15.3 (Giros, et al., 1992). The entire DAT-1 gene contains 15 exons that span approximately 60 kb (Vandenbergh,

et al., 1992). A 40-bp VNTR at the 3'UTR was first described by Vandenberg and his colleagues, with repeat copy numbers ranging from 3 to 11. The 3'UTR DAT-1 VNTR was reported to be polymorphic (more than 7 observed alleles) in certain global populations, such as African and Omania (Santovito, et al., 2008; Simsek, Al-Sharbati, Al-Adawi, Ganguly, & Lawatia, 2005). However, it has limited polymorphisms in other populations, e.g., Cambodian and Italian (Kang, Palmatier, & Kidd, 1999; Persico, Bird, Gabbay, & Uhl, 1996). It is the most commonly studied VNTR in the DAT-1 gene. Majority of the studies were associated to the role of the polymorphisms in neurological pathways and disorders (Le Couteur, Leighton, McCann, & Pond, 1997; Mignini, et al., 2012; Persico & Macciardi, 1997)

Apart from that, scientists also found another VNTR in Intron-8 region of the DAT-1 gene. It is characterized by the presence of a 30-bp repeating sequence. Like the 3'UTR VNTR, the Intron-8 VNTR harbours variable alleles that repeat up to 13 times in different human populations. As part of the DAT-1 gene sequence, the Intron-8 VNTR has been widely studied to access its relationship to various neurological disorders, such as Attention Deficit Hyperactivity Disorder and migraine (Elia, et al., 2012; McCallum, et al., 2007). In addition, neuroscientists were also interested to depict the link of the Intron-8 VNTR to nicotine and cocaine dependence, as the polymorphisms might play a key role in the recycling process of dopamine (Guindalini, et al., 2006; O'Gara, et al., 2007). Although it has not been regularly used as a marker for population study, the Intron-8 VNTR could be informative as it has varied distribution in human populations. Figure 2.7 shows the structure of the 2 VNTRs in 3'UTR and Intron-8 of the DAT-1 gene.

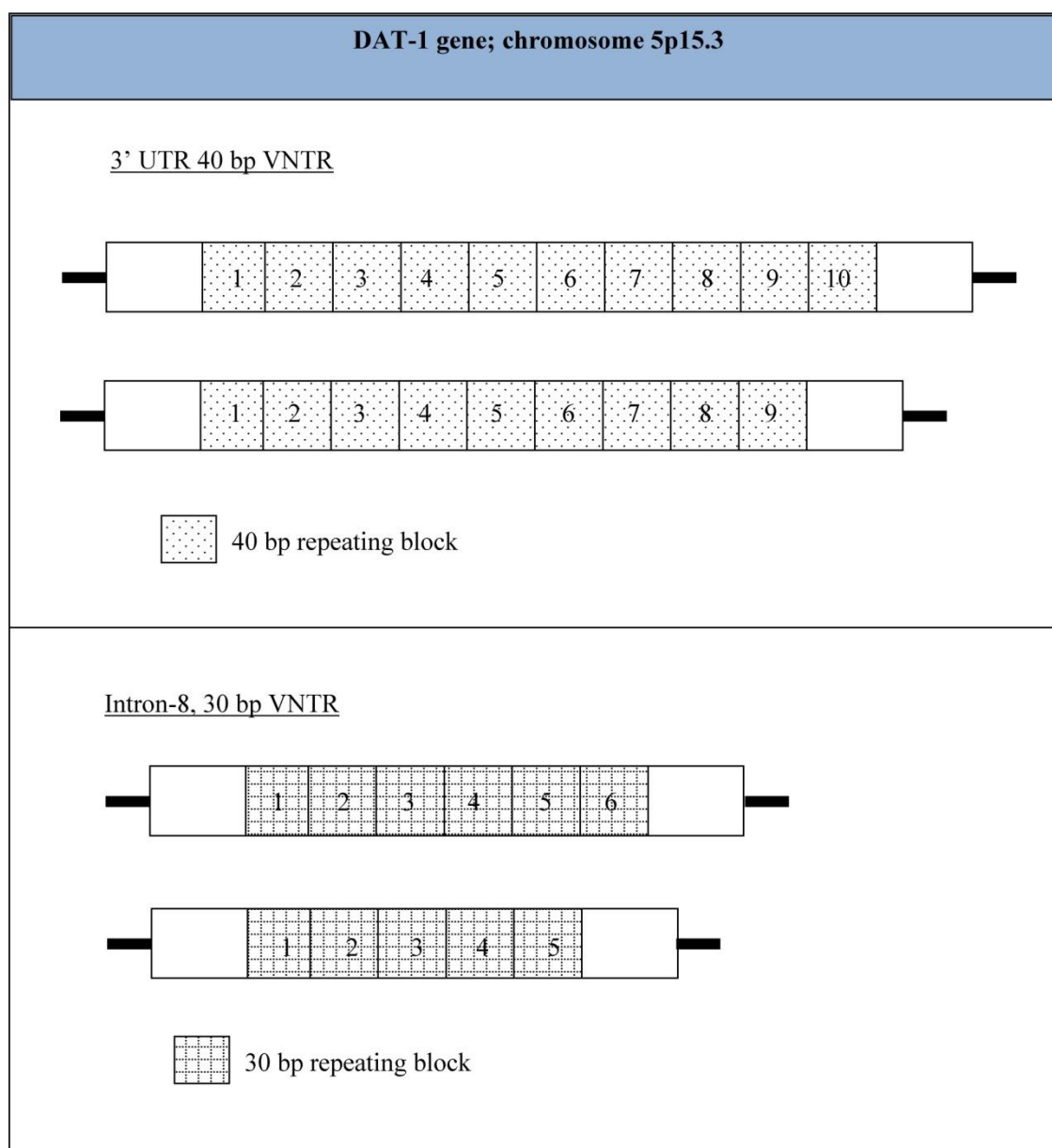
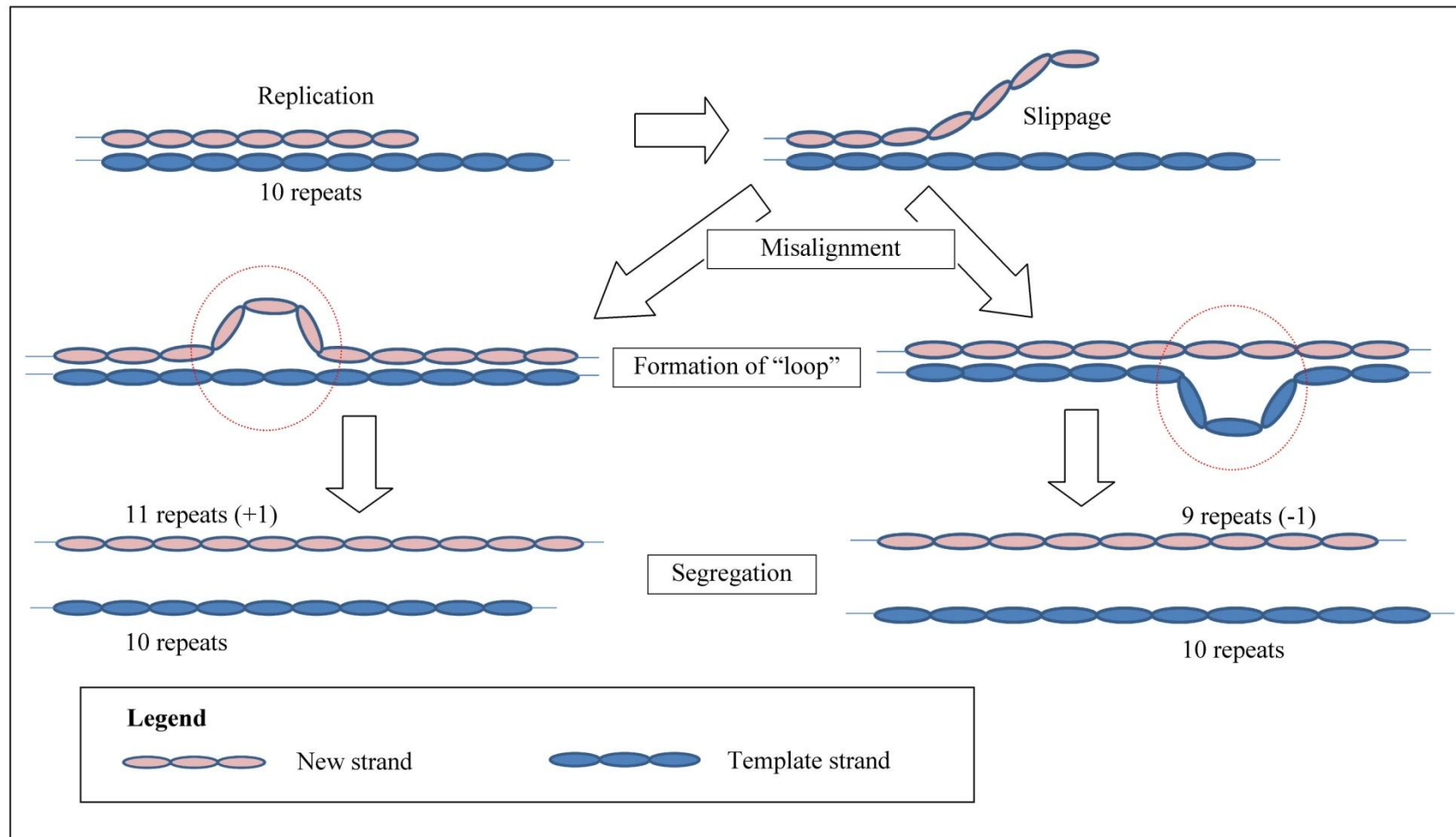


Figure 2.7 : Diagrammatic illustration of polymorphic variants of VNTRs in the 3'UTR and Intron-8 of DAT-1 gene; the presence of different number of repeating blocks within the DNA sequences gives rise to size variation within the region.

2.5.3.2 Microsatellite

Microsatellites, such as short tandem repeats (STRs), are another type of repetitive sequences found in the human genome. The characteristics of microsatellites are similar to that of minisatellites, except that microsatellites consist of repeating units of 1 to 4 bp in length. STRs occur when 1 or more nucleotides are repeated (Brown, 2010). These repeating units are located adjacent to each other. The repeating sequences can be repeated up to 100 times, thus generating high discriminating power and sensitivity (Pepinski, et al., 2001). Therefore, STRs are used as molecular markers in studies related to kinship, forensic science, and human populations (Deng, et al., 2006; Edelmann, Deichsel, Hering, Plate, & Szibor, 2002; Hering, et al., 2006). STR markers have been found scattered throughout the human genome, at an average of every 10,000 nucleotides. Microsatellites account for up to 3 % of the entire genome and the number of microsatellite loci is expected to be more than 1 million (Venter, et al., 2001). The generation of varying alleles is caused by slippage during the replication (Figure 2.8). Misalignment of the replicated and template strands leads to formation of loops which result in the production of new strands with +1 and -1 repeat, respectively (Levinson & Gutman, 1987).

Out of the vast number of STR loci available, certain loci are being selected and used as core loci in the forensic DNA laboratories, e.g., the Combined DNA Index System (CODIS), examines 13 STR loci and Amelogenin (AMEL) to determine sex, and is used in the United States. The selection of core loci is based on several factors, including high discriminating power, low stutter characteristics, low mutation rates, abilities to undergo multiplex PCR, separate chromosomal locations, and small predicted length of alleles (Butler, 2005). In order to promote improvement and standardization of discipline practice in DNA laboratories, several working groups were also organized to create standard guidelines for DNA testing (Gill, et al., 2008).



(Levinson & Gutman, 1987)

Figure 2.8 : Schematic representation of the formation of +1 and -1 repeat alleles by replication slippage.

Currently, commercial kits are also available from private research companies, such as Powerplex systems from Promega Corporation and AmpF ℓ STR[®] kits from Life Technologies. Typing with such kits is quick and convenient as all loci, 16 or more, are incorporated into a single reaction. First, extracted genomic DNA from the tested sample is added into the PCR reaction mixture, where all primers and fluorescence dyes are included, and allowed to amplify in a thermal cycler. The amplified product is then subjected to fragment analysis experiment in a genetic analyzer. Results are ready within a day. Therefore, STR typing has been prioritized as the test of choice for human identification in forensic cases. In the present study, the Promega Powerplex 16 system for the typing of multiple STR markers was employed. A total of 15 STR loci, together with the AMEL, were typed simultaneously.

2.5.3.3 InDel markers

Alongside with the completion of full human genome sequencing, scientists have unearthed various forms of natural genetic variations in humans. Unlike the SNPs, InDels markers however, received little attention in discoveries of potential genetic markers in different research aspects. In year 2001, researchers examined the distribution of InDel polymorphisms on chromosome 22 and estimated that these polymorphisms made up 18 % of the total variations in the chromosome (Dawson, et al., 2001). It was suggested that the entire human genome contains 1.6 to 2.5 million InDel polymorphisms, where approximately 21 % of the genetic variations in the human genome were InDels (Weber, et al., 2002). Most, 96 %, of these observed InDel variations consist of fragments ranging from 2 to 16 bp (Mullaney, Mills, Pittard, & Devine, 2010).

In year 2006, an initial mapping of InDel variations in the human genome was carried out via a systematic computational approach. The outcome revealed the discovery of 415,436 novel InDel polymorphisms in the human genomes (Mills, et al., 2006). The distribution of insertions and deletions was somehow quite even, with 47 % were insertions and 53 % were deletions, and the average density was 1 InDel in every 7.2 kb of DNA (Mills, et al., 2006). The data also showed that a total of 36 % of these polymorphisms present in the gene areas. It implies that InDel markers could have direct and significant impact on gene's products and subsequently leading to the development of diseases. For example, a 3-bp deletion in the CFTR gene causes Cystic Fibrosis by direct elimination of an amino acid in the protein chain (Collins, et al., 1987). On the other hand, InDels were also evaluated for their use in evolutionary and natural population assessment (Vali, Brandstrom, Johansson, & Ellegren, 2008). Recently, extensive screening of InDel variations has been established by incorporating with cutting-edge technology, where a set of 10,003 probe-based assays were developed

on microarray platform to detect small InDels in the human genomes (Mills, et al., 2011). With such large scale screening of InDel markers, the natural characteristics which include the diverse distribution of these markers can be determined and applied with confidence.

Although genetic research on InDel polymorphisms was not as established as other markers, such as SNPs and microsatellites, InDels have several advantages over the rest. Beside the considerably profuse amount of InDels scattered in the human genome, the presence of each InDel was regarded as the identity-of-descent (Shedlock & Okada, 2000). This is due to the event of every insertion or deletion at a specific location on the genome is completely independent and rarely reversed. In addition, InDel markers can be typed with ease by simple polymerase amplification and size separation, without the need of costly and sophisticated machineries. With these, InDels have gradually gained popularity among scientists, especially for those in the fields of population, evolution, and forensic investigation.

2.5.3.3(a) *Alu* insertions

The *Alu* element, 1 of the most prominent examples of repetitive DNA, is a member of short interspersed element (SINE), which is estimated to exist in more than 10 % of the entire human genome (Houck, Rinehart, & Schmid, 1979; Smith, 2005). It is named after the restriction endonuclease *AluI*. Each *Alu* element is about 300 nucleotides in length and believed to be derived from 7SL RNA (Figure 2.9) (Wiesner, Ruegg, & Morano, 1992). The *Alu* element is mobilized within the genome by a gene jumping mechanism known as retroposition, i.e., a RNA-mediated transcription process. This mechanism has contributed to the random yet wide distribution of *Alu* elements, with varied density, throughout the genome. *Alu* elements however, do not code for proteins

and depend solely on the long interspersed elements (LINEs) for replication (Kramerov & Vassetzky, 2005).

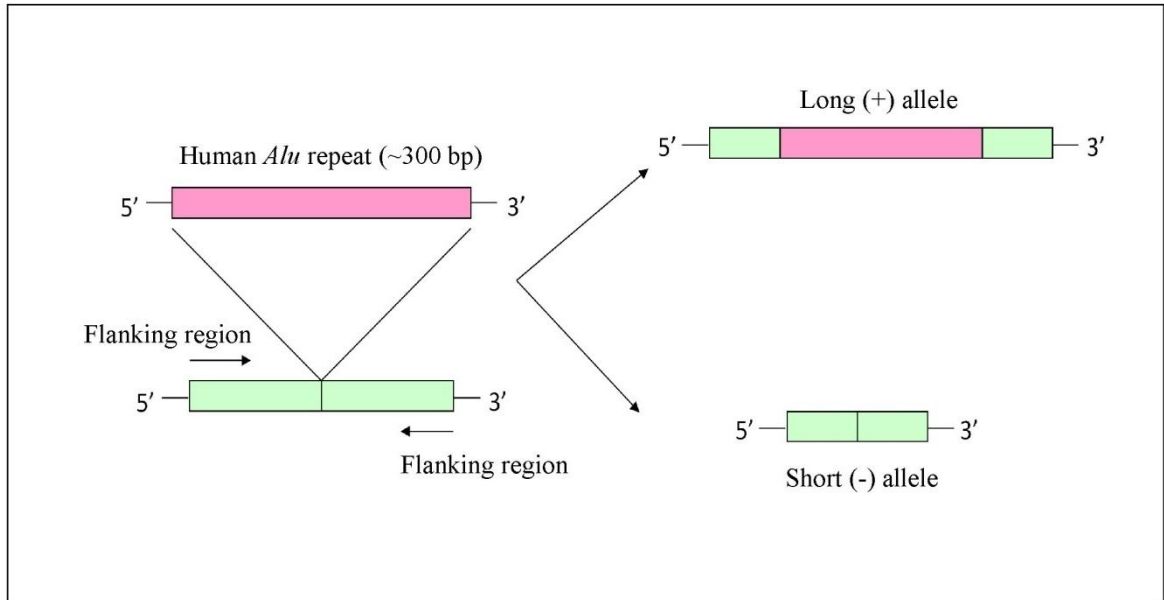


Figure 2.9 : Diagram showing the insertion of an *Alu* element into a genome sequence, resulting in the elongation of the particular stretch of sequence (“+” allele) than the original sequence without insertion (“-“ allele).

The insertion of *Alu* elements began as early as 65 Ma (Shen, Batzer, & Deininger, 1991). The current amplification rate of *Alu* insertions is found to be much lower than the past, which is merely 1 % of the previous peak rates, at 1 new insertion in 200 newborns (Deininger & Batzer, 1999). There are more than 1 million accumulated copies of *Alu* elements being reported in the human genome (Lander, et al., 2001). Most *Alu* insertions are constant in all human beings, regardless of the origin of population. However, a small number of them (~5 %) are polymorphic. These insertions arise recently during global colonization by modern humans and have great potential to reveal information on the modern human expansion and migration events.

Most of the *Alu* insertions are shared by both human and primate genomes, but about 7,000 of these insertions are unique to humans (Cheng, et al., 2005).

The function of the *Alu* elements remains controversial. Initially, *Alu* elements had been regarded as junk or parasitic DNA that did not play any role in the host's regulation and maintenance (Doolittle & Sapienza, 1980). Later, researchers discovered that some of the *Alu* elements are capable of influencing normal gene expression and involved in chromosome rearrangements (Britten, 1996). Recently, studies have shown that *Alu* elements contribute to several inherited diseases like cancer and α -thalassemia (Deininger & Batzer, 1999; Flint, et al., 1996). Scientists postulated that *Alu* elements are the major contributor to the evolution process throughout primate history, including that of human (Batzer & Deininger, 2002). Massive insertions of *Alu* elements have caused genomic instability that has facilitated the process of speciation (Challem & Taylor, 1998). Of late, the study of *Alu* elements has focused now on the search of population-specific markers. Researchers have demonstrated the possibility of inferring one's origin of population using a combination of polymorphic *Alu* markers (Ray, Xing, et al., 2005). There are several advantages of typing *Alu* insertions than other markers. The typing of *Alu* insertions is simple and rapid, yet reliable, by employing ordinary PCR reaction follow by agarose gel electrophoresis (AGE). The insertion of *Alu* element into the genome sequence is regarded as a unique event, where recombination or back-mutation is rarely occurred. Every insertion has a known state of ancestry, which is null-insertion, making evolutionary studies easier by assuming that the hypothetical ancestor has no insertion at the particular loci. Examples of some useful *Alu* insertions in population study are HS4.32, TPA25, PV92, HS3.23, B65, and APO. In our study, the screening of *Alu* insertion for all the 6 loci mentioned was carried out to infer the relationship among the indigenous populations in the Sabah state.

2.5.3.3(b) Mitochondrial 9-bp deletion

In the coding region of mtDNA, a deletion which involves a loss or gain of a 9-bp motif (CCCCCTCTA) has been characterized by scientist (Jones & Kafatos, 1982). It is located at the intergenic region of Cytochrome b Oxidase subunit II (COII) and tRNA Lysine (K) or called region V (Figure 2.10). It had been recognized as an “Asian-specific” marker due to its high prevalence in Asian and Asian-derived populations (Hertzberg, Mickleson, Serjeantson, Prior, & Trent, 1989). Being the most studied length polymorphism in mtDNA, this deletion has been included in the estimation of human mtDNA haplogroup lineages, together with other mtDNA markers. It is also used to suggest population migration routes when analyzed together with other markers, such as Y-chromosome markers. Scientists have proposed that the 9-bp deletion may originate from Central China and spread to SEA, and subsequently moved on to coastal and island populations of the Pacific (Ballinger, et al., 1992).

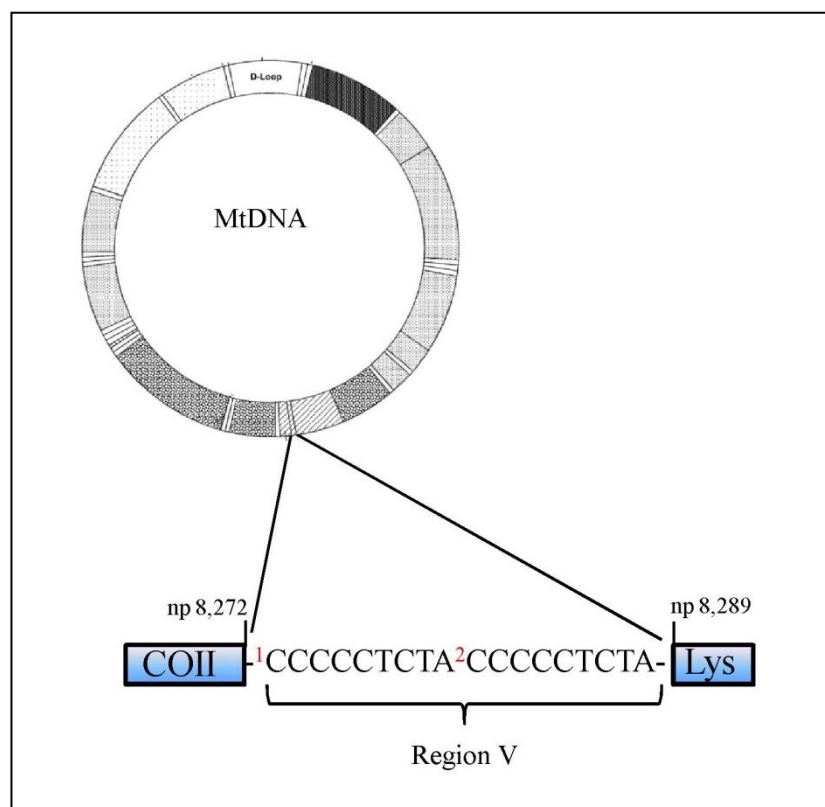


Figure 2.10 : Illustration of the position of region V and the 9-bp deletion in mtDNA.

2.5.3.4 Single nucleotide polymorphisms

A single nucleotide polymorphism (SNP) is a DNA sequence variation that occurs when a single nucleotide (e.g., A) in the genome sequence is replaced by other nucleotides (i.e., C, T, and G) and thus, is different from other members of the species. There are more than 1 million SNPs in the whole human genome, with the distribution rates of about 1 in every 1,900 nucleotides (Sachidanandam, et al., 2001). SNP can be subdivided into 2 types, i.e., transitions and transversions. A transitional SNP is characterized by the substitution of a purine (A or G) by another purine or a pyrimidine (T or C) by another pyrimidine. On the other hand, a transversion is featured by the replacement of a purine by pyrimidine or vice versa. SNPs are often reported according to their reference SNP (rs) number as assigned by the National Centre of Biotechnology Information (NCBI) (Sherry, et al., 2001). There is yet a mutual agreement on the nomenclature of SNP. However, it is always expressed with a prefix, position, and a sign of “greater than” indicates the change from the wildtype to the mutated allele. For example, c.677C > T refers to a SNP at nucleotide position-677 that involves a substitution of a Cytosine by Thymine. However, it can also be reported by the SNP’s rs number followed by the change of nucleotide, such as rs1801133C > T.

SNPs can occur in any part of the genome, be it coding or non-coding regions, and impact differently on the gene products. In the coding sequence, SNPs can lead to the production of altered proteins that have abnormal activities in the cells (Xi, Jones, & Mohrenweiser, 2004). In the non-coding regions, it can affect the regulatory mechanisms and ultimately cause irregular expression levels of particular protein structures (Lipska, et al., 2006). In both conditions, SNPs can lead to the development of diseases. However, it is hard to establish a clear and direct relationship of a SNP profile with the onset of certain disease as most diseases are multi-factorial, which could be caused by various factors other than genetic composition alone (Tiret, 2002).

Nonetheless, studies have ascertained that SNPs play a predisposing role in diseases, where individuals with particular genotypes or alleles are prone to the disease development (Scott, et al., 2007; Su, et al., 2012).

Apart from clinical scenario, SNPs are also powerful markers in forensic and evolutionary research (Morin, Luikart, & Wayne, 2004; Sobrino, Brion, & Carracedo, 2005). Despite having low discriminating power like other bi-allelic markers do, SNPs present in the mtDNA are screened in forensic cases for individual and biological parts identification when chromosomal DNA is unable to be examined, for example, in cases of highly decomposed specimens and screening of extraction DNA samples from tooth, bone, and hair (Butler, 2005). The reason being that mtDNA exists in hundreds to thousands of copies in a single cell, compared to nuclear DNA that is limited to only 1 copy per cell. In evolutionary context, SNPs in mtDNAs provide valuable clues into how modern humans had evolved and allow scientists to trace back the lineage from the maternal (Cann, Stoneking, & Wilson, 1987).

2.5.4 Polymorphisms in mtDNA

The mt control region is a non-coding section that carries the origins of transcription for both strands and replication origin for the H-strand (Taanman, 1999). The whole control region is 1,122 bp in length, spanning from nucleotide position (np) 16,024 to 576 (Anderson, et al., 1981). It is also known as “displacement loop” or “D-loop” (Figure 2.6). The mtDNA is also a hotspot for mutations, especially in the control region. The control region was found to evolve as much as 17 times faster than a single copy of gene in the nuclear DNA (Stoneking, 2000). This may occur as a result of high oxidative stress generated from the energy coupling mechanisms in the mitochondrion. The control region can be further divided into 3 hypervariable regions (HV), i.e., HV1, HV2, and HV3. These regions were reported to be highly polymorphic, harboring polymorphisms of different type, such as SNPs and InDels (Stoneking, 2000). These polymorphisms have been utilized in forensic examination and evolutionary studies (Bjork, Liu, Wertheim, Hahn, & Worobey, 2011; Gunnarsdottir, Li, Bauchet, Finstermeier, & Stoneking, 2011; Kong, et al., 2011).

2.5.4.1 Mitochondrial haplogroups and migration

The differences between mt genomes were analyzed and represented in phylogenetic tree. Haplogroups are systems employed by evolutionary scientists to represent various branches on the phylogenetic tree which helps to categorize each examined individuals according to their mt genetic content (Forster, 2004; Pakendorf & Stoneking, 2005). The mt haplogroup system allows the study of matrilineal relationship of individuals not only from the present world, but also the ability to trace the relationship back to the ancestral level (Torroni, Achilli, Macaulay, Richards, & Bandelt, 2006).

Based on the mt genome, all present living individuals are traced back to the most common recent ancestor (MCRA) lived in East Africa about 160,000 years ago. She was named the Mitochondrial Eve, whose mt DNA is the only one that has successfully passed on until present day while other parallel lineages at her time ended up to extinction (Watson, Forster, Richards, & Bandelt, 1997). Although genetic data pointed to a single ancestor, some researchers believe that the Mitochondrial Eve may most probably be a group of female dwellers in the ancient African population (Templeton, 1993).

The Mitochondrial Eve is regarded as the first element and also base of the mt phylogenetic tree – haplogroup L. The L0 is the first branch that diverged from the Mitochondrial Eve. Among all subgroups of haplogroup L, L3 played a critical role in the colonization of the world by escaping from Africa (Soares, et al., 2012). The haplogroup L3 gave rise to 2 macro-haplogroups, haplogroup M and N, which comprise of all non-African haplogroups (Figure 2.11) (Soares, et al., 2012). Both haplogroups M and N are found spanning all continents at varying frequencies. From there, modern humans went on the venture to colonize the world and various haplogroups were founded in the process.

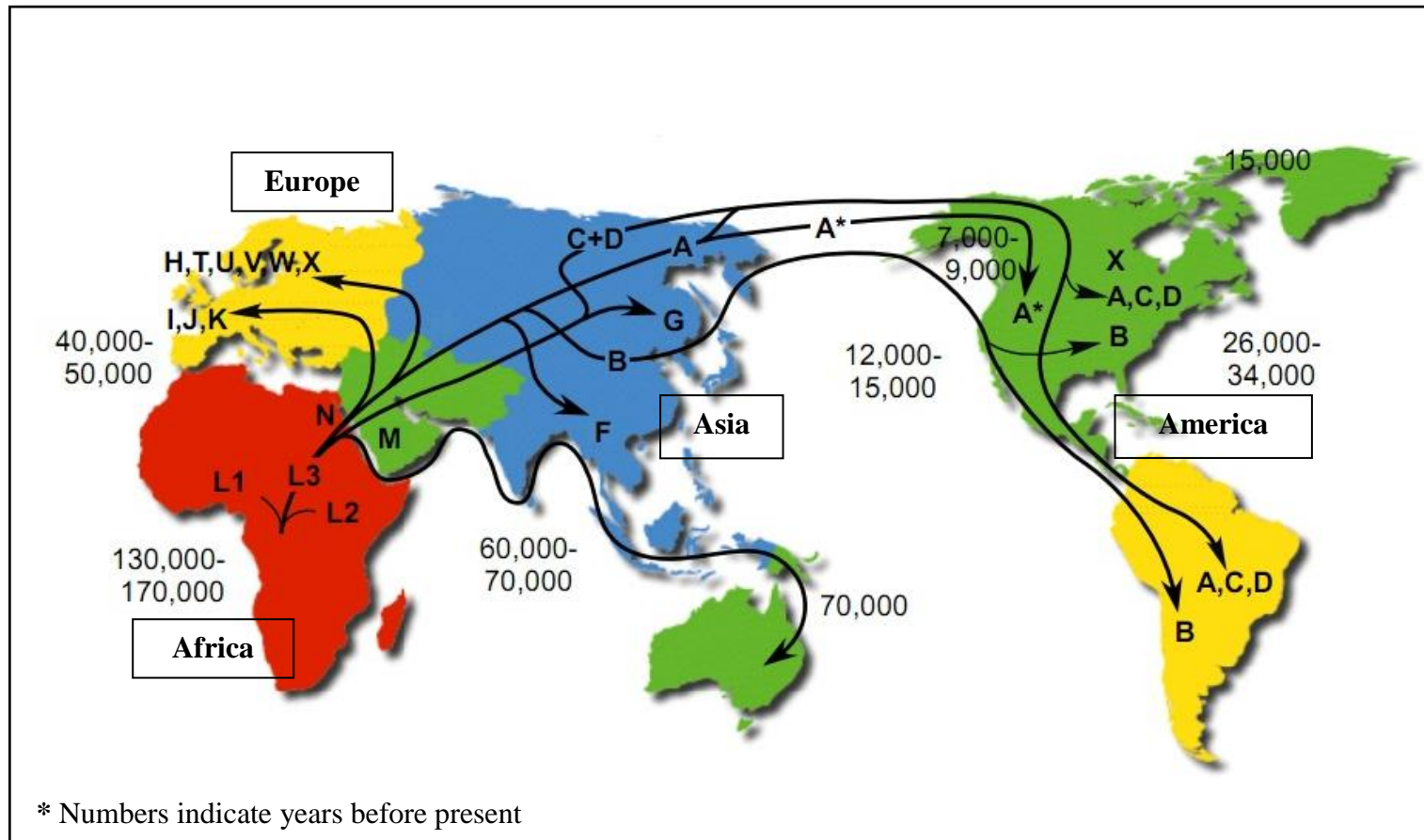


Figure 2.11 : Migration route of humans out of Africa as suggested by genetic data from mtDNA. The group of L3 settlers escaped through the horn of Africa and gradually colonized all over the world from Asia and Europe, subsequently Americas through Siberia (Adapted from MITOMAP: <http://www.mitomap.org/pub/MITOMAP/MitomapFigures/WorldMigrations.pdf>).

2.5.4.2 Nomenclature

The nomenclature of mtDNA is in accordance to recommendations compatible with the International Union of Pure and Applied Chemistry (IUPAC) codes. In year 1981, Anderson and co-researchers uncovered the first ever full mt genome sequence (Anderson, et al., 1981). Hence, this sequence has been made the reference standard in all mt genome studies. In the case of sequence differences observed between an individual and Anderson's, only the site and differing nucleotide are documented. Examples of the nomenclature are as below:

Type	Example	Nomenclature	Note
Single nucleotide change	A differing "T" at position 150 from the Anderson's "C"	150T	The designated position followed by the differing nucleotide
Insertion	An additional of "A" after position 291	291.1A	The designated position followed by a decimal point and the number of insertion and the inserted nucleotide
Deletion	A deletion of "A" at position 521	521d	The designated position followed by "d"
Heteroplasmy	Co-existence of "C" and "T" in an individual at position 16,295	16295Y	For confirmed heteroplasmy, IUCAP codes were assigned
	Suspected heteroplasmy at position 16,295	16295N	"N" is given to ambiguous heteroplasmy

CHAPTER 3

MATERIALS AND METHODS

3 MATERIALS AND METHODS

3.1 Materials

Chemicals of grades that are suitable for molecular work were purchased directly from suppliers and dealers. These chemicals were used in the preparation of the various solutions and buffers.

3.1.1 Solution, reagents, and buffer preparation

3.1.1.1 2 M NaOAc, pH 5.6

CH ₃ CooNa (MW: 82.03)	16.41 g
-----------------------------------	---------

sdH ₂ O	100 ml
--------------------	--------

NaOAc powder was weighted and dissolved in sdH₂O. Glacial acetic acid was used to adjust the solution to pH 5.6. The solution was autoclaved and kept at room temperature.

3.1.1.2 NaCl, 5 M and 6 M

NaCl (MW: 58.44)	29.22 g (5 M)
------------------	---------------

	35.06 g (6 M)
--	---------------

sdH ₂ O	Up to 100 ml
--------------------	--------------

The dedicated amount of NaCl was weighted and dissolved in sdH₂O. The solution was sterilized by autoclaving.

3.1.1.3 Ethanol, 70 % (v/v)

The 70 % (v/v) ethanol was prepared from commercially available 95 % (v/v) or absolute ethanol with appropriate dilutions.

3.1.1.4 Proteinase K buffer, 5 X strength

NaCl, 5 M	0.75 ml
EDTA, 0.5 M, pH 8.0	2.40 ml
sdH ₂ O	6.85 ml

All solutions were mixed evenly and passed through a 0.45 µm syringe filter.

The solution was then aliquoted into 1 ml working volumes and stored at -20 °C until further use.

3.1.1.5 Proteinase K solution, 10 mg/ml

Proteinase K powder	10 mg
sdH ₂ O	1 ml

The lyophilized proteinase K powder was weighted into a clean sterile eppendorf tube and dissolved in sdH₂O. The solution must be prepared fresh prior to use.

3.1.1.6 Red Cell Lysis (RCL) buffer, 10 X strength

Sucrose	548 g
Triton-X100	50 ml
MgCl ₂ H ₂ O, 1 M	25 ml
Tris-HCl, 1 M, pH 7.5	60 ml
sdH ₂ O	865 ml

All ingredients were mixed throughoutly and stored at 4 °C. The buffer was diluted to 1 X working strength when use.

3.1.1.7 Sodium Dodecyl Sulfate (SDS) solution, 20 % (w/v)

SDS powder	20 g
sdH ₂ O	100 ml

The SDS powder was dissolved into sdH₂O in a heated (45 °C) waterbath. Upon complete dissolution, membrane filtration was carried out for sterilization purpose.

3.1.1.8 MgCl₂ 6H₂O, 1 M

MgCl ₂ 6H ₂ O (MW: 203.3)	20.3 g
sdH ₂ O	100 ml

MgCl₂ 6H₂O crystals were weighted and dissolved in sdH₂O and topped up to 100 ml.

3.1.1.9 Tris-HCl, 1 M, pH 7.5

Tris-HCl (MW: 121.14)	121.14 g
sdH ₂ O	Up to 1 L

Tris base was added into 800 ml sdH₂O. The pH of the solution was adjusted to 7.5 by using concentrated HCl. The solution was then topped up to 1 L.

3.1.1.10 Ethylenediaminetetraacetic acid (EDTA), 0.5 M, pH 8.0

EDTA (MW: 292.25)	146.1 g
sdH ₂ O	Up to 1 L

EDTA powder was added into 800 ml sdH₂O. The pH of the solution was adjusted to 8.0 by adding NaOH to facilitate the dissolution of EDTA.

3.1.1.11 Double distilled water (dsH₂O)

DsH₂O was prepared by autoclaving ultrapure water obtained from Mili-Q System (Milipore Corporation, USA). This water was used in buffers and reagent preparations for molecular reactions.

3.1.2. Commercially available reagents

No.	Company	Chemical/reagent
1.	Fermentas	Loading dye, 6 X DNA reference markers, 100 bp Tris Borate EDTA (TBE) buffer, 10 X <i>Taq</i> DNA polymerase, recombinant Dream <i>Taq</i> DNA polymerase Deoxynucleotide (dNTP) mix, 25 mM
2.	Amresco	Phenol-chloroform-isoamyl alcohol MgCl ₂ 6H ₂ O NaCl NaOAc Sucrose Tris base
3.	Biorad	Sodium Dodecyl Sulfate
4.	Promega	Proteinase K Agarose powder
5.	Invitrogen	<i>Taq</i> DNA polymerase
6.	Merck	Ethidium Bromide 2-propanol-2 (Isopropanol)
7.	VWR	Ethanol, absolute
8.	John Kollin Corporation	Ethanol, 95 % (v/v)

3.1.3 Commercially available kits

No.	Company	Item	CAT. No.
1.	Qiagen	Qiaquick PCR Purification kit	28106
		Qiaquick Gel Extration kit	28706
2.	Promega	Powerplex 16 system	DC6531
		Powerplex Matrix Standards, 3100/3130	DG4650

3.1.4 Analysis software

No.	Program	Function
1.	The Excel Microsatellite Toolkit	Data checking; formatting; frequency calculation
2.	Powerstat V1.2	Calculation of PD, PE, PIC, TPI, Matching probability
3.	Genemapper V3.2	Allele calling for STR analysis
4.	Bioedit	Quality examination of electropherogram; sequence alignment
5.	Gendoc	Multiple sequence alignment
6.	ClustalW	Multiple sequence alignment
7.	Mega V5	Multiple sequence alignment; sequence data analysis, phylogenetic tree editing
8.	Arlequin V3.11	HWE analysis; AMOVA; haplotype analysis
9.	FSTAT	Population differentiation analysis
10.	XLSTAT	Principal component analysis
11.	GeneAIEx 6.5	Principal coordinate analysis
12.	POPTREE2	Genetic distance, phylogenetic tree
13.	Structure V2.3.4	Perform model-based clustering for population inference
14.	Structure Harvester	Determination and selection of best fit- <i>K</i> for Structure analysis
15.	CLUMPP V1.1.2	Alignment of member coefficient for multiple replicate analyses
16.	Distruct V1.1	Generate graphical bar plots with labels

3.1.5 Essential research services

Due to unavailability of facility in the laboratory, cooperative effort has been established with companies that provide a number of research services. These services include DNA sequencing and fragment analysis using a genetic analyzer, and customization and synthesis of oligonucleotides. Prior to selection of the service provider, several companies have been chosen for comparison of the price and quality of their services. Among all companies compared, First Base (M) Sdn. Bhd. was selected to provide all services. First Base (M) Sdn. Bhd. has the fastest result turnover rate, with reasonable price and high quality. Other than that, some of our oligonucleotides were also obtained from Research Biolabs (M) Sdn. Bhd.

3.2 Research methodologies

3.2.1 Sterilization techniques

There were 2 sterilization procedures carried out to reduce the risk of microbial contamination to minimum. Methods employed throughout the study depend on the nature of the items, i.e., heat-labile or heat-stable.

Heat-stable items, such as microcentrifuge tubes, 0.2 ml PCR tubes, pipette tips, mili-Q water, etc., were sterilized by autoclaving (Hirayama, Japan). The sterilization condition was 121 °C at 15 p.s.i. for 15 minutes. The autoclaved items were dried in an oven (Mettler, Germany) at 65 °C.

Heat-labile items, such as SDS buffer and Proteinase K buffer, were sterilized by means of syringe filtration. The solutions were passed through a 0.45 µm pore-size filter to remove any suspended particles. The filtrate was collected in sterilized tubes/bottles and stored until further use.

3.2.2 Volunteer recruitment and sample collection

In the present study, a total of 639 volunteers, from 3 of the largest indigenous groups, residing in the state of Sabah were recruited. Prior to sample collection, all volunteers were required to fill up an informed consent form, whereby personal particulars were provided by the volunteers (Appendix 1). During recruitment, only individuals with pure indigenous genetic lineage, i.e., parents and grandparents originating from the same ethnic group were selected. Blood samples were collected from 594 volunteers via venipuncture in 10 ml EDTA blood tubes. The filled tubes were mixed evenly and stored at -80 °C until extraction of the genetic materials. In addition, 45 buccal swab samples were also obtained from other Bajau individuals.

The ethnicities of each volunteer were recognized based on self-declaration. The summary of volunteers is as below:

Ethnicity	Series ID	n	Sample type
Kadazan-Dusun	C	271	Blood
Bajau	B	114	Blood
	S	45	Buccal
Rungus	R	209	Blood

Application for ethical clearance has been submitted and reviewed by Medical Ethics Committee (MEC) of University Malaya Medical Centre (UMMC) prior to the sample collection. Our research methods were approved, with the MEC reference numbers of 612.16 and 770.21 (Appendices 2 and 3).

3.2.3 Genomic DNA extraction

Blood DNA was extracted from whole blood samples via a conventional phenol-chloroform extraction method. Basically, the entire work flow was separated into 2 days: day 1 involved the purification and enzymatic digestion of the cell pellet; the isolation and ethanol purification of the genomic DNA in day 2.

Prior to extraction, EDTA tubes, which contained blood samples, were left to thaw at room temperature and mixed well. A total of 3 to 6 ml of whole blood was transferred to a sterile 50 ml falcon tube and mixed with 40 ml of 1 X RCL buffer. The mixture was left to stand for 1 minute and spun at 3,850 $\times g$ at 10 °C for 10 minutes. After centrifugation, the supernatant was discarded into a container with 5 % (v/v) Clorox for disinfection. The cell pellet was washed by resuspending in 20 ml of 1 X RCL buffer and subjected to another round of centrifugation. The supernatant was discarded and the pellet was checked. If the pellet appeared reddish in this stage, an additional washing step was performed to further purify the cell pellet. Otherwise, the falcon tube which contained the cell pellet, was inverted for 1 minute to allow the drainage of excessive buffer. The purified pellet was then subjected to enzymatic digestion by adding 40 μ l of freshly prepared 10 mg/ml Proteinase K solution, 160 μ l of 5 X strength Proteinase K buffer, and 200 μ l of sdH_2O . The mixture was evenly mixed and incubated at 37 °C in an incubator overnight to achieve complete digestion of the cell pellet.

The isolation of genomic DNA from the digested cell mixture was carried out on the second day. A total of 500 μ l of the digestion mixture was transferred to a 1.5 ml clean and sterile microcentrifuge tube. In a ventilated biological safety hood, 800 μ l of phenol-chloroform-isoamyl alcohol solution was carefully added into the microcentrifuge tube and the mixture was inverted several times until it appeared

milkish white. Subsequently, the mixture was centrifuged at 16,060 xg at 10 °C for 30 minutes. The solution settled into 3 layers at the end of the centrifugation; an aqueous layer at the top, followed by a thin cloudy interphase, and an organic phase that settled at the bottom of the solution. The DNA-containing aqueous phase was aspirated into a 1.5 ml microcentrifuge tube. The first round of DNA purification was performed by the salting-out method, where a total of 200 µl of 6 M NaCl, 40 µl of 2 M NaOAc, and 900 µl of chilled absolute ethanol were added into the aqueous phase solution. The mixture was mixed evenly and kept at -80 °C to facilitate the precipitation process.

After 2 hours in -80 °C, the mixture was thawed and centrifuged at 16,060 xg at 4 °C for 5 minutes. The supernatant was discarded and the pellet was washed with 1 ml of 70 % (v/v) ethanol. Subsequently, another round of centrifugation was performed at 16,060 xg at 4 °C for 5 minutes. At last, the supernatant was discarded and the pellet was carefully dried by inverting. The dried pellet was dissolved in 100 µl of sdH₂O. The extracted DNA solution was left to dissolve completely in a 37 °C incubator for 1 hour. The DNA sample was then stored at -20 °C until further use.

In the present study, some of the DNA samples were collected in the form of buccal swap. These swaps were subjected to genomic DNA purification by our collaborator in Universiti Malaysia Sabah (UMS). The purified samples were then used for genetic marker testing.

3.2.4 Quantity and quality assessment of DNA samples

The genomic DNA samples extracted from both whole blood and buccal cells were quantitated before subjecting to all experiments. We accessed the quality and quantity of the extracted samples by a NanoPhotometer (Implen, Germany), based on the absorbance values at wavelengths 260 nm (OD_{260}) and 280 nm (OD_{280}). Note that the OD signifies optical density. The determination of DNA concentration was calculated according to the value of OD_{260} obtained for each samples. On the other hand, the purity of the extracted DNA samples was estimated based on the ratio of OD_{260} and OD_{280} . A pure DNA sample gave a ratio between 1.8 and 2.0. Other readings, higher or lower, indicate impurities in the sample, such as protein or RNA fragments.

The assessment was conducted in the LabelGuardTM Microliter Cell using the function for double stranded (ds) DNA. Together with the cell, a 0.2 mm lid (Factor 50) was used for all samples initially. The 1 mm lid (Factor 10) was used when the amount of DNA detected in the samples was too low that the 0.2 mm lid was not able to estimate it accurately. In order to determine the concentration of DNA in each sample, the NanoPhotometer was first “blanked” with sdH_2O . Then, 1 μ l of undiluted DNA stock was dropped onto the centre of the measuring window of the Microliter Cell and covered with appropriate lids. The absorbance values were obtained for both wavelengths 260 nm and 280 nm. After each read, the DNA samples were either retrieved by pipetting back to the solution or removed by simply wiping the measuring window with KimWipes.

3.2.5 Genotyping of samples

We utilized several different techniques to analyze various types of genetic polymorphisms in our samples. The types of genetic markers examined are including single nucleotide polymorphism (SNP), insertion and deletion (InDel), short tandem repeat (STR), and variable number of tandem repeat (VNTR), as shown in the study framework in Figure 3.1.

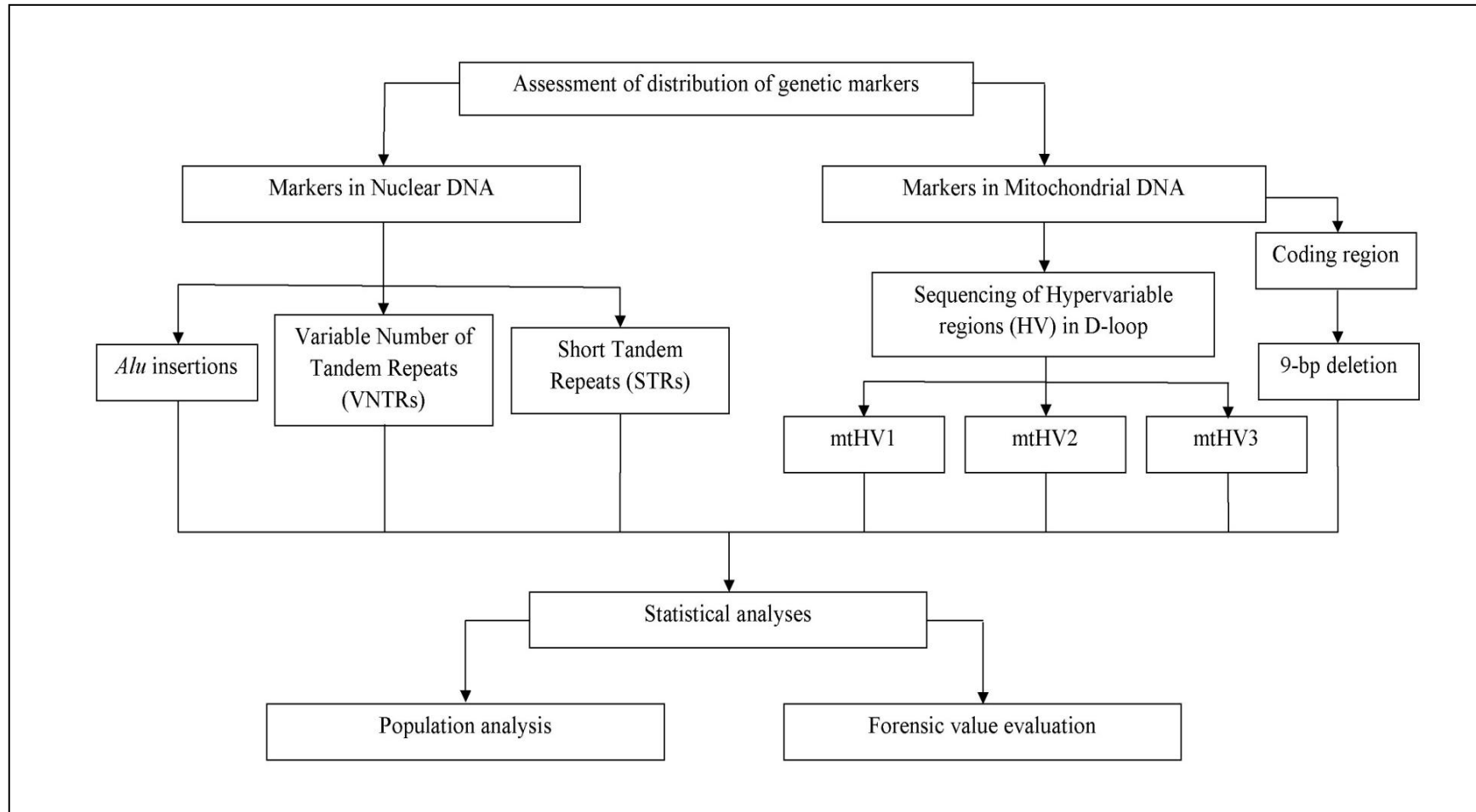


Figure 3.1 : Outline of the genetic studies carried out on the Sabahan indigenous populations, via the examination of multiple loci in different genome subsets.

3.2.5.1 Nuclear DNA markers: VNTRs

Both VNTR markers, present in 3'UTR and Intron-8 of the DAT1-1 gene, were examined via a direct PCR method in the present study. The selected markers were amplified in 15 µl reaction mixture containing 1 X power of *Taq* buffer, 2 mM of MgCl₂, 0.1 mM of dNTP mixture, 0.5 mM of each forward and reverse primers, 0.75 Unit (U) of *Taq* Polymerase, and approximately 100 ng of template DNA.

The reagent contents of PCR for both VNTRs in DAT-1 gene and their respective primer sets are listed in Tables 3.1 and 3.2. The addition of salt (MgCl₂) and additive (Betaine) in PCR mixtures was conducted for elimination of unspecific amplification. Optimization was carried out before hand to determine the appropriate ratio of salt or additive that yields the best amplification.

The mixture was placed in a thermal cycler (Mycycler, Biorad) for the amplification process under the cycling conditions of initial denaturation at 94 °C for 120 seconds, 30 cycles of 94 °C for 30 seconds, 68 °C for 20 seconds, and 72 °C for 50 seconds, followed by a final extension at 72 °C for 600 seconds. Upon completion of the amplification process, the PCR mixture was left to stand at 15 °C in the thermal cycler.

The PCR products were resolved on a 3 % (w/v) native AGE stained with Ethidium Bromide (EtBr). Migration took place under the voltage current of 130 for 30 minutes. The gel was visualized in an ultraviolet (UV) transilluminator and the genotype assignment for all samples was carried out according to the migration patterns on the agarose gel.

Statistical analyses were carried out via programs, i.e., Powerstat V1.2, Arlequin package V3.11, and FSTAT, to calculate various population and forensic parameters for these markers in the Sabahan indigenous populations.

Table 3.1 : PCR components for the amplification of VNTR markers.

Reaction component	Stock concentration	Volume (μ l)	
		3'UTR	Intron-8
10 X <i>Taq</i> buffer	10 X strength	1.50	1.50
MgCl ₂	50 mM	0.60	-
Betaine	5 M	-	1.2
Forward primer	50 mM	0.15	0.15
Reverse primer	50 mM	0.15	0.15
dNTP mix	10 mM	0.15	0.15
<i>Taq</i> polymerase	5 U/ μ l	0.15	0.15
sdH ₂ O	-	11.30	10.70
Template DNA	100 ng/ μ l	1.00	1.00

Table 3.2 : Sequences of primers used to amplify the targeted markers in the VNTR study.

VNTR	Primer	Sequence (5' – 3')	Reference
DAT-1 gene			
3'UTR	Forward	TGTGGTGTAGGGAACGGCCTGAG	Vandenbergh, et al., 2000
	Reverse	CTTCCTGGAGGTCACGGCTCAAGG	
Intron-8	Forward	GCACAAATGAGTGTTTCGTGCATGTG	
	Reverse	AGCAGGAGGGGCTTCCAGGC	

3.2.5.2 Nuclear DNA markers: InDels

The 6 *Alu* insertions (HS4.32, TPA25, PV92, HS3.23, B65, and APO) were amplified by a direct PCR method in a thermal cycler (Veriti, Applied Biosystem). The 10 μ l PCR mixture contained 100 ng of template DNA, 1 X Dream*Taq* buffer, 0.2 mM of dNTP mix, 0.4 U of Dream*Taq* DNA Polymerase, 2 mM of MgCl₂, and 0.4 mM of each forward and reverse primers. Chromosomal locations and primer sequences used for the study are listed in Table 3.3. Each reaction was subjected to a cycle of denaturation at 94 °C for 5 minutes, 35 cycles of 94 °C for 30 seconds, appropriate optimized annealing temperature (54 °C to 65 °C) for 30 seconds, 72 °C for 30 seconds, and a final extension at 72 °C for 5 minutes. Amplicons were subsequently resolved on 2 % (w/v) native agarose gels, stained with EtBr, and visualized under a UV transilluminator. Allelic scoring was carried out by the method of direct counting.

Table 3.3 : Chromosomal locations and primer sequences of the 6 *Alu* insertions investigated in the present study.

<i>Alu</i> marker (Chromosome)	Primer sequence (5' – 3')	*TA (°C)	Reference
HS3.23 (7)	^a GGTGAAGTTTCCAACGCTGT ^b CCCTCCTCTCCCTTTAGCAG	65	Arcot, et al., 1996
HS4.32 (12)	^a GTTTATTGGGCTAACCTGGG ^b TGACCAGCTAACTTCTACTTTAACC	63	
TPA25 (8)	^a GTGAAAAGCAAGGTCTACCAG ^b GACACCGAGTTCATCTTGAC	63	Tishkoff, et al., 1996
B65 (11)	^a ATATCCTAAAAGGGACACCA ^b AAAATTTATGTCATGGGTAT	54	Arcot, Wang, Weber, Deininger, & Batzer, 1995
APO (11)	^a TGTGAGCCTAGGAGTTTGAG ^b CTGGCTGATTTTAGGAGGGA	65	Batzer, et al., 1994
PV92 (16)	^a AACTGGGAAAATTTGAAGAGAAAGT ^b TGAGTTCTCAACTCCTGTGTGTTAG	60	

^a denotes forward primer

^b denotes reverse primer

* TA stands for “annealing temperature”

Analysis of Molecular Variance (AMOVA) and departure of the examined markers from the Hardy-Weinberg Equilibrium (HWE) were determined via the Arlequin package V3.11 based on exact test, with significant levels set at 5 % (Excoffier, Laval, & Schneider, 2005). Population and forensic parameters, such as heterozygosity, frequencies, polymorphism information content (PIC), power of discrimination (PD), matching probability, and power of exclusion (PE), were computed by Powerstat V1.2 (Promega, USA). Population differentiation was estimated by a set of measurements using FSTAT (Goudet, 1995).

Principal component analysis (PCA) was conducted to investigate the genetic association among the studied populations by XLSTAT (Addinsoft, USA). PCA also included available *Alu* insertion data of other global populations from the Allele Frequency Database (ALFRED) to depict relationships between the indigenous groups to global populations, which may in turn reveal information on the historical movement within this region (Rajeevan, et al., 2005). Insertion frequencies of 5 out of 6 markers (except HS3.23) in this study were utilized to construct the principal component (PC) plot.

In order to study the inter-relationship of the populations, 2 PC plots were drawn. The first PC consists of populations from various regions of the globe, which aimed to investigate the genetic relation of these populations in response to their geographic origin. On the other hand, the second PC was built via *Alu* insertion frequencies of a number of populations in the SEA and its neighboring regions. With this, it revealed the genetic association between them and portrayed the linkage among the populations.

Other than PCA, another clustering methodology has been employed in the study to infer the genetic relationship of the indigenous populations with others. Phylogenetic approach was engaged to characterize and represent each tested population in a tree-like

structure that based on the genetic similarity. The insertion frequencies of all populations were formatted to fit the input for POPTREE2 software (Takezaki, Nei, & Tamura, 2010). The software worked by converting the frequencies to a genetic distance in a pair-wise format. The Nei's D_A algorithm was selected for our analysis.

$$D_A = 1 - \frac{1}{r} \sum_j \sum_i^{m_j} \sqrt{x_{ij} y_{ij}}$$

where x_{ij} and y_{ij} are the frequencies of the i -th allele at the j -th locus in populations X and Y, respectively, m_j is the number of alleles at the j -th locus, and r is the number of loci used.

(Nei, Tajima, & Tateno, 1983)

A Neighbor-joining (NJ) tree was constructed founded on the derived pair-wise genetic distance. Resampling was conducted by bootstrapping at 10,000. The resulted unrooted tree was exported in nwk. format to Mega V5 software (Tamura, et al., 2011). Additional refining and editing of the tree structure, in terms of taxon labels, group icon assignment, flipping of branches that do not alter the initial arrangement of the tree, were carried out using Mega V5 software.

3.2.5.3 Nuclear DNA markers: STRs

A total of 16 markers were typed using Powerplex 16 system (Promega Corporation, USA) in a single amplification reaction, including 15 STR markers and a sex-determining AMEL locus. The entire typing process can be divided into 3 stages, i.e., amplification, fragment analysis, and data analysis.

The kit is supplied with reagents in 2 separated boxes for pre-amplification and post-amplification activities. The pre-amplification component box contains Gold ST★R 10 X buffer, Powerplex 16 10 X primer pair mix, and a tube of control DNA (9947A or 2800M DNAs). On the other hand, the post-amplification component box includes the marker-specific allelic ladder mix and the internal lane standard (ILS) 600.

During the experiments, precautions steps were taken according to the manufacturer's recommendation and standard good laboratory practices to minimize the possibility of cross contamination and errors that can result in false reporting data.

3.2.5.3(a) Amplification stage

Each run of the amplification process included a number of unknown samples, positive (9947A or 2800M DNAs), and non-template amplification control (NTC) reactions, in a 12 μ l reaction mixture. The reaction components are listed in Table 3.4.

Table 3.4 : Reaction components for amplification of Powerplex 16 system.

Reaction component	Volume (μ l)
Gold ST★R 10 X buffer	1.2
Powerplex 16 10 X Primer pair mix	1.2
Invitrogen <i>Taq</i> DNA Polymerase (5 U/ μ l)	0.4
Template DNA	1.5
sdH ₂ O	7.7
TOTAL	12.0

For positive and NTC control reactions, 0.5 ng of 9947A or 2800M DNAs and 1.5 μ l of sdH₂O were added respectively, to the reaction instead of DNA sample.

These reaction mixtures were then placed in a thermal cycler (C1000, Biorad) for amplification. The cycling parameters were adapted from the manufacturer's manual, as illustrated in Figure 3.2. After performing an in-house optimization, the number of cycles was reduced to 10/15, instead of 10/22 as recommended by the manufacturer, to minimize the effect of imbalanced signal peaks in the fragment analysis stage which resulted from the presence of excessive PCR amplified products.

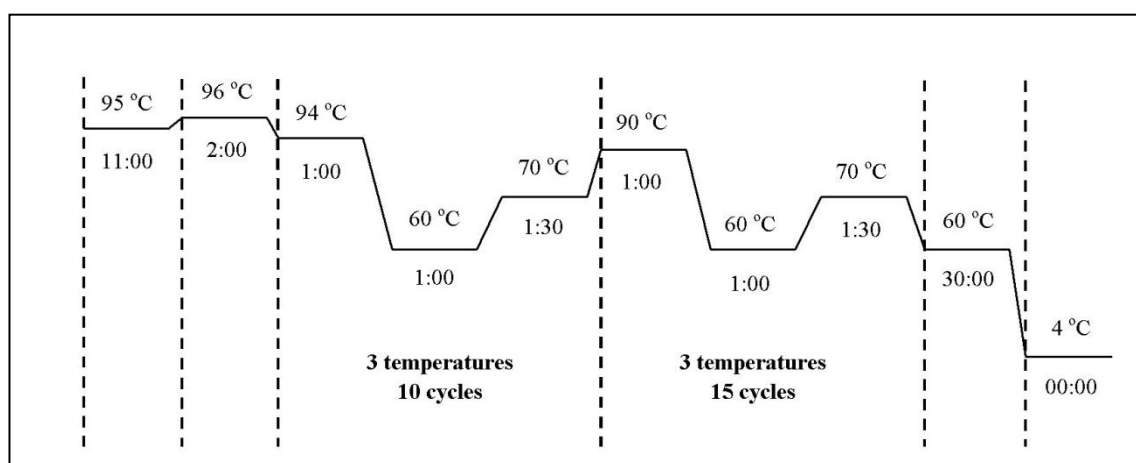


Figure 3.2 : PCR cycling parameters employed in the amplification of Powerplex 16 system (top panel displays the temperatures of each stage; bottom panel shows the incubation time in minute).

Upon completion of the amplification process, the amplified products were checked to ensure successful amplification prior to subsequent experiment. A total of 2.5 μ l of each amplified product was subjected to separation on a 2 % (w/v) native agarose gel. Each sample was mixed with 1 μ l of 6 X loading dye and loaded into separated wells on the agarose gel, followed by electrophoresis of 130 V for 20 minutes. The EtBr pre-stained gel was then visualized under a UV transilluminator. Samples with successful and satisfactory amplification were subjected to fragment analysis.

3.2.5.3(b) Fragment analysis stage

The amplified products were subjected to fragment analysis to identify fragments that were present in each sample. The process was carried out in an ABI PRISM 3100 genetic analyzer, with data collection software version 2.0, according to the manufacturer's recommendation by First Base Sdn. Bhd. in their laboratory located in Malaysia.

In order to obtain optimum and satisfactory results for the fragment analysis study, the genetic analyzer was first calibrated with Powerplex Matrix Standards, 3100/3130 (Cat.# DG4650).

In the sample preparation, each well contained 0.5 µl of ILS600, 9.5 µl Hi-Di formamide, and 1 µl of amplified sample or allelic ladder mix. The plate was briefly centrifuged to remove air bubble. Prior to loading onto the machine, samples were denatured at 95 °C for 3 minutes and immediately chilled on crushed ice for another 3 minutes.

Samples were placed in the instrument and ran according to the manufacturer's recommendations; program was set for injection time of 5 seconds, injection voltage at 3 kV, run time of 2,000 seconds.

3.2.5.3(c) Raw data collection and allele calling

The raw results obtained from the fragment analysis were subjected to allele calling by using GeneMapper ID software, version 3.2. Data analysis was conducted following the manufacturer's manuals for both PowerPlex 16 system and GeneMapper ID V3.2. The software functions by integrating the raw data from the genetic analyzer and assigning accurate alleles to each locus based on the separation of fragment in the capillary. Allelic assignment was carried out by the software with reference to the separation pattern of the allelic ladder mix, which was processed together with unknown samples in each run. The positive control reaction in each run was used as an individual quality measure for that particular run.

In order to facilitate the assignment of alleles to each sample, the manufacturer has developed proper panels and bin file, for the GeneMapper ID V3.2 software, which can be obtained from the Promega website at: www.promega.com/geneticidtools/panels_bins/. The software analyzed and categorized each result file according to their quality. Results with undesirable qualities were carefully checked for uncalled and off-ladder alleles. If necessary, these samples were repeated.

3.2.5.3(d) Non-allelic variant

With the aid of the software, the alleles of each examined individual were determined with comparison to the allelic ladder components. However, not all peaks/signals fitted perfectly. There were a small number of samples that presented with “odd” alleles, termed as “off-ladder” peaks.

In order to investigate the presence of non-allelic variant in our study, every electropherogram generated from the fragment analysis was inspected carefully. Samples with “off-ladder” peaks were further checked and repeated from amplification stage. If the peaks persisted in the repetition, the samples were subsequently subjected to cloning and DNA sequencing to identify the exact nucleotide arrangement of the “off-ladder” fragment.

3.2.5.3(d)(i) PCR and ligation reactions

A cloning strategy had been established to identify the non-allelic variants. First, a conventional PCR was carried out to amplify a 246 bp sequence, encompassing the microsatellite of interest. The entire 30 µl PCR mixture contained 1 X *Taq* buffer, 0.4 mM of each forward (5' - CAAATGCCCCATAGGTTTTG - 3') and reverse (5' - TCACGGTCTGAAATCGAAAA - 3') primers, 0.08 mM of dNTP mixture, 1.2 U of Dream*Taq* polymerase, and 100 ng of template DNA. The PCR mixture was placed in a thermal cycler (C1000, Biorad) for cycling at 94 °C for 120 seconds, followed by 30 cycles at 94 °C for 30 seconds, 63 °C for 20 seconds, and 72 °C for 50 seconds. The final extension step was allowed at 72 °C for 300 seconds and the mixture was incubated at 15 °C until being removed from the thermal cycler. The amplified products were applied on 3 % (w/v) EtBr-prestained native agarose gels and the desired

fragments were excised. The excised gels were subsequently purified via gel purification kits before being subjected to cloning reaction.

The cloning reaction was carried out via pGEM-T easy vector system II (Promega Corporation; CAT. A1380). Ligation reaction was set up by adding reagents as listed in Table 3.5.

Table 3.5 : Reagents added in the ligation mixture.

Reagent	Volume (μ l)
2 X rapid ligation buffer, T4 DNA ligase	5
pGEM-T easy vector (50 ng)	1
PCR product, purified	2
T4 DNA ligase (3 Weiss units/ μ l)	1
sdH ₂ O	1
TOTAL	10

The ligation mixture was mixed and incubated at room temperature for 1 hour. After incubation, transformation of the vectors into competent cells was carried out on LB agar plates, with Ampicillin (100 μ g/ml), IPTG (40 μ g/ml), and X-Gal (50 μ g/ml) added.

Before transformation, the JM109 high efficiency competent cells were thawed on an ice bath for about 5 minutes. Ligation reaction was briefly centrifuged and 2 μ l of the mixture was added to the bottom of a sterile microcentrifuge tube on ice. The competent cells were gently mixed after thawed. A total of 50 μ l of the cells were transferred to the tube with the ligation mixture. The tube was mixed by flicking and left standing on ice for 20 minutes. After the incubation, heat shock was conducted on a

heat block at 42 °C for 45 seconds. The tube was immediately returned to ice, without shaking, and incubated for another 2 minutes. Subsequently, 950 µl of LB broth was added into the transformation tube and incubated in a shaking incubator at 37 °C (150 rpm). After 90 minutes, the transformation tube was removed from incubator and 100 µl of the culture was used for plating. Duplication was carried out onto LB agar plates, with Ampicillin, IPTG, and X-Gal. The plates were then incubated at 37 °C for overnight.

3.2.5.3(d)(ii) Colony selection and screening

The pGEM-T easy vector contains T7 and SP6 RNA polymerase promoters flanking the multiple cloning site in the coding region of α -peptide subunit for the enzyme β -galactosidase. An active β -galactosidase cleaves the X-Gal, a colorless analog of lactose, and the by-product would be rapidly oxidized to form a blue-colored pigment. The insertion of foreign DNA fragments into the multiple cloning site would inactivate the enzyme, which in turn leads to the formation of white colonies in the culture plate. On the other hand, colonies without insert appear as blue in the culture plate. Thus, it allows quick and easy identification of colonies with insertions.

After the overnight incubation, white colonies were picked from the plate and sub-cloned onto a LB agar plate. Part of the selected colonies were also picked from the plate and suspended into 10 µl of sdH₂O for PCR screening. The suspension was then mixed with 10 µl of PCR reagents. The 20 µl of final PCR reaction contained 1 X DreamTaq buffer (with MgCl), 0.4 µM each for the universal primers T7 and SP6, 0.08 µM of each dNTP, and 0.8 U of DreamTaq DNA polymerase. The reaction tubes were then placed in a thermal cycler (Veriti, Applied Biosystems) and ran at 95 °C for 10 minutes, followed by 30 cycles at 95 °C for 30 seconds, 55 °C for 30 seconds, and 72 °C

for 30 seconds, then final extension at 72 °C for 7 minutes and soaking at 15 °C. The purpose of the prolonged denaturation step was to release the vectors from the cells. The amplified products were checked on a 2 % (w/v) native agarose gel stained with EtBr. An empty vector is represented by the fragment of approximately 180 bp, whereas successfully inserted clones produced a fragment of 800 bp.

The sub-cultured plates were incubated at 37 °C for overnight and inserted clones were selected for sub-culturing in 5 ml LB broth. The LB cultures, in universal bottles, were then placed in a 37 °C incubator for 8 hours or overnight.

3.2.5.3(d)(iii) Plasmid preparation

Plasmid preparation was carried out using Wizard Plus SV Minipreps DNA purification system (Promega Corporation; CAT. A1460). A total of 1.5 ml overnight culture was transferred into a clean and sterile microcentrifuge tube and spun for 5 minutes at 10,000 $\times g$. The clear supernatant was carefully discarded. These steps were repeated for 3 times for a total of 6 ml bacterial culture. The pellet was then resuspended with 250 μ l of cell resuspension solution and mixed thoroughly. Later, 250 μ l of cell lysis solution was added into the cell suspension and mixed by inversion for 4 times. Next, 10 μ l of alkaline protease solution was added into the tube and mixed by 4 inversions. The mixture was left stand at room temperature for 5 minutes. After incubation, 350 μ l of neutralization solution was added and mixed via 4 inversions. The solution was then centrifuged at full speed for 10 minutes at room temperature.

After centrifugation, a spin column was assembled into a collection tube. The cleared lysate from the centrifugation was decanted into the spin column. The pellet was discarded together with the microcentrifuge tube. The assembled spin column, with cleared lysate within, was spun at full speed for 1 minute at room temperature and the

flow through was discarded. A total of 750 μl of wash solution was added to the spin column and centrifuged at top speed for 1 minute. The flow through was then discarded and the washing step was repeated with additional 250 μl of wash solution. The spin column was spun for another 2 minutes at full speed to get rid of any excessive solution that might be trapped in the membrane. Subsequently, the spin column was transferred to a sterile microcentrifuge tube and 100 μl sdH_2O was added to the centre of the membrane in the spin column. The spin column was then spun at full speed for 1 minute at room temperature. The eluted plasmid was kept at $-20\text{ }^\circ\text{C}$ until further use. For quality check, the prepared plasmid was applied on a 1 % (w/v) agarose gel.

3.2.5.3(d)(iv) Plasmid sequencing

The purified plasmid contained the 246 bp inserted fragment of interest. In order to obtain the sequence arrangement of the targeted fragment, 20 μl of the prepared plasmid was sent for sequencing service. Reads by both T7 and SP6 universal primers were acquired to ensure the sequence reliability. The number of repeating a block in the samples was compared to the reference sequence of variants in STR database to determine if the “off-ladder” peaks are allelic variant of the STR locus.

3.2.5.3(e) Statistical analysis

The genotypic data of all 16 markers (15 STRs and AMEL) was exported and tabulated for further statistical analysis in an excel document. Several forensic and population parameters were estimated by Powerstat V1.2 (Promega, USA). These parameters include frequencies of alleles and genotypes, heterozygosities, matching probability, PD, PIC, and PE. The STR markers were also tested for departure from the HWE via Arlequin software V3.11 (Excoffier, et al., 2005).

Combined powers of discrimination and exclusion are effective parameters that collectively evaluate the usefulness of a set of markers, also known as “system” in forensic DNA study, in particular populations for human identification purpose. Power of discrimination measures the ability of a system to identify a particular individual in a population – individualization. On the other hand, power of exclusion serves as a parameter that dismisses a random individual as a potential match in identity testing.

The combined PD and PE are calculated by the formula below (Butler, 2005):

$$\text{Combined PD} = 1 - \prod_{i=1}^n (1 - \text{PD}_i)$$

$$\text{Combined PE} = 1 - \prod_{i=1}^n (1 - \text{PE}_i)$$

Where, n = number of markers in the system

For the assessment of genetic diversity of these STR markers in the Sabahan indigenous individuals, several population differentiation parameters were also computed using the FSTAT software (Goudet, 1995). These parameters included total gene diversity (H_T), intra-population gene diversity (H_S), inter-population gene diversity (D_{ST}), coefficient

of gene differentiation (G_{ST}), and inbreeding coefficient (G_{IS}). In addition, AMOVA test was performed with Arlequin software V3.11 as well to further illustrate the differentiation among and within these populations. Hence, the percentages of differentiation arising between the populations and within the individuals of each population were computed.

The collective genotype profile for the 15 STR loci of each tested indigenous individuals in our study was prepared in the format to fit the Structure software V2.3.4 (Pritchard, Stephens, & Donnelly, 2000). The data was subjected to admixture ancestry with a correlated allele frequencies model. The length of burn-in period was set to 50,000 and 200,000 Markov chain Monte Carlo (MCMC) repeats after burn-in. These parameters were run for K values ranging from 1 to 10, where each process of the K values was repeated for 10 times (iteration = 10). After the run, the best K value for our sample set was determined by 2 methods, i.e., log likelihood for each K , $\ln P(D)=L(K)$ and delta K (ΔK). The evaluation of best-fit K was assisted by the software – Structure Harvester (Earl & vonHoldt, 2012). The data replicates yielded from analyses by the Structure software were aligned by CLUMPP software V1.1.2 (Jakobsson & Rosenberg, 2007). Datatypes for populations and individuals were generated for K values from 2 to 6 under the Greedy algorithm, with 10,000 permutations for each K . The outputs from the CLUMPP software were subjected to the Distruct software V1.1 (Rosenberg, 2003). This software generated graphical representation of the estimated membership coefficients of each sample, symbolized by a segment of line with K colored partitions, in the K clusters.

Clustering analysis, similar to those performed for *Alu* insertions, was carried out for all 15 STR loci as well. The first analysis was the Principal coordinate analysis (PCoA). The allelic frequencies of all markers for tested population were converted to Nei's D_A genetic distance using POPTREE2 software (Takezaki, et al., 2010). The resulting

genetic distance was transferred to an Excel-macro based software – GeneAIEx 6.5 (Peakall & Smouse, 2012). The software was used to construct the plots of PCoA. STR data of different populations around the world was gathered from published articles and an online human STR database – Autosomal STR DNA Database (www.strdna-db.org). The assembled frequency data was tabulated and prepared into a format that fits the input onto POPTREE2 software for genetic distance, D_A , computation. The resulted distance matrix was used to perform PCoA via GeneAIEx 6.5. The populations were divided into 2 groups for the analysis for world and SEA-neighboring populations.

The second clustering analysis for the STR system was a phylogenetic study. The same sets of population data used for the PCoA were subjected to POPTREE2 software for the construction of NJ tree. In order to validate the confidence level of each separated branch, bootstrapping at 10,000 replications was performed. There were 2 phylogenetic trees being built. One that aimed to illustrate the inter-relation of world populations according to their geographic locality and the other tree was established to investigate the genetic similarity of the Sabahan indigenous groups to other populations in the neighboring regions. Trees built by POPTREE2 software were subjected to additional editing and labeling via MEGA V5 (Tamura, et al., 2011).

3.2.5.4 MtDNA markers: 9-bp deletion

The presence of a 9-bp deletion in the mt intergenic region of COII-K was analyzed. A total of 20 µl PCR reaction mixture was prepared, containing 2 µl 10 X PCR buffer, 2 mM MgCl₂, 0.1 mM dNTP mix, 0.5 mM each of forward and reverse primers, 1 U *Taq* DNA Polymerase, 14.9 µl ddH₂O, and 1.5 µl diluted DNA sample. The cycling parameters and primer sequences are shown in Figure 3.3. Primer sequences were adapted from previously published research (Horai, et al., 1996). The amplified products were subjected to separation on a 3.5 % (w/v) native agarose gel prestained with EtBr. The existence of 9-bp deletion in region V of mtDNA in a sample is indicated by the presence of a 91 bp PCR product. On the contrary, the presence of 100 bp fragment signifies the absence of the 9-bp deletion in an individual. The frequencies of the 9-bp deletion in each indigenous group were calculated by direct counting.

Frequencies of the 9-bp deletion in other populations were collected and compared to the studied indigenous populations. The population data was obtained from previously published articles and covered all different regions of the world, such as Africa, Europe, America, South Asia, East Asia, SEA, and Oceania. Direct comparison of the frequencies was carried out. On the other hand, the 9-bp deletion profile of each individual generated in this part of the study was incorporated in the later part of our analysis, which involved the assignment of mt haplogroups and mtDNA analysis.

Figure 3.3 : PCR cycling parameters and primer sequences for mt 9-bp deletion examination (top panel display the temperature of each stage; bottom panel displays the duration expressed in minute).

PCR cycling parameter					
Initial denaturation	Denaturation	Annealing	Extension	Final extension	Soaking
94 °C 5:00	94 °C 1:00	60 °C 1:00	72 °C 1:00	72 °C 7:00	15 °C 00:00
<p>3 temperatures 30 cycles</p>					
Primer sequence (5' – 3')					
Forward Primer	TCGTCCTAGAATTAATTCCC				
Reverse Primer	AGTTAGCTTTACAGTGGGCT				

3.2.5.5 MtDNA markers: SNPs in control region

3.2.5.5(a) PCR amplification and product purification

The exact nucleotide order of the 3 mt HV regions, i.e., HV1, HV2, and HV3, in 450 individuals (150 each from the 3 indigenous groups) was determined through a sequencing approach to identify variants present in the samples. Initially, these regions were amplified individually via PCR method. The amplification was carried out in a 0.2 ml PCR tube containing 6 μ l of 10 X PCR buffer, 1.6 mM MgCl₂, 0.1 mM dNTP mix, 0.3 mM each of forward and reverse primers, 3 U *Taq* DNA Polymerase, 46 μ l dsH₂O, and 4 μ l diluted DNA sample. The cycling parameters and primer sequences are shown in Figure 3.4. The amplified products were checked for successful amplification on a 2.0 % (w/v) native agarose gel. The EtBr pre-stained gel was visualized under UV in a transilluminator. Reactions with successful and satisfactory amplification (sharp and bright fragments without unspecific bands and primer dimers) were selected and subjected to subsequent PCR purification steps.

A total of 50 μ l of the selected PCR products were purified with Qiagen PCR Purification kit following the manual's procedures to get rid of the presence of excessive nucleotides and primers that can affect the quality of result in the subsequent experiments. This step is particularly essential to reduce the level of noise signals and ensure a fine, long, and high quality read in the sequencing reaction later. The PCR products were bound and washed in a membrane-based column and finally eluted in a clean centrifuge tube with 50 μ l of dsH₂O. The purified fragments were then checked again on a 2 % (w/v) native agarose gel before being subjected to sequencing.

Figure 3.4 : PCR cycling parameters and primer sequences for amplification of mt HV regions (For cycling parameter: the top panel displays the temperature of each stage; bottom panel displays the duration in minute).

PCR cycling parameter					
Initial denaturation	Denaturation	Annealing	Extension	Final extension	Soaking
<div><div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div></div>					

3.2.5.5(b) Sequencing reactions

The sequencing reaction was conducted by First Base Sdn. Bhd. in their laboratory using the ABI 3730 xl genetic analyzer.

Mt control region is majority made up of G and C nucleotides. Therefore, the difficult template protocol was employed to sequence these GC-rich regions.

Each fragment was read twice from both H- and L- strands to achieve double coverage for reliable sequence information. The same primers had been used for PCR amplification and sequencing reactions.

3.2.5.5(c) Use of additional primers

Despite the usefulness of the mt HV primers sets in most of the samples, additional primers were introduced to amplify or/and sequence some samples that did not work perfectly with the original sets of primers. This was largely caused by the existent of homopolymeric C-stretches in the mt control region.

Samples with homopolymeric C-stretches in the HV1, 2, and 3 regions produced an “out of phase” situation where the sequence quality drops significantly after the region, resulting in unrecognizable signal peaks. In order to recover the sequence information, additional sequencing reactions were performed on these samples using internal primers (Table 3.6) that flank the homopolymeric C-stretches. These primers provide reads from the C-stretches from both strands.

The strategy for amplification and sequencing of mt HV1, 2, and 3 regions employed in our study was illustrated by Figure 3.5.

Table 3.6 : Sequences of internal primers used for re-sequencing of samples with homopolymeric C-stretches.

Primer	Sequence (5' – 3')	Annealing °C
HV1		
mtHV1aF	TTTGATGTGGATTGGGTTT	45
mtHV1aR	CCCCATGCTTACAAGCAAGT	52
HV2 and HV3		
mtHV3aF	GGGGTTTGGTGGAAATTTTGTG	50
mtHV2aR	CCCCCGCTTCTGGCCACAGC	62
R638	GGTGATGTGAGCCCGTCTAAAC	63

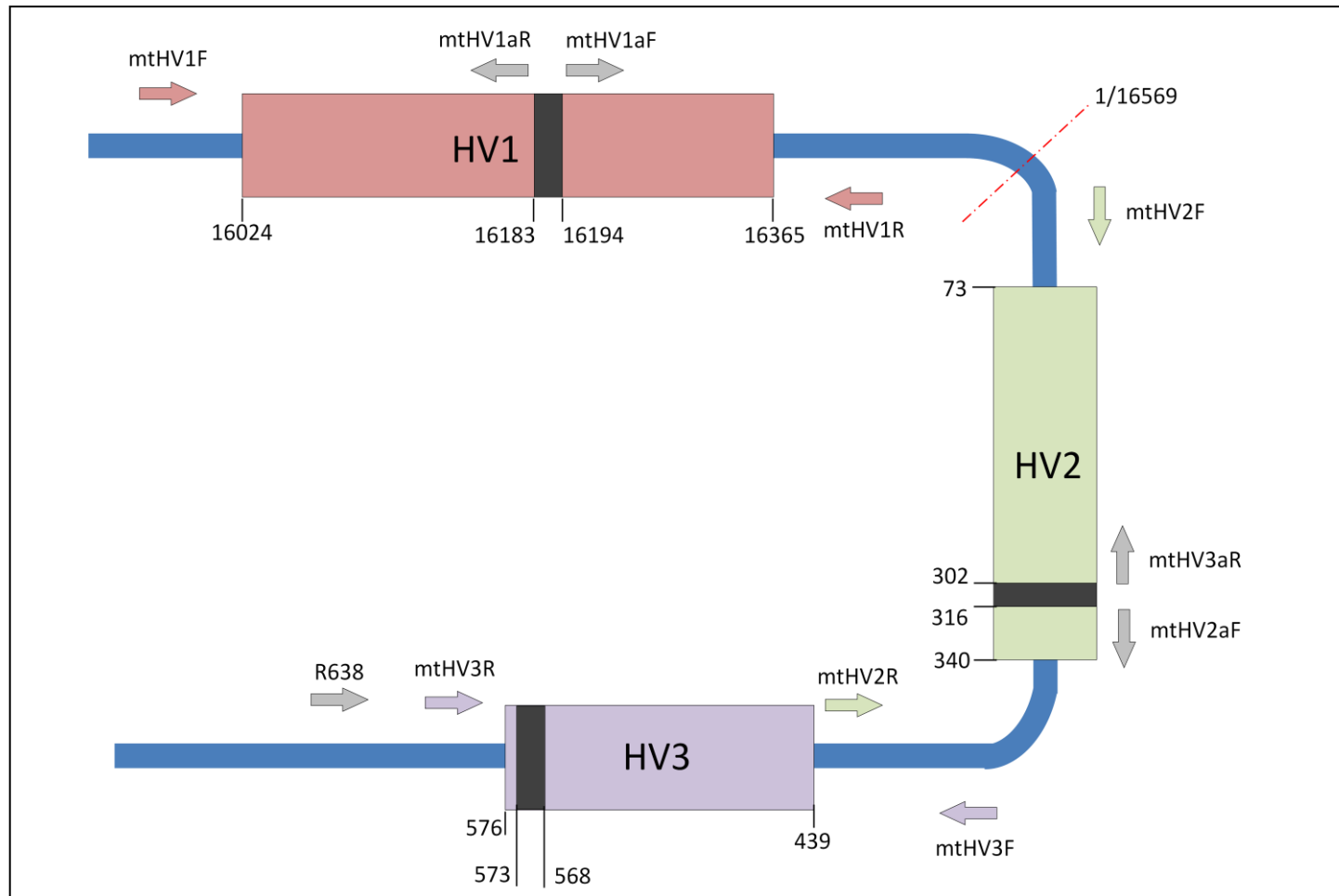


Figure 3.5 : Sequencing strategy of mt HV regions in the control region, 1,121 bp spanning from positions 16,024-576; dark boxes represent the homopolymeric C-stretches, while arrows are the primers used.

3.2.5.5(d) Cloning of problematic samples

Although additional primers used in the sequencing of the HV regions provide confident reading of the sequences from both H- and L- stands, some samples were still unable to read through. This situation happened during addition of multiple C's near the end of HV3 region at the position of 573 that results in a stretch of up to 10 C's. Unlike the poly-C stretches in HV1 and HV2 regions, this stretch prevents the reading of HV3R and results in complete signal loss.

In order to solve the problem, these samples were cloned into bacterial vectors and subjected to DNA sequencing reactions. Amplification of the HV2 and HV3 regions was carried out in a thermal cycler (Mycycler, Biorad) to yield a 651 bp sequence. The 60 µl PCR mixture was identical to the amplification of normal HV regions, except that the primers used were mtHV2F and R638. The amplified products were checked on 2 % (w/v) native agarose gels and subsequently purified via PCR purification kits prior to cloning.

The details of the entire cloning procedure has been described in the early part of the chapter [refer to section 3.2.5.3(d)].

In order to obtain the full sequence of the inserted DNA fragment, 20 µl of the prepared plasmid was sent for sequencing service for reading of both T7 and SP6 universal primers. Cloned vectors present at a higher concentration than purified PCR products, and therefore can produce better reads in the Sanger sequencing reaction. This is even more evident in samples with poly-C stretches as in some of our samples.

3.2.5.5(e) MtdNA sequence analysis

3.2.5.5(e)(i) Haplotype and polymorphism frequencies

The sequences obtained were checked for quality, in terms of peak signals and noise levels. The sequences were then analyzed with Bioedit software to construct consensus sequences of the 3 HV regions by using sequence information from both H- and L-strands (Hall, 1999).

The consensus sequences of the HV regions for each tested individuals were joined in a single FASTA file in the order of HV1, HV2, and HV3 in DNA analysis software – DNA baser (Heracle Biosoft S.R.L., Romania). The sequences were then imported into Bioedit software and aligned against the revised Cambridge Reference Sequence (rCRS). Multiple sequences were aligned under the ClustalW mode and the resulted file was exported in FASTA format. In order to spot the nucleotide differences of each sample to the rCRS, the FASTA file was accessed and converted to meg. format by MEGA V5 software (Tamura, et al., 2011). Each variation from the rCRS was highlighted in Mega software. In the Excel format, the position of each variable nucleotide in the sequenced mt HV regions was checked and marked.

The aligned sequences were also subjected to haplotype analysis in Arlequin software V3.11 (Excoffier, et al., 2005). Haplotype analysis estimates the frequency of haplotype in the sample groups and searches for the shared haplotype in the group. It also calculates the frequency of allelic variants in all loci. With the analysis, the number of unique and shared haplotypes in each indigenous group was estimated. On the other hand, all types of polymorphisms, such as insertion, deletion, transversional, and transitional SNPs, present in the examined samples and their frequencies were also documented.

2.2.5.5(e)(ii) Haplogroup assignment

Mt haplogroup assignment was completed by a fast and reliable web-oriented application developed by University of Innsbruck – Haplogrep (Kloss-Branstatter, et al., 2011). The allocation of haplogroups was based on the latest model of mt phylogenetic tree build 15, released on 4th October 2012. The most suitable haplogroup was assigned to each tested individual with regard to the polymorphisms in the mt control region. Haplogroup frequency counts were exported and tabulated.

3.2.5.5(e)(iii) Quality assurance and data deposition

The quality of the mt DNA sequences acquired in the study is essential as every change in the sequence, as compared to the reference, is analyzed as part of the genetic polymorphisms. Therefore, quality checks on the sequences are crucial to minimize the possibility to introduce phantom mutations and artificial recombination, especially during the data analysis stage.

The first line of quality check was carried out when inspecting the resulting electropherograms for all HV1, HV2, and HV3 regions. Samples with extraordinary high background noise were subjected to repetition (PCR amplification and sequencing reactions). Whenever possible, double coverage from H- and L- strands was obtained and cross-aligned to ensure no bias sequence results from any single strand read.

In addition, software and web-application were employed to assist parts of the data analysis and interpretation. One of which is the Haplogrep application that is able to determine the mt haplogroup for each sample based on their sequence patterns [refer to section 3.2.5.5(e)(ii)]. Mt haplogroup determination via the conventional manual method, by comparing each individual to the reference mt phylogenetic tree, is

sometimes subjective and often prone to mistakes. The Haplogrep application allows standardization of mt haplogroup assignment based on the latest mt phylogenetic tree across all samples and minimizes the variation caused by the manual method. Furthermore, it also decreases the bias and inconsistency especially where compilation and comparison of mt data of various populations from different literature are needed. In the analysis executed by the Haplogrep application, it automatically recognizes mutations that are odd to the determined haplogroup or never observed in the global phylogenetic tree (novel mutation point). Technical errors that are introduced by inaccurate alignment, phantom mutations, and point heteroplasmy can be detected and subsequently rectified. Thus, the raw sequencing profiles for samples with these mutations must be reviewed.

Apart from the web-based application, the mt data was also subjected to quality check via quasi-median (QM) networks, which is a quality check tool implemented by the EDNAP mtDNA population (EMPOP) database. The QM networks process and display the graphical representation of the mt data set that resembles the structure of their lineages. The analysis filters highly recurrent mutations in the data set and simplifies them in condensed and reduced haplotype forms. With that, artifacts or inconsistencies arise from sequencing data or interpretation can be detected. The nodes of the network represent condensed haplotypes and each branch represents a mutational event. Parallel branches carry the exactly identical mutation. QM is represented by a black dot in the network and it is a virtual haplotype that links haplotypes within the network, as computed by the software. Each observed virtual QM must be verified by reconfirming the sequencing data of their linked haplotypes.

Finally, the mt data was prepared and submitted to the EMPop database. The mt data was formatted as required by the database admin, which includes the haplogroup and respective polymorphisms in each sample. These data was reviewed carefully for

possible errors, such as artificial recombination and phantom mutations by the database admin. Raw data, such as electropherograms and sequencing files, would be requested for further verification upon the encounter of suspicious mutation points. After completion of the throughout examination, the mt data was deposited into the EMPOP database and accession numbers were granted.

3.2.5.5(f) MtDNA: Principal component analysis

The frequencies of mt haplogroup determined for individuals of the Sabahan indigenous populations were calculated and summarized. In order to gain insight into the genetic relevance of these groups with the surrounding populations and testify if they fit into the “out of Taiwan” migratory model, a PCA plot was constructed based on the mt haplogroup frequencies. Frequency profile of 25 populations along the “out of Taiwan” migratory path was adopted from previously published works (Hill, et al., 2007; Peng, et al., 2010; Simonson, et al., 2011; Trejaut, et al., 2005). These populations included Mainland East Asia (Hainan, Northeast, Northwest, Southwest, and Southeast China), Taiwanese Aborigines, Philippines, Mainland SEA (Vietnam, Kinh, Cham, Thailand, Orang Asli, Cambodia, Melayu Malay) and ISEA (Alor, Ambon, Sulawesi, Sumba, Java, Sumatra, Lombok, Bali, Borneo, Iban).

In order to reduce the statistical discrepancy due to different grouping methods used for the mt haplogroup determination in the published data, several pre-analysis processes have been carried out for a “clearer” comparison among the populations. First of all, haplogroup data with less than 5 % frequencies and with undefined groups were removed to retain major haplogroups that can represent the genetic structure of the populations better. Next, some “root” haplogroups were returned back to their basal haplogroup for easier comparison, as the analysis involved data from both recent and past published data. Lastly, the nomenclature of haplogroups was checked against the latest announced mt phylotree for accurate determination.

The resulted data was subjected to construction of PCA plot with XLSTAT software (Addinsoft, USA).

3.2.5.5(g) MtDNA: Phylogenetic analysis

Mt DNA sequences of the examined individuals in the study were compared with sequence data of populations in the SEA and the neighboring regions. A total of 1,036 mt sequences were subjected to the construction of phylogenetic tree via Mega software V5 (Tamura, et al., 2011). Other than 150 individuals each from the 3 studied Sabahan indigenous populations, the Reconstructed Sapiens Reference Sequence (RSRS) was included as the root of the phylogeny, as recommended by a recent publication (Behar, et al., 2012). In addition, 585 full mtDNAs were retrieved from the NCBI database (Gunnarsdottir, Li, et al., 2011; Gunnarsdottir, et al., 2011; Hill, et al., 2006; Ingman & Gyllensten, 2003; Jinam, et al., 2012; Loo, et al., 2011; Macaulay, et al., 2005; Peng, et al., 2010; Pierson, et al., 2006; Soares, et al., 2011; Soares, et al., 2008; Trejaut, et al., 2005). These included populations from East Asia (Taiwanese Aborigines, Han Chinese), Mainland SEA (Orang Asli from the Malay Peninsula, Vietnam, Thailand), ISEA (Philippines, Indonesia, Borneo), and Oceania (Polynesia, Melanesia, Micronesia, Australian Aborigines).

All mt sequences were prepared by trimming to the greatest common length of 737 nucleotides, spanning the 3 mt HV regions. Several parts of the sequence were removed, including the insertions after nps 16,193, 309, and 573, and the homopolymeric C-stretches in the HV regions (HV1: nps 16,180 to 16,193; HV2: nps 309 to 315).

Prior to the construction of phylogenetic tree, all prepared mt sequences were aligned via ClustalW algorithms. A NJ tree was built based on the Kimura 2-parameter nucleotide substitution model. Resampling of 1,000 replications was introduced for bootstrapping.

RESULTS

4 RESULTS

4.1 DNA extraction

During the entire course of study, a total of 639 DNA samples had been collected from indigenous individuals living in the state of Sabah. The collection comprised of samples from 3 of the largest indigenous tribes, i.e., Kadazan-Dusun, Bajau, and Rungus (271, 159, and 209 respectively). The isolated DNA samples were checked via spectrophotometry to ensure pure and high DNA yields. DNA samples of less desired quality were stored separately and DNA extraction was performed again using whole blood or buccal stocks. On the other hand, DNA samples of good quality were diluted to an appropriate concentration (50-200 ng/μl) for further genetic studies.

4.2 Nuclear DNA markers: VNTRs

In the genetic study of nuclear DNA markers, 2 VNTRs that are present in the DAT-1 gene were screened in all samples. The frequency data yielded in the experiments, in terms of allelic and genotypic distributions, for all populations was documented and used subsequently for computation of population parameters. The parameters were used to characterize the genetic structure of each population and evaluate their inter- and intra-relationships.

4.2.1 DAT-1 3'UTR 40-bp VNTR

The 40-bp repeats in the 3'UTR of DAT-1 gene were determined for all samples via direct PCR. Figure 4.1 displays the patterns of migration of the amplicon for all observed genotypes as viewed on a native agarose gel.

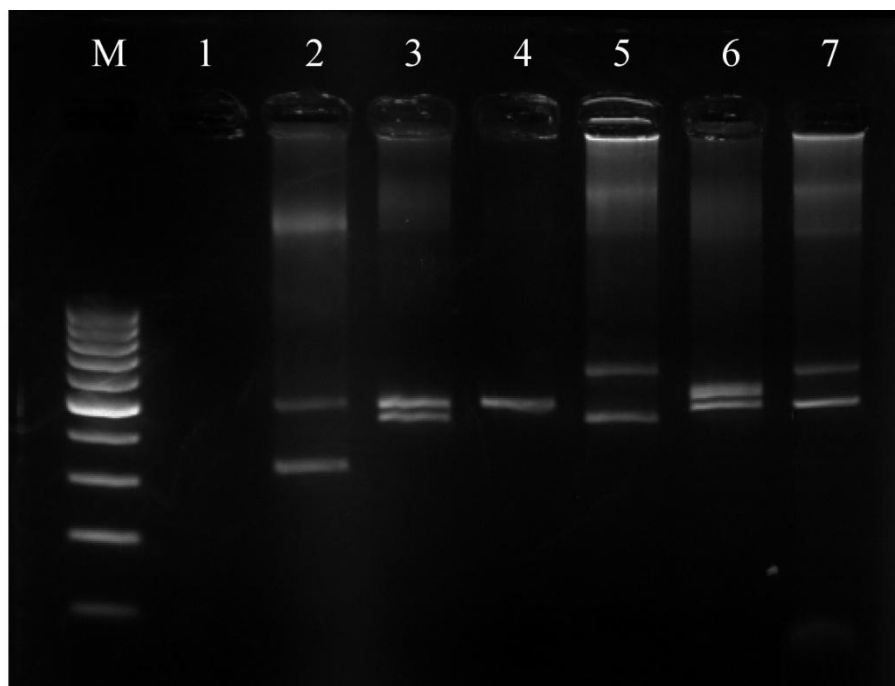


Figure 4.1 : Visualization of PCR amplicon of the DAT-1 3'UTR 40-bp VNTR on 3 % (w/v) native agarose gel stained with EtBr.

Lane M : 100 bp DNA marker

Lane 1 : Non-template control

Lane 2 : Heterozygous 7R/10R (360 bp/480 bp)

Lane 3 : Heterozygous 9R/10R (440 bp/480 bp)

Lane 4 : Homozygous 10R (480 bp)

Lane 5 : Heterozygous 9R/12R (440 bp/560 bp)

Lane 6 : Heterozygous 10R/11R (480 bp/520 bp)

Lane 7 : Heterozygous 10R/12R (480 bp/560 bp)

In the screening of DAT-1 3'UTR 40-bp VNTR, a total of 5 alleles were observed in the 3 Sabahan indigenous populations, i.e., 7R, 9R, 10R, 11R, and 12R (Table 4.1). The most prominently found allele in these populations was the 10-repeat (10R) variant. It present in more than 80 % of the individual: 97.8 %, 84.3 %, and 96.2 % in Kadazan-Dusun, Bajau, and Rungus, respectively. The 10R variant was especially abundant in Kadazan-Dusun and Rungus populations.

In terms of allele diversity, Bajau population harbored the highest number of variants when compared to the other 2 indigenous populations. We observed a total of 5 variants in the Bajau samples in our study. However, these variants only occurred in low frequencies of lesser than 2 % of the entire tested cohort, except for 9R and 10R variants. For Bajau population, the 9R variant present in 12.9 % of the population and was the second most predominant variant after the 10R variant. As for the Kadazan-Dusun group, there were only 3 variants observed, i.e., 7R, 9R, and 10R. Except for the predominant 10R, 7R and 9R variants present at relatively low frequencies. In the cohort of Rungus population, only 2 variants, 9R and 10R, were found.

For genotypic distribution, 6 genotypes were observed in the 3 populations. For Bajau population, all 6 genotypes were seen, where genotype 10R/10R was found to present at the highest frequency, 69.2 %. It follows by heterozygous 9R/10R, 25.2 % and the rest of the genotypes were less than 5 %. The Kadazan-Dusun and Rungus individuals contained 3 and 2, out of 6 observed, genotypes respectively. The homozygous 10R was seen in more than 92 % of the Kadazan-Dusun and Rungus individuals.

Table 4.1 : Distribution of alleles and genotypes of DAT-1 3'UTR 40-bp VNTR polymorphisms in Kadazan-Dusun, Bajau, and Rungus; predominant alleles and genotypes are highlighted in red.

Population	Kadazan-Dusun	Bajau	Rungus
Allelic frequency (%)			
7R	1 (0.2)	6 (1.9)	-
9R	11 (2.0)	41 (12.9)	16 (3.8)
10R	530 (97.8)	268 (84.3)	402 (96.2)
11R	-	1 (0.3)	-
12R	-	2 (0.6)	-
TOTAL (2n)	542	318	418
Genotypic frequency (%)			
7R/10R	1 (0.4)	6 (3.8)	-
9R/10R	11 (4.0)	40 (25.2)	16 (7.7)
10R/10R	259 (95.6)	110 (69.2)	193 (92.3)
9R/12R	-	1 (0.6)	-
10R/11R	-	1 (0.6)	-
10R/12R	-	1 (0.6)	-
TOTAL (n)	271	159	209

4.2.2 DAT-1 Intron-8 30-bp VNTR

The Intron-8 region of the DAT-1 gene, that contains a 30-bp VNTR, was amplified and separated on a 3 % (w/v) native agarose gel as shown in Figure 4.2.

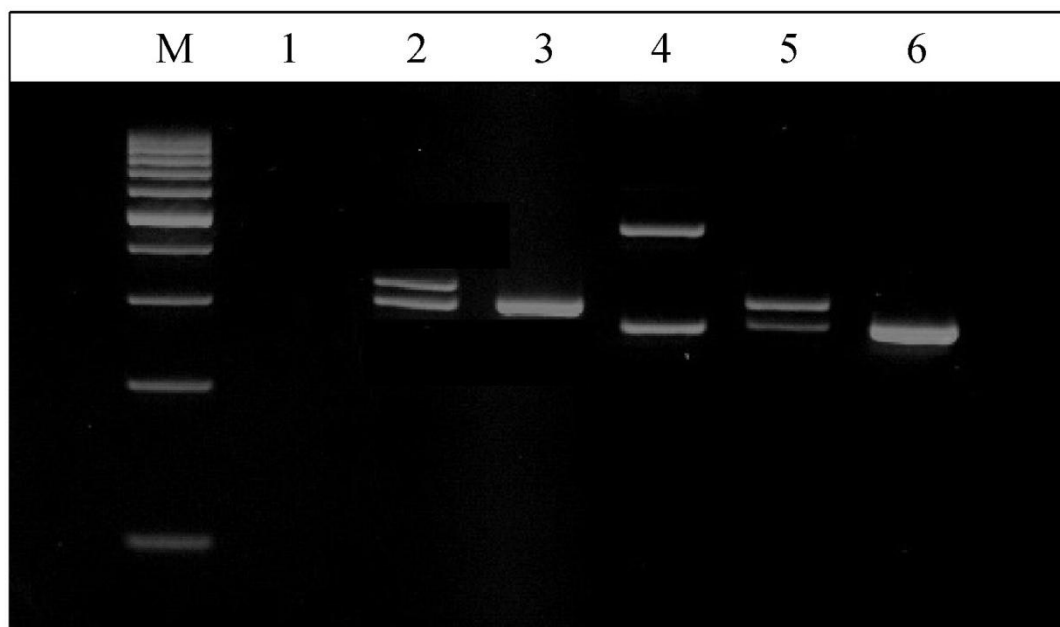


Figure 4.2 : Separation patterns of different genotypes of DAT-1 Intron-8 30-bp VNTR on 3 % (w/v) EtBr-stained native agarose gel.

Lane M : 100 bp DNA marker

Lane 1 : Non-template control

Lane 2 : Heterozygous 6R/7R (290 bp/320 bp)

Lane 3 : Homozygous 6R (290 bp)

Lane 4 : Heterozygous 5R/12R (260 bp/470 bp)

Lane 5 : Heterozygous 5R/6R (260 bp/290 bp)

Lane 6 : Homozygous 5R (260 bp)

In the examination of DAT-1 Intron-8 30-bp VNTR polymorphisms, 4 variants were seen in our sample cohort, i.e., 5R, 6R, 7R, and 12R (Table 4.2). The 6R variant was seen at the highest frequency in all the 3 populations. Amongst, Rungus individuals have the highest percentage of the variant, 91.4 %, followed by Kadazan-Dusun and Bajau (89.5 % and 79.3 %). The second most prominent variant was 5R, with highest frequency in Bajau population (20.4 %). The 7R variant was only seen in Kadazan-Dusun and Rungus groups (0.2 % and 0.5 %, respectively). Whereas, the largest repeat variant, 12R, was found solely in the Bajau population.

These variants have yielded 5 combinations of genotypes in all tested individuals. In general, the homozygous 6R was found predominantly in all populations. The percentage of homozygous 6R was even in Kadazan-Dusun and Rungus groups, approximately 80 %. Its frequency was lower in the Bajau group, 62.3 %. The second most prominent genotype was heterozygous 5R/6R. It was found with the highest frequency in the Bajau population, 34 %, followed by Kadazan-Dusun and Rungus, 15.5 % and 15.3 % respectively. The homozygous 5R was found in all 3 indigenous populations, but with low frequency of lesser than 5 %. On the other hand, like the allelic distribution, the heterozygous 6R/7R was only seen in Kadazan-Dusun and Rungus groups. In contrast, the heterozygous 5R/12R was found exclusively in the Bajau samples. However, these genotypes only present at extremely low percentage in the populations, i.e., 1 % and less.

Table 4.2 : Allelic and genotypic frequencies of DAT-1 Intron-8 30-bp VNTR polymorphisms in Kadazan-Dusun, Bajau, and Rungus; predominant alleles and genotypes are highlighted in red.

Population	Kadazan-Dusun	Bajau	Rungus
Allelic frequency (%)			
5R	56 (10.3)	65 (20.4)	34 (8.1)
6R	485 (89.5)	252 (79.3)	382 (91.4)
7R	1 (0.2)	-	2 (0.5)
12R	-	1 (0.3)	-
TOTAL (2n)	542	318	418
Genotypic frequency (%)			
5R/5R	7 (2.6)	5 (3.1)	1 (0.5)
5R/6R	42 (15.5)	54 (34.0)	32 (15.3)
5R/12R	-	1 (0.6)	-
6R/6R	221 (81.5)	99 (62.3)	174 (83.3)
6R/7R	1 (0.4)	-	2 (1.0)
TOTAL (n)	271	159	209

4.2.3 Forensic and population parameter evaluation

Several parameters, for both population characterization and correlation, were computed from the frequency data for the examined VNTRs (Table 4.3).

Table 4.3 : Forensic parameters and population differentiation study of 2 VNTRs in the DAT-1 gene (PD = Power of discrimination; PE = Power of exclusion; PIC = Polymorphism information content; TPI = Typical paternity index; HWE = Hardy-Weinberg Equation; H_T = Total gene diversity; H_S = Intra-population gene diversity; D_{ST} = Inter-population gene diversity; G_{ST} = Coefficient of gene differentiation; G_{IS} = Inbreeding coefficient).

Ethnicity	Forensic parameter	VNTR	
		3'UTR	Intron-8
Kadazan-Dusun	PD	0.085	0.310
	PE	0.002	0.020
	PIC	0.040	0.170
	TPI	0.520	0.590
	HWE	1.000	0.031
	Combined PD	0.369	
	Combined PE	0.022	
Bajau	PD	0.457	0.496
	PE	0.067	0.084
	PIC	0.250	0.280
	TPI	0.720	0.600
	HWE	0.196	0.169
	Combined PD	0.726	
	Combined PE	0.145	
Rungus	PD	0.141	0.283
	PE	0.005	0.020
	PIC	0.070	0.150
	TPI	0.540	0.600
	HWE	1.000	1.000
	Combined PD	0.384	
	Combined PE	0.025	
Population differentiation	Average	VNTR	
		3'UTR	Intron-8
H_T	0.184	0.136	0.232
H_S	0.178	0.130	0.226
D_{ST}	0.006	0.006	0.005
G_{ST}	0.030	0.043	0.023
G_{IS}	0.025	0.098	0.017
AMOVA			
Among population	4.241 %	6.454 %	3.058 %
within population	95.760 %	93.546 %	96.942 %
<i>P</i>	0.0000	0.0000	0.0000

For the power of discrimination of both markers in the Kadazan-Dusun population, the Intron-8 VNTR showed higher capacity than the 3'UTR VNTR, 0.310 vs 0.085. On the other hand, the power of exclusion computed was 0.002 (3'UTR) and 0.020 (Intron-8). The combined PD for the 2 VNTRs was 0.369 and the combined PE was 0.022. The PIC and TPI values for the 3'UTR VNTR were 0.040 and 0.520. On the other hand, the PIC and TPI values for the Intron-8 VNTR were 0.170 and 0.590. The distribution of Intron-8 VNTR, however, was found to deviate from HWE ($P = 0.031$).

In the Bajau population, the power of discrimination for both markers was even, i.e., 0.457 (3'UTR) and 0.496 (Intron-8). The power of exclusion for the 3'UTR VNTR was 0.067, whereas for the Intron-8 VNTR was 0.084. The combined power of discrimination and exclusion for both markers in the Bajau population were 0.726 and 0.145, correspondingly. The PIC value for both VNTRs was similar, 0.250 and 0.280. However, the 3'UTR VNTR had a higher TPI value than the Intron-8 VNTR (0.720 vs 0.600). For the Bajau group, distribution of the 2 VNTRs was observed to be in HW equilibrium.

For the distribution in the Rungus population, the power of discrimination differed from 0.141 (3'UTR) to 0.283 (Intron-8). The power of exclusion for both VNTRs was 0.005 and 0.020, for 3'UTR and Intron-8 VNTR, respectively. The combined PD and PE were 0.384 and 0.025. The PIC and TPI values for the 3'UTR VNTR were 0.070 and 0.540, whereas for Intron-8 VNTR were 0.150 and 0.600. The distribution of these markers in the Rungus group was also found to follow the HWE.

In the analysis of population differentiation of the 2 VNTRs in the DAT-1 gene, the Intron-8 VNTR was observed to be more diverse ($H_T = 0.232$) as opposed to the 3'UTR VNTR ($H_T = 0.136$). Majority of the variation observed in these markers were accounted by difference within population ($H_S = 0.130$; $H_S = 0.226$). The inter-

population variation was found to contribute to only 2 % and 4 % of the total observed gene diversity for these VNTRs, for Intron-8 and 3'UTR, respectively.

The AMOVA results also showed that a large proportion of the genetic variation come from within these populations, of greater than 93 %. It also indicated that the 3'UTR VNTR has higher percentage of inter-population variation (6.45 %) than the Intron-8 VNTR (3.06 %). On average, 95.76 % of the observed variation was originated within the populations, whereas 4.24 % was inter-population differences.

4.3 Nuclear DNA markers: InDels

A total of 1,278 chromosomes were successfully screened for the presence or absence of *Alu* elements in 6 nuclear DNA regions. The presence of *Alu* element was represented by the amplification of a longer product, known as “+” allele, whereas the absence of the element resulted in a shorter fragment, named the “-” allele. Heterozygotes consisted of both alleles. Figures 4.3 and 4.4 show the separation of the long and short fragments on a native 2 % (w/v) EtBr-stained agarose gel.

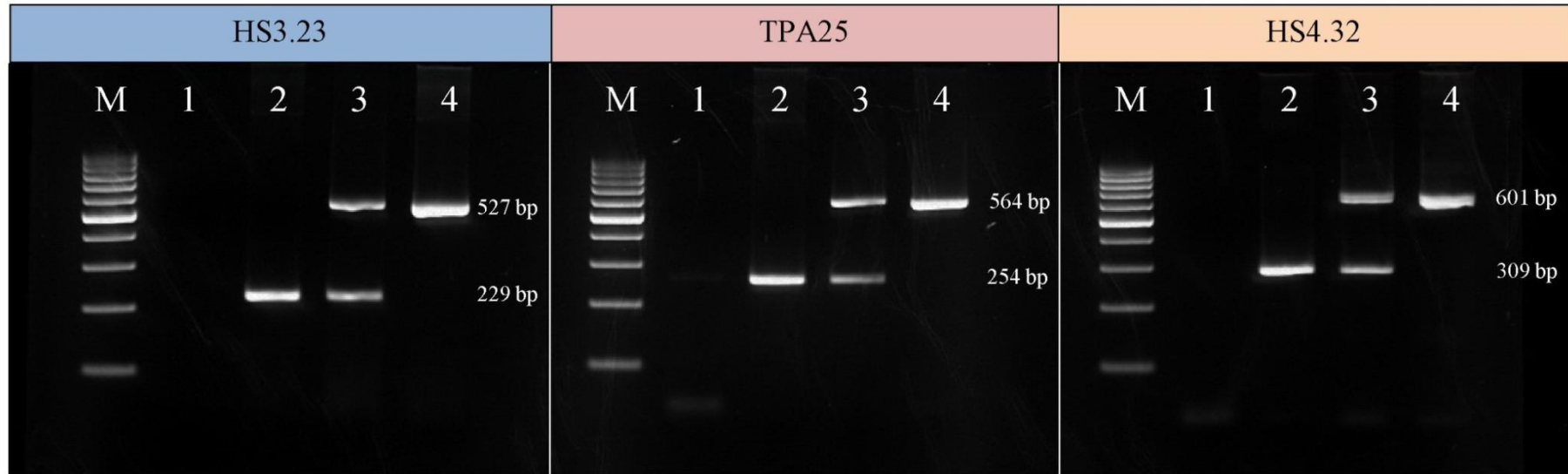


Figure 4.3 : Band patterns of various genotypes for HS3.23, TPA25, and HS4.32 on 2 % (w/v) native agarose gel, pre-stained with EtBr; the presence of an inserted *Alu* element was indicated by the long (+) allele which is approximately 300 bp larger than the short (-) allele.

Lane M : 100 bp DNA marker

Lane 1 : Non-template control

Lane 2 : Homozygous -/-

Lane 3 : Heterozygous -/+

Lane 4 : Homozygous +/+

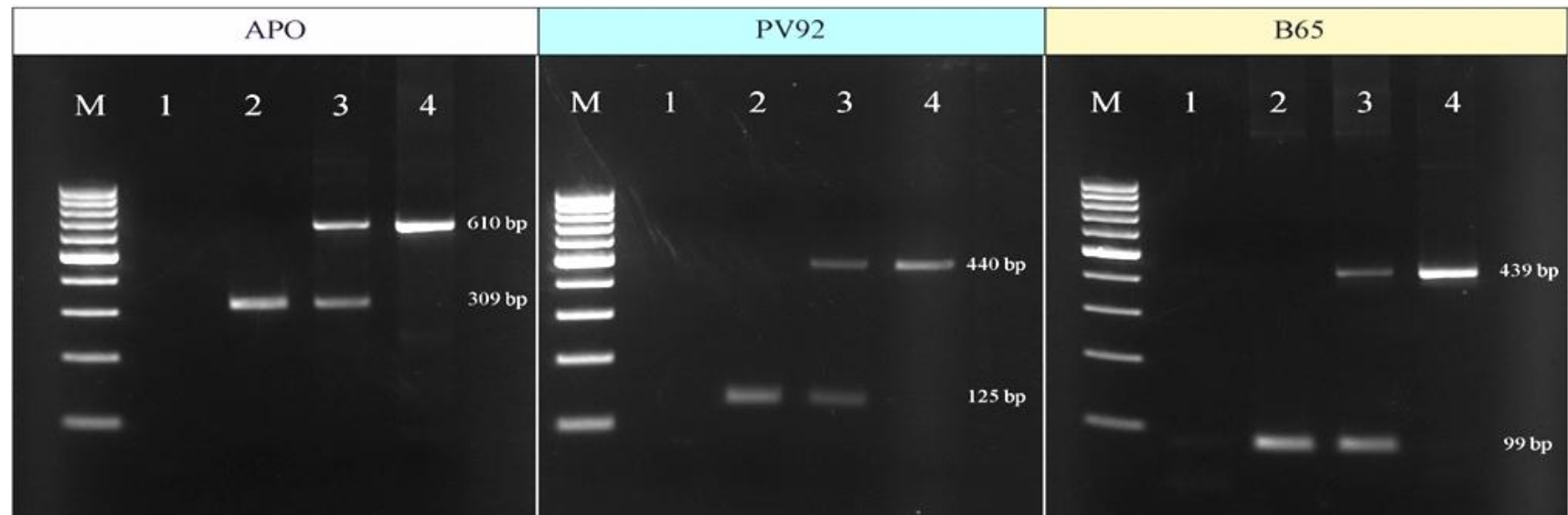


Figure 4.4 : Band patterns of various genotypes for APO, PV92, and B65 on 2 % (w/v) native agarose gel, pre-stained with EtBr; the presence of an inserted *Alu* element was indicated by the long (+) allele which is approximately 300 bp larger than the short (-) allele.

Lane M : 100 bp DNA marker

Lane 1 : Non-template control

Lane 2 : Homozygous -/-

Lane 3 : Heterozygous +/-

Lane 4 : Homozygous +/+

4.3.1 InDels: Distribution among indigenous populations

The frequencies of *Alu* insertions are outlined in Table 4.4. The highest insertion frequency was observed in the HS3.23 marker in the Rungus population, whereas the lowest frequency was found in the HS4.32 of the Kadazan-Dusun group. None of these *Alu* insertions was found to completely fix to any of the populations, despite the fact that 2 markers, APO and HS3.23, were present with high insertion frequencies, of > 0.95 . On the contrary, the HS4.32 marker was detected at the lowest frequencies in all the 3 populations, with an average of 0.2723.

For the genotypic distribution, the most common genotype (0.9378) was the homozygous insertion (+/+) for the HS3.23 marker in Rungus population. On the contrary, 2 genotypes were found to be absent in some cohorts in our study; the homozygous deletion (-/-) for the HS3.23 insertion was not observed in the Bajau population, while the homozygous deletion (-/-) for the APO was absent in the Rungus.

All populations showed high levels of genetic diversity for most of the *Alu* markers. The average heterozygosities were high in TPA25, B65, HS4.32, and PV92 insertions, 0.4869, 0.4709, 0.3640, and 0.3557 respectively, with the highest attainable heterozygosity for a biallelic markers being 0.5. However, the heterozygosity for HS4.32 and APO markers was low, with an average of < 0.1 .

Table 4.4 : Allelic and genotypic distributions of 6 *Alu* insertions in genotyped samples from Kadazan-Dusun, Bajau, and Rungus, Sabah.

Marker	Allele/ genotype	Population		
		Kadazan-Dusun	Bajau	Rungus
Allelic frequency				
HS4.32	+	0.2066	0.3113	0.2990
	-	0.7934	0.6887	0.7010
TPA25	+	0.5351	0.4843	0.6053
	-	0.4649	0.5157	0.3947
HS3.23	+	0.9465	0.9371	0.9665
	-	0.0535	0.0629	0.0335
PV92	+	0.7269	0.7358	0.7201
	-	0.2731	0.2642	0.2799
B65	+	0.5221	0.5063	0.4833
	-	0.4779	0.4937	0.5167
APO	+	0.9649	0.9371	0.9641
	-	0.0351	0.0629	0.0359
Genotypic frequency				
HS4.32	+/+	0.0590	0.1258	0.0861
	+/-	0.2952	0.3711	0.4258
	-/-	0.6458	0.5031	0.4880
TPA25	+/+	0.2952	0.2642	0.3349
	+/-	0.4797	0.4403	0.5407
	-/-	0.2251	0.2956	0.1244
HS3.23	+/+	0.8967	0.8742	0.9378
	+/-	0.0996	0.1258	0.0574
	-/-	0.0037	0.0000	0.0048
PV92	+/+	0.5461	0.5912	0.5120
	+/-	0.3616	0.2893	0.4163
	-/-	0.0923	0.1195	0.0718
B65	+/+	0.2878	0.2830	0.2344
	+/-	0.4686	0.4465	0.4976
	-/-	0.2435	0.2704	0.2679
APO	+/+	0.9336	0.8805	0.9282
	+/-	0.0627	0.1132	0.0718
	-/-	0.0037	0.0063	0.0000

4.3.2 InDels: Marker evaluation

In order to evaluate the efficiency of these loci to be employed as genetic markers in the examined populations, several parameters need to be computed based on the distribution patterns (Table 4.5).

Table 4.5 : Power of discrimination (PD), Power of exclusion (PE), Polymorphism information content (PIC), Typical paternity index (TPI), and deviation from Hardy-Weinberg Equation (HWE) of *Alu* markers in the Sabahan indigenous populations.

Ethnic group	Forensic parameter	<i>Alu</i> insertion					
		HS4.32	TPA25	HS3.23	PV92	B65	APO
Kadazan-Dusun (n = 271)	PD	0.4924	0.6321	0.1860	0.5625	0.6382	0.1245
	PE	0.0616	0.1704	0.0083	0.0922	0.1615	0.0035
	PIC	0.2741	0.3738	0.0962	0.3182	0.3745	0.0654
	TPI	0.7094	0.9610	0.5553	0.7832	0.9410	0.5335
	HWE	0.1001	0.5441	0.5458	0.1651	0.3320	0.2785
Bajau (n = 159)	PD	0.5933	0.6490	0.2199	0.5525	0.6474	0.2119
	PE	0.0973	0.1404	0.0128	0.0592	0.1449	0.0105
	PIC	0.3369	0.3748	0.1109	0.3132	0.3750	0.1109
	TPI	0.7950	0.8933	0.5719	0.7035	0.9034	0.5638
	HWE	0.0968	0.1512	1.0000	0.0018	0.2037	0.4732
Rungus (n = 209)	PD	0.5731	0.5800	0.1172	0.5595	0.6256	0.1332
	PE	0.1304	0.2256	0.0030	0.1241	0.1854	0.0045
	PIC	0.3314	0.3637	0.0626	0.3219	0.3747	0.0668
	TPI	0.8708	1.0885	0.5305	0.8566	0.9952	0.5387
	HWE	0.8707	0.0820	0.2026	0.7323	1.0000	1.0000

In the 18 HWE tests carried out for all markers in the indigenous populations, only 1 deviation from the equation was observed (PV92 insertion in Bajau population). The deviation was assumed as a random case of statistical fluctuation, as it was not observed in other populations. Furthermore, the Bajau population did not show incompliance to the HWE for other *Alu* markers.

The values for power of discrimination (PD) varied greatly, from 0.1565 (average for APO insertion) to 0.6371 (average for B65 insertion). The highest PD was found in TPA25 insertion in Bajau population, whereas the lowest PD in the HS3.23 marker in Rungus individuals. HS4.32, TPA25, PV92, and B65 insertions were of high PD (> 0.5). However, 2 markers were observed to have average PDs of < 0.2 , i.e., HS3.23 and APO insertions. In order to access the joint discriminating power of these markers, the combined PD was calculated. The values were 0.9789, 0.9862, and 0.9774 for Kadazan-Dusun, Bajau, and Rungus populations, respectively. When expressing in matching probability (MP), the combined power of these *Alu* markers were 1 in 47, 72, and 44 individuals for Kadazan-Dusun, Bajau, and Rungus, respectively.

The power of exclusion (PE) ranged from 0.003 to 0.2256. The highest PE was observed in TPA25 insertion in the Rungus population, and the lowest was found in HS3.23 insertion, in the Rungus population as well. The combined PEs for all *Alu* markers were 0.4144, 0.3902, and 0.5232 for Kadazan-Dusun, Bajau, and Rungus, respectively. The polymorphism information content (PIC) differed from 0.0626 to 0.3750. All markers have high typical paternity index (TPI), of no less than 0.7, except for HS3.23 and APO insertion (~ 0.5).

4.3.3 InDels: Population differentiation analysis

Although Kadazan-Dusun, Bajau, and Rungus are 3 indigenous groups that have lived in close proximities in the state of Sabah at least for the past few decades, they appear to have distinctive characteristics, be it cultural or linguistic, that are unique to their own group. In order to access the degree of differences in the genetic structure of these populations, statistical tests had been employed to evaluate their association (Table 4.6).

Table 4.6 : Insertion frequency, population differentiation analysis, and AMOVA of 6 *Alu* markers in Kadazan-Dusun, Bajau, and Rungus populations.

<i>Alu</i> insertion	HS4.32	TPA25	HS3.23	PV92	B65	APO
Population						
Kadazan-Dusun (n = 271)	0.207	0.535	0.947	0.727	0.522	0.965
Bajau (n = 159)	0.311	0.484	0.937	0.736	0.506	0.937
Rungus (n = 209)	0.299	0.605	0.967	0.720	0.483	0.964
Differentiation analysis						
H_T	0.397	0.497	0.095	0.397	0.500	0.085
H_S	0.393	0.493	0.095	0.397	0.500	0.085
D_{ST}	0.004	0.004	0.000	0.000	0.000	0.000
G_{ST}	0.009	0.008	0.002	0.002	0.001	0.002
G_{IS}	0.074	0.012	0.006	0.105	0.060	0.030
AMOVA differentiation						
Among populations	1.616	1.086	0.203	0.000	0.000	0.295
Within populations	98.384	98.914	99.797	100.000	100.000	99.705
P	0.001	0.003	0.139	0.893	0.497	0.125

H_T : Total gene diversity

H_S : Intra-population gene diversity

D_{ST} : Inter-population gene diversity

G_{ST} : Coefficient of gene differentiation

G_{IS} : Inbreeding coefficient

H_T is accounted for by 2 fractions of genomic variability that arise within and among populations, which are represented by H_S and D_{ST} . The most diverse *Alu* insertion was B65, with H_T value of 0.5. On the other hand, the least diverse marker was APO, with H_T value of 0.085. Among the 6 examined *Alu* insertions, only 2 (HS4.32 and TPA25) were observed to exhibit gene differences between the 3 indigenous populations. However, the D_{ST} value was low, 0.04 or 0.8 % to 1 % of the H_T values. G_{ST} varied from 0.001 (B65 insertion) to 0.009 (HS4.32 insertion). Generally, these insertions showed low levels of differentiation, with an average G_{ST} value of 0.004. The highest G_{IS} was observed in PV92 insertion ($G_{IS} = 0.105$). The average G_{IS} value was 0.048.

In the AMOVA evaluation, 2 markers were shown to present with significant genetic variations among populations, i.e., HS4.32 (1.616 %; $P < 0.005$) and TPA25 (1.086 %; $P < 0.005$). In contrast, the other insertions were found to present with less than 0.3 %. The average fraction that is contributed by inter-population variations, as computed by AMOVA, for all insertions was 0.5 %. In other words, 99.5 % of variations observed in these populations were attributed within individuals of the populations.

4.3.4 InDels: Principal component analysis

Other than investigating the inter-relationship of the Sabahan indigenous groups, statistical tests were also conducted to establish the genetic linkage of these groups to other populations. PC plots were constructed based on the frequencies of *Alu* insertions (Appendix 4). Frequency data was obtained from the online database - ALFRED. The database is developed and maintained by researchers of Yale University (Rajeevan, et al., 2005). Genetic data was compiled to form a comprehensive collection, covering a wide range of genetic polymorphisms such as SNPs, VNTRs, STRs, InDels, etc. The data in ALFRED is contributed by researchers from both their published and unpublished works. On the other hand, some *Alu* frequencies that have been included in the PCA were adopted from reports of *Alu* insertions in other populations (Mamedov, et al., 2010; Watkins, et al., 2003).

4.3.4.1 PCA of world populations

PC plot was constructed with insertion frequencies of 5 *Alu* markers (B65, TPA25, PV92, HS4.32, and APO) (Figure 4.5). A total of 34 global populations allocated into 5 major groups that represent their geographical localities, i.e., African, European, SEA or East Asian, South Asian, and the Sabahan indigenous groups, were included in the analysis.

The 2 main PCs represented a total of 64.64 % of the entire variation. The first component (D1) contributed to 39.57 % and the second component (D2) scored 25.07 %. Other than the main body of the PC plot, the biplot (top left of Figure 4.5) revealed constructive information on how the “parameters” (markers) affect the dispersal of data points across the plot.

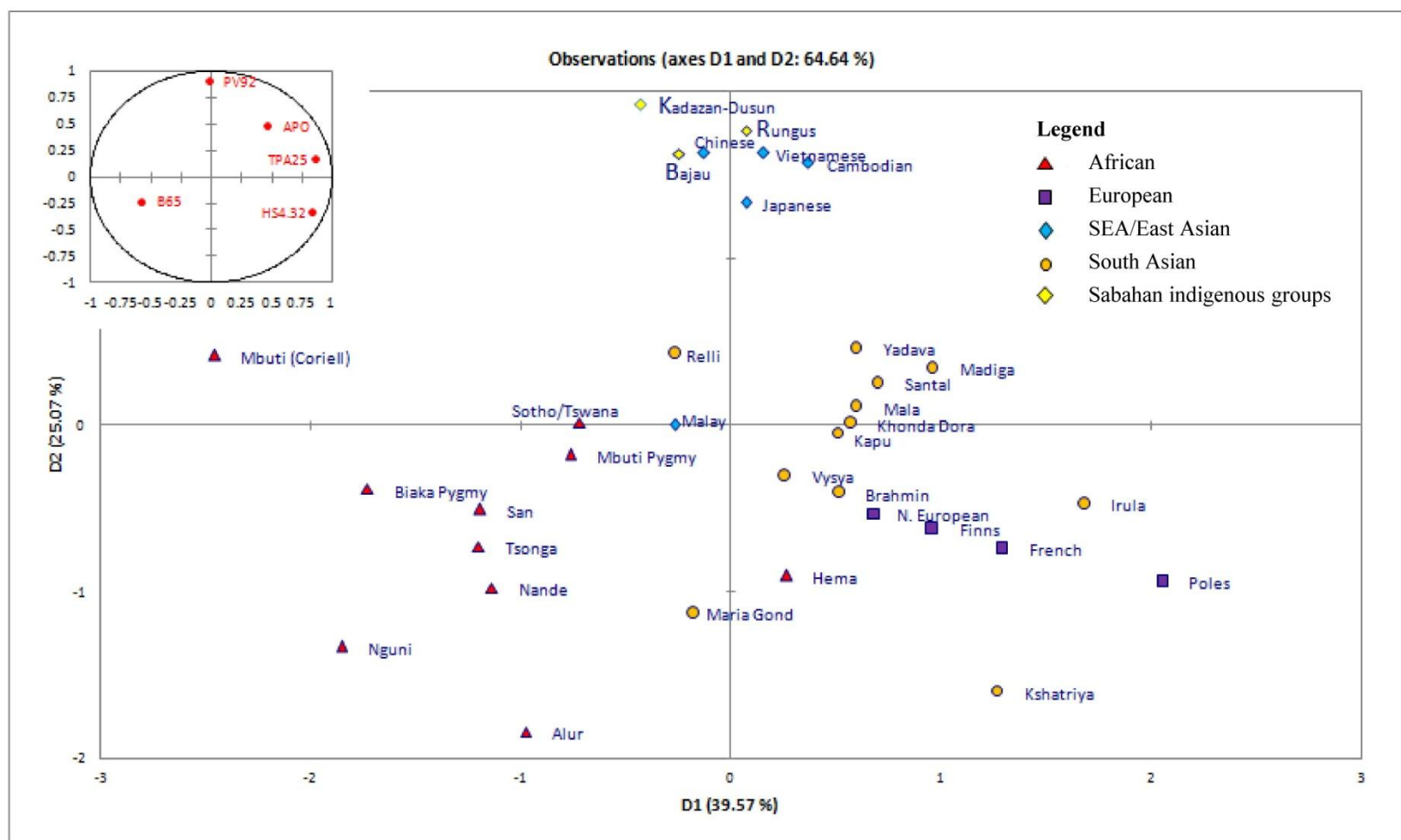


Figure 4.5 : PC plot based on *Alu* insertion frequencies of the 3 Sabahan indigenous and world populations; on the top left is the biplot diagram of the PCA.

From the distribution of data points on the PCA, all African populations, except Hema group, can be found on the left side of the plot, concentrating at the lower end. Based on the biplot, the distribution of African groups was mainly contributed by the B65 insertion. On the other hand, the European populations were located at the lower corner on the right side of the plot. The South Asian populations were positioned slightly top of the European populations, spreading across to the right of the midline. The dispersal of the South Asian populations was the sum effect of 2 markers, i.e., TPA25 and HS4.32 insertions. Interestingly, the SEA and East Asian (SEA/EA) populations split away from other world populations. The SEA/EA populations huddled closely together with the Kadazan-Dusun, Bajau, and Rungus groups in our study, in the centre at the top of the PC plot. As shown on the biplot, the distinctive factor for the cluster was primarily due to the high insertion frequency of the PV92 loci in both the SEA/EA and Sabahan indigenous populations. Worthy to note, an outlier to this cluster has been observed, whereby the Malay population was not seen within the SEA/EA cluster, but at the mid-point of the African and South Asian clusters.

4.3.4.2 PCA of populations in the neighboring regions

The first PCA of the Sabahan indigenous groups against other world populations showed that the 3 groups displayed strong affinity to Asian populations, specifically the East Asian and other SEA groups, signaling a higher degree of genetic resemblance. We thus narrowed the search to focus on the genetic trends in populations of neighboring regions. Insertion frequencies of *Alu* markers for as many East Asian and SEA populations were collected from various publications and databases. Due to the limited published data on these populations, PCA was executed based on 4 *Alu* insertions (B65, APO, TPA25, and PV92).

Data from 20 populations were included in the PCA. These populations were divided into 5 fractions, i.e., Taiwanese Aborigines, Mainland SEA, ISEA, East Asian, and Sabahan indigenous groups (Figure 4.6).

The 2 main PCs in the plot consisted of 62.91 % of the variation. The first and second components attributed to 37.64 % and 25.27 % of the differences, respectively. The first dimension separated the ISEA populations from the rest. All ISEA populations, except Filipino, were situated to the left of the PC plot. Majority of the Mainland SEA, East Asian, and Taiwanese Aborigines fitted at the lower right corner. Again, all 3 Sabahan indigenous groups were observed to cluster together at the upper right of the plot. The Han (China, East Asia) and Filipino (ISEA) groups were seen in close proximity to the Sabahan indigenous groups. The Ami tribe, a Taiwanese Aboriginal group, was located at the far end to the top right, signifying its genetic difference from other groups.

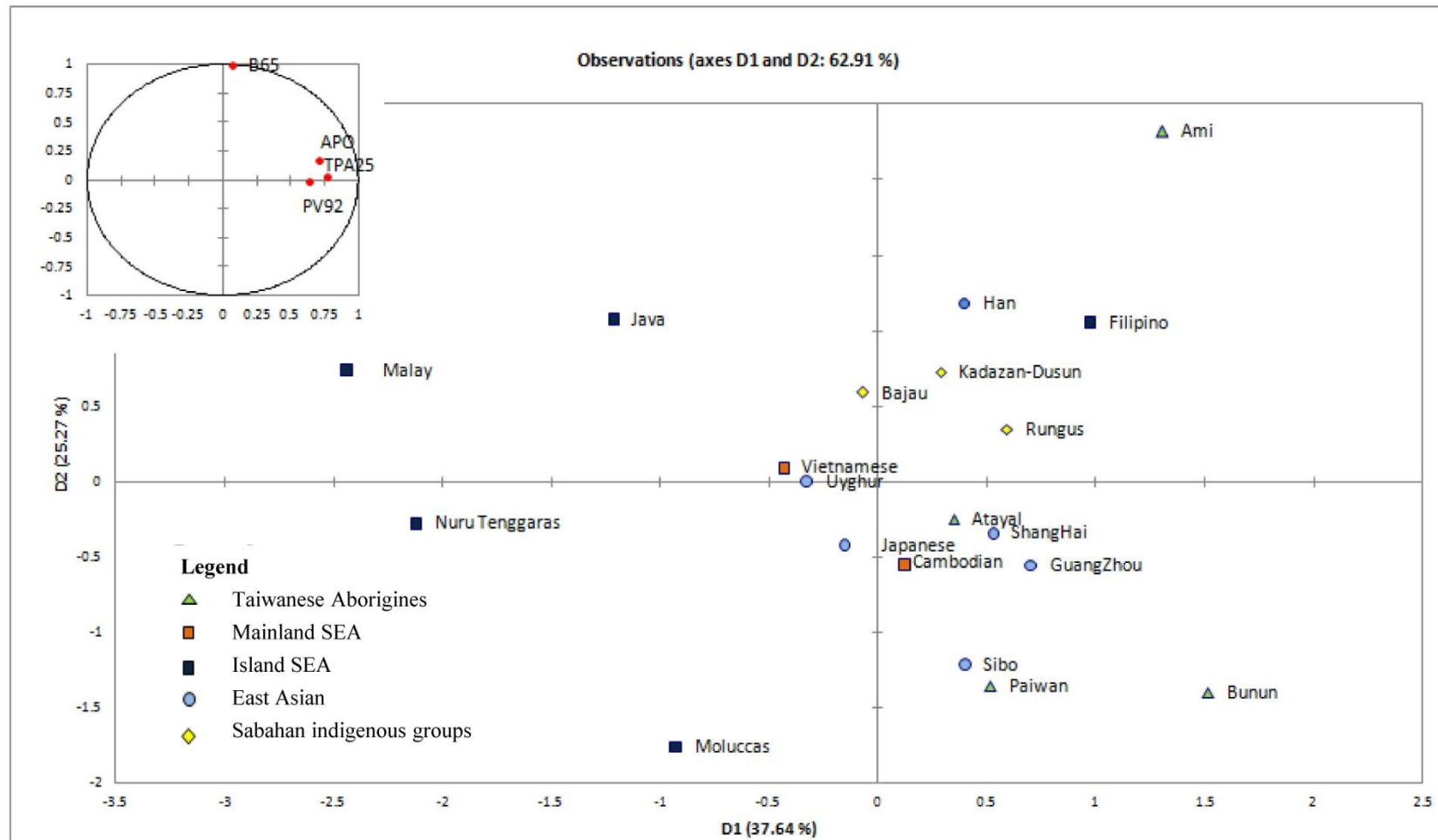


Figure 4.6 : Diagram showing PCA and biplot constructed using insertion frequencies of 4 *Alu* markers of the Sabahan indigenous groups and populations in the neighboring regions.

4.3.5 InDels: Phylogenetic assessment

The genetic association of populations can be represented diagrammatically by incorporating their genetic information into a phylogenetic exercise. The “distance” method was employed in the assessment based on the insertion frequencies of *Alu* markers in various populations. Genetic distance was calculated for each population pair as the core material for building of the phylogenetic tree. The same cohort of samples has been used to examine inter-population association of the Sabahan indigenous groups in both global and neighboring contexts. A total of 10,000 random replications of data were conducted for each data set for bootstrapping purpose. The phylogenetic trees were built by using the NJ algorithm.

4.3.5.1 Phylogenetic assessment with world populations

In the unrooted NJ tree, African groups exhibited similar levels of genetic differentiation from other populations, forming a cluster at one end. Along the tree, the tested populations were subsequently split into “South Asia/Europe” and “East Asia/SEA” clusters. In the former cluster, South Asia groups branched off earlier and Europeans were found further into the cluster. In the East Asia/SEA cluster, East Asia and the Sabahan indigenous populations were observed to diverge from South Asian. The Sabahan indigenous populations were seen to form a distinct clade from East Asian and Mainland SEA populations. In the bootstrap resampling method, the values varied greatly from 7.6 % to 68 %. Higher bootstrap values were observed in the African and East Asia/SEA populations (Figure 4.7).

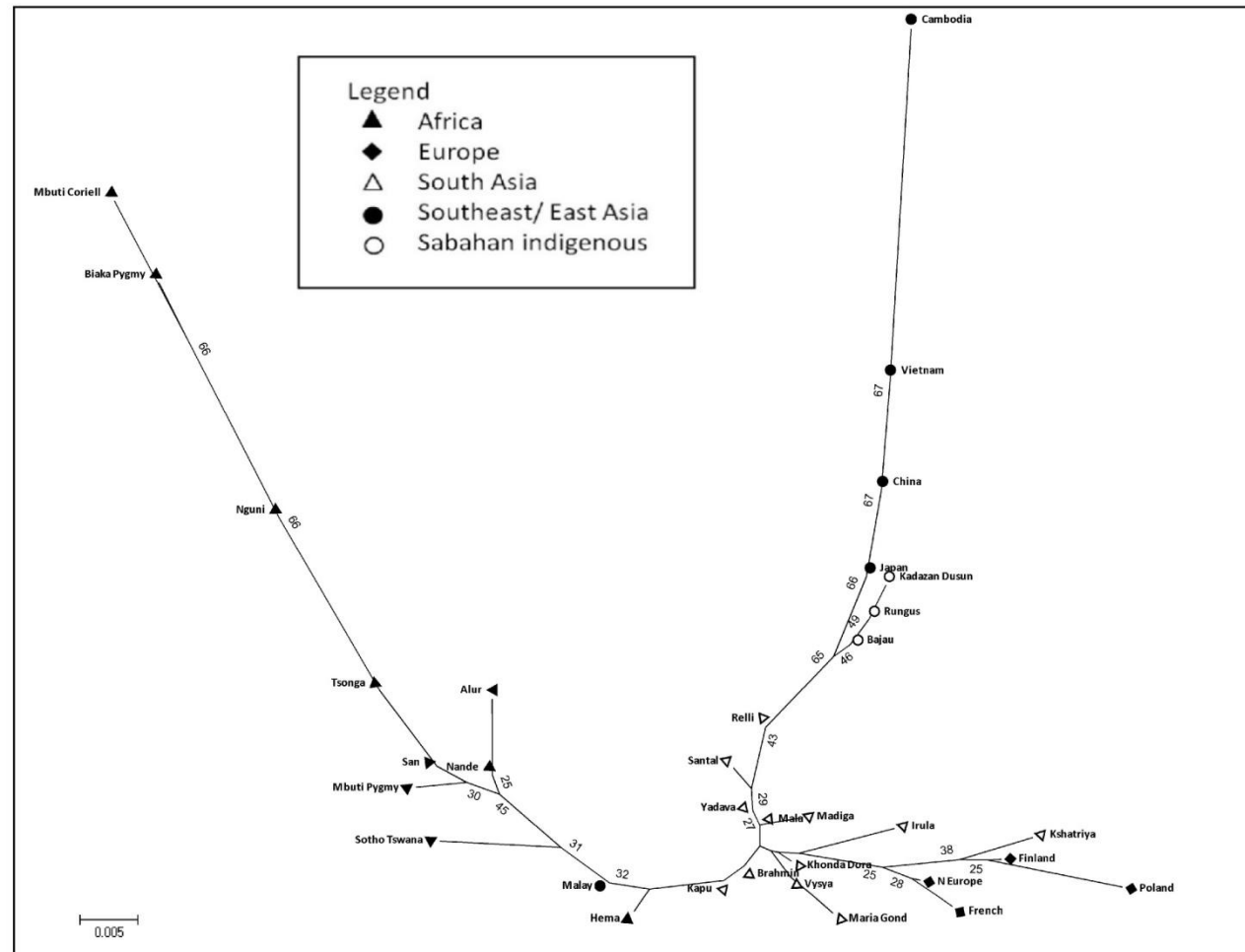


Figure 4.7 : Radiated and unrooted NJ tree constructed based on insertion frequencies of *Alu* markers from the world populations; number at the node is the confidence bootstrap value for the branch, only values higher than 25 % were shown.

4.3.5.2 Phylogenetic assessment in neighboring regions

A total of 3 distinct clusters were observed in the unrooted NJ phylogenetic tree for populations in SEA and its neighboring regions (Figure 4.8). The first cluster comprised mainly of East Asian and Mainland SEA population, i.e., Han, Sibon, Japanese, Vietnamese, and Cambodian, apart from the Atayal and Bunun groups, Taiwanese Aborigines, as well as Javanese. The second cluster consisted of ISEA/Pacific populations (Nusa Tenggara, Malay, and Moluccas). The Bajau were also included in this cluster, though they were placed very near to the divergence point of the cluster. The last cluster is made up of Kadazan-Dusun, Rungus, Philippines, and Ami group. The bootstrap values varied from 9.1 % to 68 %.

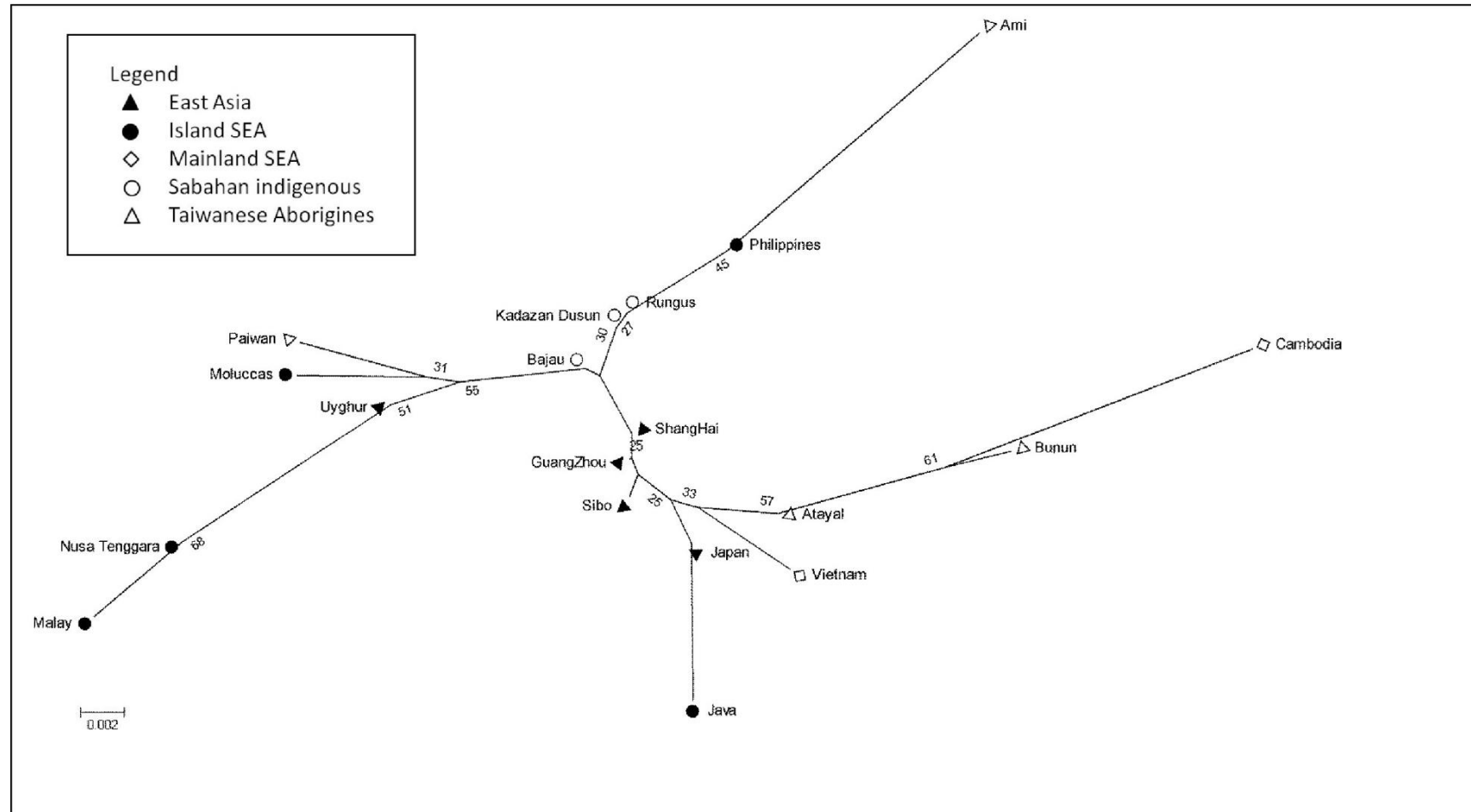


Figure 4.8 : Unrooted phylogenetic tree depicts relationships of Kadazan-Dusun, Bajau, and Rungus with their neighboring populations; bootstrap values, higher than 25 %, were displayed at the node of each branch.

4.4 Examination of autosomal STR markers

4.4.1 STR distribution in the Sabahan indigenous populations

The amplification of 15 STR loci and AMEL marker for all DNA samples from Kadazan-Dusun, Bajau, and Rungus was successful using the Promega Powerplex 16 system. The amplified fragments were subsequently resolved on a genetic analyzer and sizes for all amplicons were determined. Allelic calling was carried out using a bioinformatic software, specialized for fragment recognition – Genemapper ID V3.2. Alleles for each STR loci for every sample were assigned in accordance to their separation on the genetic analyzer, in relative to the sizing standard - ILS600 and allelic ladders. Quality checks were performed by verifying the STR profile generated by the known positive control (9947A or 2800M DNAs) that was included in each round of amplification. Unsatisfactory and suspicious samples, such as those with odd alleles (off-ladder), were re-amplified and subjected to another round of fragment analysis.

Each allele of varying lengths (from 106 bp to 474 bp) gave rise to peaks that appeared either in blue, green, or black colors in correspondence to the dye that was incorporated at 1 of the primers utilized.

Figures 4.9 to 4.11 illustrate the electropherograms for allelic ladders (upper lane) and positive controls (lower lane), captured under Fluorescein (FL), Carboxy-tetramethylrhodamine (TMR), and 6-carboxy-4',5'-dichloro-2',7'-dimethoxyfluorescein (JOE) dye channels.

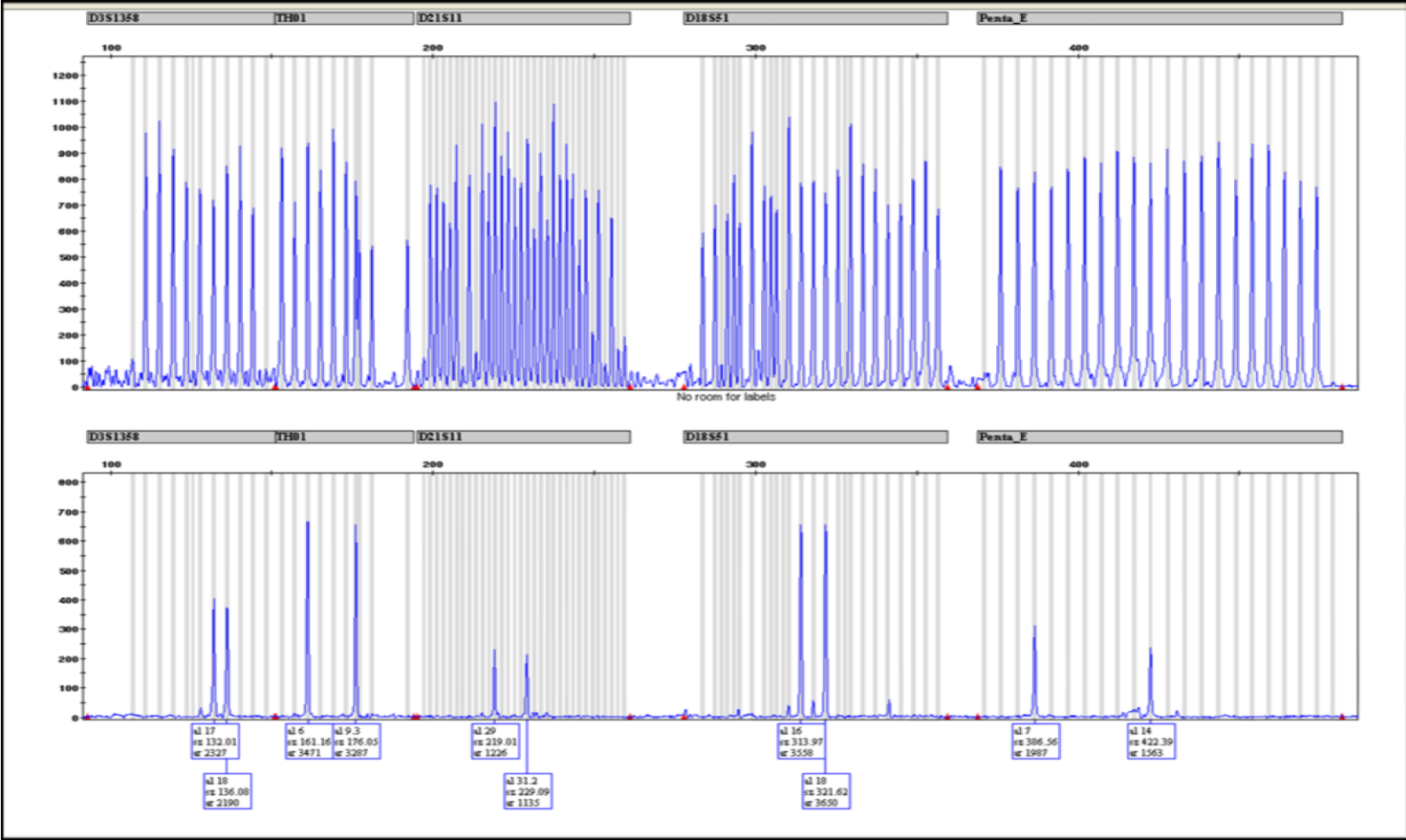


Figure 4.9 : Diagram showing the electropherogram of fluorescein-labeled STR loci (D3S1358, TH01, D21S11, D18S51, and Penta E).

Upper lane : Allelic ladder components

Lower lane : Control reaction of 2800M

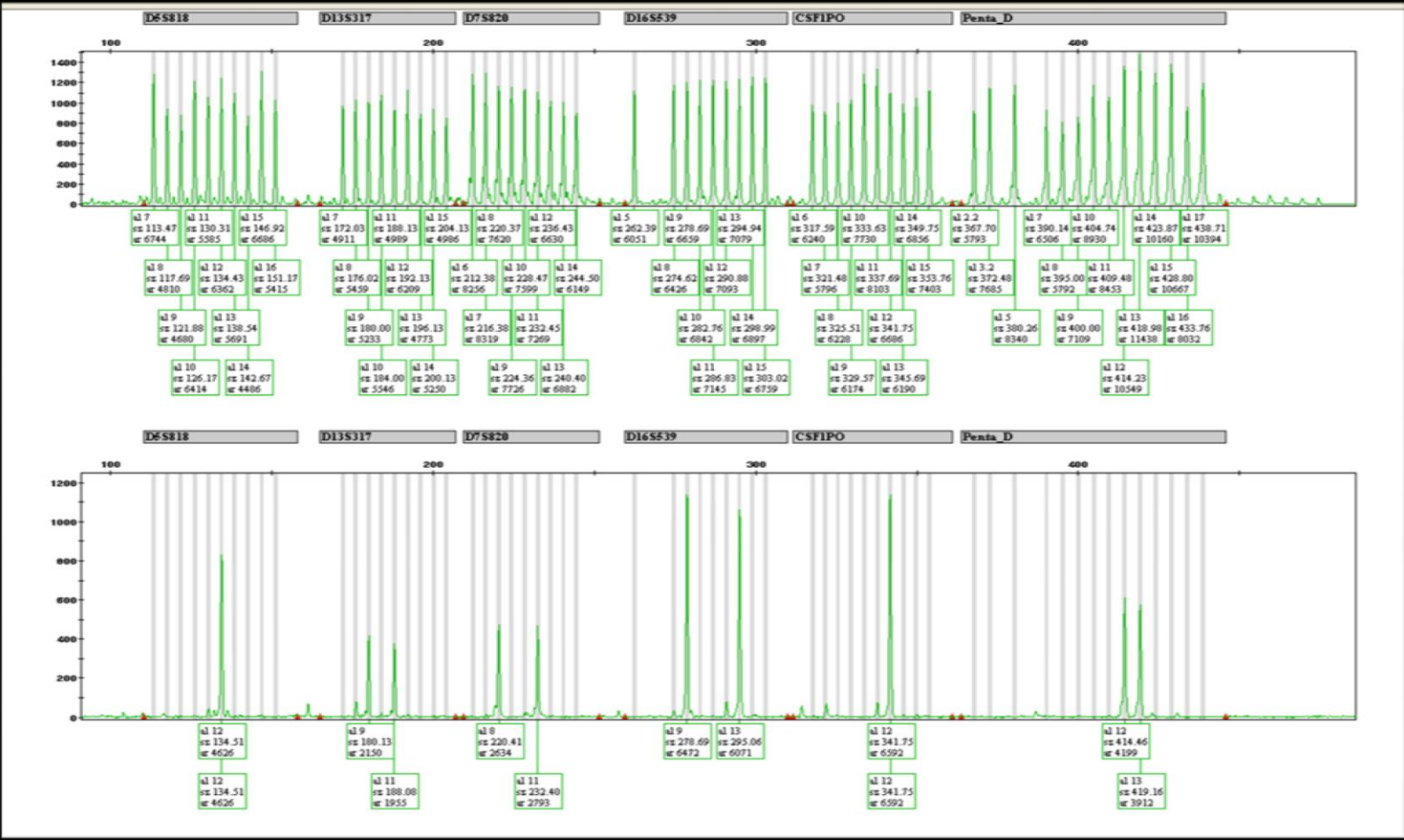


Figure 4.10 : Diagram showing the electropherogram of JOE-labeled STR loci (D5S818, D13S317, D7S820, D16S539, CSF1PO, and Penta D).

Upper lane : Allelic ladder components

Lower lane : Control reaction of 2800M

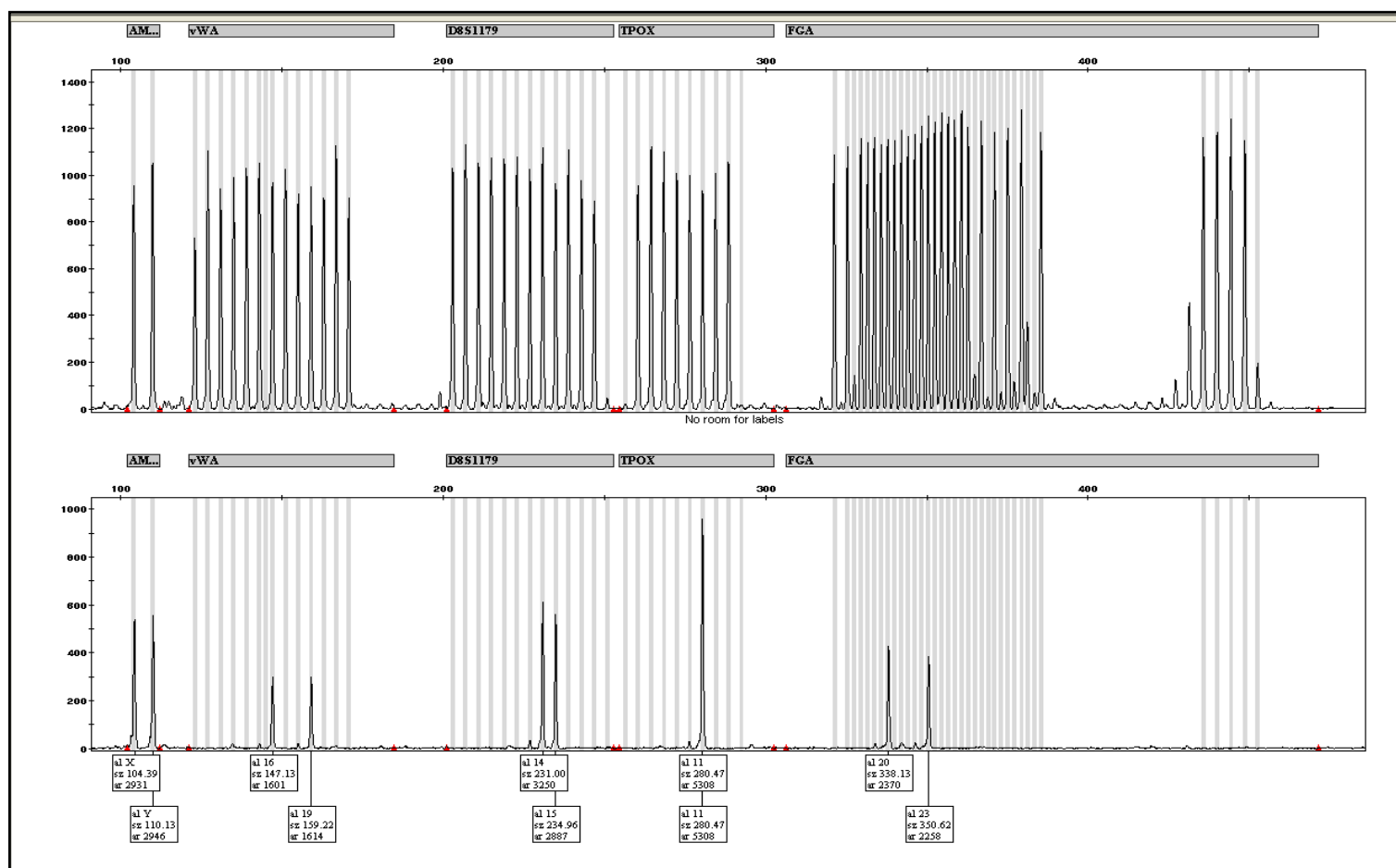


Figure 4.11 : Diagram showing the electropherogram of TMR-labeled STR loci (sex-determining Amelogenin, vWA, D8S1179, TPOX, and FGA).

Upper lane : Allelic ladder components

Lower lane : Control reaction of 2800M

4.4.1.1 Kadazan-Dusun

The frequencies of alleles for 15 STR markers in the Kadazan-Dusun are shown in Table 4.7. The numbers of alleles differed from as few as 5 (TPOX marker) to as many as 18 (Penta E marker). Allele 11 for D7S820 marker was the most prominent allele among the Kadazan-Dusun individuals, with an allelic frequency of 0.504. All examined markers exhibited high levels of diversity. The heterozygosity of markers ranged from 0.561 (TPOX) to 0.9 (Penta E) with an average of 0.7508. In addition, all STR markers also presented with substantially high degrees of PD. The highest single locus PD was observed in Penta E (0.978), whereas the lowest value was found in TPOX (0.745). The PIC values varied from 0.504 to 0.890. On the other hand, the PE values ranged from 0.247 (TPOX) to 0.796 (Penta E). All markers, except D21S11, were examined to have no deviation from the HWE.

Table 4.7 : Allelic distribution of 15 autosomal STR markers in the Kadazan-Dusun population; the most prominent alleles are in red.

Allele	TH01	Penta E	D13S317	D7S820	CSF1PO	Penta D	TPOX
5	-	0.105	-	-	-	-	-
6	0.055	-	-	-	-	-	-
7	0.244	-	-	-	-	0.013	-
8	0.214	-	0.220	0.162	0.002	0.048	0.421
9	0.367	0.022	0.028	0.031	0.030	0.386	0.090
9.3	0.055	-	-	-	-	-	-
10	0.065	0.024	0.105	0.135	0.229	0.135	0.002
11	-	0.177	0.397	0.504	0.310	0.076	0.474
12	-	0.085	0.249	0.155	0.410	0.155	0.013
13	-	0.039	0.002	0.006	0.018	0.188	-
14	-	0.096	-	0.007	0.002	-	-
15	-	0.175	-	-	-	-	-
16	-	0.041	-	-	-	-	-
17	-	0.057	-	-	-	-	-
18	-	0.039	-	-	-	-	-
19	-	0.028	-	-	-	-	-
20	-	0.018	-	-	-	-	-
21	-	0.022	-	-	-	-	-
22	-	0.022	-	-	-	-	-
23	-	0.037	-	-	-	-	-
24	-	0.007	-	-	-	-	-
25	-	0.006	-	-	-	-	-
H _{Obs}	0.782	0.900	0.705	0.653	0.697	0.738	0.561
H _{Exp}	0.751	0.900	0.722	0.678	0.684	0.767	0.591
PD	0.892	0.978	0.875	0.853	0.832	0.915	0.745
PIC	0.711	0.890	0.674	0.638	0.622	0.735	0.504
PE	0.567	0.796	0.436	0.360	0.424	0.489	0.247
HWE	0.322	0.290	0.453	0.149	0.719	0.560	0.069

Table 4.7 : continuation.

Allele	D3S1358	D21S11	D18S51	D5S818	D16S539	vWA	D8S1179	FGA
9	-	-	-	0.028	0.221	-	0.011	-
10	-	-	-	0.321	0.275	0.002	0.129	-
11	-	-	-	0.367	0.323	0.002	0.085	-
12	-	-	0.076	0.194	0.098	-	0.024	-
13	0.002	-	0.015	0.087	0.074	-	0.290	-
14	0.007	-	0.262	0.004	0.009	0.116	0.212	-
15	0.408	-	0.334	-	-	0.059	0.148	0.009
16	0.299	-	0.155	-	-	0.107	0.094	-
17	0.164	-	0.066	-	-	0.279	0.007	0.002
18	0.101	-	0.024	-	-	0.288	-	-
18.2	-	-	-	-	-	-	-	0.033
19	0.018	-	0.066	-	-	0.103	-	0.061
19.2	-	-	-	-	-	-	-	0.004
20	-	-	-	-	-	0.042	-	0.113
21	-	-	-	-	-	0.002	-	0.149
21.2	-	-	-	-	-	-	-	0.004
22	-	-	0.002	-	-	-	-	0.221
22.2	-	-	-	-	-	-	-	0.002
23	-	-	-	-	-	-	-	0.179
24	-	-	-	-	-	-	-	0.087
25	-	-	-	-	-	-	-	0.096
26	-	-	-	-	-	-	-	0.028
27	-	-	-	-	-	-	-	0.011
28	-	0.037	-	-	-	-	-	0.002
29	-	0.196	-	-	-	-	-	-
30	-	0.227	-	-	-	-	-	-
30.2	-	0.020	-	-	-	-	-	-
31	-	0.157	-	-	-	-	-	-
31.2	-	0.057	-	-	-	-	-	-
32	-	0.028	-	-	-	-	-	-
32.2	-	0.133	-	-	-	-	-	-
33.2	-	0.137	-	-	-	-	-	-
34.2	-	0.009	-	-	-	-	-	-
H _{Obs}	0.708	0.797	0.779	0.712	0.742	0.797	0.845	0.845
H _{Exp}	0.708	0.845	0.782	0.718	0.757	0.800	0.817	0.863
PD	0.864	0.954	0.920	0.867	0.902	0.932	0.937	0.964
PIC	0.657	0.824	0.750	0.666	0.716	0.771	0.792	0.846
PE	0.442	0.594	0.560	0.447	0.496	0.594	0.685	0.685
HWE	0.574	0.039	0.867	0.492	0.976	0.120	0.788	0.237

4.4.1.2 Bajau

Table 4.8 shows the distribution of alleles for 15 STR markers in the Bajau population. Similar to the Kadazan-Dusun group, TPOX was present with the least number of alleles, with only 5 being scored in the Bajau group. FGA and Penta E markers were observed to harbor the most number of alleles, i.e., 18 and 17, respectively. The most prominent allele was allele 8 of TPOX (49.7 %). All STR markers were found to present with high degrees of heterozygosity, from 0.591 (TPOX) to 0.906 (Penta E). The averaged heterozygosity across all 15 STR markers was 0.7816. The PD values of markers were high, ranged from 0.803 to 0.978. The average PD for all markers in the Bajau population was 0.9158. The PIC values differed from 0.562 to 0.899. Penta E was observed to have the highest PE value among all markers (0.807), while the TPOX marker present with the lowest PE value of 0.280. There was only 1 marker, Penta E, found to deviate from the HWE.

Table 4.8 : Allelic distribution of 15 autosomal STR markers in the Bajau population; the most prominent alleles are in red.

Allele	TH01	Penta E	D13S317	D7S820	CSF1PO	Penta D	TPOX
5	-	0.057	-	-	-	-	-
6	0.119	-	-	-	-	-	-
7	0.289	-	0.003	-	-	0.009	-
8	0.116	-	0.296	0.258	0.003	0.031	0.497
9	0.336	0.044	0.135	0.035	0.066	0.393	0.113
9.3	0.060	-	-	-	-	-	-
10	0.079	0.069	0.142	0.126	0.239	0.176	0.028
11	-	0.201	0.270	0.340	0.280	0.119	0.333
12	-	0.088	0.135	0.204	0.314	0.189	0.028
13	-	0.063	0.019	0.035	0.082	0.038	-
14	-	0.110	-	0.003	0.016	0.013	-
15	-	0.075	-	-	-	0.031	-
16	-	0.069	-	-	-	-	-
17	-	0.031	-	-	-	-	-
18	-	0.057	-	-	-	-	-
19	-	0.047	-	-	-	-	-
20	-	0.019	-	-	-	-	-
21	-	0.047	-	-	-	-	-
22	-	0.013	-	-	-	-	-
23	-	0.003	-	-	-	-	-
25	-	0.006	-	-	-	-	-
H _{Obs}	0.774	0.906	0.780	0.755	0.761	0.730	0.591
H _{Exp}	0.768	0.909	0.785	0.761	0.757	0.763	0.630
PD	0.909	0.978	0.916	0.898	0.887	0.912	0.803
PIC	0.731	0.899	0.749	0.719	0.713	0.730	0.562
PE	0.551	0.807	0.562	0.518	0.529	0.475	0.280
HWE	0.467	0.048	0.383	0.464	0.122	0.352	0.415

Table 4.8 : continuation.

Allele	D3S1358	D21S11	D18S51	D5S818	D16S539	vWA	D8S1179	FGA
7	-	-	-	0.009	-	-	-	-
9	-	-	-	0.035	0.173	-	-	-
10	-	-	-	0.299	0.176	-	0.145	-
11	-	-	0.009	0.283	0.374	-	0.082	-
12	-	-	0.060	0.214	0.176	-	0.057	-
13	0.006	-	0.063	0.154	0.094	-	0.255	-
14	0.031	-	0.170	0.006	0.006	0.151	0.198	-
15	0.270	-	0.296	-	-	0.082	0.138	-
16	0.362	-	0.167	-	-	0.126	0.113	-
17	0.248	-	0.069	-	-	0.270	0.013	0.003
18	0.069	-	0.060	-	-	0.242	-	0.003
18.2	-	-	-	-	-	-	-	0.003
19	0.013	-	0.053	-	-	0.113	-	0.123
20	-	-	0.041	-	-	0.013	-	0.063
21	-	-	0.003	-	-	0.003	-	0.170
21.2	-	-	-	-	-	-	-	0.009
22	-	-	0.003	-	-	-	-	0.208
22.2	-	-	-	-	-	-	-	0.025
23	-	-	-	-	-	-	-	0.101
23.2	-	-	-	-	-	-	-	0.003
24	-	-	0.003	-	-	-	-	0.101
24.2	-	-	-	-	-	-	-	0.006
25	-	-	-	-	-	-	-	0.094
26	-	-	0.003	-	-	-	-	0.047
27	-	0.009	-	-	-	-	-	0.019
28	-	0.053	-	-	-	-	-	0.019
29	-	0.201	-	-	-	-	-	0.003
30	-	0.245	-	-	-	-	-	-
30.2	-	0.025	-	-	-	-	-	-
31	-	0.119	-	-	-	-	-	-
31.2	-	0.085	-	-	-	-	-	-
32	-	0.050	-	-	-	-	-	-
32.2	-	0.164	-	-	-	-	-	-
33.2	-	0.035	-	-	-	-	-	-
34.2	-	0.006	-	-	-	-	-	-
35.2	-	0.006	-	-	-	-	-	-
H _{Obs}	0.742	0.855	0.855	0.755	0.723	0.799	0.855	0.843
H _{Exp}	0.731	0.846	0.838	0.762	0.761	0.813	0.836	0.879
PD	0.871	0.953	0.948	0.902	0.907	0.936	0.948	0.969
PIC	0.681	0.826	0.818	0.719	0.724	0.784	0.812	0.864
PE	0.496	0.705	0.705	0.518	0.465	0.597	0.705	0.681
HWE	0.296	0.432	0.492	0.438	0.902	0.383	0.572	0.211

4.4.1.3 Rungus

Table 4.9 summarizes the screening of 209 Rungus individuals for a total of 15 STR markers in the present study. The number of alleles observed in all STR markers varied from 5 to 17. As for other Sabahan indigenous individuals in our study, there were only 5 alleles present in the TPOX marker. In contrast, 17 alleles were seen in both Penta E and FGA markers. The most commonly observed allele being the 8-repeat of the TPOX marker, with a frequency of 0.455. The average observed heterozygosity of all 15 markers was 0.7678, with the lowest value from 0.612 to the highest of 0.890. PD values were high, with the D21S11 marker reaching 0.951. The lowest PD value was scored in the TPOX marker (0.767). The average PD of all markers was 0.9007. The PIC values differed from 0.524 (TPOX) to 0.900 (Penta E). However, PE values of markers varied from 0.306 to 0.775. In the HWE context, the vWA marker was found to deviate significantly. Conversely, other markers were found in full compliance.

Table 4.9 : Allelic distribution of 15 autosomal STR markers in the Rungus population; the most prominent alleles are in red.

Allele	TH01	Penta E	D13S317	D7S820	CSF1PO	Penta D	TPOX
5	-	0.031	-	-	-	-	-
6	0.045	-	-	-	-	-	-
7	0.301	-	-	-	-	0.014	-
8	0.170	-	0.287	0.239	0.002	0.124	0.455
9	0.364	0.026	0.048	0.029	0.026	0.292	0.103
9.3	0.062	-	-	-	-	-	-
10	0.057	0.026	0.077	0.177	0.261	0.134	0.017
11	-	0.184	0.352	0.397	0.318	0.057	0.421
12	-	0.110	0.237	0.146	0.342	0.211	0.005
13	-	0.017	-	0.007	0.050	0.163	-
14	-	0.105	-	0.002	-	-	-
15	-	0.100	-	0.002	-	0.005	-
16	-	0.045	-	-	-	-	-
17	-	0.091	-	-	-	-	-
18	-	0.072	-	-	-	-	-
19	-	0.043	-	-	-	-	-
20	-	0.036	-	-	-	-	-
21	-	0.050	-	-	-	-	-
22	-	0.033	-	-	-	-	-
23	-	0.002	-	-	-	-	-
24	-	0.026	-	-	-	-	-
H _{Obs}	0.732	0.890	0.761	0.751	0.722	0.809	0.612
H _{Exp}	0.741	0.909	0.731	0.733	0.712	0.809	0.607
PD	0.892	0.980	0.874	0.883	0.852	0.933	0.767
PIC	0.697	0.900	0.682	0.689	0.655	0.780	0.524
PE	0.480	0.775	0.528	0.512	0.464	0.615	0.306
HWE	0.917	0.430	0.878	0.961	0.085	0.493	0.904

Table 4.9 : continuation.

Allele	D3S1358	D21S11	D18S51	D5S818	D16S539	vWA	D8S1179	FGA
7	-	-	-	0.005	-	-	-	-
9	-	-	-	0.091	0.297	-	0.024	-
10	-	-	-	0.321	0.258	-	0.227	-
11	-	-	0.002	0.361	0.325	-	0.091	-
12	-	-	0.077	0.144	0.067	-	0.041	-
13	-	-	0.055	0.077	0.050	-	0.301	-
14	0.038	-	0.213	0.002	0.002	0.141	0.148	-
15	0.421	-	0.337	-	-	0.065	0.065	0.007
16	0.215	-	0.127	-	-	0.129	0.098	-
17	0.184	-	0.105	-	-	0.297	0.005	0.002
18	0.129	-	0.060	-	-	0.261	-	0.002
18.2	-	-	-	-	-	-	-	0.038
19	0.012	-	0.014	-	-	0.081	-	0.115
19.2	-	-	-	-	-	-	-	0.002
20	-	-	0.010	-	-	0.026	-	0.036
21	-	-	-	-	-	-	-	0.158
21.2	-	-	-	-	-	-	-	0.007
22	-	-	-	-	-	-	-	0.309
23	-	-	-	-	-	-	-	0.201
23.2	-	-	-	-	-	-	-	0.002
24	-	-	-	-	-	-	-	0.060
25	-	-	-	-	-	-	-	0.038
26	-	-	-	-	-	-	-	0.007
26.2	-	-	-	-	-	-	-	0.005
27	-	0.014	-	-	-	-	-	0.010
28	-	0.022	-	-	-	-	-	-
29	-	0.294	-	-	-	-	-	-
30	-	0.165	-	-	-	-	-	-
30.2	-	0.026	-	-	-	-	-	-
31	-	0.163	-	-	-	-	-	-
31.2	-	0.089	-	-	-	-	-	-
32	-	0.048	-	-	-	-	-	-
32.2	-	0.103	-	-	-	-	-	-
33.2	-	0.069	-	-	-	-	-	-
34.2	-	0.007	-	-	-	-	-	-
H _{Obs}	0.727	0.861	0.775	0.732	0.708	0.818	0.809	0.809
H _{Exp}	0.726	0.835	0.803	0.734	0.734	0.798	0.813	0.820
PD	0.879	0.951	0.936	0.879	0.883	0.924	0.938	0.941
PIC	0.683	0.814	0.777	0.688	0.684	0.767	0.787	0.797
PE	0.472	0.717	0.554	0.480	0.441	0.633	0.615	0.615
HWE	0.563	0.249	0.321	0.062	0.914	0.032	0.455	0.403

4.4.2 Identification of non-allelic variant

In the present study, a total of 8 samples with “off-ladder” fragments were observed, 5 in the Kadazan-Dusun and 3 in the Rungus individuals (Figure 4.12). All “off-ladder” fragments were seen in the region of FGA loci, with the size around 317 bp.

Validation by DNA resequencing has confirmed that these fragments are FGA alleles, with 15 repeats of the tetra-nucleotide (allele 15).

4.4.3 STRs: Population differentiation analysis

The degree of genetic differentiation in the Sabahan indigenous populations was estimated by calculation of diversity parameters and AMOVA analysis (Table 4.10). The average value for H_T for all 15 markers was 0.776. The highest H_T value was 0.910 in Penta E, whereas the lowest was TPOX (0.613). In average, all markers demonstrated substantial diversity. A vast majority of the differences arise within the populations. Variations within populations (H_S) scored an average of 0.772 across all markers ($H_T = 0.776$). In other words, the inter-population differences (D_{ST}) contributed only 0.004 to the value. Penta D and D3S1358 were observed to yield the highest D_{ST} value, 0.009. On the other hand, the total gene diversity for vWA was solely accounted by intra-population diversity ($H_T = 0.803$; $H_S = 0.803$; $D_{ST} = 0.000$). The G_{ST} of all markers was low (average of 0.006) and G_{IS} ranged from 0.001 (vWA) to 0.036 (D16S539).

When AMOVA was performed for all loci as a whole, 99.13 % of the variation was found to come from within the populations. There was only < 1 % of the variation seen among populations. In locus to locus AMOVA, most loci showed > 99 % of intra-population variations. The highest inter-population variation was only 1.62 %, in D3S1358 marker. In concordance to the degree of population diversity as estimated by G_{ST} value, vWA marker exhibited 0 % difference among the indigenous populations. The AMOVA evaluation was significant for all tested loci ($P < 0.05$), except vWA marker ($P = 0.471$).

Table 4.10 : Summary of the genetic differentiation and AMOVA analyses performed on 15 autosomal STR markers in Kadanzen-Dusun, Bajau, and Rungus individuals (H_T = Total gene diversity; H_S = Intra-population gene diversity; D_{ST} = Inter-population gene diversity; G_{ST} = Coefficient of gene differentiation; G_{IS} = Inbreeding coefficient).

STR loci	TH01	Penta E	D13S317	D7S820	CSF1PO	Penta D	TPOX
Differentiation analysis							
H_T	0.756	0.910	0.754	0.730	0.719	0.788	0.613
H_S	0.753	0.906	0.746	0.724	0.718	0.780	0.609
D_{ST}	0.002	0.004	0.008	0.006	0.002	0.009	0.004
G_{ST}	0.003	0.004	0.011	0.009	0.003	0.011	0.006
G_{IS}	0.013	0.008	0.003	0.006	0.013	0.027	0.034
AMOVA differentiation							
Among populations	0.439	0.702	1.506	1.385	0.410	1.513	0.845
Within populations	99.561	99.298	98.494	98.615	99.590	98.487	99.155
P	0.007	0.000	0.000	0.000	0.013	0.000	0.004

STR loci	D3S1358	D21S11	D18S51	D5S818	D16S539	vWA	D8S1179	FGA
Differentiation analysis								
H_T	0.731	0.847	0.810	0.741	0.757	0.803	0.825	0.860
H_S	0.722	0.842	0.808	0.738	0.751	0.803	0.822	0.854
D_{ST}	0.009	0.005	0.003	0.003	0.006	0.000	0.003	0.006
G_{ST}	0.013	0.006	0.003	0.004	0.008	0.000	0.004	0.007
G_{IS}	0.006	0.005	0.006	0.007	0.036	0.001	0.017	0.026
AMOVA differentiation								
Among populations	1.621	0.905	0.521	0.583	1.059	0.000	0.649	1.023
Within populations	98.379	99.095	99.479	99.417	98.941	100.000	99.351	98.977
P	0.000	0.000	0.001	0.002	0.000	0.471	0.000	0.000

4.4.4 STRs: Population structure analysis

The proportion of ancestral components in each tested individual was estimated by clustering analyses via Structure and CLUMPP softwares. The Structure program is a model-based clustering method that works by assuming a number of K populations in the sample set. Each individual is assigned to the populations according to the allelic frequencies of all loci. The “admixture” model employed in our analysis presumes that there is mix-ancestry in each individual, whereby the genome is made up of different fractions of ancestry from K populations (Evanno, Regnaut, & Goudet, 2005).

Inference of the true K value is essential to ensure reliability of the reflected genetic structure of the studied populations. In our study, the K values were evaluated via 2 approaches [L(K) and ΔK as shown in Figure 4.13]. In the L(K) method, the L(K) values plateau or decrease minutely when approaching a true K value (Rosenberg, et al., 2001). In addition, the variance between runs is high. In our analysis, the L(K) values approaching plateau for K were from 1 to 3. The value decreased dramatically from $K = 4$ to 10.

Delta K measures the second order rate of change of the likelihood (Evanno, et al., 2005). The L(K) is average of $\text{LnP}(D)$ of replicates, 10 in this case, of a tested K . The formula to derive ΔK is as below:

$$\begin{aligned} L'(K) &= L(K)_n - L(K)_{n-1} \\ L''(K) &= L'(K)_n - L'(K)_{n-1} \\ \Delta K &= [L''(K)] / \sigma \end{aligned}$$

where σ is the standard deviation of replicates of $\text{LnP}(D)$

The best-fit K , as inferred by ΔK calculation, shows a clear peak on the plot and it was $K = 3$ in sample set.

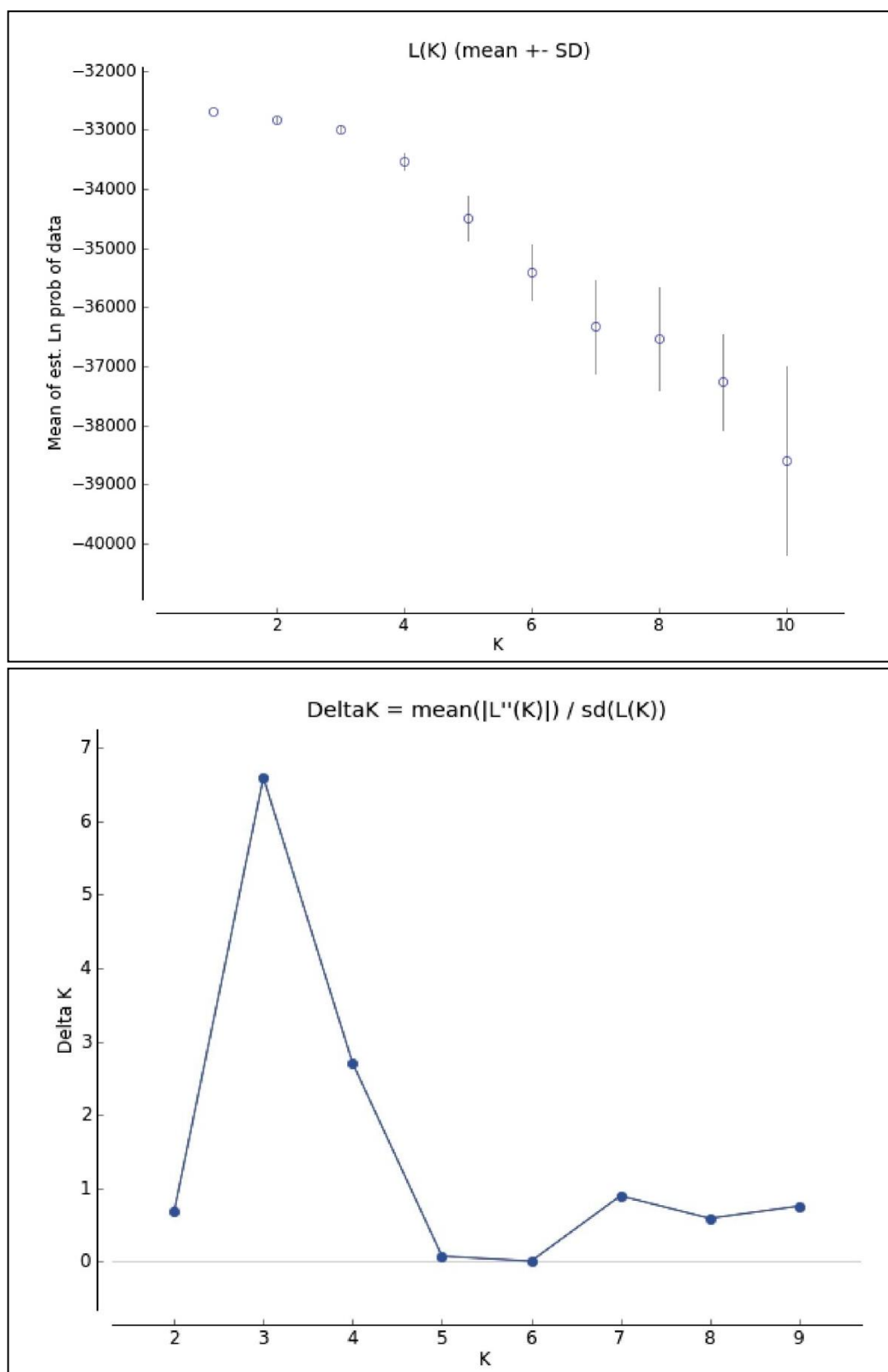


Figure 4.13 : Two approaches for the inference of best-fit K ; upper plot displays the $L(K)$ versus K ; lower diagram shows the changes of ΔK with respective K ; the true K , as determined by these methods was $K = 3$.

In order to minimize the differences of independent runs for each K , alignment of the membership coefficients, or Q matrix, of $K = 2$ to $K = 6$ was performed using the CLUMPP software. The resulted Q matrix of each K was used to produce bar plots as shown in Figure 4.14. The contribution of each “presumed ancestor” was represented by different colors, i.e., red, green, purple-blue, pink, brown, and orange for ancestors 1 to 6. On the other hand, the 639 indigenous individuals were represented by vertical lines (separators not shown for individuals within the populations but only to separate the respective indigenous groups).

At $K = 2$, Kadazan-Dusun and Rungus showed more-or-less similar patterns of ancestral components. However, the patterns for Bajau individuals were obviously different from those in Kadazan-Dusun and Rungus. Bajau peoples were shown to harbor a higher proportion of the “green” element, which correspond to the second ancestral population.

At $K = 3$, all 3 populations were seen to display different patterns, although Kadazan-Dusun and Rungus populations still shared a certain degree of similarity. The Bajaus were characterized by high fraction of ancestor 2 (green) and almost equal contribution for the other 2 ancestral components. Although the distribution of Kadazan-Dusun and Rungus was similar (low fraction of ancestor 2), Kadazan-Dusun was observed to harbor higher proportion of ancestor 1 (red), while Rungus has higher share of ancestor 3 (purple-blue).

At $K = 4$, the contribution of ancestor 4 (pink) was found to be even in all indigenous populations. The patterns were homogenous for higher K 's, $K > 5$, where no distinctive pattern was seen in the distribution of these populations.

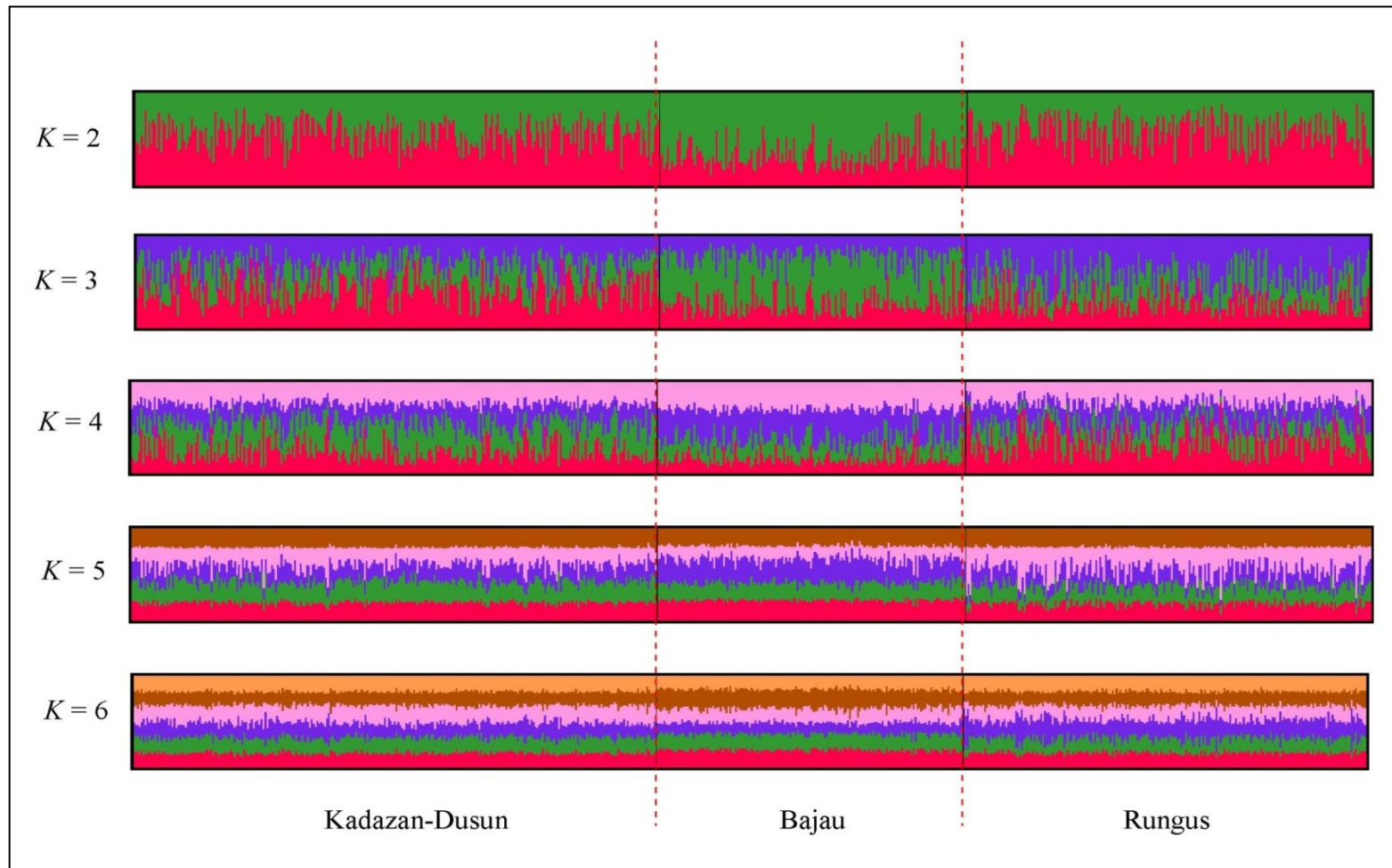


Figure 4.14 : Bar plot displays individual ancestry estimates of K structure of 2 to 6 of the studied populations; each individual is represented by a vertical line and colors represent the probable partitions of inferred ancestral from the clustering analyses.

4.4.5 STRs: Cluster analysis – Principal coordinate analysis

Apart from the genetic differentiation analysis carried out to examine the inter-relationship among the Sabahan indigenous populations, cluster analysis was also conducted to determine their association with other populations around the world.

4.4.5.1 PCoA of world populations

PCoA was employed for the clustering of populations involving STR markers, instead of PCA in the case of *Alu* insertion examination. This is because STRs involve a large amount of data which cannot be directly input and analyzed by PCA. PCA searches for direct patterns of observations and variables. However, genetic data generated by STR markers involves the consideration of various factors, such as multiple alleles for a single locus. Therefore, a pre-treatment or translation of the data must be performed prior to cluster analysis. One of which is to convert the genetic data of all populations into a distance format in a triangle matrix.

In order to depict the association of world populations with Kadazan-Dusun, Bajau, and Rungus groups in Sabah, STR data of a total of 45 populations from all major regions of the world were included (Appendix 5). These populations were selected and generally representing several major geographical regions, which include Africa (6), America (11), Europe (3), South Asia (3), East Asia (7), SEA (15), and Oceania (3). STR data was either adopted from previously published articles or from database – Autosomal STR DNA Database (<http://www.strdna-db.org>). After extraction of data from various sources, only 13 out of 15 loci (excluded Penta E and Penta D) were used for the maximum coverage of population that represent certain regions of the world. Allelic frequencies of all markers were arranged and converted to a pair-wise genetic distance

based on Nei's algorithm (D_A). The genetic data was subsequently subjected to building of the PCoA.

The first 2 axes explained cumulatively 59.64 % of the variation (Figure 4.15). In the PCoA, African populations, together with African-derived groups such as Afro-Ecuadoran and African-American, were found at the left side of the plot. Next to the African group is European population. It was not surprise to see that the Australian Caucasian was situated within the European cluster as it is in concordance to their origin, when they travelled from Britain in the 17th century. To the right of the Europeans are the Oceania populations, Australian Aborigines and Samoa. The bottom of the plot was occupied by populations from North and South Americas. Whereas the right side of the plot was the cluster of all Asian groups, including of East Asia, South Asia, and SEA. All Asian populations were found clustered close to each other, and there was no distinctive separation observed.

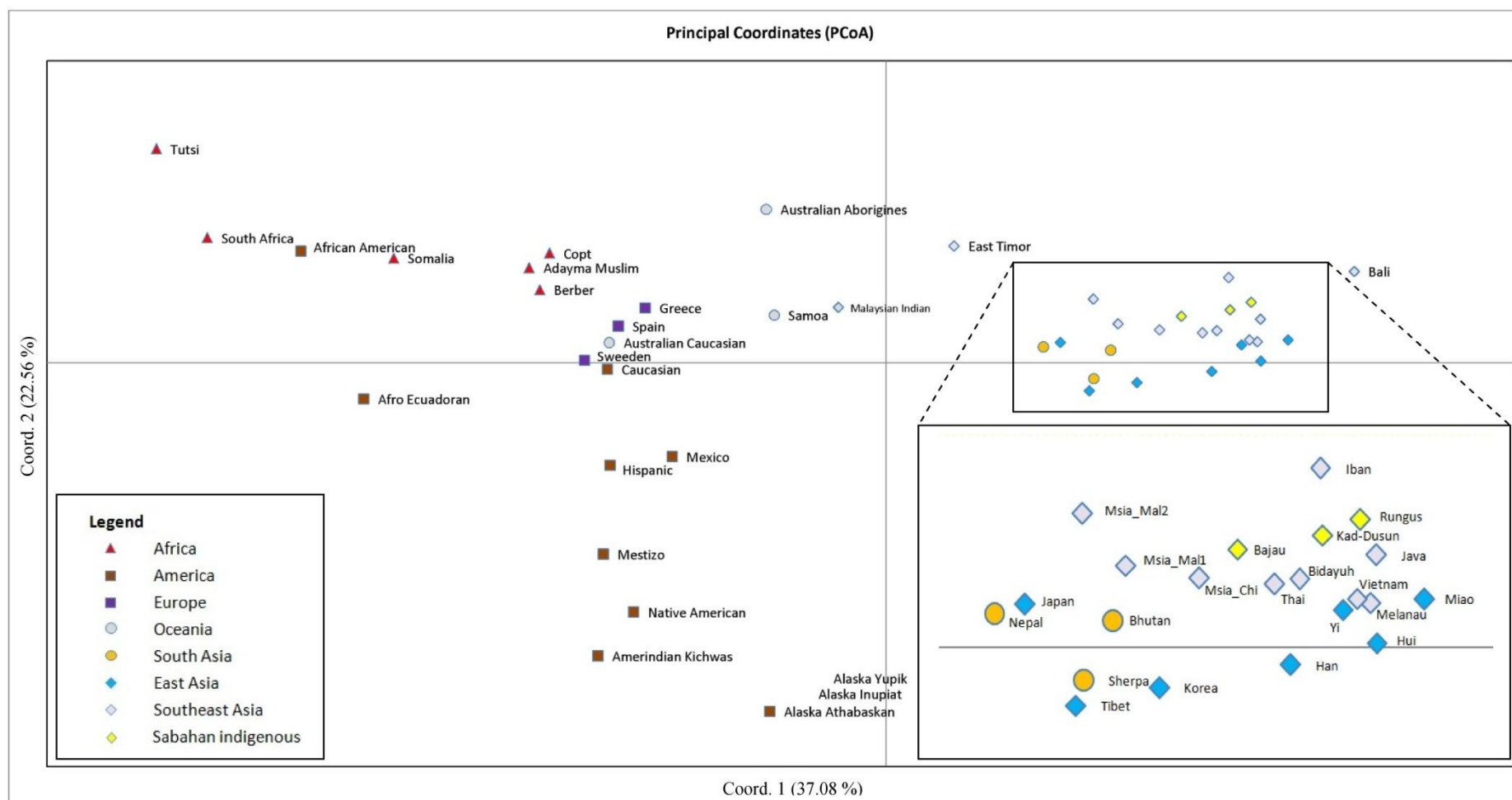


Figure 4.15 : Distribution of populations on the PCoA plot built from allelic frequencies of 13 STR loci; the box in the lower right corner shows the enlarged view of clusters for East Asia and SEA groups.

4.4.5.2 PCoA of neighboring populations

Although a strong genetic affinity of the Sabahan indigenous groups with other populations in East Asia and SEA was found, the diversity of these populations within the region remains ambiguous. Hence, another PCoA was performed and included a number of populations in this region (Appendix 6). According to Bellwood's model, SEA has been repopulated by migrants, who had travelled across the sea from South China through Taiwan about 5,000 years ago. Therefore, STR data sets from 10 Taiwanese Aboriginal populations, i.e., Bunun, Paiwan, Saisiat, Tao, Ami, Rukai, Atayal, Tsou, Pazeh, and Puyumah, were included in the second PCoA. Other than that, 2 populations of Chinese inhabitants living in Taiwan, Han and Hakka, were also included in the analysis.

Beside Bellwood, an alternative route of spread has been proposed following the coastal path from southern China through Vietnam, Thailand, etc. The migrants eventually arrived at SEA and populated the entire region. Thus, STR data was collected for few populations from the southern part of China, the Han, Miao, Yi, and Hui groups, together with populations along the coastal path, i.e., Vietnam and Thailand. It would be interesting to also include Japanese and Korean populations in the PCoA, as these islanders are situated not further below southern China. It may yield informative assumption on the distribution of SEA populations.

Also included were 3 populations (Sherpa, Bhutan, and Nepal) from South Asia. South Asia was suggested as a significant "station" during the dispersal of modern humans after the migration out of Africa.

In order to access the genetic affinity of the Sabahan indigenous groups to other population within SEA, 3 indigenous groups from Sarawak – the sister state of Sabah in East Malaysia (Iban, Melanau, and Bidayuh), 2 Indonesian (Bali and Java), 4

populations of the 3 main Malaysian ethnicities (Malay, Chinese, and Indian), and East Timor were included. In addition, several populations in the Pacific region were also added into the analysis, which included Australian (Aborigines and Caucasian) and Samoa.

Altogether, there were 40 sets of STR population data being inspected. A total of 12 markers were selected for each population, i.e., CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, TPOX, and vWA.

The PCoA plot explained a total of 67.32 % variations in the data set (Figure 4.16). The variations represented by both axes 1 and 2 were 52.86 % and 14.46 %, respectively. In general, the populations were separated into 3 main clusters in the plot. The first cluster was located on the left. It consisted of populations from East and South Asia. This cluster also included the Japanese and Korean. The Australian Aborigines were located away from other populations at the top of the plot.

To the right of the plot, 2 clusters were clearly parted by the midline. At the lower part, the cluster was mainly structured by the 10 Taiwanese Aboriginal groups. Among them, the Kadazan-Dusun and Rungus, from our study, were found sitting tightly in the cluster. Also, a group of indigenous in the state of Sarawak, the Melanau, was also observed in this cluster.

In the upper right of the plot, a cluster was formed by mainly SEA, both mainland and island. The Bajau group was detected within this cluster.

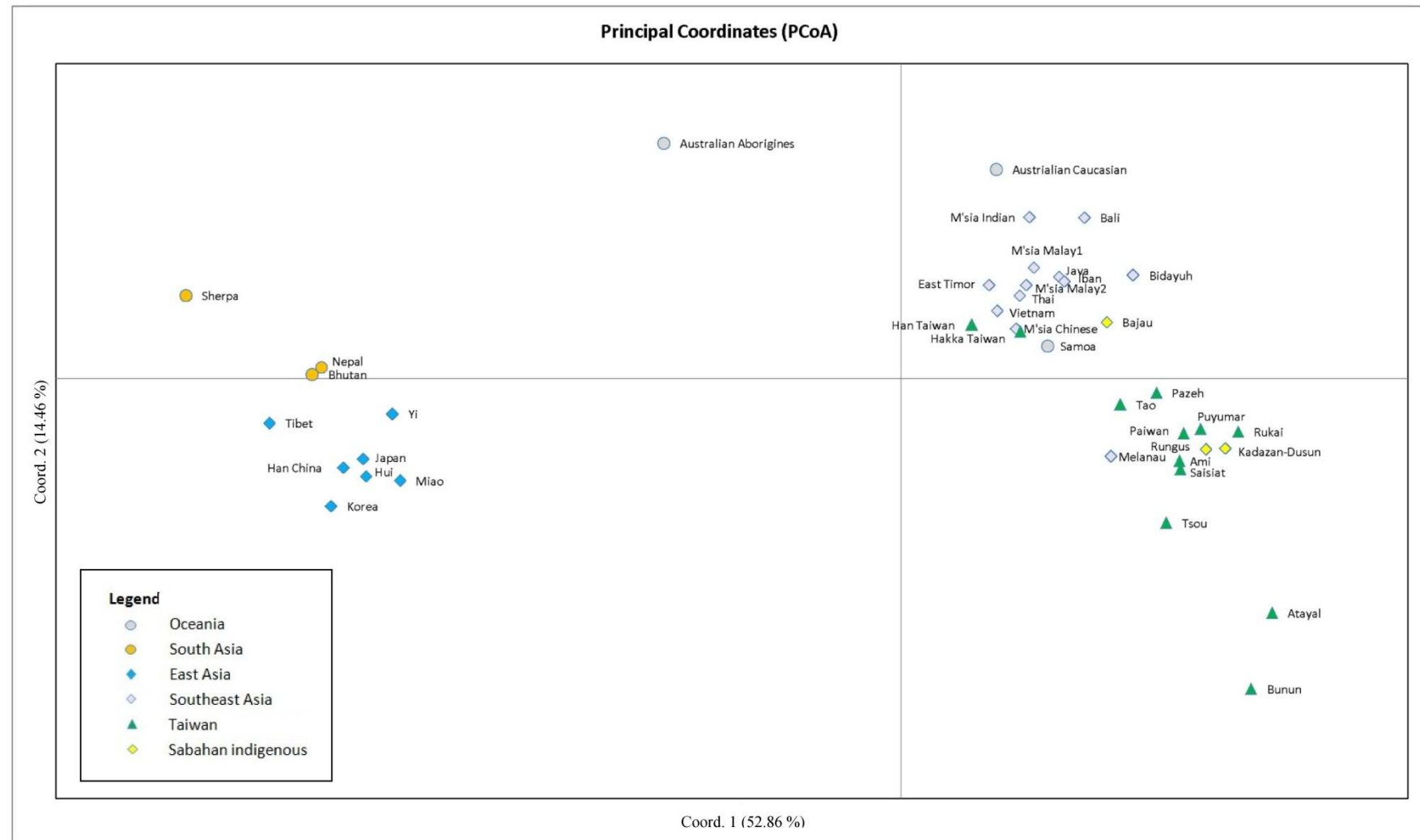


Figure 4.16 : PCoA plot of 40 populations in SEA and the neighboring regions (South Asia, East Asia, and Oceania).

4.4.6 STRs: Cluster analysis – Phylogenetic analysis

In addition to visualizing the distribution of various populations on the 2D plot, another cluster analysis was also conducted to generalize and represent all tested populations in a tree-like diagram, i.e., phylogenetic study. In the phylogenetic study, the genetic distance of each population corresponds to the distance between the “nodes” (symbolize the populations). Groups that have greater level of similarity are put under a cluster, whereas those that differ would be placed apart.

4.4.6.1 Phylogenetic analysis of world populations

The same cohort of populations, as included in the PCoA, was tested in the phylogenetic study. Genetic distance, expressed as in D_A , was calculated from allelic frequencies of 13 STR markers for all 48 populations from various regions, comprising of Africa, America, Europe, South Asia, East Asia, SEA, and Pacific.

Figure 4.17 shows the tree-structure of the world populations in the phylogenetic study.

The very first branched-off populations were the African and African-derived groups (African-American and Afro-Ecuadoran). It was subsequently followed by the Egyptians, i.e., Copts and Adayma Muslim.

The next groups that separated from the tree were the Europeans, together with Australian Caucasians. Americans were grouped under a cluster, where there were 2 distinct divisions between Natives and Hispanics. Under the native clade, Amerindian and Alaskan peoples were separated into 2 clusters (bootstrap value of 80).

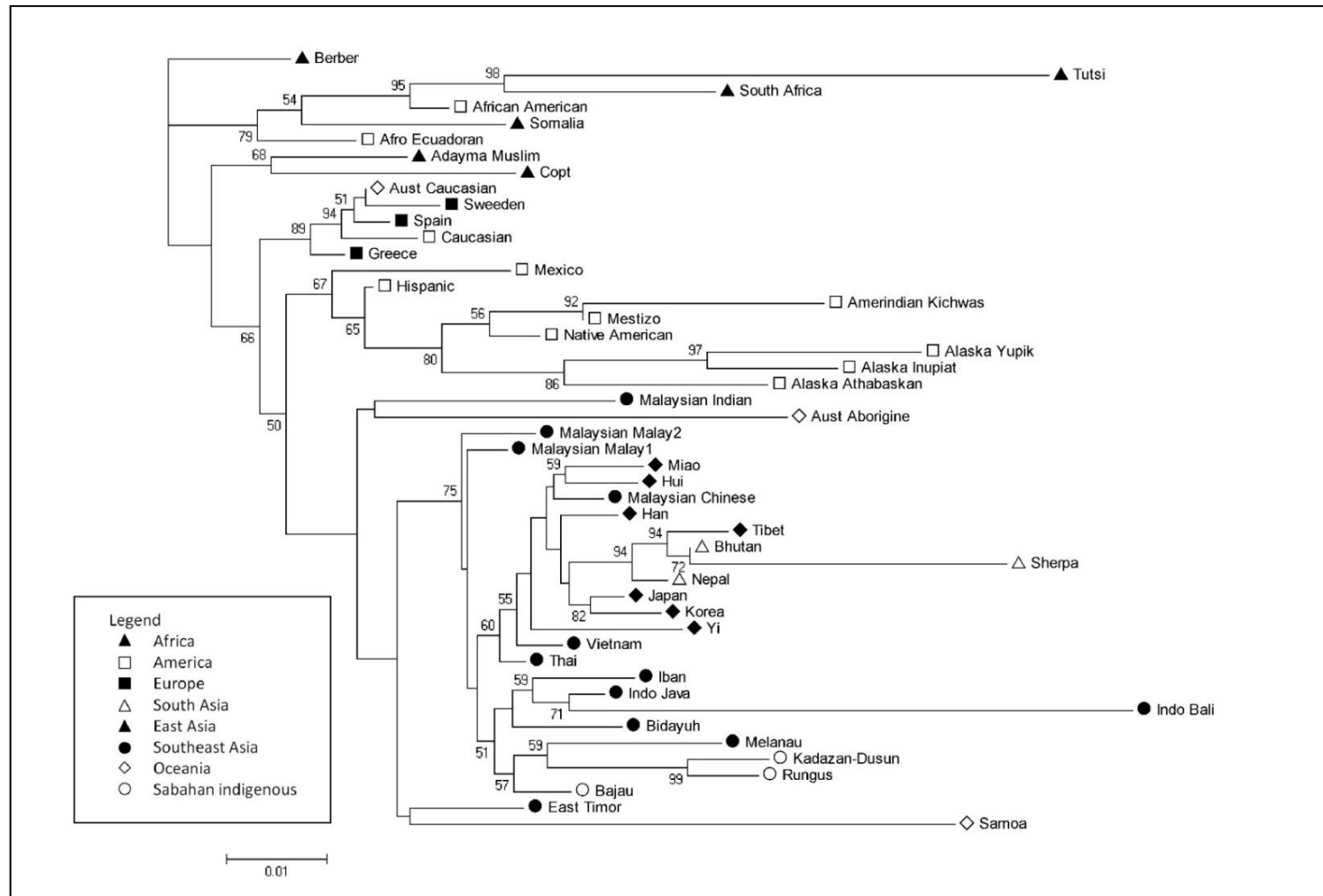


Figure 4.17 : NJ tree showing the clustering pattern of the world populations; number at the node denotes the bootstrap value for the branches, only values higher than 50 % were shown.

The Asian populations formed a complex and wide cluster. Interestingly, the Malaysian Indian was observed to group together with the Australian Aborigine. On the other hand, East Timor population was found to huddle closer to the Samoan population.

Basically, there were 2 major clades in the cluster. One is mainly constructed by populations from East Asia (Han, Yi, hui, Miao, Tibet, Japanese, Korean) and its derivative (Malaysian Chinese), Mainland SEA (Vietnam, Thailand), and South Asia (Nepal, Bhutan, Sherpa). The other clade was built by populations in the ISEA region. There were 2 sub-clades observed in the ISEA groups. The Iban, Bidayuh (indigenous groups from Sarawak), and Indonesian (Java, Bali) formed a discrete partition from the Sabahan indigenous groups. The Kadazan-Dusun, Bajau, and Rungus were found to have closer relationship among each other and formed a clade. The Melanau, another indigenous group from Sarawak, was seen in this clade, which showed that it has greater genetic similarity to the Sabahan indigenous group. Interestingly, it was also found to be closer to Kadazan-Dusun and Rungus than the Bajau group.

4.4.6.2 Phylogenetic analysis of neighboring populations

In order to investigate the relationship of the Sabahan indigenous groups with other neighboring populations, the phylogenetic analysis on 30 populations was conducted. Pair-wise genetic distance was derived from allelic frequencies of 12 STR markers and used for the NJ tree construction (Figure 4.18).

Among the first to branch off from the tree were Iban (Sarawak indigenous) and Indonesian (Java and Bali), suggesting that they have the most dissimilar genetic contours among all populations in the neighboring regions. It was then followed by Bidayuh. The rest of the populations in the test fell into 2 major clusters. The smaller cluster was made up of 2 clades. The first clade consisted of Malaysian Malays, followed by 2 sub-clades, i.e., Samoa and East Timor; Australian and Malaysian Indian. This observation was in agreement with the previous phylogenetic tree drawn for the world populations. The second clade was made up of populations with affinity to East Asian populations, i.e., Malaysian Chinese, Taiwanese Han and Hakka, Vietnamese, and Thai.

In the other major cluster, 2 clades were observed. The first clade was built solely by Taiwanese Aboriginal populations (Saisiat, Ami, Puyumah, Rukai, and Paiwan). There were 2 obvious sub-clades formed in the second clade, where the first sub-clade consisted of another group of Taiwanese Aboriginal populations (Tao, Tsou, Atayal, and Bunun) and the second sub-clade was Kadazan-Dusun and Rungus. The observation showed that the Sabahan indigenous groups were genetically closer to the Taiwanese Aboriginal populations. Nonetheless, the Bajau group was the first to separate from the cluster. Kadazan-Dusun and Rungus were shown to have higher genetic affinity to the Taiwanese Aborigines than the Bajau, especially to the Bunun, Atayal, Tsou, and Tao groups.

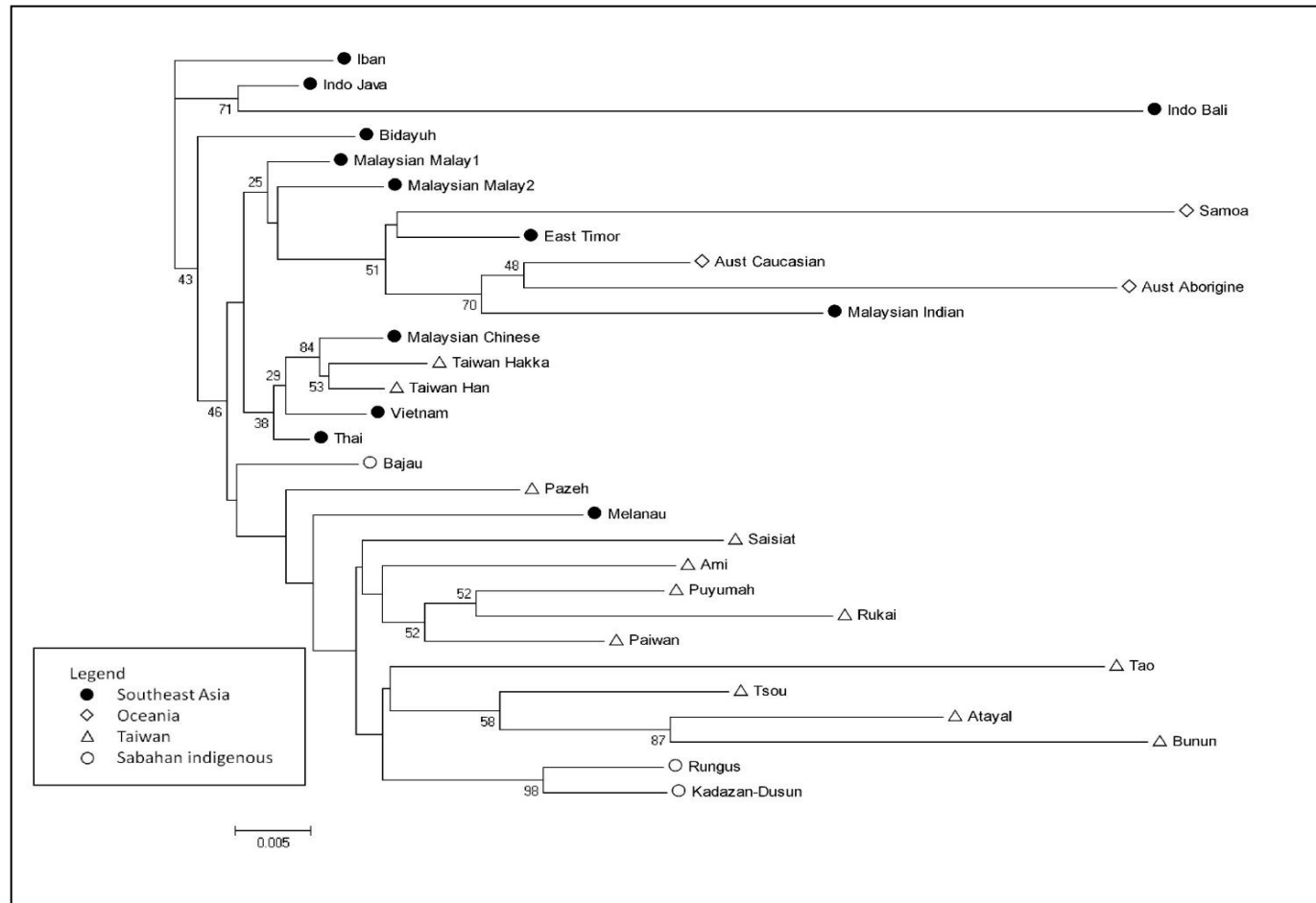


Figure 4.18 : NJ tree constructed based on the genetic distance derived from 12 STR loci for the SEA-neighboring populations; bootstrap values, higher than 50 %, were displayed at the node of each branch.

4.5 Assessment of mtDNA

Unlike the nuclear genome, mtDNA is inherited exclusively from the mother, i.e., maternal inheritance. It is not subjected to recombination with the other set of mt genome (the father's). In addition, the mt genome also mutates at high rates as compared to the nuclear genome. Together, it makes the mtDNA extremely useful to trace the ancestry of maternal lineage in many species (Alvarez, et al., 2012; Dulik, et al., 2012; Niemi, et al., 2013).

In our study, genetic polymorphisms in 2 functional parts of the mt genome were examined. The presence of a 9-bp deletion in the intergenic region of mt coding region was screened. In addition, sequences of the 3 HV regions in the mt control region were determined.

4.5.1 Intergenic 9-bp deletion marker

All indigenous samples were subjected to PCR to examine for the 9-bp deletion in the intergenic region of COII-K in mt coding region (nps 8,271 – 8,279), as shown in Figure 4.19. The frequency of deletion scored in each population was tabulated (as below).

Sabahan indigenous group	n	Frequency (%)
Kadazan-Dusun	271	50 (18.45)
Bajau	159	46 (28.39)
Rungus	209	19 (9.09)
Average		18.64 %

The highest deletion frequency was observed in the Bajau population and the lowest in the Rungus.

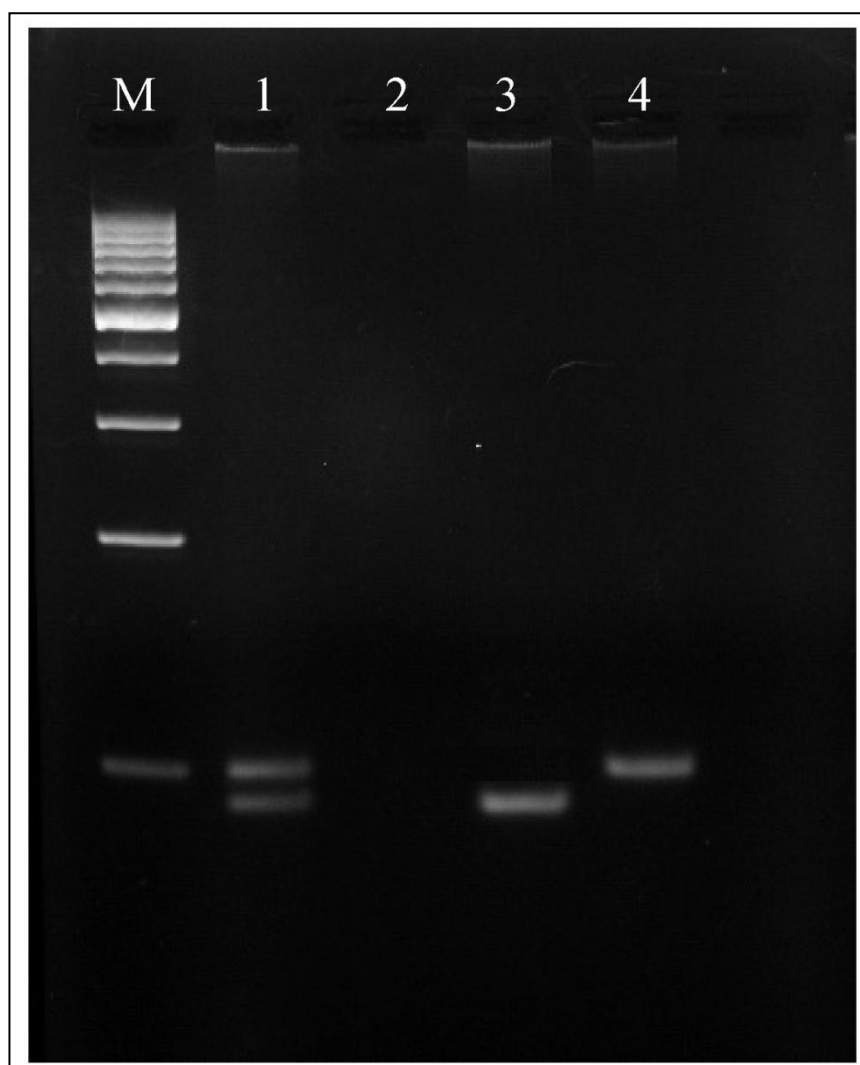


Figure 4.19 : A 3.5 % (w/v) EtBr-stained native agarose gel showing PCR amplicon for the screening of mt 9-bp deletion in region V.

Lane M : 100 bp DNA marker

Lane 1 : Self-customized DNA fragment mixture (91 bp/100 bp)

Lane 2 : Non-template control

Lane 3 : Sample with the 9-bp deletion (91 bp)

Lane 4 : Sample without the 9-bp deletion (100 bp)

4.5.2 SNPs in the control region

All 3 HV regions in 450 individuals from the Kadazan-Dusun, Bajau, and Rungus populations were sequenced. The position and length of each HV region are summarized below:

Region		Position on mtDNA (np)	Length (bp)
HV	1	16,024 – 16,365	342
	2	73 - 340	268
	3	439 - 576	138
TOTAL			748

A total of 3 different types of genetic polymorphisms were observed, the most common being SNPs (> 75 %), followed by insertions and deletions.

4.5.2.1 Sequence quality check

Peaks of all electropherograms were checked carefully and samples with less satisfactory results were repeated.

Odd mutation points, as regarded as “global private mutation” by the Haplogrep application, were verified by rechecking the raw sequencing data or/and repeated, if necessary.

QM networks were drawn from mt data for each data set in accordance to the populations. There were 2 types of graphical representation for interpretation by QM analysis, i.e., network and torso. Torso is a very condensed network, where all pendant subtrees are collapsed into the base nodes. It is useful for a quick check of the mt data. On the other hand, network provides more details on the lineages and therefore it is

more informative than the torso (Brandstatter, Klein, Duftner, Wiegand, & Parson, 2006; Zimmermann, et al., 2011).

Example of the network and torso is shown in Figure 4.20. These networks were established from the mt data of 150 Kadazan-Dusun individuals from our study. There was no over-complex star-like structure observed in the networks of all 3 indigenous populations.

Mt sequences and data generated from our study were submitted to the EMPOP for deposition into the database (Accession numbers: EMP00648, EMP00649, and EMP00650).

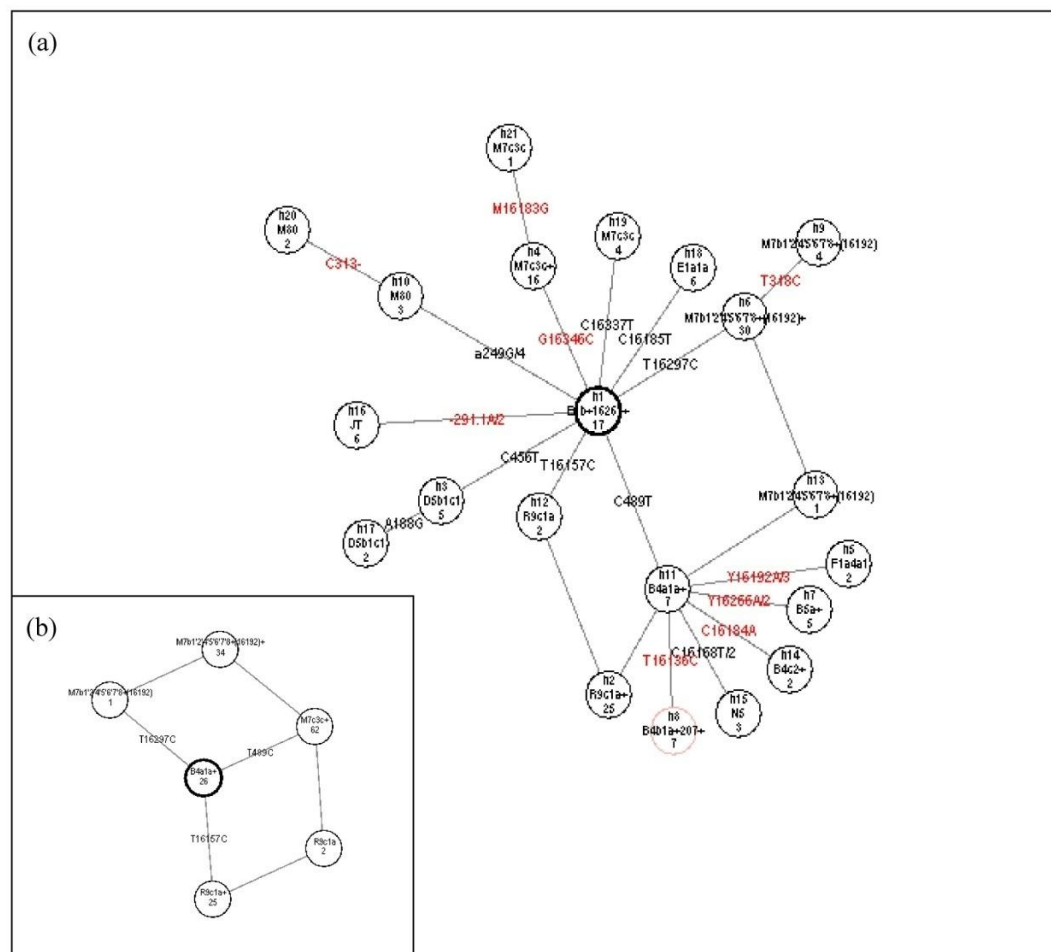


Figure 4.20 : Drawings of mt lineages among 150 Kadazan-Dusun individuals as represented by (a) network (b) torso; bold circle in the middle of the network is the root node; transitional mutations are labeled as black and other mutations are red.

4.5.2.2 Distribution of genetic polymorphisms

In the total of 748 nucleotides that were sequenced and compared to the rCRS, 10 % to 17 % of the sites were found to be variable. The Bajau population contained the most number of polymorphic sites (130 or 17.4 %). In contrast, the Kadazan-Dusun population harbored the least sites, only 76 or 10.2 %. However, the Rungus tribe was shown to be variable at 94 sites, or 12.6 %. In all populations, some sites were also detected to harbor more than 1 variant in the HV1 and HV2 regions (Table 4.11).

Table 4.11 : MtDNA sites with more than 1 observed variant in the Sabahan indigenous individuals.

Ethnic group	Region	Position	Variant	Frequency	
				No. observed	%
Kadazan-Dusun	HV1	16,183	A > C	42	28.00
			A > G	1	0.67
	HV1	16,192	C > A	2	1.33
			C > T	33	22.00
	HV2	249	A-DEL	27	18.00
			A > G	5	3.33
Bajau	HV1	16,181	A > C	1	0.67
			A > G	1	0.67
	HV1	16,192	C > T	6	4.00
			C > A	1	0.67
Rungus	HV1	16,183	A > C	45	30.00
			A-DEL	1	0.67
	HV2	249	A-DEL	25	16.67
			A > G	1	0.67

Tables 4.12 to 4.14 show the distribution of genetic polymorphisms observed in the control region for Kadazan-Dusun, Bajau, and Rungus.

Nucleotide substitution is the most frequently seen polymorphism in the control region of mtDNA. The percentages of nucleotide substitution in the Kadazan-Dusun, Bajau, and Rungus in our study were 82.3 %, 84.1 %, and 78.1 %, respectively. There were 2 types of SNPs observed, i.e., transtitional and transversional SNPs. Majority of the SNPs, up to 92.8 % as in Bajau population, detected in our study were transitional SNPs, which can be found in all 3 HV regions in each population. The most abundant transitional SNP is the T > C variant, following by the C > T variant. On the other hand, transversional SNPs were only found limited to the HV1 region. The most frequently observed transversional SNPs were the A > C and C > A variants.

Most of the insertional and deletional mutations were found to concentrate at the HV2 and HV3 regions, except an insertion was seen in HV1 in the Rungus population. The highest number of insertion, 10, was observed in the HV2 region of Rungus population, whereas the Bajau population was shown to harbor the most number of deletions in the HV3 region.

In summary, the Bajau population was determined to harbor the most number of variable sites, with reference to the rCRS, among the 3 Sabahan indigenous populations examined in our study. While comparing the 3 HV regions, HV1 was found to carry the most polymorphisms and the HV3 the least.

Table 4.12 : Summary of the distribution of genetic polymorphisms present in all 3 mt HV regions in 150 Kadazan-Dusun individuals.

Characteristic	Mitochondrial HV region			TOTAL
	HV1	HV2	HV3	
Variable site	43	24	9	76
Polymorphism	45	25	9	79
Site > 1 polymorphism	2	1	0	3
Nucleotide substitution	45	16	4	65
Transitional SNP				
G > A	3	3	1	7
A > G	5	6	0	11
T > C	15	6	1	22
C > T	16	1	2	19
TOTAL	39	16	4	59
Transversional SNP				
A > C	2	0	0	2
C > A	3	0	0	3
C > G	0	0	0	0
G > C	1	0	0	1
A > T	0	0	0	0
TOTAL	6	0	0	6
Insertion	0	4	4	8
Deletion	0	5	1	6

Table 4.13 : Summary of the distribution of genetic polymorphisms present in all 3 mt HV regions in 150 Bajau individuals.

Characteristic	Mitochondrial HV region			TOTAL
	HV1	HV2	HV3	
Variable site	78	33	19	130
Polymorphism	80	33	19	132
Site > 1 polymorphism	2	0	0	2
Nucleotide substitution	79	24	8	111
Transitional SNP				
G > A	5	6	2	13
A > G	12	9	0	21
T > C	27	8	3	38
C > T	27	1	3	31
TOTAL	71	24	8	103
Transversional SNP				
A > C	5	0	0	5
C > A	3	0	0	3
C > G	0	0	0	0
G > C	0	0	0	0
A > T	1	0	0	1
TOTAL	8	0	0	8
Insertion	0	4	7	11
Deletion	0	5	4	9

Table 4.14 : Summary of the distribution of genetic polymorphisms present in all 3 mt HV regions in 150 *Rungus* individuals.

Characteristic	Mitochondrial HV region			TOTAL
	HV1	HV2	HV3	
Variable site	49	36	9	94
Polymorphism	50	37	9	96
Site > 1 polymorphism	1	1	0	2
Nucleotide substitution	49	22	4	75
Transitional SNP				
G > A	4	2	0	6
A > G	4	10	0	14
T > C	17	8	1	26
C > T	18	2	3	23
TOTAL	43	22	4	69
Transversional SNP				
A > C	3	0	0	3
C > A	1	0	0	1
C > G	1	0	0	1
G > C	1	0	0	1
A > T	0	0	0	0
TOTAL	6	0	0	6
Insertion	1	10	2	13
Deletion	0	5	3	8

4.5.2.3 Sequencing of problematic samples

Samples with additional C's at the end of HV3 region, which prevent the read from reverse primers (HV3R and R638), were cloned into plasmids. The selected vectors were sequenced using universal primers, i.e., SP6 and T7.

Direct resequencing of the PCR amplicon of samples that harbor the homopolymeric stretch using sequencing primers returned unsatisfied results. For the forward primers, HV2F and HV3F, peak signals were excellent since from the beginning until it reached the homopolymeric region. The signal quality dropped dramatically after the poly-C tract and resulted in an “out-of-phase” registry. This condition is caused by enzyme slippage, after which the growing strand does not stay paired with the template, and repetition of varying lengths of the same template results in signals of n-1, n-2, n-3, etc. populations. Thus, peak signals after the homopolymeric region appear as wave-like patterns.

Figure 4.21 shows the sequencing results obtained from the PCR amplicon and cloned vector. For PCR amplicon, the signals after the poly-C's regions were pure noise and not interpretable.

In contrast, results from the cloned vector were satisfactory from both forward and reverse reads. When compared to the rCRS, our samples contained 6 additional nucleotides after np 573 (-CCCCAC-). This insertion resulted in the formation of a chain of 10 C's, only 6 in most of the samples and rCRS, that caused slippage during the polymerization process during PCR or sequencing.

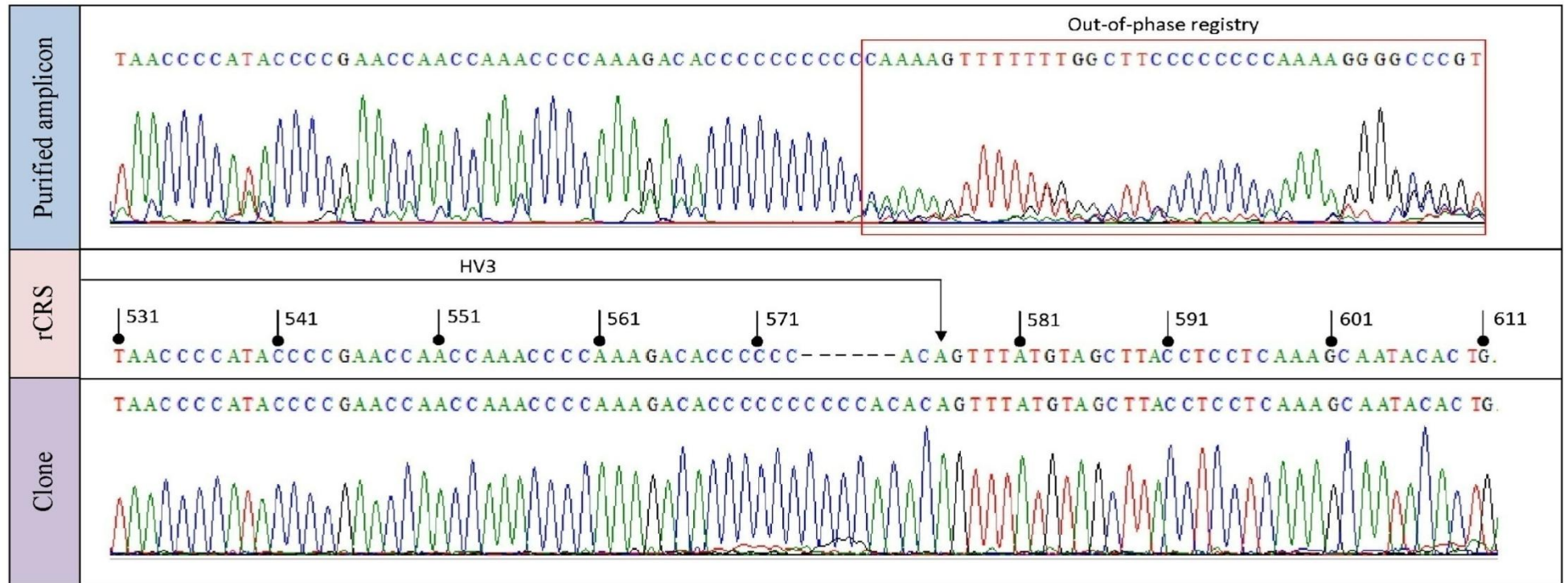


Figure 4.21 : Electropherograms of DNA sequencing via forward primer for purified amplicon and cloned vector, in comparison to rCRS; direct sequencing of the amplicon resulted in a wave-like pattern downstream to the polymeric region due to enzyme slippage; the results from cloned vector revealed the addition of 6 nucleotides after position 573 in the sample.

4.5.2.4 Haplogroup determination

Haplogroup of each tested individuals from the Sabahan indigenous populations was assigned with the aid of the Haplogrep application. The assignment was conducted based on the observed variants, as opposed to the rCRS, within the mt control region that fit to the latest phylogenetic tree build (Appendices 7, 8, and 9).

As a result, all 450 individuals were assigned to a total of 62 mt haplogroups (Table 4.15 and Figure 4.22). Without surprise, all individuals fell into 2 macro-haplogroups, M and N, which descend from the L3 – the haplogroup that gives rise to all non-African haplogroups. The distribution of haplogroups M and N in the indigenous groups was fairly even, where they covered about half the populations tested. On average, 54.7 % of these individuals belong to the macro-haplogroup M, and the remaining individuals are descendants of the macro-haplogroup N.

The macro-haplogroup M covers a number of haplogroups, besides its own sub-groups. Haplogroups C, Z, D, E, G, and Q are all derived from the macro-haplogroup M. Haplogroup M is most profusely observed in the Sabahan indigenous, where 27.8 % of these individuals fitted in the haplogroup M. Furthermore, up to 40 % of the Kadazan-Dusun people were also found to belong to this haplogroup, making it the most common haplogroup in this tribe. On the other hand, 22.7 % and 20.7 % of Bajau and Rungus individuals, respectively, were also descended from haplogroup M. The second most common M-derived haplogroup is the haplogroup E, with an average frequency of 18.2 %. The frequency of haplogroup E was especially high in the Rungus people (27.3 %).

The haplogroup N comprises of several subgroups, including another macro-haplogroup - R, and haplogroups A, I, S, W, X, and Y. Macro-haplogroup R gives rise to few haplogroups – B, F, P, U, and subgroups – H, V, J, T, K. A total of 94 (out of 150)

individuals from the Bajau were assigned under the haplogroup N. In other words, 62.7 % of Bajau individuals from our study were descendants of haplogroup N. On the other hand, 34 % of Rungus and 39.3 % of the Kadazan-Dusun people were determined to descend from the macro-haplogroup N as well.

Among the N-derived haplogroups, haplogroup B was most abundantly observed in the indigenous individuals, with an average frequency of 16.7 %. Haplogroup B was also the most frequently seen haplogroup among the Bajau people, 27.4 %. It was followed by the sub-groups of macro-haplogroup R, with an average frequency of 16 %. The contribution of macro-haplogroup R's subgroups was evenly observed in all 3 indigenous populations, where it was seen at 17.3 % in Kadazan-Dusun group and 15.3 % in both Bajau and Rungus. The frequency of haplogroup F was particularly high in the Bajaus (13.4 %), but fairly low in the Kadazan-Dusun and Rungus groups.

Table 4.15 : Distribution of mt haplogroups determined in 450 Sabahan indigenous individuals (n = number of individuals; % = frequency in percentage).

No.	Haplogroup	Kadazan-Dusun		Bajau		Rungus	
		n	%	n	%	n	%
1)	B4a1a	5	3.33	9	6.00	12	8.00
2)	B4a1a1a	0	0	1	0.67	0	0
3)	B4a2b	0	0	3	2.00	0	0
4)	B4b1	2	1.33	0	0	0	0
5)	B4b1a+207	5	3.33	8	5.33	0	0
6)	B4c1b2a2	0	0	11	7.33	0	0
7)	B4c2	1	0.67	4	2.67	0	0
8)	B4h	0	0	1	0.67	0	0
9)	B4j	1	0.67	0	0	0	0
10)	B4'5	1	0.67	0	0	0	0
11)	B5a	3	2.00	1	0.67	2	1.33
12)	B5ald	2	1.33	0	0	0	0
13)	B5b	0	0	3	2.00	0	0
14)	C	0	0	1	0.67	0	0
15)	D4b1	0	0	0	0	1	0.67
16)	D4s	2	1.33	0	0	0	0
17)	D5b	0	0	1	0.67	0	0
18)	D5b1c1	7	4.67	0	0	23	15.33
19)	D6	0	0	1	0.67	0	0
20)	D6c	0	0	1	0.67	0	0
21)	E1a1a	12	8.00	11	7.33	28	18.67
22)	E1b	0	0	1	0.67	2	1.33
23)	E1b+16261	7	4.67	1	0.67	9	6.00
24)	E2	3	2.00	3	2.00	5	3.33
25)	F1a	0	0	7	4.67	0	0
26)	F1a1a	0	0	1	0.67	0	0
27)	F1a3a	0	0	1	0.67	0	0
28)	F1a4a1	2	1.33	1	0.67	0	0
29)	F3b	0	0	1	0.67	0	0
30)	F3b1	0	0	9	6.00	3	2.00
31)	G3	0	0	1	0.67	0	0
32)	HV2	1	0.67	0	0	0	0
33)	JT	6	4.00	1	0.67	4	2.67
34)	M5a1	0	0	0	0	1	0.67
35)	M7b1'2'4'5'6'7'8	2	1.33	0	0	0	0

Table 4.15 : continuation.

No.	Haplogroup	Kadazan-Dusun		Bajau		Rungus	
		n	%	n	%	n	%
36)	M7b1'2'4'5'6'7'8 +(16192)	33	22.00	3	2.00	3	2.00
37)	M7b3	0	0	2	1.33	0	0
38)	M7b4	0	0	0	0	1	0.67
39)	M7c3c	19	12.67	15	10.00	23	15.33
40)	M9	0	0	3	2.00	1	0.67
41)	M20	0	0	3	2.00	0	0
42)	M21c	0	0	1	0.67	0	0
43)	M21d	0	0	1	0.67	0	0
44)	M31a2	1	0.67	0	0	0	0
45)	M33a1b	0	0	0	0	1	0.67
46)	M43b	0	0	2	1.33	0	0
47)	M44	0	0	1	0.67	0	0
48)	M51	0	0	1	0.67	0	0
49)	M74	0	0	1	0.67	0	0
50)	M74b1	0	0	1	0.67	0	0
51)	M80	5	3.33	0	0	1	0.67
52)	N5	4	2.67	2	1.33	4	2.67
53)	N8	0	0	3	2.00	0	0
54)	N21+195	0	0	0	0	3	2.00
55)	Q1	0	0	1	0.67	0	0
56)	R14	0	0	2	1.33	0	0
57)	R22	0	0	2	1.33	0	0
58)	R9b1a1a	0	0	1	0.67	0	0
59)	R9b2	0	0	3	2.00	0	0
60)	R9c	0	0	1	0.67	2	1.33
61)	R9c1a	26	17.33	14	9.33	21	14.00
62)	Y2	0	0	4	2.67	0	0
TOTAL		150	100	150	100	150	100

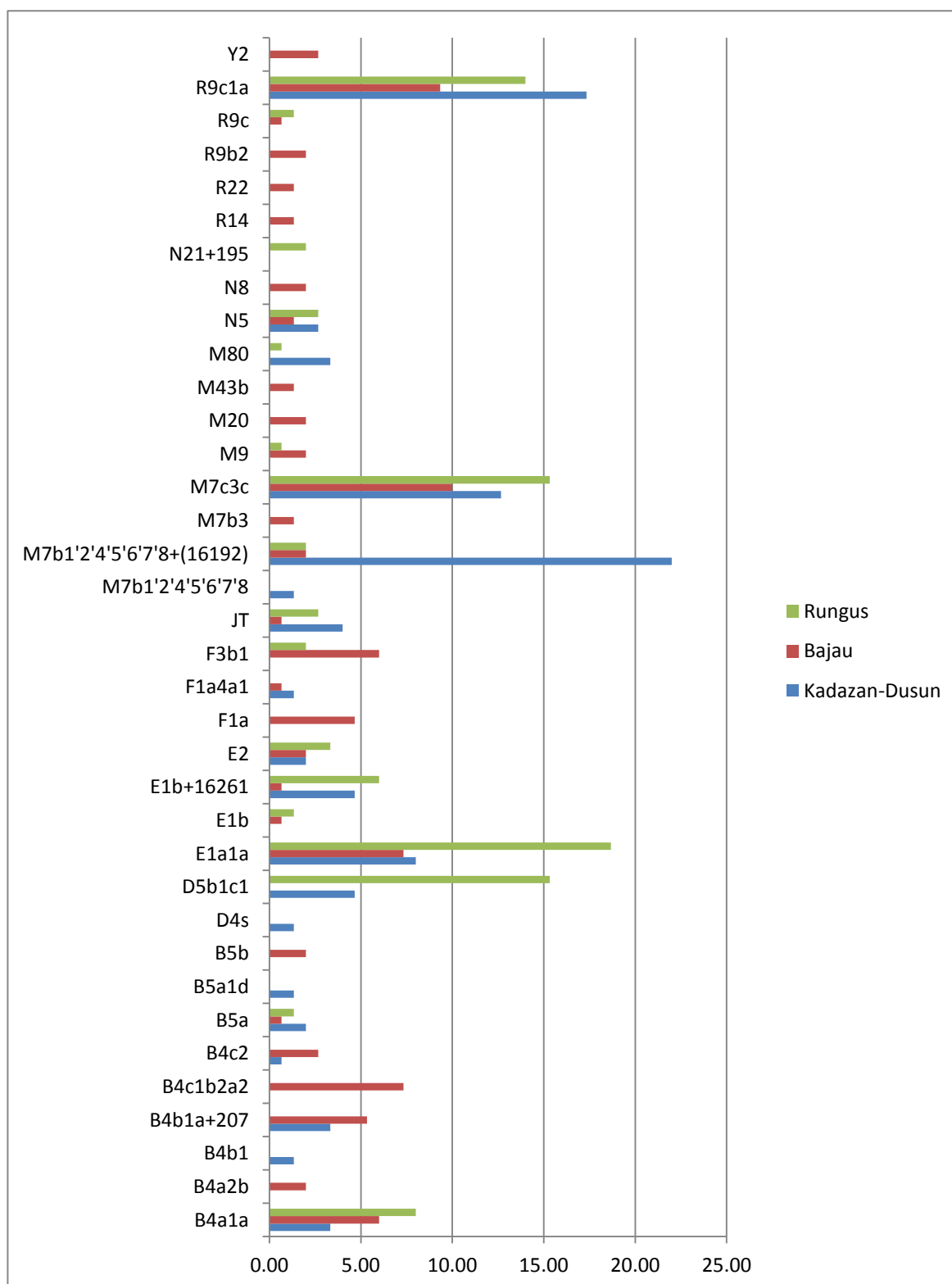


Figure 4.22 : Bar chart showing the frequencies of mt haplogroups in Kadazan-Dusun, Bajau, and Rungus populations; only haplogroups with frequency higher than 1 % in any of the populations were included.

When looking into the 62 subgroups individually, the frequency of each subgroup was seen to give additional and more detailed implications. The R9c1a haplogroup was most abundant in all the tested individuals, with frequency of 13.6 %. It was comprised of 17.3 % of the Kadazan-Dusun population, followed by 14 % in the Rungus population and 9.3 % in the Bajau population. The second most profuse haplogroup was the M7c3c, average of 12.7 %. It was highest in the Rungus group, 15.3 %, subsequently 12.7 % and 10 % in the Kadazan-Dusun and Bajau individuals. Some of the haplogroups also present predominantly in certain population, such as haplogroup M7b1'2'4'5'6'7'8 + (16192). It was as high as 22 % in the Kadazan-Dusun population, but low (2 %) in Bajau and Rungus individuals. Similar observations were also seen in the distribution of haplogroup D5b1c1. It was exclusively high (15.3 %) in the Rungus population, but low (4.7 %) in the Kadazan-Dusun and absent in the Bajau group. Furthermore, some haplogroups were also noted to restrict to a fixed population. For example, the haplogroup B4c1b2a2 present in the Rungus individuals, with frequency of 7.3 %, but not in the other 2 Sabahan indigenous groups.

4.5.2.5 Haplotype analysis - individual population

Apart from the mt haplogroup determination, haplotype analysis was also performed on all Sabahan indigenous samples sequenced in our study (Table 4.16). Haplotype analysis is conducted by assigning each sample into their respective “group” based on their absolute variations among each other. Unlike the mt haplogroup assignment that allocates the group based on certain specific variants at particular positions [which serves as the landmark (cumulatively or not) to a group], haplotype analysis determines the “group” by taking all variants that are present in the sequence into account. In other words, only sequences that are 100 % identical get to be assigned into the same “group” or haplotype.

First, all populations were tested individually. The Bajau group yielded the most number of haplotypes (89) as compared to its counterparts (Kadazan-Dusun – 60; Rungus – 64). In addition, unique haplotypes in each population were investigated. The “unique” haplotype is referred to a particular haplotype that only presents once in the entire population. Haplotypes that occurred more than once, were coined as “shared” haplotypes within the population. In fact, a majority of the observed haplotypes were unique in their respective populations. The Bajau had the most number of unique haplotypes (66.3 %), while the Kadazan-Dusun had the least (56.7 %).

We also observed the Bajau population to contain the most number of shared haplotypes. However, most (18 out of 30) of these haplotypes were only shared by 2 samples ($n = 2$). The Kadazan-Dusun population was shown to have 3 haplotypes that were shared by more than 10 individuals. On top of that, one of the haplotypes was shared by 22 individuals from the population. For the Rungus group, the highest number of individuals that shared a haplotype was 13.

Table 4.16 : Distribution of haplotypes observed within respective examined indigenous populations.

Characteristic within population	Kadazan-Dusun	Bajau	Rungus
No. of haplotype observed	60	89	64
Unique haplotype	34	59	42
Shared haplotype	26	30	22
No. of haplotype shared by n individuals			
n = 2	7	18	5
n = 3	9	5	5
n = 4	4	2	2
n = 5	1	1	3
n = 6	2	1	1
n = 7	0	3	1
n = 8	0	0	2
n = 9	0	0	2
n = 10	2	0	0
n = 13	0	0	1
n = 22	1	0	0

4.5.2.6 Haplotype analysis - combined study

The sequence data of mt control region for all tested individuals were subjected to an overall haplotype analysis. The 450 indigenous samples resulted in a total of 181 distinctive haplotypes. Notably, none of these samples shared the same sequence as the rCRS within the examined mt regions. The Bajau population harbored the most unshared haplotypes among the indigenous groups, with as many as 72 haplotypes that were only observed within Bajau individuals. On the hand, only 43 and 39 haplotypes were found to be conserved in the Rungus and Kadazan-Dusun populations, respectively (Figure 4.23). Amongst, 108 haplotypes (59.67 %) were detected once in the entire sample cohort.

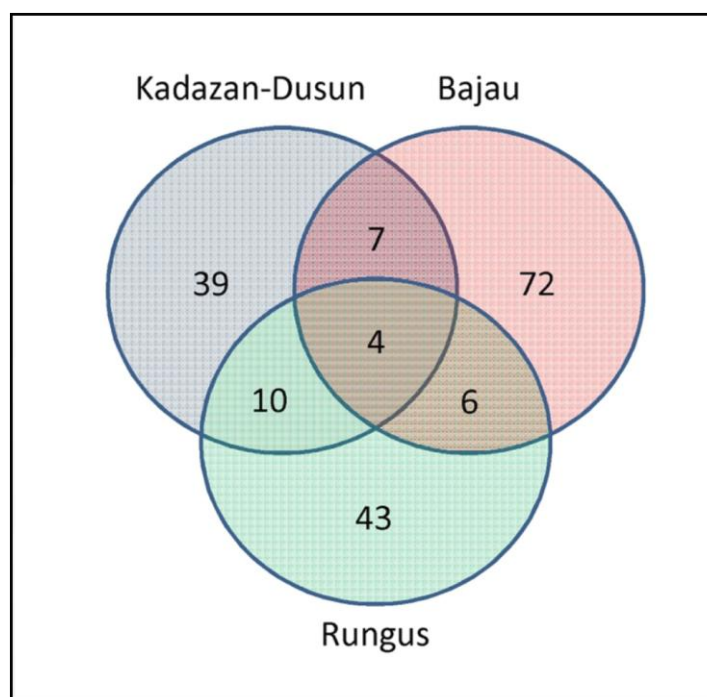


Figure 4.23 : Diagram illustrating the division of haplotypes for the Kadazan-Dusun, Bajau, and Rungus populations.

A total of 27 (14.92 %) haplotypes were shared by 2 or more populations. Kadazan-Dusun and Rungus have the most number of shared haplotypes, i.e., 10 % or 37.04 % of the shared haplotypes. There were 7 shared haplotypes between Kadazan-Dusun and Bajau and only 6 for Bajau and Rungus. There were 4 haplotypes found to present in all 3 Sabahan indigenous populations. These haplotypes were determined to belong to 2 haplogroups, R9c1a (hap6 and hap13) and E1a1a (hap28 and hap43).

The frequencies of the most common haplotypes are summarized in Table 4.17. Hap4 was observed most frequently in the studied populations, at a frequency of 5.11 %. However, hap4 was not found in the Bajau population. Based on the variants in the control region, hap4 was determined to belong to the lineage of mt haplogroup M7c3c.

Hap7 was the second most commonly observed haplotype, present at frequency of 4.89 % or in 22 individuals. The haplotype was found solely restricted to the Kadazan-Dusun group, belonging to the haplogroup M7b1'2'3'4'5'6'7'8 (+16192). Following that was the hap6, which was the third most prominent haplotype among the indigenous individuals. As discussed in the early part, hap6 was also the most frequently observed haplotype shared by all 3 populations. Next in the line were hap6 and hap27 (belonging to haplogroup E1a1a), that were seen in 13 and 12 individuals, respectively.

Table 4.17 : Numbers of haplotypes that are shared and most frequently found in the 3 indigenous populations in Sabah.

Haplotype	Haplogroup	Total number observed	Number observed in individual population		
			Kadazan-Dusun	Bajau	Rungus
Haplotype shared by all populations					
Hap6	R9c1a	18	10	2	6
Hap13	R9c1a	11	3	7	1
Hap28	E1a1a	8	6	1	1
Hap43	E1a1a	12	1	4	7
Most frequently observed haplotype in Kadazan-Dusun, Bajau, and Rungus					
Hap4	M7c3c	23	10	-	13
Hap7	M7b1'2'3'4'5'6'7'8 (+16192)	22	22	-	-
Hap6	R9c1a	18	10	2	6
Hap27	E1a1a	13	5	-	8
Hap43	E1a1a	12	1	4	7
Hap3	D5b1c	12	3	-	9
Hap13	R9c1a	11	3	7	1
Hap9	D5b1c	10	1	-	9
Hap40	M7c3c	9	1	-	8
Hap26	B4a1a	8	3	-	5
Hap28	E1a1a	8	6	1	1
Hap2	R9c1a	8	4	-	4
Hap15	E1b + 16261	7	6	-	1
Hap82	M7c3c	7	-	7	-
Hap96	F3b1	7	-	7	-
Hap101	R9c1a	6	-	1	5
Hap80	M7c3c	6	-	6	-

4.5.2.7 Principal component analysis

The frequencies of mt haplogroups of 28 populations residing in East Asia and SEA regions were included to construct the PCA plot (Appendix 10). The 2 main PCs attributed to 28.45 % of all variations ($D1 = 14.69\%$; $D2 = 13.75\%$).

All populations were segregated into 2 distinctive clusters (Figure 4.24), separated by the central axis of the plot. The lower part was occupied by East Asian populations, including the Chinese from Mainland China. These populations (NE China, SE China, NW China, SW China, and Hainan) huddled closely together, which representing genetic homogeneity among them. The mt haplogroups that corresponded to the position of these populations were mostly root haplogroups, i.e., F1a1*, D*, C*, and M7b*. Some Mainland SEA populations (Vietnam, Thailand, Kinh, Cham) were found near to the East Asian cluster, indicating high genetic resemblance between them.

At the upper part of the plot, the SEA populations were scattered from the centre to the right portion. The Indonesian populations were found separated according to their geographic localities. In general, they can be grouped into 2 less obvious clusters, consisting of East and West Indonesians. The West Indonesian populations were situated at the right side of the plot, with Sumatra, Java, and Lombok forming a sub-cluster and Bali population was situated slightly distant from the cluster. The Orang Asli and Melayu Malay were seen within the West Indonesian populations, beside the Bali people. The East Indonesian populations (Alor, Sulawesi, Sumba, and Ambon) were located close to the central axis. The Philippines and Sabahan indigenous populations were included in the East Indonesian cluster. The 2 Taiwanese Aboriginal populations formed a distinctive cluster from other populations, situated along the central axis at the left part of the plot. The position of these Taiwanese Aborigines were due to the high frequencies of haplogroups F4, B4b1, B4a2, and M7b3.

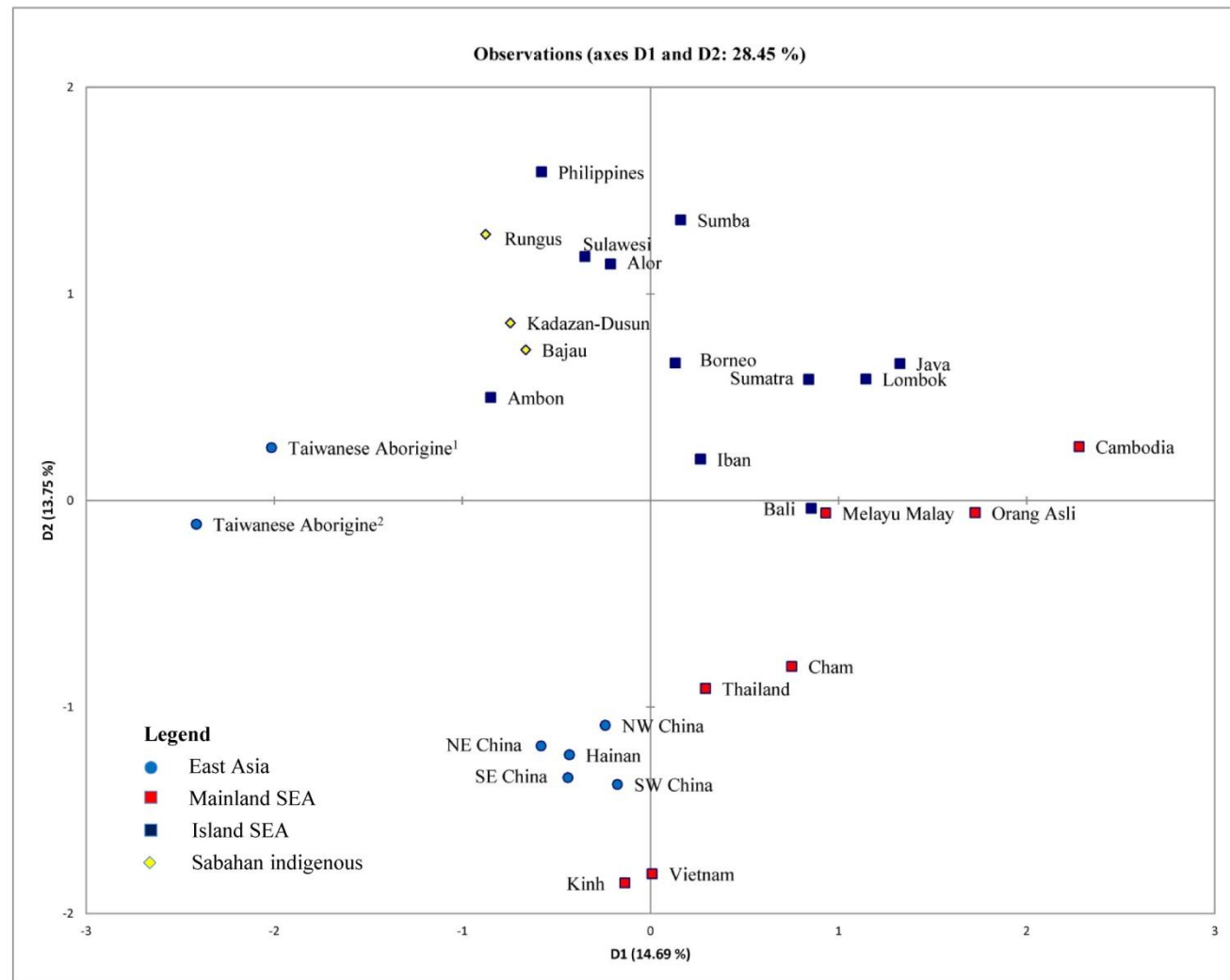


Figure 4.24 : PCA plot illustrating clusters of populations based on mt haplogroup frequencies in East Asia and SEA regions.

4.5.2.8 Phylogenetic analysis

Based on 678 analyzed sites (out of a total of 737) in 1,036 sequences of the mt HV regions, a rooted NJ phylogenetic tree was constructed for the Sabahan indigenous groups and populations within and nearby SEA region (Figure 4.25).

The RSRS served as the outgroup in the constructed NJ tree. Unlike the rCRS that was originated from a contemporary European mt lineage – H2a2a1, the RSRS was built based on mtDNAs from *H. neanderthalensis*. Therefore, it could be a better reference point in the mt phylogenetic analysis.

In general, the Sabahan indigenous individuals were found to disperse into different sub-clades in the phylogenetic tree. Majority of their lineages were shared by other populations, e.g., Chinese, Taiwanese Aborigines, Filipinos, Indonesians, Orang Asli, Vietnamese, and Oceania.

When corresponding the distribution of the Sabahan indigenous individuals with their predetermined mt haplogroups, the old (e.g., M80 and R14) and young (e.g., Y2, M7c3c, and M7b1) lineages were found scattering throughout the tree.

The distribution of Bajau individuals was even, where they were found to present in vast number of clades in the phylogenetic tree. They shared a number of their lineages with other ISEA populations (Filipino and Indonesian) and Taiwanese Aborigines. These lineages included Y2, B4c1b2a2, F1a, F3b, R9b2 and M20. On the other hand, some lineages were shared with the Indochina populations, e.g., R14, R22, M21d, and B4c2. The Bajau group exhibited extremely diverse lineages comprising of both old and recent haplogroups that were also seen in other populations within and nearby SEA region. However, none of these lineages formed distinctive clade in the resulted phylogenetic tree.

Figure 4.25 : Condensed NJ tree constructed via 1,036 mtDNAs; detailed phylogeny of clusters (in box) are available in Chapter 5.

On the other hand, the distributions of the Kadazan-Dusun and Rungus people were found concentrated to each other, unlike their Bajau counterpart. The Kadazan-Dusun and Rungus populations were seen to be related closely as they shared most of their mt lineages. The Kadazan-Dusun and Rungus were associated to the Bajau group by several mt lineages that were also shared by other ISEA populations, i.e., M7c3c, M7b1, B4a1a, E1b, Elala, and R9c. Although less diverse, the Kadazan-Dusun and Rungus populations were found to harbor distinct lineages that were not found in other SEA populations, i.e., haplogroups N5 and JT.

DISCUSSION

5 DISCUSSION

5.1 VNTRs in the Sabahan indigenous populations

When examining and comparing the distribution of 2 VNTRs in the DAT-1 gene in the Kadazan-Dusun, Bajau, and Rungus individuals, the observed allelic and genotypic frequencies were similar for Kadazan-Dusun and Rungus. In contrast, there was a higher degree of genetic diversity in the Bajau population, in terms of allele variants and heterozygosity. All these indicate a closer genetic relationship between the Kadazan-Dusun and Rungus groups compared to the Bajaus. The population differentiation study and AMOVA revealed that the 3'UTR VNTR expressed higher inter-population variation (4 % to 6 %) than the Intron-8 VNTR (2 % to 4 %). Majority of the genetic variability of the 2 VNTRs was due to differences within the populations, rather than inter-population. However, it is sufficient to show that the 3 Sabahan indigenous populations are distinctively separated from each other, in view of their genetic content. The combined PD for these VNTRs was highest in the Bajau population (0.726) (refer to Table 4.3), but only about half the value in the Kadazan-Dusun and Rungus people.

5.1.1 Distribution in world and neighboring populations

Table 5.1 shows the distribution of VNTR in the 3'UTR of DAT-1 gene in populations living in various regions of the world, i.e., Africa, Europe, America, Asia, and Oceania. There were 2 predominant alleles observed in all populations, i.e., alleles 9 and 10. The allele 10 was the most prominent allele, with more than 50 % of all observed alleles in each population. Other alleles of varying sizes were seen at relatively low frequencies.

The frequencies of allele 10 in the African populations ranged from 0.543 to 0.769. Noteworthy, other alleles, 3, 7, 8, and 9, were also present constantly. For the European populations, the frequencies of the major allele, allele 10, were similar to the Africans, ranging from 0.691 to 0.784. However, the frequencies of other alleles were very low, < 0.006 , with exception of the allele 9. Perhaps these alleles were either eliminated or reduced in the ancient migrants after a series of founder effects before the group(s) reached and colonized the Europe. Thus, this resulted in the low frequencies of other alleles in the present European as opposed to the African populations.

For populations in America, the Hispanic Americans showed similar distribution to the African and European groups. The native people in this region, however, exhibited different patterns of distribution. The frequencies of allele 10 were high in the natives, with a frequency of more than 0.9. In the Karitiana group, an indigenous population of Brazil, the allele 10 was fixed.

In Asia, the Arabs (Kuwaiti and Omani) have similar distribution to the Europeans, where alleles 9 and 10 present at much higher frequencies than other variants. It could be suggested that the Arabs have a closer genetic ancestry to the Europeans. Kuwait and Oman lie at the exit point of Africa, between Asia and Europe.

Table 5.1 : Global allelic distribution of VNTR in the DAT-1 3'UTR; frequencies of the most prominent alleles in each population are in red.

Region	Population	2n	Allelic frequency of DAT-1 3'UTR VNTR								
			3	6	7	8	9	10	11	12	13
Africa	Biaka	138	0.014	-	0.167	0.043	0.232	0.543	-	-	-
	Yoruba	156	0.038	-	0.013	0.032	0.128	0.769	-	0.019	-
	African-American	310	0.010	-	0.030	0.044	0.174	0.725	0.010	-	0.007
Europe	European-American	324	0.003	-	-	-	0.287	0.704	0.003	-	0.003
	Hungarian	356	-	-	-	-	0.303	0.691	0.006	-	-
	Irish	204	-	-	-	0.005	0.304	0.691	-	-	-
	Russian	1004	-	-	0.001	-	0.215	0.784	-	-	-
North America	Hispanic American	162	0.020	-	0.007	-	0.277	0.682	0.007	-	0.007
	Amerindian	86	-	-	-	-	0.012	0.988	-	-	-
	Maya	104	-	-	-	-	0.067	0.923	0.010	-	-
South America	Karitiana	108	-	-	-	-	-	1.000	-	-	-
	Ticuna	130	-	-	-	-	0.069	0.931	-	-	-
Asia	Kuwaiti	164	0.012	-	-	0.018	0.323	0.622	0.024	-	-
	Omani	220	0.009	0.018	0.009	0.005	0.332	0.609	0.018	-	-
	Mongolian	156	-	-	0.026	-	0.051	0.904	0.013	-	0.006

Table 5.1 : continuation.

Region	Population	2n	Allelic frequency of DAT-1 3'UTR VNTR								
			3	6	7	8	9	10	11	12	13
East Asia/ Southeast Asia	Ami	76	-	-	-	-	0.132	0.868	-	-	-
	Atayal	84	-	-	-	-	0.060	0.940	-	-	-
	Bunun	112	-	-	-	-	0.150	0.850	-	-	-
	Paiwan	68	-	-	-	-	0.040	0.960	-	-	-
	Han	432	-	0.007	0.021	-	0.037	0.907	0.028	-	-
	Taiwanese	508	-	0.002	0.010	-	0.059	0.915	0.014	-	-
	Hakka	68	-	-	-	0.015	0.088	0.868	0.029	-	-
	Japanese	230	-	-	0.009	-	0.039	0.939	0.013	-	-
	Cambodian	50	-	-	-	-	0.200	0.800	-	-	-
	Kadazan-Dusun	542	-	-	0.002	-	0.020	0.978	-	-	-
	Bajau	318	-	-	0.019	-	0.129	0.843	0.003	0.006	-
	Rungus	418	-	-	-	-	0.038	0.962	-	-	-
Oceania	Nasioi	48	-	-	-	0.063	-	0.938	-	-	-
	Micronesian	74	-	-	-	-	0.135	0.865	-	-	-
	YoIngu	36	-	-	0.028	-	-	0.972	-	-	-

Therefore, it is possible that Arabs and Europeans share the similar ancestry lineage from the early migratory wave out of Africa or during a later recolonization. The Mongolians, on the other hand, showed a different distribution. The allele 10 was highly prevalent, at 0.904, as in the native people in America.

Distribution of the VNTR in East Asia, SEA, and Oceania was similar. In all populations, allele 10 was highly predominant over the other variants, with frequencies higher than 0.8. The distribution pattern of populations in these regions was similar to the Mongolians and American natives. It has been postulated that the world colonization by modern humans occurred in several tiers since the first exodus from Africa about 70,000 years ago (Richards, Bandelt, Kivisild, & Oppenheimer, 2006). According to various genetic findings, the colonizing process did not occur in a single migration out of Africa. There were several waves that gave rise to the different genetic composition of modern human populations today. The observation based on the examined VNTR seems to match this postulation. The Native Americans may have originated from the earlier migratory wave from Africa to Asia following the “southern coastal route”, which also spread in southward manner to East Asia, SEA, and Oceania (Richards, et al., 2006). Along the route of dispersal, these ancient settlers passed on the genome, with the prominent allele 10 to their descending groups. Conversely, the Europeans and Arabs may be descendants of another wave of migrants, the later “northern route”, who travelled northward scattering around Europe (Richards, et al., 2006).

The Sabahan indigenous groups in our study showed similar distribution pattern as in the populations of East Asia, SEA, and Oceania. Therefore, they may be traced back to the ancestry of the “southern coastal route” migration.

The distribution of the VNTR in the Intron-8 of the DAT-1 gene for world populations is outlined in Table 5.2. In contrast to the 3'UTR VNTR, the Intron-8 VNTR was not as diverse, where a majority of the populations only consisted of 2 allelic variants. The 2 most dominant alleles were the 5- and 6-repeat, in which the latter was found most prominent in all populations, except in the African groups. In the African populations (Biaka, Mbuti, and Yoruba), allele 5 was observed in higher frequencies, ranging from 0.628 to 0.744. However, the alleles 5 and 6 existed evenly in the African-American. Allele 4 was found to present exclusively in the African groups.

The distribution pattern of the VNTR in the African populations differs from populations in other parts of the world. This could be described by the founder effect, where scientists proposed that all non-African populations today are descendants from a small group of eastern African colonists, causing the drastic genetic variations between the modern African and the world populations (Forster, 2004). Allele 6 was found predominantly in all European groups, with frequencies above 0.75, except in the Samaritans (0.59). The distribution of alleles 5 and 12 in the Samaritans was fairly equal. The deviation of Samaritans from other Europeans may due to the fact they live in a close community and adopt a endogamous marriage system (Bonne-Tamir, et al., 2003). They are thus, stranded from other populations and have very limited interaction even with their neighboring groups in close geographic proximity.

The Native Americans present with high frequencies of allele 6. The VNTR was monomorphic in Karitiana and Ticuna populations. Similar distribution was also observed in the Oceania groups. The frequencies of alleles 5 and 6 in the East Asia and SEA populations were comparable to other populations, except the African. On the other hand, a few variants (7-, 8-, and 9-repeat) present exclusively in these regions.

Table 5.2 : Global distribution of alleles for DAT-1 Intron-8 VNTR in populations from different regions; the 2 most abundant alleles in each population are in red.

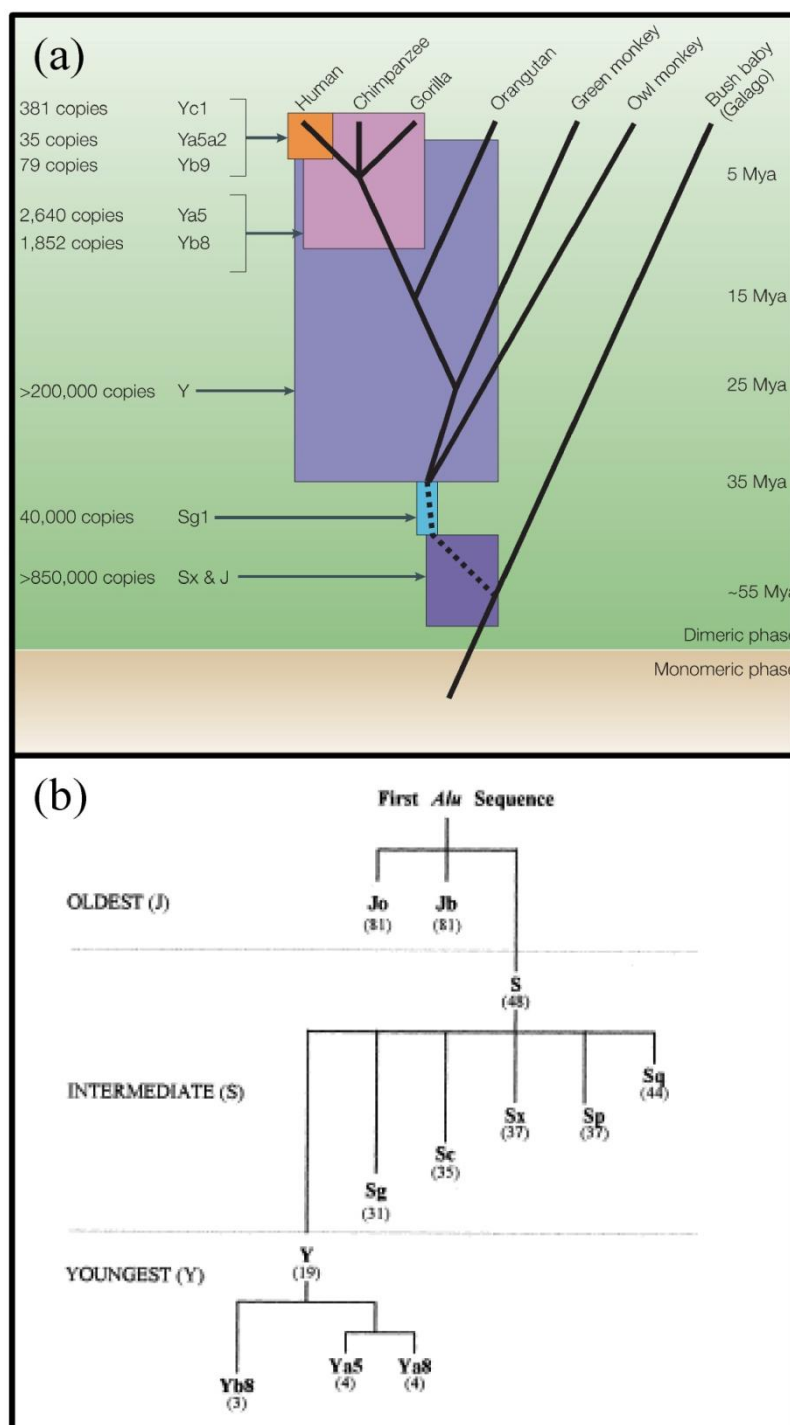
	Allelic frequency of DAT-1 Intron-8 VNTR							
Population (2n)	4	5	6	7	8	9	12	13
Africa								
Biaka (138)	-	0.667	0.333	-	-	-	-	-
Mbuti (78)	0.064	0.744	0.192	-	-	-	-	-
Yoruba (156)	0.006	0.628	0.365	-	-	-	-	-
African – Americans (182)	-	0.500	0.495	-	-	-	0.005	-
Europe								
Samaritan (78)	-	0.218	0.590	-	-	-	0.192	-
Danish (100)	-	0.150	0.820	-	-	-	0.010	0.020
Finn (72)	-	0.097	0.889	-	-	-	0.014	-
Russian (96)	-	0.240	0.750	-	-	-	-	0.010
North America								
Pima – Mexico (106)	-	0.038	0.962	-	-	-	-	-
Maya (98)	-	0.061	0.939	-	-	-	-	-
South America								
Karitiana (110)	-	-	1.000	-	-	-	-	-
Ticuna (130)	-	-	1.000	-	-	-	-	-
Oceania								
Nasioi (46)	-	0.087	0.913	-	-	-	-	-
Micronesian (74)	-	0.270	0.730	-	-	-	-	-
East Asia/Southeast Asia								
Ami (80)	-	0.175	0.812	0.012	-	-	-	-
Atayal (84)	-	0.071	0.929	-	-	-	-	-
Japanese (98)	-	0.163	0.827	-	-	0.010	-	-
Cambodian (50)	-	0.060	0.940	-	-	-	-	-
Hakka (70)	-	0.114	0.871	-	0.014	-	-	-
Kadazan-Dusun (542)	-	0.103	0.895	0.002	-	-	-	-
Bajau (318)	-	0.204	0.792	-	-	-	0.003	-
Rungus (418)	-	0.081	0.914	0.005	-	-	-	-

In the Sabahan indigenous populations, the Kadazan-Dusun and Rungus groups were found to harbor the allele 7, which was also seen in the Ami tribe, a Taiwanese Aboriginal population. Scientists have suggested the role played by the Taiwanese Aboriginal groups in the dispersal of human population in the SEA region (Meacham, 1988). Therefore, allele 7 may represent the shared genetic content between the Borneo people and the Taiwanese Aboriginal groups. On the other hand, allele 12 that was observed in the Bajau group is also found in Europeans. Although it is uncommon in the Bajau, it may be the result of gene flow between populations of these regions. However, it could also be an independent event whereby the new allele could have emerged due to an event of replication slippage.

5.2 *Alu* elements

Ever since the insertion of the first *Alu* element in the primate lineage, the evolution of these elements continues and gives rise to a variety of subfamilies. Often, these inserts accumulate new sequence variation that eventually result in the formation of distinctive classes (Shen, et al., 1991). Generally, all *Alu* can be sub-divided into 3 clades, i.e., *AluJ*, *AluS*, and *AluY* (Batzer, et al., 1996). The *AluJ* and *AluS* represent the 2 oldest families (Figure 5.1). The *AluJ* was believed to occur 81 Ma, when the rodents and primates parted for separated lineages (Kapitonov & Jurka, 1996). *AluS* denotes the most active phase of *Alu* insertion throughout the evolution of *Alu* element about 35 to 55 Ma, where the amplification rate was peaked. However, both the *AluJ* and *AluS* were found to have lost their capability to amplify in modern human lineages (Lee flank, Liu, Hashimoto, Choudary, & Schmid, 1992; Matera, Hellmann, Hintz, & Schmid, 1990). Although rare, *Alu* elements that retain the ability to produce new *Alu* insertions are limited to certain members in the youngest derivatives of the *AluY* subfamily, particularly the Ya5/8 and Yb8 (Batzer, et al., 1995).

The *AluY* subfamily has begun to integrate into the human genome about 4 to 6 Ma, after the divergence of humans from the African ape lineage (Batzer, et al., 1996). Most of these insertions are monomorphic in the modern human groups, as they were inserted before the radiation of the African population. Nevertheless, almost a quarter of these elements (approximately 1,200) happened so recently that they appear to be polymorphic in the modern human populations (Arcot, Fontius, Deininger, & Batzer, 1995). Therefore, the examination of these human specific elements and their variants are handy to access the genetic profile of various human groups that could reveal valuable information on the human evolution. *Alu* insertions confer a more balanced and comprehensive interpretation of genetic information from both parents, unlike uniparental markers that are often prone to selection and drift caused by sexual bias.



(Batzer & Deininger, 2002; Mighell, Markham, & Robinson, 1997)

Figure 5.1 : Paths of evolution of *Alu* elements (a) the expansion of *Alu* insertions throughout the primate lineages, which includes the active involvement of various subfamilies at different timeframes (b) evolutionary relationship of the 12 major subfamilies of *Alu* elements from the 3 clades; the numbers indicate the time of occurrence in mya (Ma).

5.2.1 *Alu* insertions in the Sabahan indigenous populations

In the present study, the genetic relevance among 3 major indigenous groups, i.e., Kadazan-Dusun, Bajau, and Rungus in the state of Sabah, Malaysia, were examined by utilizing genetic data obtained from 6 polymorphic *Alu* insertions. We also compared the distribution with populations in other regions of the globe. These markers were evaluated for their efficiency, individually and collectively, and suitability as markers for forensic and human identification investigations in the local context. To date, there is no previous publication reporting on the characterization of genetic association of these indigenous groups based on the distribution of polymorphic *Alu* insertions. Thus, the present study is the first to report the genetic structure of the 3 major Sabah indigenous populations in terms of the *Alu* elements distribution.

High expected heterozygosity values were observed in most of the markers examined, except in the HS3.23 and APO insertions. This is because the insertion is highly dominant in all the 3 populations, where the frequencies of the inserted allele were greater than 0.93. Heterozygosity was comparable for all the 3 populations, suggesting that these populations may share a common source of genetic ancestry.

Gene diversity analysis reveals that all markers were present with considerably high degree of diversity ($H_T > 0.39$), except in the HS3.23 and APO insertions ($H_T = \sim 0.1$). A majority of the genetic variability is contributed by differences between individuals of a population (average $H_S = 0.327$), whereas only a small portion of the variability comes from genomic diversity between populations ($G_{ST} = 0.004$; $D_{ST} = 0.001$). The low G_{ST} implies that there is minimal degree of genomic differentiation between the Kadazan-Dusun, Bajau, and Rungus populations. Concordantly, the AMOVA study also showed that these indigenous populations are genetically homogenous as only 0.533 % of the differences are attributed by the variations among the populations. Only

2 out of 6 markers (HS4.32 and TPA25) had exhibited significant heterogeneity. Noteworthy, PV92 and B65 insertions were shown to have no difference among the populations (from both differentiation analysis and AMOVA). Hence, these markers are not useful to illustrate the disparity of genetic structure for the tested populations, in terms of their insertion frequency.

5.2.2 World distribution of *Alu* insertions

Figure 5.2 displays the line chart that represents the trend of *Alu* insertion frequencies of modern human populations globally using data obtained from the database and previously published works. Populations adopted in the chart can be generally grouped according to their geographical homeland, i.e., Africa, East Asia and SEA, Europe, and South Asia. The insertion frequency varied across all populations. In essence, the insertion frequencies of these *Alu* markers might correlate with their geographical distribution, although some markers exhibited even distribution in all populations. The study of *Alu* elements in human populations may shed light on the dispersal route of the prehistoric colonization, but the data should be interpreted carefully because the insertion frequencies could be affected by selection and drift during expansion.

The frequency of APO insertion was high (0.70 to 0.97), except for the Nguni from Africa, and it did not show substantial variation in all populations. The insertion frequencies of several populations were near to fixation (> 0.93), i.e., Africa (Zaire Pygmy), East Asia/SEA (Vietnamese, Kadazan-Dusun, Bajau, Rungus), Europe (Finnish, French, Northern European, Polish). Similar observations were also seen in the distribution of the B65 marker, where there was no specific recognizable pattern for populations in the same continent. The insertion frequencies of the B65 marker fluctuated among populations in all continents, except the East Asia/SEA, with a consistent distribution of 0.4 to 0.5.

In terms of the HS4.32 insertion, the rate was high in European populations, but low in most African groups. The insertion frequency was comparable in South Asian and some populations in East Asia/SEA. Interestingly, the Sabahan indigenous populations expressed much lower insertion frequencies than their East Asian counterparts (Cambodian, Chinese, Japanese, Malay, and Vietnamese). This implies that the

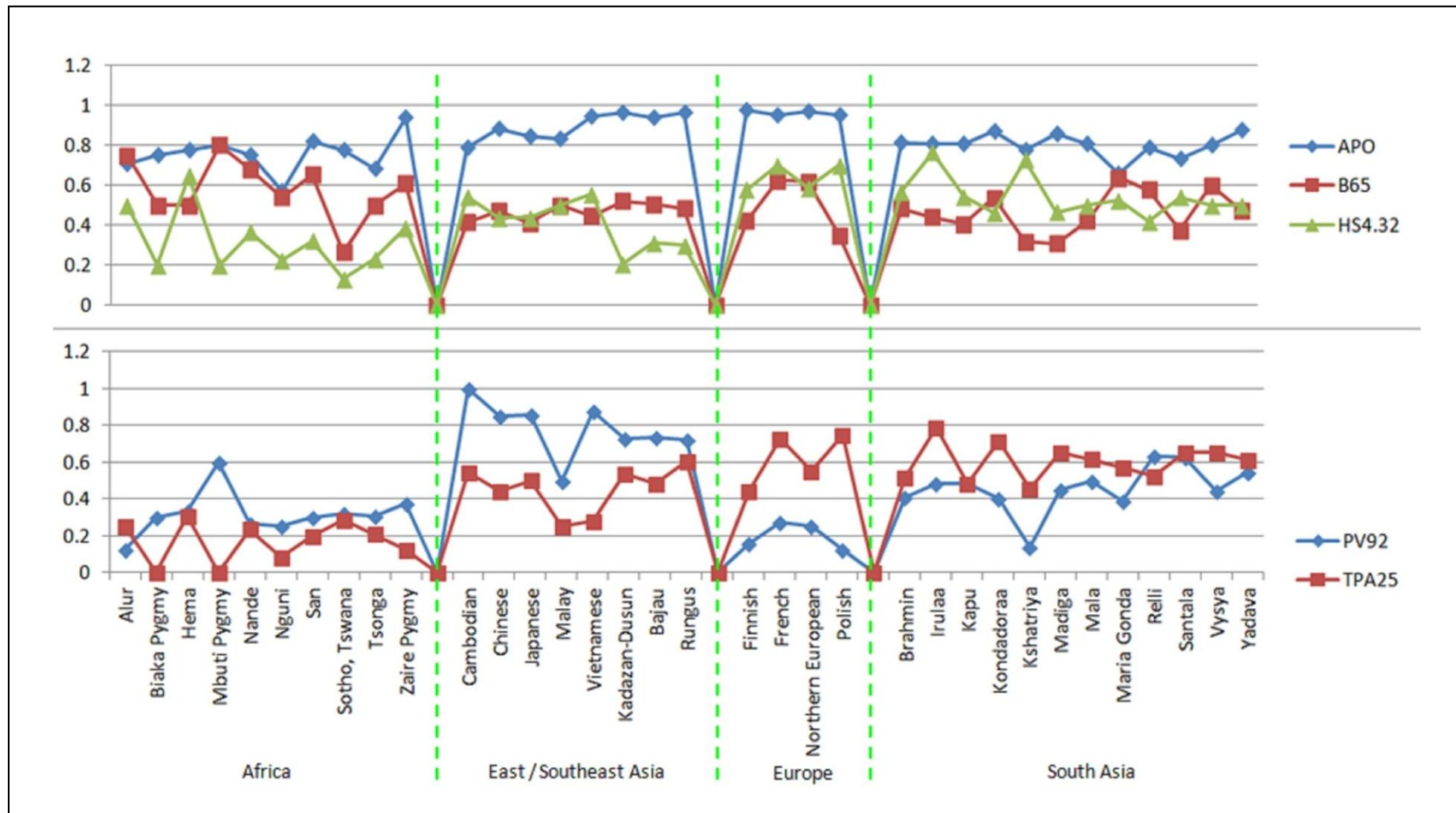


Figure 5.2 : Line chart shows insertion frequencies of *Alu* markers (APO, B65, HS4.32, PV92, and TPA25) in modern human populations from different continents.

Sabahan indigenous and East Asian groups might have descended from a different ancestry, or the low insertion rates in the Sabahan indigenous individuals may be a result of founder/bottleneck effects when these populations migrated away from the parent population.

On the other hand, the PV92 insertion was generally lower in all populations from Africa, Europe, and South Asia, ranging from 0.2 to 0.5. In contrast, the insertion rates were high for populations in East Asia and SEA, with an approximate average of 0.8 (with the exception in the Malay population). The insertion was also observed to be fixed in the Cambodian population. Hence, the PV92 insertion could serve as a population inference marker for East Asia/SEA region. Furthermore, it reflects a possible isolation of these settlers into restricted populations perhaps due to catastrophic events (such as Toba eruption about 70,000 years ago). The recovered populations then continued to colonize the world in different directions, resulting in the genetic discrepancies in modern human populations.

A different scenario was seen in the case of the TPA25 marker. This insertion was predominant in populations outside of Africa, with rates doubling those of the African groups. This observation may indicate that the ancient group(s) that emigrated from Africa 80,000 years ago may carry high proportion of the insertion, which is then passed on to their descendents during world colonization. In addition, the TPA *Alu* insertion, which is located in the Tissue Plasminogen Activator (PLAT) gene, may favor the survival of these ancient settlers against harsh environmental challenges during the great migration.

5.2.3 PCA plots and NJ trees

The genetic relevance of the indigenous groups with global populations from different regions was illustrated by PCA plots. In the first PCA that compared the Sabahan indigenous groups to the world populations, the 2 main PCs contributed to substantial variations (89.71 %) among the tested populations. All populations were found to clump into a few clusters according to geographic continents. The African groups were scattered in the left portion of the plot, clearly separated from other populations. This observation supports the African-origin hypothesis, as the parent population(s) is expected to have greater diversity than the derived groups (Jorde & Wooding, 2004). It is especially true when migration occurs in a small group; the hypothesized “out of Africa” movement only involved 2,000 to 5,000 individuals (Zhivotovsky, Rosenberg, & Feldman, 2003). Due to the limited number of genetic variants carried by the migrant population, the new population establishes a new genetic pool that is less diverse compared to the origin. Although the Europeans and South Asians were situated in close proximity in the PCA, they were separated unambiguously, whereby the South Asians were situated above the European cluster. While separated from each other, the Europeans and South Asians have higher genetic similarity, as exhibited by the *Alu* markers, than with populations from other regions. This suggests that they may have descended from a shared ancestry or have greater degree of genetic admixture during the prehistoric migration. Genetic studies have implied a mix-ancestry for populations in South Asia, whereby the influence from both west and east sides was thought to be involved in the shaping of their genetic structure (Metspalu, et al., 2011). The Sabahan indigenous populations were placed indistinctively within the East Asia/SEA cluster. This cluster was isolated from the African, Europeans, and South Asians, suggesting a close relationship among populations within the cluster. The 3 indigenous populations were clustered close to each other in the PC plot, showing that these populations share a

high degree of genetic similarity among them. This is in agreement with our interpretation obtained from the population differentiation analysis.

Similar interpretations were also observed from the NJ phylogenetic tree analysis. Based on the NJ tree, all populations segregated into 3 major clades – the first clade by African populations, second by South Asians and Europeans; and the last clade consisted of populations from East Asia and SEA. Interestingly, the Malay group was seen to cluster within the African clade. In the second clade, the Europeans branched off after the South Asians. Again, this observation indicates a close genetic relevance of the South Asians and Europeans, in agreement with the PCA results. There were 2 branches in the East Asia/SEA clade; the first branch comprised of the Japanese, Chinese, Vietnamese, and Cambodian; whereas the second branch consisted of the 3 Sabahan indigenous groups. This shows that the Sabahan indigenous groups are genetically distinct from the populations in the first branch, despite their close relationship as indicated by the NJ tree.

The second PCA plot was constructed to assess the genetic association of various populations in SEA and its neighboring regions, comprising of populations from Taiwanese Aborigines, Mainland SEA, ISEA, East Asia, and Sabahan indigenous. A total of 88.18 % of variation was summarized by the first 2 PCs. The ISEA groups were situated at the left side of the plot, showing that these populations were distinct from other populations, except the Filipinos. The Sabahan indigenous groups remained closely clustered, signifying the level of their genetic similarity. These indigenous groups were grouped within the Mainland SEA and East Asia clusters, together with the Han Chinese, Japanese, Cambodian, and Vietnamese. This is in concordance with the “out of Taiwan” hypothesis, where the ISEA was postulated to be recolonized by agriculturists from southern China through Taiwan and Philippines (Bellwood, 2007). The Filipino group was found next to the Sabahan indigenous groups. It was believed

that the Sabahan indigenous populations, particularly the Bajau tribe, were recent migrants from the Philippines (Nimmo, 1968). The Sabahan indigenous populations showed higher affinity to the Mainland SEA and East Asian populations than the ISEA counterparts. A possible explanation regarding the discrepancies of the ISEA populations with the Filipino and Sabahan indigenous populations is that these populations may descend from different waves of expansion as migrations occurred continuously in the region. In another investigation of population genetic structure via mtDNA, Y-chromosomes, and autosomal markers, Ibans (an indigenous group residing in the neighboring state of Sabah, Sarawak) was also shown to have high genetic similarity to the Mainland SEA populations (Simonson, et al., 2011). It is therefore, indicative that the indigenous populations in the Borneo Island may be descendants of migrating settlers from Mainland SEA or East Asia. Similar remarks have been observed in a genetic study based on SNP arrays, whereby SEA populations were found to have significant influence on the genetic composition of the East Asian populations (Abdulla, et al., 2009). It was suggested that all populations in SEA and East Asia were direct descendants of the migrants from the primary wave into the continent, expanding from southern Asia (Abdulla, et al., 2009; Xing, et al., 2010). On the other hand, a genetic study on Andaman Islanders had revealed a recent migration from SEA into the region that occurred some 18,000 years ago (Thangaraj, et al., 2005). This has again highlighted the influence of SEA populations to the peopling patterns of their neighboring regions. Despite extensive and in-depth studies conducted by scientists, it is still difficult to understand and conclude the routes taken by the ancient man during the prehistoric migration. This is largely accounted by several factors, such as the interpretations that were only based on a limited number of sex-biased markers and analysis methods (Stoneking & Delfin, 2010). In studies determining the peopling

pattern in a sub-region (e.g., ISEA), the population genetic data is often complicated by admixture among the groups.

The Taiwanese Aborigines were included in the analysis because Taiwan is one of the key “stations” involved in the dispersal out of Taiwan. Furthermore, it has been shown that the colonization of ISEA is tightly associated with these aboriginal populations (Diamond, 1988; Oppenheimer & Richards, 2001b). Our analysis, however, did not see a very close relationship of the Taiwanese Aborigines to the Sabahan indigenous groups. Although linguists believed that the Austronesian languages spoken widely by ISEA inhabitants branched off from Taiwanese Aborigines, it could also happen in a reverse manner (Blust, 1999; Meacham, 1988). Thus, it is possible that a reverse migration could have started in the Borneo Island and expanded into Taiwan.

There were 3 components observed in the NJ tree of the Sabahan indigenous groups with the neighboring populations. The ISEA groups, except Java and Philippines, were clustered with Paiwan tribe, suggesting the role of the Taiwanese Aborigines in the human expansion in the ISEA. The Kadazan-Dusun, Bajau, and Rungus populations were situated closely, together with the Philippines and Ami. Again, this observation is in agreement with our analysis via PCA. The Mainland SEA and Taiwanese Aborigines (Bunun and Atayal) branched off after the East Asian populations. The result indicates that the dispersal from East Asia may occur in multiple directions, which includes both coastal and maritime routes into the SEA region.

In our study, the 6 *Alu* markers showed moderate degree of discriminating power in the indigenous populations. The average combined PD of these markers was 0.980833. Although *Alu* markers do not generate discriminating power that is as high as multi-allelic markers (such as STR), they are known for their ability to infer population origin owing to their identical-by-descent state (Shedlock, Takahashi, & Okada, 2004).

5.3 STRs within the Sabahan indigenous populations

Malaysia is a multi-racial country, consisting of people from different ethnicity, and hence multiple differences in their genetic backgrounds. There have been a number of genetic studies on the different Malaysian sub-populations, based on the STRs in autosomal and sex chromosomes, mainly on the 3 largest ethnic groups, i.e., Malay, Chinese, and Indian, as they represent the majority of the Malaysian society (Chang, Perumal, Keat, & Kuehn, 2007; Seah, et al., 2003). However, only a handful of studies extended to have covered the indigenous populations in East Malaysia and “Orang Asli” in the Peninsular Malaysia, who are the minority groups (Bekaert, Zainuddin, Hadi, & Goodwin, 2006; Kee, Lian, Lee, Lai, & Chua, 2011; Simonson, et al., 2011; Suadi, et al., 2007).

We present the STR data of Sabahan indigenous populations, to complement the existing data of other Malaysian sub-populations, and to create a broader representation of the genetic structure in Malaysia.

All 15 markers present with great diversity ($H_{Obs} > 0.65$) in the Kadazan-Dusun, Bajau, and Rungus populations, except the TPOX loci. This marker was detected with only 5 alleles, much lesser than other STR markers in the system, and has an average H_{Obs} of 0.588. In terms of forensic efficiency, all populations showed remarkably high level of discriminatory and exclusion powers within the STR system (Table 5.3). The highest efficiency was observed in the American cohort, as the markers were intentionally selected and produced for application in the American populations. Nevertheless, it can still be used efficiently in the local populations.

Table 5.3 : Comparison of the efficiency of the 15 STR markers in different populations; combined PD is expressed as the probability of 2 random unrelated individuals to present with the exactly identical genetic profile as generated by the STR system.

Population	Combined PD (1 in X individual)	Combined PE	Reference
African-American	1.41×10^{18}	0.999999 <u>6</u>	Promega Powerplex 16 kit manual
Caucasian-American	1.83×10^{17}	0.999999 <u>4</u>	
Hispanic-American	2.93×10^{17}	0.999999 <u>83</u>	
Asian-American	3.74×10^{17}	0.999999 <u>8</u>	
Kadazan-Dusun	6.08×10^{15}	0.999999 <u>21</u>	Present study
Bajau	9.97×10^{16}	0.999999 <u>86</u>	
Rungus	9.08×10^{15}	0.999999 <u>59</u>	
Chinese (Singapore)	1.03×10^{17}	0.999999 <u>95</u>	Yong, Aw, & Yap, 2004
Malay (Singapore)	1.42×10^{17}	0.999999 <u>58</u>	
Indian (Singapore)	3.66×10^{17}	0.999999 <u>96</u>	

The population differentiation study indicated that nearly all variations in the tested Sabahan indigenous groups arise within the populations, suggesting that these populations are in tight genetic association. These observations indicate that these populations may have originated from a recent common ancestral lineage and thus retain a large portion of genetic similarity. It could also be explained that the genetic similarity could be a result of constant gene flow among these populations as they inhabited in close proximity in the SEA region. The inbreeding index (G_{IS}) revealed a low inbreeding rate in the populations. The AMOVA results have shown a generally high similarity across the STR markers. Again, it reflects the high genetic affinity of the examined Sabahan indigenous populations.

5.3.1 STRs: Non-allelic variant

All STR multiplex systems cover a range of common alleles for each examined loci in the allelic ladder mix. Nevertheless, non-allelic variants could also present in a sample cohort. These variants are regarded as “off-ladder” fragments as their sizes do not fit for any of the known alleles when compared to the allelic ladder.

All 8 “off-ladder” fragments in the present study were identified as the 15-repeat allele of the FGA locus. The variant was identified and published in British Caucasian and Afro-Caribbean populations (Barber, McKeown, & Parkin, 1996). The observed variant in our study bear exactly identical nucleotide arrangement as in Barber’s study - [TTTC]₃ TTTT TTCT [CTTT]₇ CTCC [TTCC]₂.

The allele 15 was only present in the Kadazan-Dusun and Rungus populations, but not in the Bajau. This observation points that the Kadazan-Dusun and Rungus may have a closer genetic relationship, where the allele could have existed in their common ancestor. On the other hand, their common ancestor with the Bajau is more distant, most probably before the emergence of allele 15.

We did a frequency search for the allele on a database – Earth Human Short Tandem Repeat Allele Frequencies Database (EHSTRAFD). Over 451 sets of stored human population STR data, only 8 populations were reported to harbor the allele (Table 5.4).

These populations were residing in different parts of the globe, including South Asia, South America, Europe, and Africa. The highest frequency was found in the Lai people, a tribe in the Northeast India. In the remaining of the populations, the frequencies were consistently low (0.01 or lower).

Table 5.4 : Frequencies of FGA allele 15; comparison among the Sabahan indigenous groups and populations residing in different parts of the world, including Africa, SEA, South Asia, Europe, and South America.

Population	Frequency	Place of collection	Country	Reference
Lai	0.05400	Mizoram	India	Maity, Nunga, & Kashyap, 2003
Hmar	0.01200	Mizoram	India	
Lusei	0.01100	Mizoram	India	
Kadazan-Dusun	0.00900	Sabah	Malaysia	Present study
Tutsi	0.00880	Rwanda	Africa	Regueiro, et al., 2004
Rungus	0.00700	Sabah	Malaysia	Present study
Equatorial Guinean	0.00400	Equatorial Guinea	Africa	Alves, et al., 2005
Caucasian	0.00200	Austria	Austria	Steinlechner, Berger, Scheithauer, & Parson, 2001
Colombian	0.00050	Valle del Cauca	Colombia	Gomez, Reyes, Cardenas, & Garcia, 2003
Brazilian	0.00006	South-central Brazil	Brazil	Whittle, Romano, & Negreiros, 2004

With regard to the limited presence of the allele 15 in human populations, this allele could have present in low frequency in the ancestor groups. Along the colonization, it was carried and maintained in the migrant's gene pool. During the spread, the allele was eradicated in most groups, as indicated by its absence in most of the modern human groups. On the other hand, it could have gotten amplified, as in the Lai people of India, possibly by natural selection and founder effect. Otherwise, it may be maintained in a closed gene pool and get passed on from generation to generation.

Although higher frequencies of the variant were seen in Africans, South Asians, and Sabahan indigenous groups (Kadazan-Dusun and Rungus) compared to the Europe-America counterparts, no direct link between the distribution and population movement in these regions was seen. The Lai, Hmar, and Lusei people are inhabitants in the Northeast India. They are believed to descend from Tibeto-Burman speaking ancestors originating from East Asia (Cordaux, Weiss, Saha, & Stoneking, 2004; Su, et al., 2000). The presence of allele 15 in both Kadazan-Dusun and Rungus individuals in our study may indicate the influence of South Asian lineage in the shaping of genetic structure of the SEA populations.

It is also possible that the variant may not be absent in the human populations, but under-reported. This is because the allele 15 was not included as a standard allele in STR typing kits and, therefore, it was recognized as “off-ladder” signal and technically dropped out from the tested sample cohort. Moreover, due to its persistently low frequency, the allele may be under-detected, especially in the studies where the sample size is low or there is uneven sampling.

5.3.2 STRUCTURE of the Sabahan indigenous populations

The genetic compositions of the 3 Sabahan indigenous populations were subjected to test by the model-based clustering method implemented by the STRUCTURE program. At lower K numbers, distinct clusters were identified. At $K = 2$, the Kadazan-Dusun and Rungus formed an undifferentiated cluster from the Bajau group. This observation suggests that the Kadazan-Dusun and Rungus shared a greater extent of their genetic content, than the Bajau group. The Kadazan-Dusun and Rungus were separated at $K = 3$, forming 3 clusters. At higher K 's however, no sub-structure was detected in any of the indigenous groups, emphasizing the homogeneity of these populations. Despite the distinct clusters formed at lower K 's, it is noteworthy that the structure of all indigenous groups were constructed by varying proportions of same ancestry memberships. This dictates that there is substantial amount of admixture or genetic similarity in these populations, which in turn suggests a close ancestry between the 3 indigenous populations.

5.3.3 STRs: Clustering of world populations

Clustering analysis of the Sabahan indigenous populations with populations in other continents of the world would help in understanding their inter-relationships. It would give us an insight on the movement of modern humans in the past.

Every population was assigned into 5 distinctive clusters according to the geographic location, i.e., Africa, Europe, America, Oceania, and Asia, in the PcoA plot (Figure 4.15). Within the African cluster, 2 discrete groups were formed. The first group consisted of Tutsi people, South Africans, Somalians, and African-Americans. This group was rather isolated from other populations, signifying high degree of genetic dissimilarity. The second group was made up by Copts, Adayma Muslims, and Berbers. These populations were shown to have higher affinity to the European populations and were situated adjacent to the European cluster. This may due to the fact these populations resided in close proximity to the Europe continent, at the North of Africa. By having a geographical advantage, gene flow is easily facilitated between the African and European populations. Studies have shown that a large portion of the European gene pool (40 % to 66 %) originated from Near Eastern populations nearly 10,000 years ago (Belle, Landry, & Barbujani, 2006). The Europeans formed a tight cluster next to the African cluster. Australian Caucasians were seated within the cluster as it was well known that they are recent migrants from Europe. Similar clustering patterns were also observed in the NJ phylogenetic tree constructed via the STR markers for Africans and Europeans.

The American populations formed a large cluster of 3 groups. The first was made up by the Afro-ecuadorian, closer to the African cluster. The NJ tree has revealed better grouping of the Afro-ecuadorian, where they were included in the African clade. However, the Afro-ecuadorian group was among the earliest to separate out in the

cluster. This is because they have a mix genetic pool of both African and local inhabitants. The second group consisted of Mexican, Mestizo, Hispanic, Native American, and Amerindian. The Alaska Natives were isolated from other American populations, forming the third group. The vast differences among the American populations are in concordance to genetic findings reported previously that suggested multiple migration events taking place and resulting in the complex peopling structure in the region (Greenberg, Turner, & Zegura, 1986; Lell, et al., 2002; Santos, et al., 1999).

On the other hand, the Oceania populations (Australian Aborigines and Samoan) were located between the European and Asian clusters. Scientists justified that the genetic discrepancies between Oceania and Asian populations were evidence of multiple dispersals of modern humans in the regions. The Australian Aborigines, together with several Oceania populations, were shown to carry ancestral lineages of modern humans who reached as early as 75,000 BP., making them one of the oldest contemporary non-African populations (Rasmussen, et al., 2011). On the other hand, all Asian populations were descendants of migrants from a later wave that happened about 35,000 BP, where the remnants from earlier migration were replaced or assimilated (Macaulay, et al., 2005).

Within the Asian cluster, there were 3 main groups consisting of South Asians, East Asians, and SEA. The South Asian cluster was made up of Indian populations and their neighbor, i.e., the Tibetans from East Asia. Remarkably, the Japanese and Koreans were also included in this cluster, despite their geographic distance from the South Asian populations. The rest of the East Asian populations grouped nearer to the SEA cluster. Apart of the Han Chinese, the minority ethnic groups in China, i.e., Yi, Hui, and Miao people, were seen to have affinity to the SEA cluster, especially to the Vietnamese

and Melanau people from Sarawak. Therefore, they may be a link of genetic contribution or flow between these East Asian minority groups and the SEA people.

We have seen a closer relationship between the Kadazan-Dusun and Rungus populations than to the Bajau population on the PCoA plot. Although it may not be absolute, most SEA populations have shown to have a higher affinity to the East Asian than the South Asian, implying that the peopling patterns of SEA may receive or confer influence on the East Asian populations. The Y-chromosome study on various human populations had concluded that the SEA populations were more diverse than the East Asian populations, giving hints that the initial settlement could have occurred in the Mainland SEA before the migrants moved further north into East Asia (Jin & Su, 2000). In addition, the migration may have continued southward, meanwhile heading north, into the SEA region and populated the region. Although there was no strong affinity observed between the SEA populations and the South Asian cluster, researchers have postulated inter-migration between populations in these regions that facilitated gene exchange (Thangaraj, et al., 2005). The clustering of Asian populations was clearer as illustrated via the NJ tree method (Figure 4.17). Apart from several populations that branched out early in the Asian clade, there were 2 major sub-clades observed. The first sub-clade consisted of most populations from East Asia, South Asia, and Mainland SEA. The second sub-clade comprised of ISEA populations, included the Sabahan indigenous groups. The observation emphasizes that the peopling pattern of ISEA may not be as simple as direct gene flow from East Asia or/and South Asia, but it could involve complicated genetic ancestry and insular development that ultimately leads to the establishment of its complex gene pool.

5.3.4 STRs: Clustering of SEA neighboring populations

The second PCoA plot, constructed via frequency data of 12 STR markers, zoomed in and further elaborated the genetic association of populations within and surrounding SEA region. By including more populations in the regions and restricting populations from other continents of the world, the plot showed better and resolved illustration of various human groups (Figure 4.16).

Populations from South and East Asia were found within their respective clusters, situated next to each other. The East Asian groups included Tibetan, Han Chinese, Japanese, Korean, Yi, Hui, and Miao people. Derivatives of the Chinese population residing outside of China, such as Han Taiwanese, Hakka Taiwanese, and Malaysian Chinese, were not grouped under the East Asian cluster. It may be explained that these recently migrated populations could have different ancestry from the Han Chinese in this analysis, as it is well accepted that the former Han derivatives were originated from southern China (Fujian and GuangZhou) (Trejaut, et al., 2005). Otherwise, it may be caused by founder effect and admixture with the local populations that ultimately shaped a distinctive gene pool from the origin.

The populations on the right part of the PCoA plot segregated into 2 clusters. The first cluster consisted mainly of SEA groups and the second cluster was made up of Taiwanese Aborigines. From the view, it is obvious that the Taiwanese Aborigines harbor considerable genetic variations from populations in the SEA region. Our observation is in concordance with genetic survey conducted via paternal markers (Li, et al., 2008). Li and colleagues concluded that the Taiwanese Aborigines and SEA populations were both descendant of Daic speakers that derived independently. It was suggested that SEA populations migrated from Gulf of Tonkin and spread through the

coastal route from Vietnam. On the other hand, Taiwanese Aborigines evolved and migrated from southern China (Li, et al., 2008).

The Sabahan indigenous populations were found in both SEA (Bajau) and Taiwanese Aborigines (Kadazan-Dusun and Rungus) clusters. In view of this, it is possible that there are multiple separated origins in the Sabahan indigenous people, whereby they settled down in Sabah at different time points in the past. The Bajau people may descend from the same ancestry with other SEA populations, possibly originating from Mainland SEA. Moreover, the Bajau people were recent migrant to the Borneo Island from the Philippines in 1950s, indicating that they may have different lineages from the originally existing population on the island (Nimmo, 1968). Aside from the Bajau, 2 indigenous groups residing in the Sarawak state (Iban and Bidayuh) were also found occupying within the SEA cluster, which may suggest a common origin of these populations with the Bajau people in Sabah. In contrast, the Kadazan-Dusun and Rungus people may navigate from southern China as proposed by the “out of Taiwan” model. The Melanau, an indigenous tribe from Sarawak, was observed in the Taiwanese Aborigines cluster. Again, this stresses on the prospective multiple introduction of modern humans from different directions, into the SEA region. In order to further testify the validity of this observation, more genetic data from SEA populations, especially indigenous tribes from East Malaysia and “Orang Asli” from Peninsular Malaysia, should be included for future study to generate better remarks on the genetic structure of these populations.

Within the Taiwanese Aborigines cluster, the Kadazan-Dusun, Rungus, and Melanau were situated between the southern (Tao, Paiwan, Puyuma, and Rukai) and northern/eastern populations. The “intermediate” genetic identity of these groups infers the common ancestral lineage among the Taiwanese Aborigines and Borneo Island’s

indigenous people about 3,000 to 5,000 years ago before the split of each sub-population that established inhabitation at different geographic localities.

5.4 Examination of mtDNA

5.4.1 Mt intergenic 9-bp deletions

In order to compare our data to other populations around the world, 9-bp deletion frequency data was collected from previously published works (Table 5.5). The accumulated data of 47 populations was analyzed, including the 3 Sabahan indigenous groups studied in the present research.

Table 5.5 : Frequency review of the intergenic 9-bp deletion within the mt coding region at nps 8,271 to 8,279 in world populations.

Population	n	Frequency (%)	Reference
Africa			
Namibia (South)	361	0.8	Soodyall, Vigilant, Hill, Stoneking, & Jenkins, 1996
Algeria (North)	50	0.0	Ivanova, et al., 1999
Kenya (East)	60	1.6	Soodyall, et al., 1996
Gambia (West)	48	0.0	Soodyall, et al., 1996
Europe			
Italy	56	1.7	Torroni, et al., 1995
Portugal	96	2.1	Alves-Silva, Guimaraes, Rocha, Pena, & Prado, 1999
America			
Canada (North)	42	2.4	Merriwether, Hall, Vahlne, & Ferrell, 1996
US (North)	147	0.0	Merriwether, et al., 1996
Brazil (South)	245	8.6	Alves-Silva, et al., 1999
East Asia			
China, Han	813	15.1	Yao, Watkins, & Zhang, 2000
Mongolia	42	2.4	Merriwether, et al., 1996
Japan	62	18.0	Harihara, Hirai, Suutou, Shimizu, & Omoto, 1992
Korea	64	8.0	Harihara, et al., 1992
Taiwan Ab, Atayal	109	11.0	Trejaut, et al., 2005
Taiwan Ab, Saisiat	63	17.5	Trejaut, et al., 2005
Taiwan Ab, Tsou	60	40.0	Trejaut, et al., 2005
Taiwan Ab, Bunun	89	41.6	Trejaut, et al., 2005
Taiwan Ab, Paiwan	55	32.7	Trejaut, et al., 2005
Taiwan Ab, Rukai	50	24.0	Trejaut, et al., 2005
Taiwan Ab, Puyuma	52	21.2	Trejaut, et al., 2005
Taiwan Ab, Amis	98	49.0	Trejaut, et al., 2005
Taiwan Ab, Tao	64	50.0	Trejaut, et al., 2005

Table 5.5 : continuation.

Population	n	Frequency (%)	Reference
South Asia			
Iran	152	0.0	Alemohammad, Farhud, Hooshmand, & Sanati, 2003
Pakistan	76	0.0	Melton, et al., 1995
India	898	0.6	Clark, et al., 2000
Nepal	107	8.0	Passarino, Semino, Modiano, & Santachiara-Benerecetti, 1993
Southeast Asia			
Vietnam	50	20.0	Ivanova, et al., 1999
Thailand	195	25.1	Fucharoen, Fucharoen, & Horai, 2001
Philippines	176	40.0	Melton, et al., 1995
Nusa Tenggara	96	24.0	Redd, et al., 1995
Moluccas	50	16.0	Redd, et al., 1995
Java	98	25.5	Melton, et al., 1995
Malay	81	25.9	Melton, et al., 1995
Orang Asli, Semai senoi	30	36.7	Melton, et al., 1995
Nicobar	33	15.0	Prasad, et al., 2001
Proto-Malay	89	3.0	Lim, Ang, Mahani, Shahrom, & Md-Zain, 2010
Kadazan-Dusun	271	18.0	Present study
Bajau	159	29.0	Present study
Rungus	209	9.0	Present study
Oceania			
PNG, Coastal	55	40.0	Redd, et al., 1995
PNG, Highland	64	0.0	Redd, et al., 1995
Samoa	24	100.0	Redd, et al., 1995
Fiji	24	66.0	Hagelberg, et al., 1999
Australia, Aborigine	290	1.4	Betty, Chin-Atkins, Croft, Sraml, & Easteal, 1996
Polynesia	1178	94.0	Sykes, Leiboff, Low-Beer, Tetzner, & Richards, 1995
Hawaii	25	92.0	Lum, Rickards, Ching, & Cann, 1994
New Zealand	30	100.0	Hertzberg, et al., 1989

In general, the populations were selected based on the availability of data and their geographic localities. Populations that had distributed and originated from various regions were chosen to cover most major geographic provinces. All selected populations can be grouped under 7 clusters, i.e., Africa (4), Europe (2), America (3), East Asia (13), South Asia (4), SEA (13), and Oceania (8). Other than the frequency of each individual population, the averaged frequency of the deletion in all populations within every cluster was calculated. It is represented by the red line in Figure 5.3 as the “mean frequency”.

On average, the deletion was extremely low, near to absent, in the African populations. The mean frequency of the deletion was 0.6. Low frequencies were also seen in populations from Europe and America. While these populations harbor slightly higher numbers of the 9-bp deletion as compared to the African populations, the frequencies still did not exceed 3 %, except for Brazil (8.6 %). Similar observations were found in the South Asian populations, where the mean frequency was 2.2 %.

The deletion was found to increase in East Asia, > 10 %, except in the Mongolians and Koreans. The mean frequency was 25.4 % among the East Asian populations. The deletion was especially prevalent in the Taiwanese Aborigines, where it was seen in almost half the population (Tsou, Bunun, Amis, and Tao).

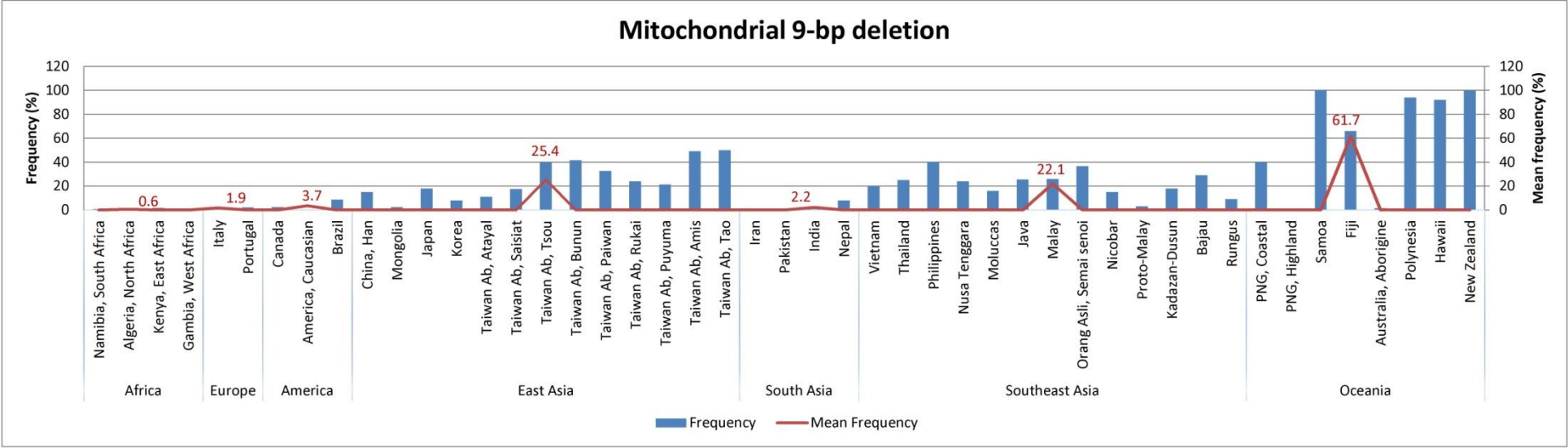


Figure 5.3 : Diagrammatic representation of mt 9-bp deletion frequencies as seen in various modern human populations around the world. Each population is categorized according to their geographic locality to 7 groups, i.e., Africa, Europe, America, East Asia, South Asia, SEA, and Oceania. The Sabahan indigenous populations examined in our study are included in the SEA cluster. Blue bars correspond to the frequency of the 9-bp deletion, while the red line represents the averaged frequency in the geographic groups.

In the context of SEA, the mean frequency of the deletion was 22.1 %, similar to its East Asian counterpart. All populations displayed frequencies of higher than 15 %, except the Proto-Malay. The highest frequency of deletion was seen in the Philippines, with 40 %. Among the Sabahan indigenous groups, the 9-bp deletion was not evenly distributed. Bajau people carried the highest frequency of the deletion at 29 %, followed by Kadazan-Dusun and Bajau (18 % and 9 %, respectively).

Among all populations from various regions, the 9-bp deletion was most prominent in the Oceania. The frequency of deletion in the population from this part of the world was extremely high, mean frequency of 61.7 %. In majority of the populations, the frequencies were > 60 %. The deletion was also found to be fixed in some populations, such as Samoa and New Zealand. However, the deletion was low in the Australian Aborigines, 1.4 %.

5.4.2 Mitochondrial haplogroups

Based on the “out of Africa” theory, all contemporary humans who live outside of Africa are descendants of a small group of leavers who exited Africa 70,000 to 80,000 years ago (Richards, et al., 2006). These ancient migrants belong to the L3 branch of the African haplogroup L, whose distribution is restricted to the region (Watson, et al., 1997). Along the journey off-Africa, new haplogroups were raised from the L3 stem, i.e., haplogroups M and N, which comprise of all non-African lineages today. The birth of both haplogroups M and N was estimated to occur some 60,000 to 70,000 years ago (Watson, et al., 1997). However, the region of origin for the emergence of these haplogroups remains elusive. This is due to their presence in and out Africa. It has been suggested that these haplogroups may have emerged soon after the L3’s divergence in Africa and migrated across the Red Sea, subsequently spreading all over the world. On the contrary, some scientists postulate that haplogroups M and N could have emerged in Asia and back-migrated to Africa (Cruciani, et al., 2002; Forster & Romano, 2007). Both these haplogroups are found spanning all continents with varying frequencies. About 66,000 years ago, another major branch of the mt haplogroup emerged from haplogroup N in South Asia, i.e., haplogroup R (Soares, et al., 2009). This haplogroup is another much extended macro-haplogroup that can be observed in many regions, such as Eurasia, Oceania, and America.

Alongside the spread of modern humans, the emergence of new haplogroups and sub-branches continued. The distribution of various mt haplogroups is summarized in Table 5.6.

Table 5.6 : Tree of global mt haplogroups; showing all major branches and their respective ages, origin of birth, and region(s) of distribution.

Mitochondrial haplogroups	Age (years ago)	Origin of birth	Distribution
Mitochondrial Eve			
└─ L0	112,200 – 188,000	East Africa	Sub-Saharan Africa
└─ L1-6			
└─ L1	107,600 – 174,500	Central Africa	Central Africa, West Africa
└─ L2	68,000 – 111,100	East Africa	All over Africa
└─ L4		East Africa	East Africa, Horn of Africa
└─ L5	114,200 – 126,200	East Africa	East Africa
└─ L6	81,500 – 129,800	East Africa	Yamen, Ethiopia
└─ L3	84,000 – 104,000	East Africa	East Africa
└─ M	60,000	Asia/Africa	All regions, most common in Asia
└─ CZ			
└─ C	60,000	Central Asia	Northeast Asia, Siberia
└─ Z		Central Asia	Korea, northern China, Central Asia
└─ D	40,000 – 60,000	East Asia	Northeast Asia, Siberia
└─ E	16,400 – 39,000	Indonesia/Taiwan	Southeast Asia, Malay Peninsula, Sabah, Papua New Guinea
└─ G	35,700	East Asia	East Asia, northeastern Siberia, Japan
└─ Q	50,000		New Guinea, Melanesia, Australian Aborigine
└─ N	71,000	Asia/Africa	All regions, most common in Europe & Oceania
└─ A	50,000	Asia	Northern & Central Asia, Siberia, North & Central America
└─ S			Australian Aborigine
└─ R	66,000	South Asia	Western Eurasia, Oceania, America
└─ B	50,000	East Asia	East Asia
└─ F	43,400	Asia	Eastern China, Japan
└─ R0			
└─ HV			
└─ H	25,000 – 30,000	Southwest Asia	Europe, North Africa, Middle East
└─ V	13,600	Western Mediterranean	Scandinavia, western & northern Africa
└─ Pre-JT			
└─ JT			
└─ J	45,000	West Asia	Near East, Europe
└─ T	45,000 – 50,000	Syria & Turkey	Eastern & northern Europe
└─ P			Papua, Melanesia, Australian Aborigine
└─ U	55,000	West Asia	Europe, India, Africa, Arab
└─ K	22,700 – 40,400	West Asia	Europe & Near East
└─ I	26,300	West Asia	Europe
└─ W	23,900	West Asia	Europe, West & South Asia
└─ X	30,000	West Asia	Wide spread
└─ Y	11,800 – 33,300		Japan & western Indonesia

New nucleotide changes occur due to many factors, such as genetic drift, natural selection, genetic bottleneck effect, founder effect, *in situ* evolution after long-standing isolation, and admixture with archaic humans. Some of the mt haplogroups represent deep rooting of ancestral composition, whereas some denote a recent establishment by their newly derived motifs. The geographic distribution of mt haplogroups varies among each other; some are widely spread across all major continents, while some are localized. By capturing the information of mt haplogroups in human populations, scientists can scrutinize both temporal and spatial distributions of modern humans during the pre-historic migration.

5.4.3 Haplogroups in the Sabahan indigenous populations

In terms of haplogroup diversity, the Kadazan-Dusun and Rungus groups were seen relatively more conserved than the Bajau. There were twice as many haplogroups in the Bajau (47) compared to the Kadazan-Dusun and Rungus, 23 and 21 respectively, thus suggesting a diverse genetic background of the Bajau tribe, which could have resulted from the nomadic lifestyle that encourages gene flow with neighboring populations. Nonetheless, it could also mean that the Bajau tribe came from an ancient ancestral population with rich genetic diversity. In contrast, the Kadazan-Dusun and Rungus tribes could have just separated from the ancestral populations more recently, hence they have not recovered the lost diversity due to the founding effects. Or perhaps it may be caused by prolonged isolation, culturally or geographically, of these populations that limit their interaction with other local tribes, which favors intra-marriage within their community.

The Sabahan indigenous groups consist of various mtDNA lineages. The Kadazan-Dusun and Rungus populations comprise of a higher composition of macro-haplogroup M, which make up more than 60 % of the tested subjects in these populations. The Bajau group, in contrast, bears more lineages descended from the macro-haplogroup N, in as much as 63 % of the individuals. Although the distribution of both macro-haplogroups M and N spans most regions of the modern world and does not correspond exclusively to any demographic event, their frequencies in a particular ethnic groups may reflect characteristics of population movement in the past. The frequencies of both these macro-haplogroups differ in ethnicities in different geographic regions. The highest frequency of haplogroup M was seen in Asia, while N was in Europe and Oceania. Based on the components of macro-haplogroups in the indigenous populations, it seems that the Kadazan-Dusun and Rungus may share a common

ancestry originating from Asia. On the other hand, the genetic structure of the Bajau people may have more influence from the Oceania counterparts than the Asia Mainland.

Broadly speaking, the Kadazan-Dusun, Bajau, and Rungus populations examined in the present study shared 4 mt haplogroup lineages, namely B, E, M7, and R9. These haplogroups and their descending branches covered 64 % to 81.3 % of the indigenous individuals tested, which suggest a strong common genetic background for these populations. It is possible that the old age DNA content passed on along the human lineage before they parted into their respective sub-populations. Apart from R9 haplogroup that has a fairly even distribution in all 3 indigenous groups, haplogroups B, E, and M7 were also predominantly in Bajau, Rungus, and Kadazan-Dusun, respectively. Hapogroup F was also observed to be particularly frequent in the Bajau population, 13.33 %, whereas its frequency was limited to 2 % in the other indigenous populations. While 16 % of the Rungus population was assigned into haplogroup D, only 6 % and 2.01 % were noted in the Kadazan-Dusun and Bajau, respectively. These haplogroups may shed light on how the population movement occurred in this region and elucidate the probable origin of these populations.

5.4.3.1 Haplogroup B4a

The haplogroup B is characterized by the presence of the 9-bp deletion in the mt region V and is most commonly seen in the ISEA populations. It consists of 2 main clades, i.e., B4 and B5. Majority of the B lineages in the ISEA falls in the B4 clade, as displayed by the 3 selected indigenous populations. B4 is postulated to have emerged from eastern Asia or Mainland SEA some 44,000 years ago and spread along the coastal Vietnam to Japan (Achilli, et al., 2008). The B4a lineages arose about 24,000 years ago on the mainland and can be presently found especially in highest frequency in the Polynesians, particularly in the remote and coastal Oceania region. It is, however, absent in the Australian Aborigines and New Guinea highlanders (Soares, et al., 2011). The lineage is also shared by most ISEA populations and Mainland East Asian, including Taiwanese Aborigines. On the other hand, the successor of the B4a, B4a1a, is restricted to the Taiwanese Aborigines, ISEA, and Polynesia, and it is not found in the Mainland East Asian. It is postulated that the B4a1a lineage emerged in Taiwan, or nearby regions that may have submerged under the sea in the present days, during the terminal Pleistocene dated 6,000 to 10,000 years ago (Trejaut, et al., 2005). The B4a1a makes up a considerable portion of the maternal lineages of the Sabahan indigenous people, from 3.33 % to 8 %. After the B4a1a made its entry into the ISEA, it continued to evolve and derived a new branch, characterized by the change at np 16,247, which is highly dominant in Polynesia. It is called the “Polynesian motif” and was believed to have emerged in eastern ISEA or near Oceania some 6,200 years ago (Pierson, et al., 2006). This motif is completely absent in the Taiwanese Aborigines and most ISEA populations. Interestingly, this motif was seen in 0.67 % of Bajau individuals, indicating a possible gene flow from the eastern ISEA or Polynesia. The other sister clades of the B4a, i.e., B4b, B4c, B4j, and B4h, were also found in the Kadazan-Dusun and Bajau populations, but none observed in the Rungus group. Hence, Kadazan-

Dusuns and Bajaus have a greater diversity, which may be the result of longer inhabitation and interaction with the neighboring populations. Bajau individuals harbored a good share of B4c1b lineage, at 7.33 %. The B4c1b was found commonly in the southern Taiwan, but present at low frequencies in China, Java, Vietnam, Philippines, Japan, and Malaysia (Lum & Cann, 2000; Melton, et al., 1995; Tajima, et al., 2004; Tanaka, et al., 2004). Haplogroup B4c2 was regarded as a potential marker for the post-glacial dispersal in the SEA region (Peng, et al., 2010). It is widely distributed in southern China and SEA, reaching its highest frequency in Cambodia, Thailand, and Vietnam (Peng, et al., 2010). The B4c2 was observed at 0.67 % and 2.67 % in the Kadazan-Dusun and Bajau, respectively, but absent in the Rungus population.

5.4.3.2 Haplogroup R9

The haplogroup R is an extended subgroup nested within the macro-haplogroup N lineages. It is among the earliest mtDNA lineages that emerged outside of Africa and gives rise to a number of significant haplogroups, such as B and F. The ancestral R root was suggested to originate in South Asia, the major route of dispersal of the earliest settlers into ancient Asia and Europe, about 52,600 to 64,600 years ago (Soares, et al., 2009). From there, the R lineage went on spread and diverged in East Asia and West Eurasia some 45,000 years ago (Soares, et al., 2009).

Among the mt R lineage, the R9 subgroups, R9b and R9c1a, present most abundantly in the Sabahan indigenous groups approximately 12.7 % to 17.3 %. The ancestral R9c reflects ancient lineages deep within the SEA populations. The R9c1a lineages persisted in East Asia since 30,000 to 37,000 years ago, and believed to have dispersed into SEA 9,000 years ago (Derenko, et al., 2012). Similar to the Kadazan-Dusun, Bajau, and Rungus, the R9c1a was observed at high frequencies (5 % to 11.1 %) in Philippines and Alor, but it was relatively uncommon in East Asian, Taiwanese, Mainland SEA, and ISEA (< 5 %) (Hill, et al., 2007; Tabbada, et al., 2010).

Another R9 subgroup, R9b, was only seen in the Bajau population (2.7 %). The R9b has very limited distribution in the SEA and this may indicate traces of an early settlement of its bearers in this region. It has been found in Chinese, Malays, Sumatrans, Javanese, and Thais (< 5 %), with appreciably high frequencies in Aboriginal Malays (Semelai: 28 %; Temuan: 21 %) (Hill, et al., 2007; Hill, et al., 2006). A phylogenetic study conducted on R9b lineages in various regions suggested the presence of an ancient ancestor, pre-R9b, in Indochina, with an estimated age of 29,000 years old (Hill, et al., 2006). The R9b lineages seen in Vietnam and South China were derived from this ancestral group about 19,000 years ago. The divergence of more recent R9b

lineages occurred later, about 9,000 years ago, and gave rise to R9b lineages in Thais, Aboriginal Malays, and Indonesians in a southward direction (Hill, et al., 2006). The R9b lineages seen in the Bajau population indicates the continuity or interaction of ancient Indochina mtDNA lineages into their genome.

R14 and R22 are nested within the R lineage and have been found in the Bajau. These old lineages descended directly from the basal root of the R haplogroup. The R14 was also observed in the Papua New Guinea (PNG) populations, one of the earliest founder groups after exodus from Africa (Hudjashov, et al., 2007). The sharing of the R14 subgroup by the Bajau and PNG populations depicts the possibility of gene exchange between the Bajau and Melanesians along the colonization of modern humans in these regions. Unlike the R14, the R22 has wider distribution covering various neighboring regions around SEA, as it has been previously reported in populations of Indonesia (Sumatra, Java, Bali, Sulawesi, Lombok, Sumba, and Alor), Mainland SEA, and Nicobar Islands (Hill, et al., 2007; Trivedi, et al., 2006). Haplogroup R22 represents an old mt lineage persisting from the earliest settlers and has a likely origin in the SEA, dated 28,300 years ago (Hill, et al., 2007). The R22 lineage is the most diverse in ISEA and is rooted in the Alor and Lombok populations (Hill, et al., 2007). Therefore, it could possibly serve as an indigenous marker for that particular region.

5.4.3.3 Haplogroup E

The haplogroup E emerged from its immediate predecessor, subgroup M9 of the macro-haplogroup M, about 30,000 years ago in the northeastern region of ancient Sundaland, or the modern Sulawesi and Sulu seas (Soares, et al., 2008). Haplogroup E is found prevailing in populations within ISEA and serves as a characteristic marker for the region. On the other hand, its sister clade, M9a, is reported in Mainland East Asia (Peng, et al., 2011). Among the Sabahan indigenous groups, the Rungus present with the highest proportion of E lineages, i.e., 29 %, followed by the Kadazan-Dusun and Bajau (14.7 % and 10.5 %, respectively). Haplogroup E has a shallower time depth as compared to the basal roots of other haplogroups. The haplogroup E can be divided into 2 sub-clades, i.e., E1 and E2, aged 17,000 and 9,500 years old, respectively (Soares, et al., 2008). Further stratification of the haplogroup E reveals 4 principal sub-clades, aged 4,300 to 11,700 years, namely E1a, E1b, E2a, and E2b. All 4 sub-clades were proposed to originate from within the ISEA and have very distinctive geographical distribution in different parts of the region (Soares, et al., 2008). Thus, they can be highly informative with regards to the demographics of these regions. The E1a and E2b can be seen in both ISEA and Taiwanese Aborigines, whereas the E1b and E2a are largely restricted to populations in the ISEA only.

The lineage diversity and coalescent time of E1b, E2a, and E2b decrease in Taiwanese Aborigines, as compared to ISEA populations (Sulawesi and Philippines), indicating that the likely origin of these lineages in the ISEA (Soares, et al., 2008). However, the E1 lineages, particularly the E1a1a, do not follow such a trend, where a higher diversity has been reported in the Taiwan than ISEA (Tabbada, et al., 2010). Like other populations in ISEA, the E1a1a represents a considerably large proportion of mt lineages in the Sabahan indigenous individuals, recording up to 18.7 % in the Rungus

tribe alone. Other branches of E1a lineage were not seen in the ISEA populations. Therefore, it is probable that the root of E1a1a evolved in Taiwan, instead of ISEA.

The E1b lineages, E1b and its direct descending variant - E1b + 16261, were found as high as 7.3 % in the Rungus individuals, and 4.7 % in the Kadazan-Dusun. However, it was less frequent in the Bajau population, < 2 %. Within the ISEA, these lineages were found mostly in Indonesia and Philippines. The distribution of E1b has also extended into New Guinea and the Malay Peninsula, suggesting that these sub-clades arose in ISEA and dispersed eastwards and westwards recently into the neighboring regions (Friedlaender, et al., 2007; Soares, et al., 2008).

Meanwhile, the E2 made up the least amount of mt E lineages in the Sabahan indigenous samples, only 2 % to 3.3 %. All of these mtDNAs were determined as the root type of E2, which reflects an ancient E2 lineage in these indigenous groups.

5.4.3.4 Haplogroup M7

The haplogroup M7 represent a significant fraction of mtDNAs in the Sabahan indigenous population. The M7 lineages are predominant especially in the Kadazan-Dusun population, with over one-third of the mt lineages belonging to this haplogroup. The 2 predominant subgroups seen in this study are the M7b1'2'4'5'6'7'8 (including both variants with and without a mutation at np 16,192) and M7c3c.

The M7b1'2'4'5'6'7'8 is a new branch in the latest global mt phylogeny builds. Based on the diagnostic mutation points in previous mt phylogeny, these mtDNAs are reported as a M7b1 subgroup (Hill, et al., 2007; Kayser, et al., 2003; Simonson, et al., 2011). This haplogroup is found to have a “spotty” distribution across the wide geographic range in SEA and East Asia (Trejaut, et al., 2005), despite its complete absence in the Oceania population. Low frequencies of the haplogroup, < 5 %, have been recorded in northern Han Chinese, Japanese, Koreans, Filipinos, and Indonesians. On the other hand, higher frequencies, of ~10 %, were seen in Taiwanese, southern Han Chinese, Thai, and Vietnamese people. Although the haplogroup is fairly common in the Iban and Kadazan-Dusuns (13 % and 23 %, respectively), it is uncommon (2 %) in the Bajau and Rungus populations (Simonson, et al., 2011). This haplogroup made up 5 % of the mt lineages in the Amis group, a Taiwanese Aboriginal tribe residing at the East coast of Taiwan, but is obsolete in most of the other aboriginal tribes (Trejaut, et al., 2005). The M7b1 lineages in Amis differs from those in the East Asians by a substitution at the np 16,126, which signifies a clear division of M7b1 lineages between the Chinese Han and Taiwanese Aborigine. It is thus unlikely that the lineages were picked up recently through admixture of the aborigines and Han Chinese in Taiwan. All 41 mtDNAs belonging to the lineage in the Sabahan indigenous populations posed the mutation point at np 16,126, as for the Amis lineage. This observation suggests a possible genetic link of the Taiwanese Aborigines, particularly the Amis, to the Sabahan

indigenous groups through the sharing of the M7b1-16,126 variant. The ancestral type of M7b1, without the mutation, could have emerged in the Mainland East Asia and was brought into Taiwan by settlers in the southern coastal region. The substitution of nucleotide at np 16,126 occurred in these Taiwanese settlers before they dispersed to various regions in the ISEA. The varying frequencies of these lineages in the ISEA populations may be the result of series of founder effects during the spread. Although the age estimated for the M7b1 lineages in the Amis tribe was 20,200 years old, the standard error for the estimation was as high as 13,400 years (Tabbada, et al., 2010). This could be due to the small number of samples involved in the statistical estimation. Therefore, it is possible that the mutation at np 16,126 happened in a more recent timeframe (< 10,000 years ago). The M7b1-16,126 lineage could be useful to explain the demic event during the Neolithic expansion from Taiwan.

The M7c3c, another prominent M7-subgroup, presents evenly among the indigenous populations of Sabah, with frequencies ranging from 10 % to 15 %. The M7c3c was previously reported as “M7c1c” in mt studies that included haplogroup assignment based on earlier versions of phylotree builds (Hill, et al., 2007; Trejaut, et al., 2005). The ancestral type of the M7c3c, but not the M7c3c itself, was found commonly in China. It can be found predominantly in Taiwanese Aborigines and ISEA populations. The lineage is most diverse in the Taiwanese Aborigines and Borneo Island. In addition, it may present in these regions since early Holocene and spread to the neighboring areas in the mid of Holocene (Hill, et al., 2007). The frequency of M7c3c concentrates in populations in Philippines, Borneo, Sulawesi, and Taiwanese Aborigines (Tabbada, et al., 2010; Trejaut, et al., 2005). An age estimation of the lineage shows that it has an older age in the Taiwanese Aborigines (14,000 years) compared to Filipinos (11,400 years) and Sulawesians (4,400 years). Therefore, the distribution of M7c3c lineage is seen to support the path of prehistoric movement of modern human as proposed by the

“out of Taiwan” model. The M7c3c is absent in the Mainland SEA and Oceania populations, indicating that its distribution was limited by ocean boundaries during the post-glacial period.

The M7b3 is another mt lineage that may illustrate the mid Holocene demic movement into ISEA from Taiwan. The lineage is reported with highest frequencies in the Taiwanese Aborigines, especially in populations residing in the northern region (Atayal and Saisiat). It follows a decreasing pattern along the path suggested by the “out of Taiwan” model. Different variants of the M7b3 lineage have been observed in the Taiwanese Aborigines, whereas there is only a single predominant variant seen in populations in the ISEA (Tabbada, et al., 2010). Therefore, it is most likely that the lineage emerged in Taiwan, about 10,300 years ago, and one of the local variants was carried by the migrating settlers into ISEA. In the present study, the M7b3 lineage was only found in 1.3 % of Bajau individuals and it was completely absent in other Sabahan indigenous groups. It was also reported in 1.2 % of the Iban people (Simonson, et al., 2011). The low frequencies of the lineage in the ISEA populations could be explained by the “washing off” effect, whereby the lineage is “diluted” and gradually decreased during the migration in small groups. Hence, while the M7b3 lineage may show the genetic association between Taiwanese Aborigines and some populations in the ISEA region, it may not be used as an exclusive “out of Taiwan” marker as it could be under-detected due to its low frequency in some populations.

Apart from the predominant M7 lineages (M7c3c and M7b1'2'4'5'6'7'8), other branches of the M-macrohaplotype were observed in the Sabahan indigenous individuals. Despite the low frequency, these M-subgroups reveal the existence of ancient mtDNA lineages. The Bajau population comprises of a larger spectrum of M-subgroups, where many of these subgroups branched directly from the root of the M haplogroup. A majority of these M-subgroups were found exclusively in the Bajau of

our study, reflecting a strong and ancient backbone of M haplogroups in the Bajau population (M20, M21, M43, M44, M51, and M74). M20 and M74 are rare in SEA populations (Bodner, et al., 2011; Jinam, et al., 2012). In the present study, 2 Bajau individuals were identified to carry the M74 lineages, one each for the root type M74 and variant M74b1. The M74 was thought to have emerged in East Asia and derivatives of the root type were found in various regions of East Asia, including South China (Kong, et al., 2011). The estimated age of the M74 in ISEA corresponds to the early colonization of the region by modern humans (34,100 to 39,300 years ago) (Jinam, et al., 2012). The M74b1 subtype, which may have evolved *in situ*, was also seen in the Philippines and Sumatra (Gunnarsdottir, Li, et al., 2011; Gunnarsdottir, Nandineni, et al., 2011). The presence of M74 lineages in the ISEA illustrates the dispersal of early modern humans from South China during the end of Pleistocene epoch. On the other hand, parts of the M lineages in the Bajau reflect the maternal component of populations in ISEA receiving influence from Mainland SEA and South Asia, other than East Asia. The M43b and M44 have been found in northeastern and western India, respectively, and is estimated to have emerged in South Asia at least 26,000 years ago (Chandrasekar, et al., 2009). The SEA-exclusive M51, aged 22,100 years, has been reported in Mainland SEA populations (Vietnamese Austronesian and Cambodian) (Hartmann, et al., 2009; Peng, et al., 2010). M21 is an ancient “relict” lineage, with a time depth of 57,000 years, and is localized to SEA and South China (Macaulay, et al., 2005). Its 2 subtypes, M21c and M21d, have been sampled in Indochina and ISEA regions, including Vietnam, Laos, Thailand, Peninsular Malaysia (Aboriginal Malays), and Philippines (Bodner, et al., 2011; Dancause, Chan, Arunotai, & Lum, 2009; Peng, et al., 2010).

Similar to the Bajau population, the Kadazan-Dusun and Rungus individuals were seen to receive influence from the surrounding regions. Parts of their genetic composition

were shown to associate with the Indian continent as illustrated by the M5a1, M31a2, and M33a1b lineages. These mt lineages originated from the root type and are highly restricted to India (Chandrasekar, et al., 2009; Dubut, Murail, Pech, Thionville, & Cartault, 2009; Endicott, et al., 2006; Malyarchuk, et al., 2008). Other than that, an East Asia-specific lineage was also observed in the Rungus population, M7b4 (Kong, et al., 2003).

It is obvious that movement between continents is common before the last glacial maximum (LGM) because these regions were highly accessible by walking. In view of the M lineages, the Bajau people consists of an extensive spectrum of maternal genetic components, suggesting that they could have persisted in the ISEA region for a substantially long period, where *in situ* mutations have developed. They also interacted consistently with populations in the neighboring regions, such as East Asia, Mainland SEA, Oceania, and South Asia, which is shown by the presence of region-specific lineages. Kadazan-Dusun and Rungus have less diverse haplogroup M lineages than the Bajau people, it may because that these populations were brought in to the Borneo Island by different waves of migratory event, whereby the Bajaus were introduced first followed by the Kadazan-Dusun and Rungus populations in more recent times.

5.4.3.5 Haplogroup F

The haplogroup F is one of the direct descending branches of haplogroup R9c, sister clade of R9c1. The haplogroup F contains a small subset of lineages, namely F1, F2, F3, and F4. The Bajau population was made up by an appreciable proportion of F lineages (13.3 %), whereas it was found less commonly in the Kadazan-Dusun and Rungus groups (~2 %). In the Bajau population, the F1a and F3b1 were seen with higher frequencies, 4.7 % and 6 %, respectively.

Among the subgroups of the haplogroup F, F1a is most frequently observed among individuals in the ISEA region, dated 33,900 years ago (Hill, et al., 2007). F1a, together with its sister clades F1b and F1c, can be found with substantial frequencies in East Asia. Unlike its sister clades, the F1a lineages also present commonly throughout the SEA, suggesting that South China could be the place of origin for F1a and its derivatives in SEA (Hill, et al., 2007). F1a1a, the subgroup of F1a, dates back 7,300 years, is found most commonly in the western and southern regions of ISEA. It presents at considerably high frequencies in the Indochina (Thailand and Laos) and Malay Peninsula (Senoï people) (Bodner, et al., 2011; Hill, et al., 2006). The statistical test of various sequences of the F1a1a subtypes in these regions revealed that the most recent common ancestor of the lineages appears in the northern region of Indochina about 7,000 to 10,000 years ago (Hill, et al., 2006). The F1a1a lineage in the ISEA region may represent the dispersal wave from Mainland SEA in the early Holocene. The lack of F1a1a lineage in the Sabahan indigenous populations (only one individual in the Bajau group) shows that the effect of F1a1a dispersal is minute to these populations. In addition, the higher frequency of F1a type than the derived F1a1a in the Bajau people, indicates that the population had arrived, most probably from South China, before the F1a1a dispersal from the northern Indochina. On the other hand, the F1a3 and F1a4 were proposed as candidate markers of “out of Taiwan” lineages (Hill, et

al., 2007; Tabbada, et al., 2010). However, these lineages were not commonly observed in our Sabahan indigenous groups.

The F3b haplogroup is another common lineage in the mt gene pool of ISEA populations. It can be found mainly in the Philippines and Borneo Island. The F3a (sister clade of F3b) on the other hand, is found to have restricted distribution to the mainland region. The F3b is observed to present at low frequencies in the South Chinese populations, whereas high frequencies are found in Taiwanese Aboriginal tribes in the southern region (Kong, et al., 2003; Trejaut, et al., 2005). F3b lineages have also been reported in the Malagasy populations in Madagascar, where the founding haplotype is observed in all Malagasy populations (Razafindrazaka, et al., 2010). The F3b lineage in Melanesia has been determined to originate from ISEA (Dubut, Cartault, Payet, Thionville, & Murail, 2009). The F3b lineage may be useful as the marker to illustrate the late dispersal wave from Taiwan to Oceania, through the Philippines and East Indonesia before the sea levels rose. The migration could occur rapidly along the path, without much interaction with the surrounding populations. The increased number of F3b lineages in the Bajaus suggests its origin to be in the eastern region of ISEA.

5.4.3.6 Haplogroup D

The haplogroup D (sister clade of haplogroup M80), is the most frequent mt lineage found in Eastern Eurasia. It has a pre-LGM time depth (~35,000 years) and was believed to have an East Asian origin (Derenko, et al., 2010; Metspalu, et al., 2011). The haplogroup D4 is most prevalent in East Asia, concentrating in central Asia and southern Siberia. In China, the highest frequency was reported in the northern region and it decreases in the southern China. Representation of the lineage is even lower in regions further west and south of China, including the Indochina and ISEA (Metspalu, et al., 2011). The presence of D4 lineages (D4b1 and D4s1) in the Kadazan-Dusun and Rungus populations represents dispersal from East Asia. The D4b1 is found peaking in the Tibet (18 %), followed by populations in Siberia (Schurr, Sukernik, Starikovskaya, & Wallace, 1999; Tanaka, et al., 2004). Lower frequencies have been detected in the China and Japan, while none reported thus far in Thailand, Indonesia, or Borneo Island (Tanaka, et al., 2004; Yao, et al., 2000).

The D5b subgroups make up a good amount of mt lineages in the Kadazan-Dusun and Rungus populations, 4.7 % and 15.3 %, respectively. In contrast, only one Bajau individual was found to descend from the D5b lineage. The haplogroup D5 is suggested to play a role in the mid-Holocene dispersal out of Taiwan. The root type of D5 is most common in China and Taiwan (Hill, et al., 2007). The 4,000 years old D5b1c, previously determined as D5d1, is seen in many individuals in ISEA (Hill, et al., 2007). Although the D5b root type was not found in the Taiwanese Aborigines, 3 derived groups were reported in Taiwan. Therefore, it is very likely that the D5b1c evolved in Taiwan. In the present study, its root type is only seen in the Bajau group, without any derived subgroup. Hence, it may be explained that the root type existed in the group before the mid-Holocene dispersal. On the other hand, the root type is absent in the Kadazan-Dusun and Rungus, but the derived D5b1c present prominently among these

groups, indicating a recent contribution of the Taiwanese lineages to the gene pool of populations in Borneo Island.

Among the Sabahan indigenous groups examined in our study, the D6 branch is only observed in 2 Bajau individuals. Within SEA, the haplogroup D6 is also observed in several populations residing in Philippines (Luzon, Visayas, and Mindanao) (Tabbada, et al., 2010). Thus, the observation indicates a possible association of the Bajau people with the Filipino population.

5.4.3.7 Haplogroups N and Y

Although the descending branches make up a large proportion of mt lineages of indigenous populations in our study, the haplogroup N was not common. It only contributed less than 5 % of all tested mtDNAs. The N5 was seen in all 3 indigenous populations, with frequencies of nearly 2 %. The N5 lineage, together with the N8 subgroup found in the Bajau group, represents ancient maternal components in these populations. The N5 is distributed almost exclusively in the contemporary South Asian populations, although with low frequencies. It is proposed to have arisen in the region as early as the root of haplogroup R, about 65,000 years ago (Palanichamy, et al., 2004). The consistent existence of the N5 in the Sabahan indigenous populations reflects the persistence of such ancient lineage in the modern human, without having been removed by drifts. On the other hand, it could also depict genetic links between the Sabahan indigenous groups and South Asian established by gene flow.

The N21 subgroup was regarded as insular haplogroup within SEA, where it is almost reported exclusively for populations in the region. The N21 haplogroup displays evidence of early settlement of the SEA. It can be found in both mainland and island regions of the SEA. However, its diversity is higher in the ISEA (Indonesia) than the Mainland SEA, suggesting an insular origin (Hill, et al., 2007). The mainland lineage was believed to arrive from Sumatra about 43,000 years ago (Hill, et al., 2007). The variant (N21 + 195) observed in the Rungus population was also found in the Chamic speakers in Vietnam and Aboriginal Malay in the Malay Peninsula (Hill, et al., 2006). The N21 lineage in the Aboriginal Malay is suggested to have derived from the ancestral type in the Chamic speaker, with an estimated age of 21,000 years old (Jinam, et al., 2012). Thus, it is probable that the N21 lineage in the Rungus people may have descended from the Malay Peninsula following the spread from Indochina.

The haplogroup Y is a relatively smaller group in the mt phylotree, with only 2 principal subgroups, i.e., Y1 and Y2. The Y2 is proposed to participate in the mid-Holocene dispersal out of Taiwan southward and westward into Borneo and Sumatra from Philippines (Hill, et al., 2007). The Y2 is found distributed widely within ISEA, including Malaysia, Philippines, Indonesia, and Sumatra (Hill, et al., 2007; Hill, et al., 2006; Tabbada, et al., 2010). In the Borneo Island, the Y2 was seen in Bajaus (2.7 %) and Ibans (12.4 %) (Simonson, et al., 2011). The overall age of Y2 lineage in the ISEA is 3,400 years old (Hill, et al., 2007). Apart from the ISEA, the Y2 is also observed in northern Asia, Korea, Japan, and Taiwan, with an older age estimated at 13,000 years (Derenko, et al., 2007; Tanaka, et al., 2004). The spread of Y2 lineages may have started from North Asia to Taiwan, and further into ISEA through Philippines as suggested by the “out of Taiwan” model. However, the Y2 lineage was not observed in any of our Kadazan-Dusun and Rungus samples.

5.4.4 Mitochondrial haplogroups: PCA

The PCA plot was constructed based on the frequencies of mt haplogroup of populations in East Asia and SEA. These populations were selected according to the proposed migratory path as in the “out of Taiwan” model, where the movement was initiated in the southern China, and subsequently spread throughout the SEA through Taiwan and Philippines. The Chinese cohort was represented by mainland populations from 4 major regions of China, i.e., Northeast (NE), Southeast (SE), Northwest (NW), and Southwest (SW). The Taiwanese Aborigines were suggested to have a close genetic relationship to the populations in SEA region. Genetic data of 2 Taiwanese Aboriginal populations was obtained from previous publications for comparison, and named Taiwanese Aborigine¹ (Hill, et al., 2007) and Taiwanese Aborigine² (Peng, et al., 2010). These cohorts consisted of a mixed number of individuals from different Taiwanese Aboriginal tribes. The Philippines Island is believed to bridge the dispersal from Taiwan into SEA. Therefore, it is crucial to include the Philippines and populations in the ISEA (Indonesia, Sabahan indigenous, Borneo, and Iban). Populations from the Mainland SEA (Vietnam, Thailand, Cambodia, Kinh, Cham, Melayu Malay, and Orang Asli) were also included for illustration of genetic association among populations in the SEA region.

The first 2 PCs made up 28.45 % of the variations observed among these populations. Based on the mt haplogroup frequencies, all populations were found to fall within 2 main clusters. The contemporary Chinese populations were clearly separated from others, showing strong genetic homogeneity among themselves and discrepancy from other populations.

Several SEA populations from the mainland were also found grouping adjacent to the East Asian cluster, i.e., Kinh Vietnam, Cham, and Thailand. The Cham people showed

high affinity to the Thai people, than the major ethnic in their Vietnam homeland - the Vietnamese/Kinh. The Cham people are remnants of the annihilated Kingdom of Champa that lasted more than one millennium (192 - 1832 AD) (Peng, et al., 2010). The Cham and Moken (a nomad minority group) are the only groups of Austronesian speakers in the mainland of SEA. It has been suggested that the Austronesian language was brought along by ancestors of the Chamic people from the ISEA during their migration about 2,500 years ago, possibly initiated in the Southwest of Borneo Island (Bellwood, 2007; Thurgood, 1999). However, as seen in the PCA, the Sabahan indigenous was situated remotely from the Cham people, indicating a distant genetic link of them. Despite the origin in ISEA, the Chamic genetic lineages could be drifted from other ISEA populations due to long term assimilation with the mainland local indigenous populations, especially with the Mon-Khmer speakers in Thailand and South Vietnam (He, et al., 2012; Peng, et al., 2010).

The Indonesian archipelago consists of more than 17,000 islands and hosts widely diverse populations. Despite that, the eastern Indonesia is known to harbor the most diverse communities than the western counterpart (Tumonggor, et al., 2013). In our PCA plot, the western Indonesians formed a cluster and distant from the Taiwanese Aborigines. On the other hand, the eastern populations were positioned closer to the Taiwanese Aborigines, indicating a closer genetic relationship of these groups. Studies have revealed large Neolithic movements into eastern Indonesia, possibly from Taiwan that involved the spread of Austronesian languages to the entire archipelago (Tumonggor, et al., 2013; Xu, et al., 2012). It also explained the observation of lesser basal mt haplogroups in the eastern Indonesia, as compared to the western parts.

On the other hand, 2 populations from the Mainland SEA (Cambodian and Orang Asli) were found located further right to the western Indonesian cluster, which could signify the influence of lineages from the ISEA in these populations. The Melayu Malay and

Iban (Sarawak indigenous) clustered within the Western Indonesians, denoting a possible origin in the ISEA.

The Sabahan indigenous populations in the present study did not show very high affinity to the Taiwanese Aborigines, neither did the Philippines. They were grouped within the Eastern Indonesian cluster. It showed a close genetic association among the Sabahan indigenous populations and eastern Indonesian with the Philippines, which could be the results of gene flow and assimilation of the Neolithic migrants. The matrilineal discrepancy of these populations from the Taiwanese Aborigines could be explained by founder effects and genetic drift of migrating populations. However, it could also due to the mixed number of tribes in the represented Taiwanese Aboriginal populations, where it may dismiss or mask the distinctive genetic characteristics in a particular aboriginal tribe.

5.4.5 Phylogenetic analysis of mtDNAs

Based on sequences of the mt HV regions, the Sabahan indigenous populations were segregated into various clusters in the phylogenetic tree (Figure 4.25). They were shown to share their mt lineages, in varied extent, with different populations in SEA and the neighboring regions. Higher genetic similarity was observed in ISEA populations, followed by East Asian and Mainland SEA. The Sabahan indigenous groups were comparatively least similar with the populations in the Oceania.

From the phylogenetic tree, there could be multiple episodes of human movements into the ISEA region, which contributed to different fractions of mt lineages in the Sabahan indigenous populations. These demic events could have occurred as early as 50,000 years ago to recent times of < 5,000 years ago, as evidence from the presence of both ancient and young mt lineages in these indigenous populations. It is also likely that these movements happened at varying scales and directions, resulting in discrepancies in the genetic materials of the studied populations.

The Bajau population exhibited a greater level of diversity compared to their Kadazan-Dusun and Rungus counterparts. It is especially evident from the wide distribution of the Bajau individuals in the clades of the constructed phylogenetic tree.

The oldest mt lineages present in the Bajau population was represented by haplogroups M9, G3, D6, and M74, which were found to cluster closely with ISEA populations (i.e., Filipino and Indonesian). These lineages could be brought into the ISEA region as early as 50,000 years ago, most probably by the first wave of migrating humans from Africa, and are regarded as the genetic relics of some populations in the ISEA region.

Apart from genetic remnants from the earliest migrants into the region, the Bajau population was also found to harbor several mt lineages that have high affinity to the East Asian origin (R14 and M51). The haplogroups R14 and R22 bearers from the Bajau group formed a distinctive clade with Thai and Cham people from the Vietnam [Figure 5.4(a)]. The R14 and M51 were believed to emerge in East Asia about 25,000 years ago, whereas R22 was shown to have a likely origin in ISEA (Hill, et al., 2007). Younger East Asian lineages, with approximate ages of 10,000 to 15,000 years old, were also observed in the Bajau populations, i.e., B4c2, M20, M21c, M21d, M43b, and R9b. These lineages may have dispersed into SEA through the ancient Indochina by multiple demic events. Haplogroups M21c and M43b were also reported in South India, suggesting a continuous migration from the Indochina or SEA regions. On the other hand, gene flow from South India was also seen in the Bajau individuals, where the N8 lineage bearers formed a unique clade in the phylogenetic tree [Figure 5.4(b)]. Thus, it is suggested that there were two-way movements between the SEA and South India.

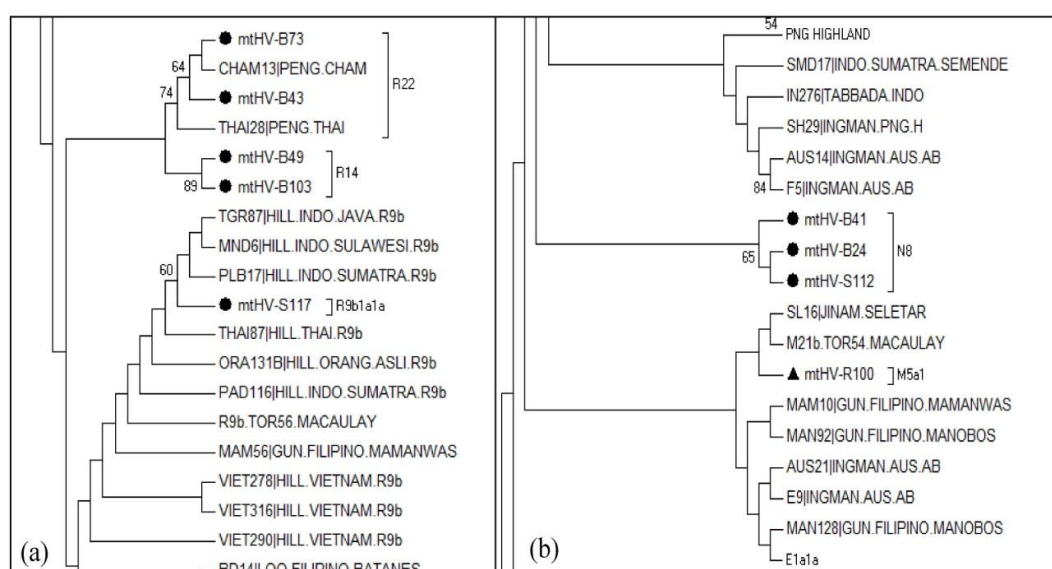


Figure 5.4 : Clustering of (a) haplogroups R14 and R22 (b) N8 in the NJ tree; the number at the node denotes the bootstrap value for the branches (only values > 50 % were shown); the dark circles denote individuals from the Bajau population.

Several mt lineages in the Bajau individuals were also present predominantly in the Taiwanese Aborigines, Indonesians, Filipino, and Oceania. These lineages included B4a2b and Y2, aged between 5,000 to 10,000 years old, which may represent the “out of Taiwan” traits in the Bajau population (Figure 5.5). The Polynesian motif, B4a1a1a, clustered among the Oceania and ISEA populations.

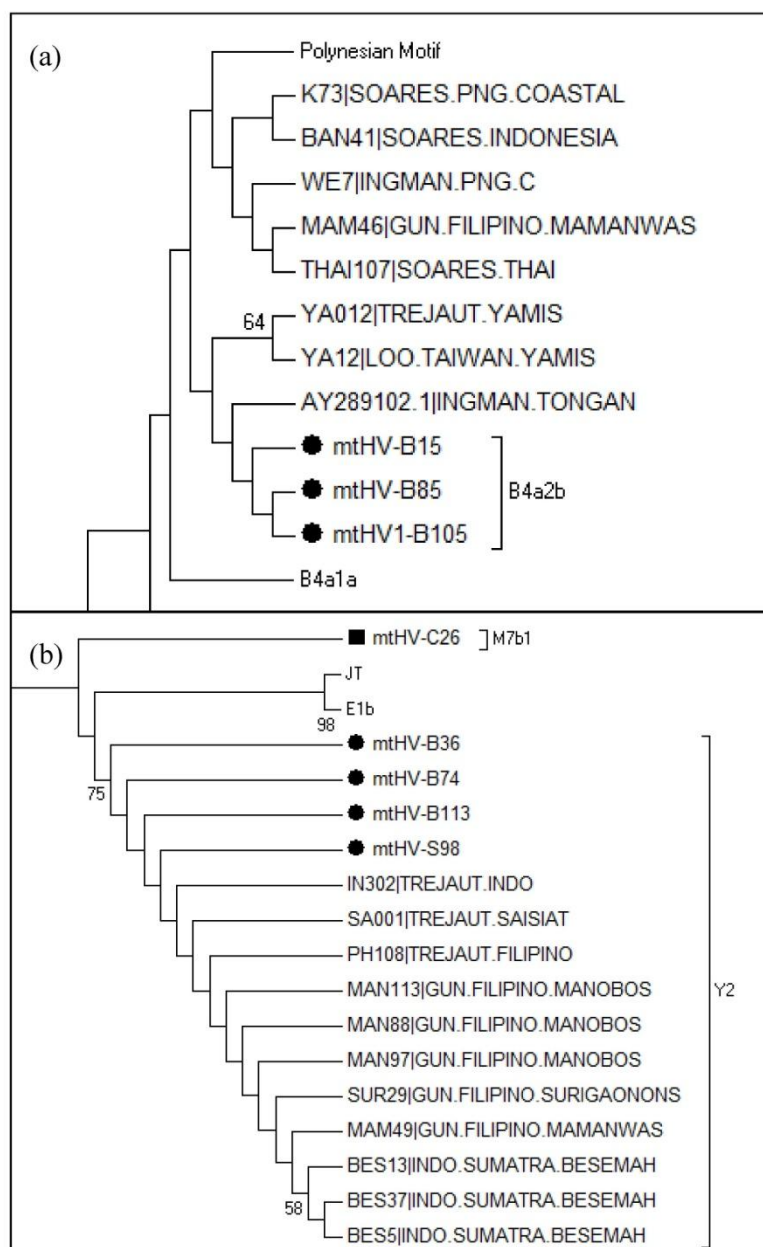


Figure 5.5 : Segregation of Bajau individuals in the clusters of haplogroups (a) B4a2b and B4a1a1a (Polynesian motif) (b) Y2; the number at the node denotes the bootstrap value for the branches (only values > 50 % were shown); the dark circles denote individuals from the Bajau population.

Bajau individuals with haplogroup B5b grouped closely as a cluster with the Filipino (Figure 5.6). The B5b has been reported with sporadic frequencies and is believed to originate in East Asia, with an estimated age of 35,000 year old (Wen, et al., 2005). It was also found in populations in various regions of Indonesia (Mona, et al., 2009). Its variant (B5b1) however, is suggested to have emerged from eastern Indonesia about 11,000 years ago and spread northward to Philippines about 4,700 years ago (Tabbada, et al., 2010). This variant is absent in Taiwanese Aborigines and Bajau population (Loo, et al., 2011). Thus, it is clear that the Bajau received influence from the earlier dispersal of ancestral B5b from East Asia, but not the variant derived in the eastern Indonesia.

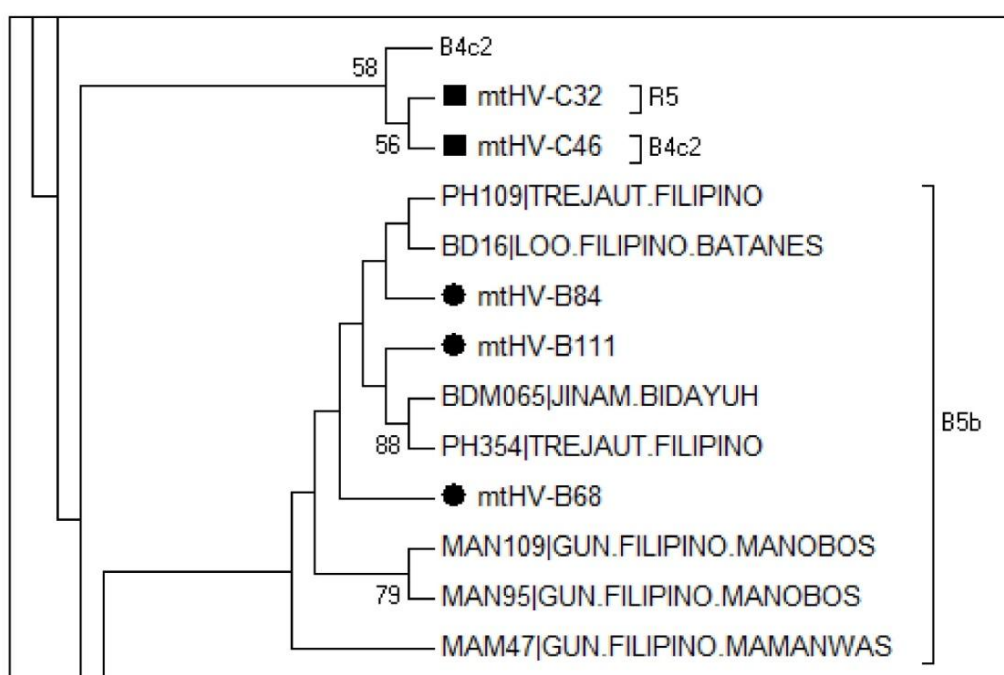


Figure 5.6 : B5b individuals of the Bajau group; the number at the node denotes the bootstrap value for the branches (only values > 50 % were shown); the dark circles denote individuals from the Bajau population.

Although haplogroup B4b1a + 207 was shared by Bajau and Kadazan-Dusun populations, the lineage bearers from these populations were grouped into 2 separated sub-clades (Figure 5.7). The Bajau clade did not exhibit affinity to any other populations in the ISEA and East Asia, whereas the Kadazan-Dusun clade consisted of lineages from the Philippines, with a mix of the ancestral B4b1 haplogroup. The B4b1 lineage was estimated to spread to Philippines about 8,800 years ago. Therefore, the B4b1a variant may have emerged in that region and subsequently dispersed to other ISEA region. The presence of only B4b1a variant in the Bajau population may be the result of founding effect when the ancestors of Bajau people migrated from southern Philippines.

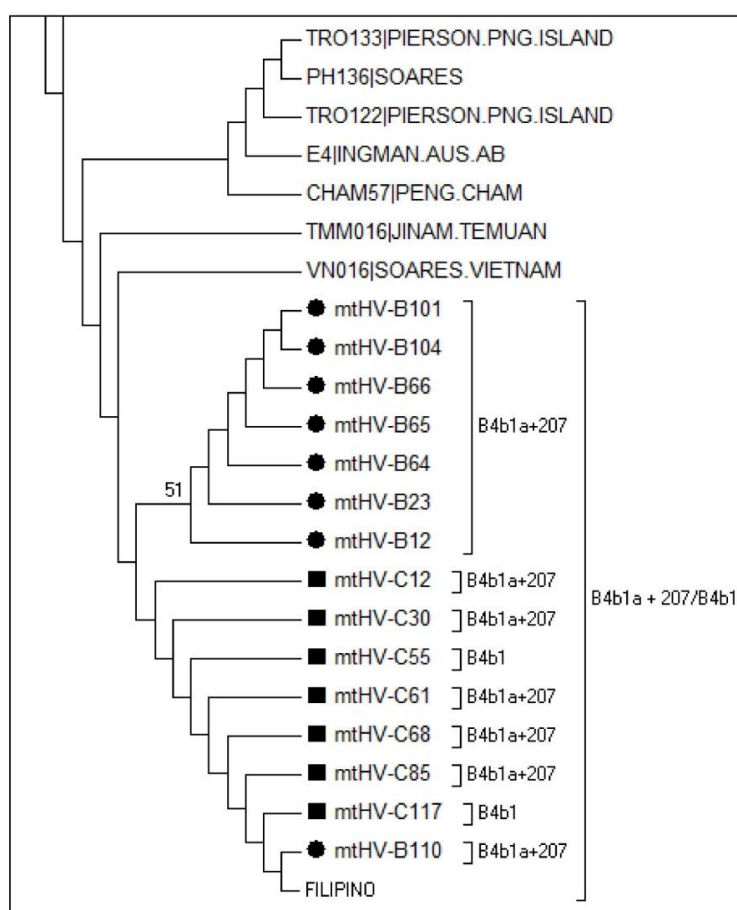


Figure 5.7 : The division of Kadazan-Dusun and Bajau individuals in the B4b1 cluster; the number at the node denotes the bootstrap value for the branches (only values > 50 % were shown); dark rectangles and dark circles denote individuals from Kadazan-Dusun and Bajau populations, respectively.

Similar observations were also seen in the F3b1 lineage that was shared by the Bajau and Rungus populations (Figure 5.8). The F3b1 and its predecessor (F3b) were found in the Taiwanese Aborigines and ISEA populations, indicating that the F3b1 may have emerged in Taiwan. The F3b lineage in the Bajau population formed a separated cluster from the Rungus, Indonesian, and Taiwanese Aborigines, which may also be a result of the founding effect.

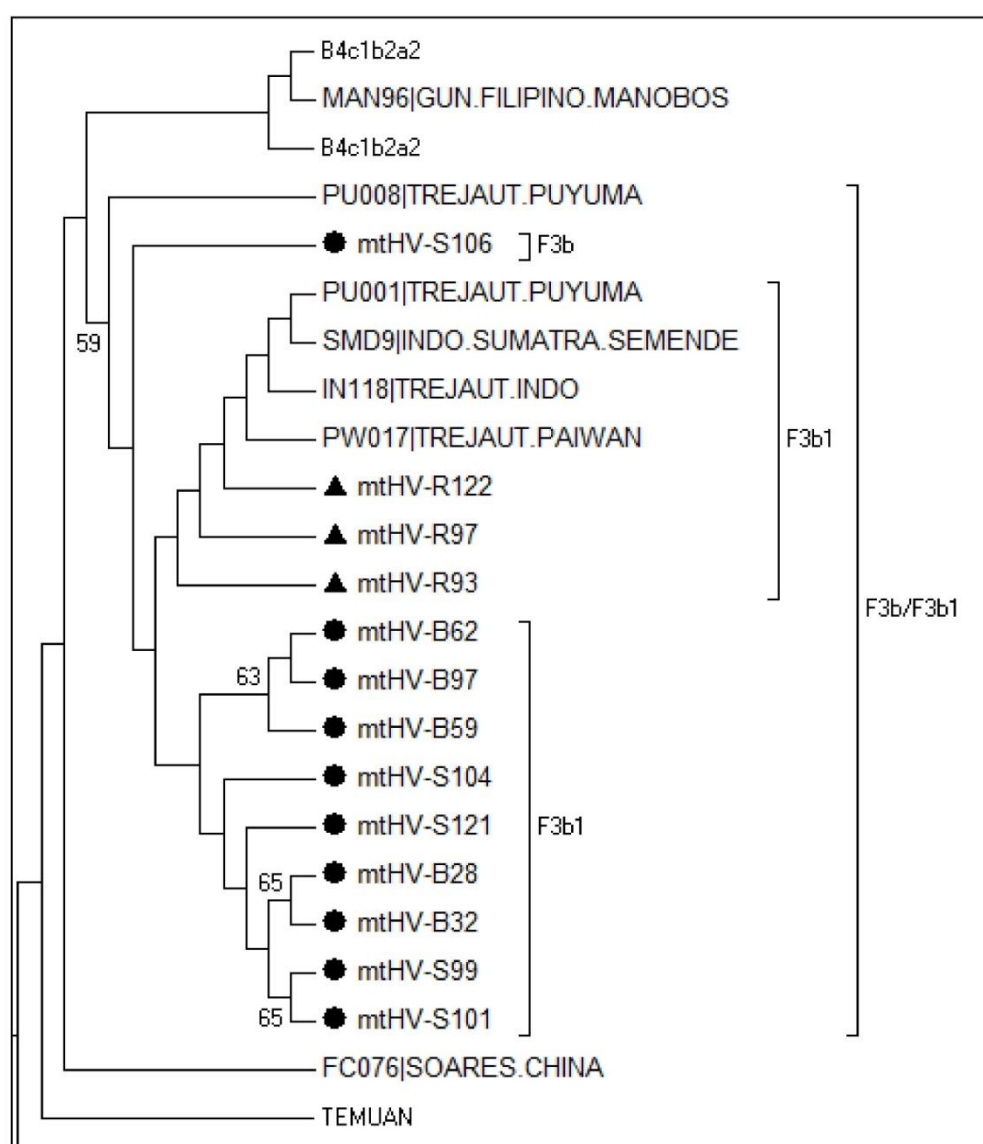


Figure 5.8 : Sub-clades of haplogroups F3b and F3b1 for Bajau and Rungus individuals in the NJ tree; the number at the node denotes the bootstrap value for the branches (only values > 50 % were shown); dark circles and dark triangles denote individuals from Bajau and Rungus populations, respectively.

Haplogroup D5b1c1 represented a very recent demic event from Taiwan, with an estimated age similar to the Y2 lineage, i.e., about 4,000 years old. The D5b1c1 lineage in the Rungus individuals segregated into 2 distinctive clusters, first clade consisted exclusively of the Rungus individuals and the second clade was built of Rungus and Kadazan-Dusun people (Figure 5.9). The root type may be lost due to drift, as in the Taiwanese Aborigines (Hill, et al., 2007). It is possible that the Kadazan-Dusun and Rungus were yet to separate during the dispersal of D5b1c1 into ISEA and the unique variant in the Rungus people developed afterwards via *in situ* evolution.

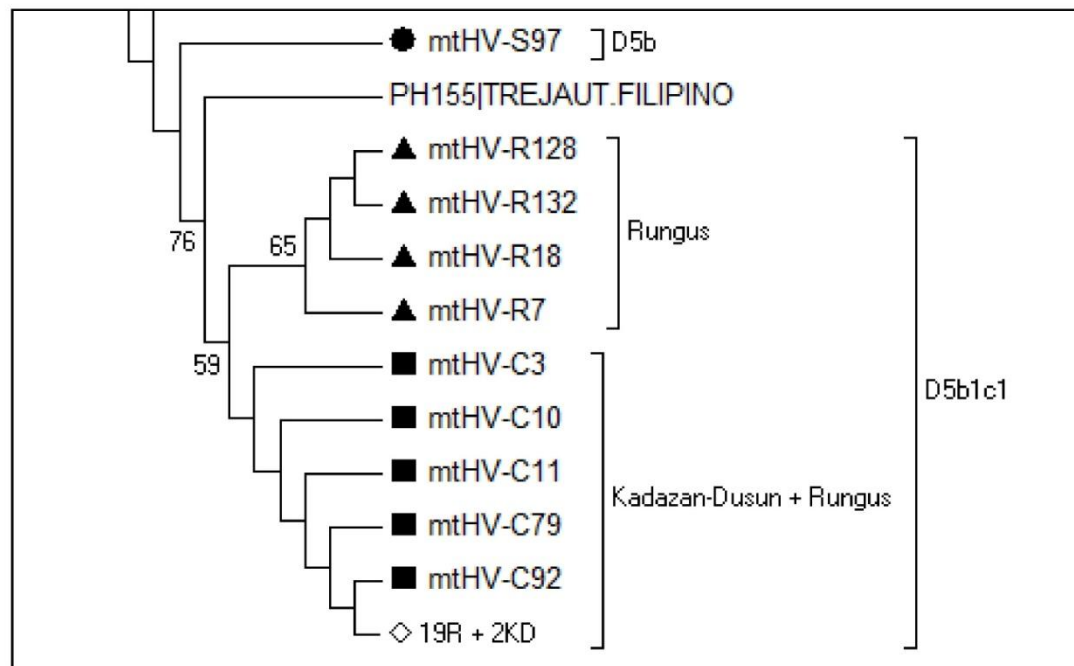


Figure 5.9 : D5b1c1 clusters for Kadazan-Dusun and Rungus individuals; the number at the node denotes the bootstrap value for the branches (only values > 50 % were shown); dark rectangles, dark triangles, and dark circles denote individuals from Kadazan-Dusun, Rungus, and Bajau populations, respectively; empty diamond represents condensed clade with Kadazan-Dusun (KD) and Rungus (R) individuals, the number before the alphabet indicates the amount of individual.

Several mt haplogroups were found to be shared by the Sabahan indigenous populations in the present study, each making up varying fractions of the genetic components of these populations. Based on the phylogenetic analysis, the B4a1a lineages in the Sabahan indigenous populations were grouped into 2 distant clusters. The first cluster consisted of Bajau individuals and various populations in the ISEA region and Oceania [Figure 5.10(a)]. The second cluster on the other hand, was made up by Kadazan-Dusun, Bajau, Rungus, Indonesian, Seletar (Orang Asli), and Bidayuh individuals [Figure 5.10(b)].

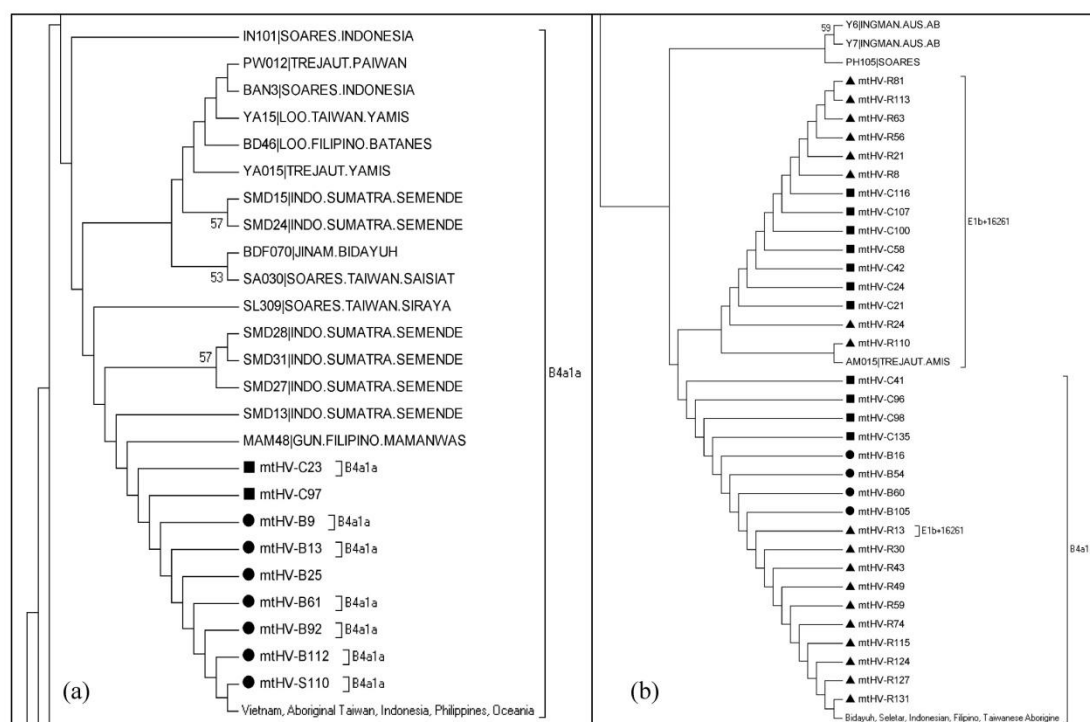


Figure 5.10 : Haplogroups B4a1a and E1b; (a) First variant of B4a1a (b) Second variant of B4a1a and E1b; the number at the node denotes the bootstrap value for the branches (only values > 50 % were shown); dark rectangles, dark triangles, and dark circles denote individuals from Kadazan-Dusun, Rungus, and Bajau populations, respectively.

From the observation, it is likely that the B4a1a lineage dispersed along the route as proposed by the “out of Taiwan” model. It was picked up by the Bajaus, or their ancestral population, in the southern region of Philippines. The first variant was also expected to expand into Indonesia and Oceania through South Philippines. The second variant was seen in Taiwanese Aborigines and Filipino. Thus it may have emerged in Taiwan and dispersed quickly following the same route to the South Philippines, without much interaction with the local populations along the dispersal as explained by the “Express Train” model (Diamond, 1988). From there, it went on to spread to Borneo Island and Malay Peninsula in the Northeast direction and eastward into Indonesia. However, the distribution of the second variant is restricted from the Oceania populations. On the other hand, it could also have emerged as the result of *in situ* evolution in the South Philippines and back migrated to Taiwan.

The Elb clade was situated close to the second variant of B4a1a, where 2 of the individuals identified to carry the Elb lineage fell within the B4a1a cluster [Figure 5.10(b)]. It is because these lineages share a high degree of nucleotide similarity in the control region. The Elb lineage is suggested to have emerged in the ISEA and its limited distribution in this region reflects its recent emergence (Soares, et al., 2011).

The E1a1a, sister lineage of E1b, was found to have greater and wider distribution. The E1a1a lineage made up a large cluster in the phylogenetic tree, with at least 8 observed sub-clades, each of them consisted of individuals from various populations in the ISEA region, including the Sabahan indigenous populations and Oceania (Figure 5.11). There was no unique clade formed by any of the ISEA populations, which reflects a homogenous distribution of the E1a1a variants in this region. Similar observations have been reported and the diversification of the lineage was explained as the impact of several rapid episodes of raising sea level due to global warming on the coastal dwelling populations during the Holocene. It also facilitated the spread of the E1a1a lineage to Taiwan and Oceania (Soares, et al., 2008).



Figure 5.11 : Haplogroup E1a1a; the number at the node denotes the bootstrap value for the branches (only values > 50 % were shown); dark rectangles, dark triangles, and dark circles denote individuals from Kadazan-Dusun, Rungus, and Bajau populations, respectively.

The M7c3c lineage formed another large cluster in the NJ tree, where several sub-clades were observed (Figure 5.12). Despite the observation of a small unique sub-clade, majority of the Bajau individuals clustered close to other ISEA population, such as Filipino, Indonesian, and Oceania. Unlike the Bajaus, the Kadazan-Dusun and Rungus populations formed 2 distinctive clades that did not group together with other ISEA populations. The second clade can be further divided into 2 unique sub-clades by Kadazan-Dusun and Rungus, respectively. These distinctive clades may have derived *in situ* due to long stand evolution in these populations. Based on the distribution patterns, the M7c3c lineage in the Sabahan indigenous population may have arrived in different migratory routes. Following the exodus from Taiwan into North Philippines, the migrating populations may have diverged into 2 separated directions at the Luzon area, where parts of them took the western way into Borneo Island through Palawan Islands and eventually gave rise to the M7c3c lineage in the Kadazan-Dusun and Rungus populations. On the other hand, the remaining of the migrants continued to move southward and reached East Indonesia and Oceania. The southern M7c3c lineage was picked up by ancestors of Bajau people residing in the southern region of Philippines.

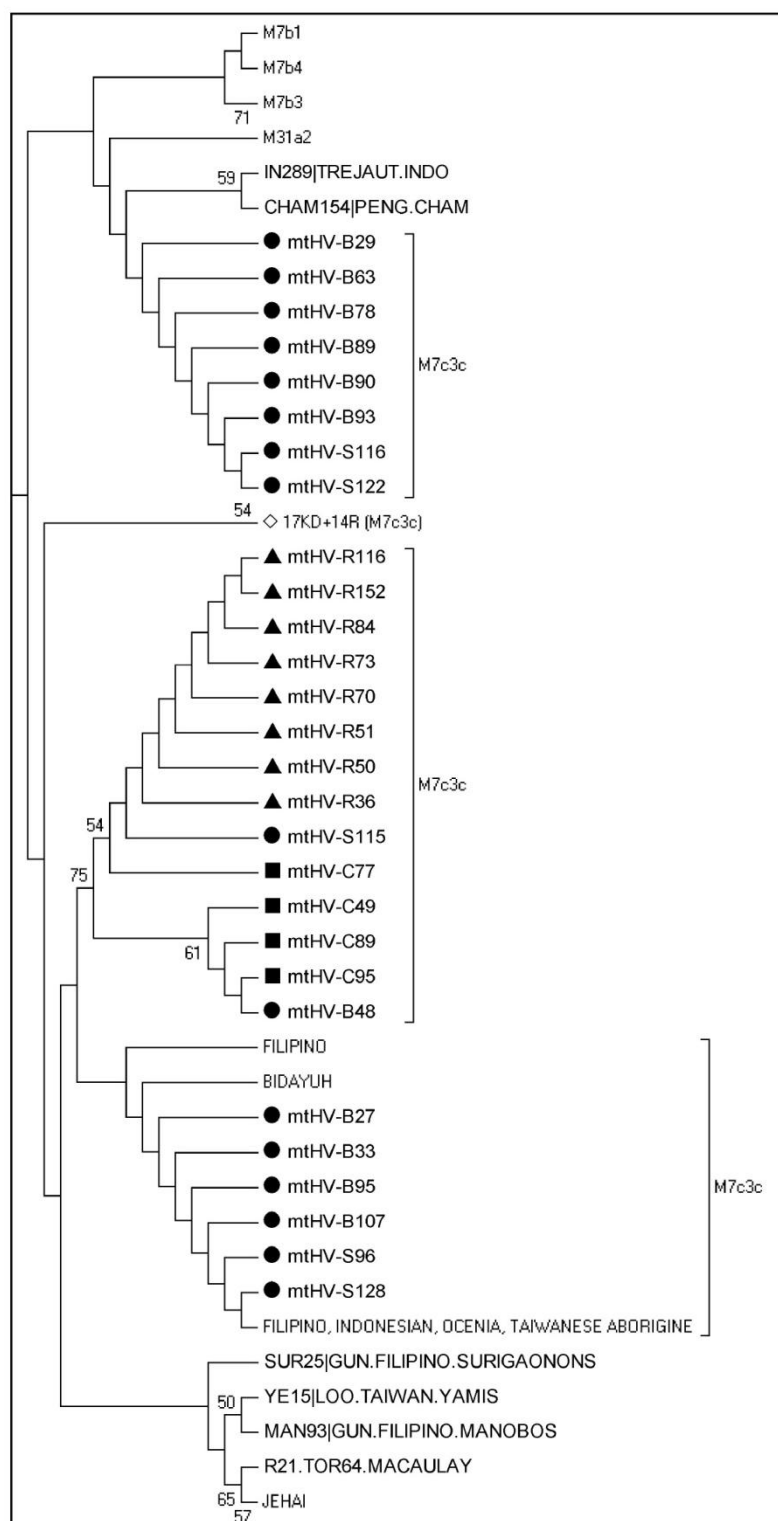


Figure 5.12 : Clustering of M7c3c individuals in the Sabahan indigenous populations; the number at the node denotes the bootstrap value for the branches (only values > 50 % were shown); dark rectangles, dark triangles, and dark circles denote individuals from Kadazan-Dusun, Rungus, and Bajau populations, respectively; empty diamond represents condensed clade with Kadazan-Dusun (KD) and Rungus (R) individuals, the number before the alphabet indicates the amount of individual.

This postulation is further fortified by the distribution patterns of R9c and M7b1 lineages in the ISEA region. The R9c haplogroup was usually reported < 5 % in ISEA populations (i.e., Filipinos and Indonesians). However, it made up a considerably high portion of the mtDNAs in the Sabahan indigenous populations, 17.3 %, 15.3 %, and 10 % in Kadazan-Dusun, Rungus, and Bajau, respectively. It was also found in as much as 60 % of the mt lineages in the Negrito group in Palawan Islands. Therefore, the R9c lineage could have been carried prominently by the migrants who traveled westward into Borneo Island.

Similar to the R9c lineage, higher frequencies of the haplogroup M7b1 was reported in Taiwanese Aborigines and southern Han Chinese than their ISEA counterparts. In the present study, this lineage was especially common in the Kadazan-Dusun population (23 %). Both M7b1 lineages in the Amis (Taiwanese Aborigines) and Sabahan indigenous populations contained a mutation point at np 16,126, which was absent in the East Asians. Therefore, this variant of the M7b1 lineage could have emerged in Taiwan and spread to the Sabahan indigenous groups following the westward route, without dispersing further south. This is evident from the clustering of M7b1 lineages on the NJ tree, where the cluster was made up dominantly by Sabahan indigenous individuals and 2 Filipinos (Figure 5.13).

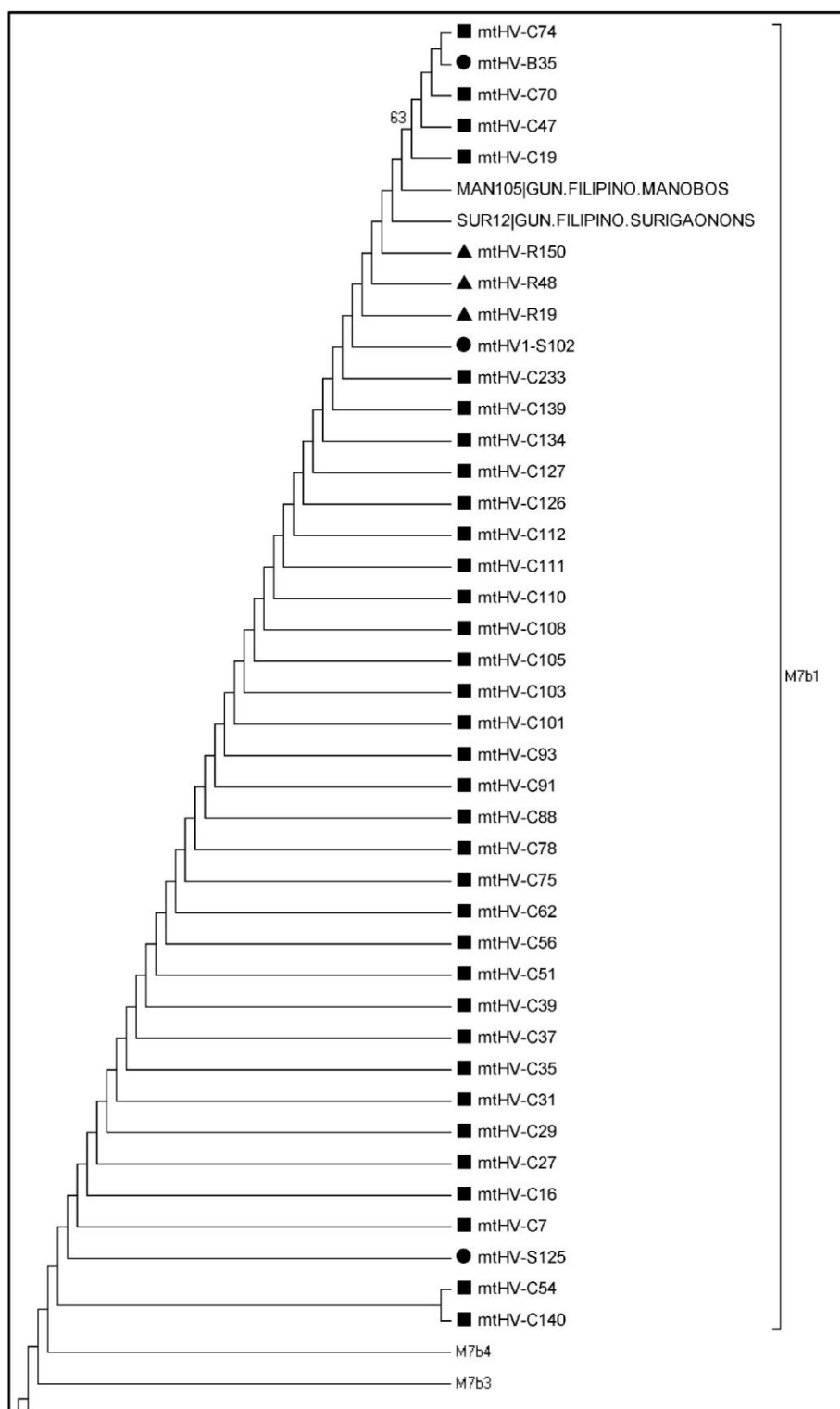


Figure 5.13 : Clade of haplogroup M7b1; the number at the node denotes the bootstrap value for the branches (only values > 50 % were shown); dark rectangles, dark triangles, and dark circles denote individuals from Kadazan-Dusun, Rungus, and Bajau populations, respectively.

5.5 SUMMARY AND CONCLUSION

We have examined 4 subsets of genetic polymorphism (VNTRs, *Alu* insertions, STRs, and mtDNAs) and their distribution in 3 Sabahan indigenous populations (Kadazan-Dusun, Bajau, and Rungus). The advantages and limitations of these marker types are summarized in Table 5.7.

In terms of discriminatory power of each individual marker per se, there was an increasing trend for VNTRs < *Alu* insertions < STRs, with the highest observed values of 0.496, 0.649, and 0.980, respectively. Although majority of the genetic variability of these populations were expressed from within the populations, they can be distinguished as separated populations based on the population differentiation and AMOVA tests. The Kadazan-Dusun and Rungus populations were found to have closer genetic relationship than with the Bajau population.

5.5.1 Variable number of tandem repeat

In the study of 2 VNTRs in the DAT-1 gene, we observed higher diversity in the Bajau population than the Kadazan-Dusun and Rungus groups. The Sabahan indigenous populations displayed similar distribution of the 3'UTR VNTR to other populations in East Asia, SEA, and Oceania, which may indicate the possibility of a common ancestry of these populations following the “southern coastal route” during early migration. On the other hand, the co-existence of the allele 7 of Intron-8 VNTR in Taiwanese Aborigine (Ami tribe) and the Sabahan indigenous populations (Kadazan-Dusun and Rungus people) may represent the gene flow from Taiwan into SEA during the “out of Taiwan” expansion.

Table 5.7 : Advantages and limitations of the 4 subsets of genetic markers used in the study.

Advantage	Limitation
VNTR	
<ul style="list-style-type: none"> • Capability for simple and fast examination by direct PCR and native AGE 	<ul style="list-style-type: none"> • Laborious and time-consuming • Testing requires relatively large amount of high-molecular weight intact DNA sample • Dropout of large-size allele through preferential amplification
<i>Alu</i> insertion	
<ul style="list-style-type: none"> • Insertion represents unique event along the evolution • Stable; insertion is unlikely to revert; identity-by-descent • Known ancestral state, i.e., absence of the <i>Alu</i> insertion • Detection by rapid and simple method 	<ul style="list-style-type: none"> • Bi-allelic marker; display less variation • Low tendency for multiplexing due to the formation of heteroduplexes
STR	
<ul style="list-style-type: none"> • Small-size repeating block • Compatible with degraded DNA • Multiplex amplification with high power of discrimination • Commercially availability and core loci for standard DNA profile 	<ul style="list-style-type: none"> • Relatively high mutation rate that could lead to complication in identification test • Expensive typing assays; high-resolution detection of alleles by costly equipment
MtDNA	
<ul style="list-style-type: none"> • Present in high copy number; increase typing efficiency in highly degraded sample • Useful for tracing evolution • Stable; encapsulated within a double membrane organelle 	<ul style="list-style-type: none"> • Uniparental marker; not individual specific resulting in low discriminating power • Examination requires expensive assay and equipment

5.5.2 *Alu* insertion

The examination of *Alu* insertions in the indigenous populations showed little differentiation among them (AMOVA = 0.5333; G_{ST} = 0.004). Higher frequency of TPA25 insertion was observed in populations outside of Africa, it may indicate that the migrants who left Africa could have carried high proportion of this insertion. The insertion frequency of PV92 marker was higher in populations in East Asia/SEA region, which may serve as a genetic indicator for populations in this region.

Although the Sabahan indigenous populations showed higher affinity to populations in East Asia and SEA regions based on clustering analyses, they retain high degree of genetic similarity among each other.

5.5.3 Short tandem repeat

All STR markers showed sufficiently high degree of diversity ($H_{Obs} > 0.65$) in the Kadazan-Dusun, Bajau, and Rungus populations, except the TPOX loci due to limited number of alleles. In addition, the system displayed high level of discriminatory and exclusion powers, therefore it can be used efficiently for human identification purposes (combined PD = 6.08 to 9.97×10^{16} ; combined PE = 0.9999921 to 0.9999986).

STRUCTURE analysis showed that all 3 Sabahan indigenous populations formed clear separation, representing their distinctive genetic structures. Despite that, greater genetic similarity was observed between the Kadazan-Dusun and Rungus populations than with the Bajau population.

Based on the analysis from the study of VNTRs and *Alu* insertions, these Sabahan indigenous populations were shown to have close genetic resemblance to other populations in the East Asia and SEA regions than populations in other continents.

However, these markers could not provide a more detailed association of populations in this region, which could be caused by the use of limited number of loci. Nevertheless, based on the STR study, the Bajau was shown to group closely with other ISEA populations such as Filipinos and Indonesians. On the other hand, the Kadazan-Dusun and Rungus people have higher affinity with various Taiwanese Aboriginal tribes. This observation may suggest multiple dispersal waves into ISEA/Borneo Island that give rise to different genetic fractions in this region.

5.5.4 Mitochondrial DNA

The study of mtDNAs in these indigenous populations revealed that the Bajaus harbor higher degree of diversity, which highlights the diverse genetic background of the Bajau people in this region. They carried various mt lineages that can be traced to originate from the surrounding regions, i.e., East Asia, South Asia, Oceania, and Indochina. In addition, these mtDNAs consisted of both old and young lineages, aged from 1,500 to 50,000 years old. The genomes of Kadazan-Dusun and Rungus were relatively conserved and less diverse as compared to the Bajaus. Majority of their mt lineages were derived recently, with ages from 2,000 to 30,000 years old, lacking of ancient lineages. They were also observed to harbor more distinct lineages from other ISEA populations, which could be the result of *in situ* evolution for long isolation.

The Bajaus could have arrived in the SEA region in much earlier times than the Kadazan-Dusun and Rungus groups. They could have descended from one of the earlier exodus from Africa that reached as far as Australia about 50,000 years ago. Or, they could have migrated from East Asia more recently, about 20,000 years ago. In the recent “out of Taiwan” expansion, the Bajaus, residing in the southern Philippines, could have assimilated with the migrants before they travelled to Borneo Island through the

Sulu Islands, carrying distinctive mt lineages such as Y2 and B4a2b (Figure 5.14). On the other hand, the Kadazan-Dusun and Rungus people could have derived from the migrants who travelled following the westward route into Borneo Island through the Palawan Islands, as demonstrated by high frequency of D5b1c1, R9c, and M7b1 lineages in the populations along the route. The Kadazan-Dusun and Rungus groups were then separated in Borneo Island and developed into distinctive sub-tribes.

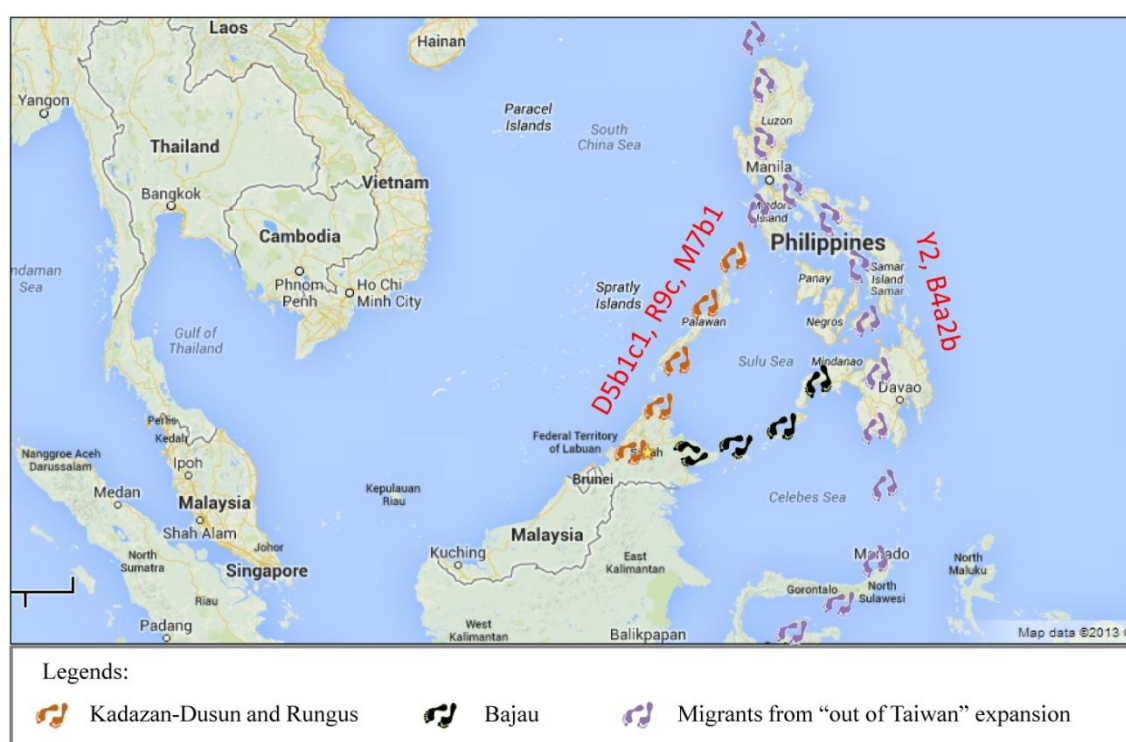


Figure 5.14 : “Out of Taiwan” expansion and the routes of migration taken by the ancestors of Kadazan-Dusun, Rungus, and Bajau populations into the Borneo Island (Source: edited from Google map).

LIMITATIONS
AND
FUTURE STUDIES

LIMITATIONS AND FUTURE STUDIES

While attempted to construct the genetic structure of the Sabahan indigenous populations to the best of our ability, i.e., by including various types of polymorphic markers from different subsets of the human genome, there are still several limitations in our study. First, the genetic data of other human populations in our comparison studies was depending on their public availability, and a majority of them were retrieved from published datasets or online databases. Often, publications in scientific journals only include characteristic data that favor their discussion and these data are incomplete and are unsuited for comparison in our study. Although of date, many specific databases have been established for deposition and storage of various genetic data, not all researchers opt to make their data available, particularly for data sets in older publications. Other limitations include partial data which do not represent the entire population cohort. Thus, our comparison studies were only conducted using complete datasets from limited populations that are currently available. We were not being able to include data from some of the populations of interest, such as other East Malaysian indigenous groups and populations in Sulu and Palawan islands.

Second, the genetic analysis of human populations in our study was based mainly on the variant frequencies of markers, and populations that share similar distribution patterns were determined to have a greater level of genetic relevance. It is true that descended groups would carry subsets from the ancestral genetic pool and inherit some of the characteristic mutation points. However, the frequencies of these genetic variants do not necessarily correspond to the genetic relatedness of the observed populations, especially after these groups had parted for a long time. In general, human migration usually occurs in small groups and the colonists' gene pool would be altered and stabilized due to several phenomena, i.e., founder and bottleneck effects. In the end,

some of the variants could have amplified, be maintained, or even varnished from the descending population.

On top of that, data produced from various genetic studies should be interpreted carefully, especially in statistical analyses with low confidence interval. For example, the distribution of a marker that violates HWE could imply several biases, including high inbreeding, mutation, selection, and small population size in the sampling cohort. Whereas, low bootstrapping value in phylogenetic analysis suggests low confidence level of the tree-structure representation.

For future more conclusive work, elaborative analysis/experiments should be conducted. Although screening of the mt control region is currently sufficient for most genetic analysis, as technology advances, full mt sequencing could be carried out by high throughput and deep coverage platforms for all sample collected. The mt lineages in the Sabahan indigenous populations, such as haplogroups M7b1, Elala, and R9, would especially be of interest. These selected lineages are distributed all over SEA and present a high degree of diversity among SEA populations and their neighboring regions. Complete mt data of these lineages could yield more in-depth information and shed light on the migration patterns. Furthermore, additional parameters can be computed more reliably with the complete mt sequence, such as coalescent age, which could be helpful to determine the estimated ages of these lineages in various populations.

Apart from markers in autosomal and mt DNAs, examination of these indigenous individuals for genetic polymorphisms in the sex chromosomes could provide additional genetic information, to further complement to data in our present study. Both X- and Y-chromosomes harbor an abundant number of genetic markers (*Alus*, STRs, and SNPs), which have been proven useful in many research aspects. Similar to the mtDNA that is

inherited maternally, the Y-chromosome represents uni-parental lineage in the paternal line.

Last but not least, more DNA samples could be collected from the other indigenous populations residing in the Malay Peninsula, Borneo, Sulu, and Palawan islands. This would give us a better overall encompassing view of their genetic markers and the genetic data generated, which could also be used to justify the migratory direction within ISEA.

REFERENCES

-
- Abdulla, M. A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S. K., Calacal, G. C., et al. (2009). Mapping human genetic diversity in Asia. *Science*, 326(5959), 1541-1545.
- Achilli, A., Perego, U. A., Bravi, C. M., Coble, M. D., Kong, Q. P., Woodward, S. R., et al. (2008). The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. *PLoS One*, 3(3), e1764.
- Aiello, L. C., & Wells, J. C. K. (2002). Energetics and the Evolution of the Genus Homo. *Annu Rev Anthropol*, 31, 323-338.
- Alemohammad, S., Farhud, D., Hooshmand, M., & Sanati, M. (2003). Distribution of mitochondrial DNA intergenic COII/tRNALYS 9 bp deletion in Iranian populations. *Iranian J Publ Health*, 32(2), 1-5.
- Alvarez, I., Fernandez, I., Lorenzo, L., Payeras, L., Cuervo, M., & Goyache, F. (2012). Founder and present maternal diversity in two endangered Spanish horse breeds assessed via pedigree and mitochondrial DNA information. *J Anim Breed Genet*, 129(4), 271-279.
- Alves-Silva, J., Guimaraes, P. E., Rocha, J., Pena, S. D., & Prado, V. F. (1999). Identification in Portugal and Brazil of a mtDNA lineage containing a 9-bp triplication of the intergenic COII/tRNALys region. *Hum Hered*, 49(1), 56-58.
- Alves, C., Gusmao, L., Lopez-Parra, A. M., Soledad Mesa, M., Amorim, A., & Arroyo-Pardo, E. (2005). STR allelic frequencies for an African population sample (Equatorial Guinea) using AmpFI STR Identifiler and Powerplex 16 kits. *Forensic Sci Int*, 148(2-3), 239-242.
- Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806), 457-465.
- Anton, S. C. (2003). Natural history of Homo erectus. *Am J Phys Anthropol*, Suppl 37, 126-170.
- Arcot, S. S., Adamson, A. W., Lamerdin, J. E., Kanagy, B., Deininger, P. L., Carrano, A. V., et al. (1996). Alu fossil relics - distribution and insertion polymorphism. *Genome Res*, 6(11), 1084-1092.
-

-
- Arcot, S. S., Fontius, J. J., Deininger, P. L., & Batzer, M. A. (1995). Identification and analysis of a 'young' polymorphic Alu element. *Biochim Biophys Acta*, 1263(1), 99-102.
- Arcot, S. S., Wang, Z., Weber, J. L., Deininger, P. L., & Batzer, M. A. (1995). Alu repeats: a source for the genesis of primate microsatellites. *Genomics*, 29(1), 136-144.
- Armitage, S. J., Jasim, S. A., Marks, A. E., Parker, A. G., Usik, V. I., & Uerpmann, H. P. (2011). The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia. *Science*, 331(6016), 453-456.
- Ballinger, S. W., Schurr, T. G., Torroni, A., Gan, Y. Y., Hodge, J. A., Hassan, K., et al. (1992). Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. *Genetics*, 130(1), 139-152.
- Barber, M. D., McKeown, B. J., & Parkin, B. H. (1996). Structural variation in the alleles of a short tandem repeat system at the human alpha fibrinogen locus. *Int J Legal Med*, 108(4), 180-185.
- Batzer, M. A., & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nat Rev Genet*, 3(5), 370-379.
- Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M., et al. (1996). Standardized nomenclature for Alu repeats. *J Mol Evol*, 42(1), 3-6.
- Batzer, M. A., Rubin, C. M., Hellmann-Blumberg, U., Alegria-Hartman, M., Leeftang, E. P., Stern, J. D., et al. (1995). Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *J Mol Biol*, 247(3), 418-427.
- Batzer, M. A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D. H., Shaikh, T. H., et al. (1994). African origin of human-specific polymorphic Alu insertions. *Proc Natl Acad Sci U S A*, 91(25), 12288-12292.
- Behar, D. M., van Oven, M., Rosset, S., Metspalu, M., Loogvali, E. L., Silva, N. M., et al. (2012). A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from Its Root (vol 90, pg 675, 2012). *Am J Hum Genet*, 90(5), 936-936.
-

-
- Bekaert, B., Zainuddin, Z., Hadi, S., & Goodwin, W. (2006). A comparison of mtDNA and Y chromosome diversity in Malay populations. *Int Congr Ser*, 1288(0), 252-255.
- Belle, E. M., Landry, P. A., & Barbujani, G. (2006). Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. *Proc Biol Sci*, 273(1594), 1595-1602.
- Bellwood, P. (2004). *First Farmers: The Origins of Agricultural Societies*: Wiley-Blackwell.
- Bellwood, P. (2007). *Prehistory of the Indo-Malaysian archipelago*: ANU E Press.
- Betty, D. J., Chin-Atkins, A. N., Croft, L., Sraml, M., & Easteal, S. (1996). Multiple independent origins of the COII/tRNA(Lys) intergenic 9-bp mtDNA deletion in aboriginal Australians. *Am J Hum Genet*, 58(2), 428-433.
- Bjork, A., Liu, W., Wertheim, J. O., Hahn, B. H., & Worobey, M. (2011). Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. *Mol Biol Evol*, 28(1), 615-623.
- Blust, R. (1999). Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics. *Symp Ser Inst Linguist Acad Sinica*, 1, 31-94.
- Bodner, M., Zimmermann, B., Rock, A., Kloss-Brandstatter, A., Horst, D., Horst, B., et al. (2011). Southeast Asian diversity: first insights into the complex mtDNA structure of Laos. *BMC Evol Biol*, 11, 49.
- Bonne-Tamir, B., Korostishevsky, M., Redd, A. J., Pel-Or, Y., Kaplan, M. E., & Hammer, M. F. (2003). Maternal and paternal lineages of the Samaritan isolate: mutation rates and time to most recent common male ancestor. *Ann Hum Genet*, 67(Pt 2), 153-164.
- Boutin, M., & Boutin, A. (1985). *Indigenous Groups of Sabah: An Annotated Bibliography of Linguistic and Anthropological Sources: Supplement 1*: Summer Institute of Linguistics.
- Bowler, J. M., Johnston, H., Olley, J. M., Prescott, J. R., Roberts, R. G., Shawcross, W., et al. (2003). New ages for human occupation and climatic change at Lake Mungo, Australia. *Nature*, 421(6925), 837-840.
-

-
- Brandstatter, A., Klein, R., Duftner, N., Wiegand, P., & Parson, W. (2006). Application of a quasi-median network analysis for the visualization of character conflicts to a population sample of mitochondrial DNA control region sequences from southern Germany (Ulm). *Int J Legal Med*, 120(5), 310-314.
- Britten, R. J. (1996). DNA sequence insertion and evolutionary variation in gene regulation. *Proc Natl Acad Sci U S A*, 93(18), 9374-9377.
- Brown, T. (2010). *Gene cloning and DNA analysis: an introduction*: Wiley-Blackwell.
- Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A., Ahounta, D., et al. (2002). A new hominid from the Upper Miocene of Chad, central Africa. *Nature*, 418(6894), 145-151.
- Butler, J. M. (2005). *Forensic DNA typing: biology, technology, and genetics of STR markers*: Academic Press.
- Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, 325(6099), 31-36.
- Challem, J. J., & Taylor, E. W. (1998). Retroviruses, ascorbate, and mutations, in the evolution of Homo sapiens. *Free Radic Biol Med*, 25(1), 130-132.
- Chandrasekar, A., Kumar, S., Sreenath, J., Sarkar, B. N., Urade, B. P., Mallick, S., et al. (2009). Updating phylogeny of mitochondrial DNA macrohaplogroup m in India: dispersal of modern human in South Asian corridor. *PLoS One*, 4(10), e7447.
- Chang, Y. M., Perumal, R., Keat, P. Y., & Kuehn, D. L. (2007). Haplotype diversity of 16 Y-chromosomal STRs in three main ethnic populations (Malays, Chinese and Indians) in Malaysia. *Forensic Sci Int*, 167(1), 70-76.
- Chen, F. C., & Li, W. H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet*, 68(2), 444-456.
- Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., et al. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, 437(7055), 88-93.
-

-
- Clark, V. J., Sivendren, S., Saha, N., Bentley, G. R., Aunger, R., Sirajuddin, S. M., et al. (2000). The 9-bp deletion between the mitochondrial lysine tRNA and COII genes in tribal populations of India. *Hum Biol*, 72(2), 273-285.
- Collins, F. S., Drumm, M. L., Cole, J. L., Lockwood, W. K., Vande, W. G. F., & Iannuzzi, M. C. (1987). Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science*, 235(4792), 1046.
- Cordaux, R., Weiss, G., Saha, N., & Stoneking, M. (2004). The northeast Indian passageway: a barrier or corridor for human migrations? *Mol Biol Evol*, 21(8), 1525-1533.
- Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., et al. (2002). A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet*, 70(5), 1197-1214.
- Dancause, K. N., Chan, C. W., Arunotai, N. H., & Lum, J. K. (2009). Origins of the Moken Sea Gypsies inferred from mitochondrial hypervariable region and whole genome sequences. *J Hum Genet*, 54(2), 86-93.
- Darwin, C. (1871). *The descent of man*. London: Gibson Square Books.
- Dawson, E., Chen, Y., Hunt, S., Smink, L. J., Hunt, A., Rice, K., et al. (2001). A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. *Genome Res*, 11(1), 170-178.
- Deininger, P. L., & Batzer, M. A. (1999). Alu repeats and human disease. *Mol Genet Metab*, 67(3), 183-193.
- Demeter, F., Shackelford, L. L., Bacon, A. M., Durringer, P., Westaway, K., Sayavongkhamdy, T., et al. (2012). Anatomically modern human in Southeast Asia (Laos) by 46 ka. *Proc Natl Acad Sci U S A*, 109(36), 14375-14380.
- Deng, Z. H., Li, Q., Li, D. C., Wang, D. M., Gao, S. Q., & Wu, G. G. (2006). [The genetic polymorphisms of nine Y-STR loci with short fragment size alleles in southern Chinese Han population and its application in forensic science]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi*, 23(4), 470-474.
-

-
- Derenko, M., Malyarchuk, B., Denisova, G., Perkova, M., Rogalla, U., Grzybowski, T., et al. (2012). Complete mitochondrial DNA analysis of eastern Eurasian haplogroups rarely found in populations of northern Asia and eastern Europe. *PLoS One*, 7(2), e32179.
- Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Dambueva, I., Perkova, M., et al. (2007). Phylogeographic analysis of mitochondrial DNA in northern Asian populations. *Am J Hum Genet*, 81(5), 1025-1041.
- Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Rogalla, U., Perkova, M., et al. (2010). Origin and post-glacial dispersal of mitochondrial DNA haplogroups C and D in northern Asia. *PLoS One*, 5(12), e15214.
- Detroit, F., Dizon, E., Falgueres, C., Hameau, S., Ronquillo, W., & Semah, F. (2004). Upper Pleistocene *Homo sapiens* from the Tabon cave (Palawan, The Philippines): description and dating of new discoveries. *Comptes Rendus Palevol*, 3(8), 705-712.
- Diamond, J. M. (1988). Express train to Polynesia. *Nature*, 336, 307-308.
- Doolittle, W. F., & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284, 601-603.
- Duarte, C., Mauricio, J., Pettitt, P. B., Souto, P., Trinkaus, E., van der Plicht, H., et al. (1999). The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. *Proc Natl Acad Sci U S A*, 96(13), 7604-7609.
- Dubut, V., Cartault, F., Payet, C., Thionville, M. D., & Murail, P. (2009). Complete mitochondrial sequences for haplogroups M23 and M46: insights into the Asian ancestry of the Malagasy population. *Hum Biol*, 81(4), 495-500.
- Dubut, V., Murail, P., Pech, N., Thionville, M. D., & Cartault, F. (2009). Inter- and extra-Indian admixture and genetic diversity in reunion island revealed by analysis of mitochondrial DNA. *Ann Hum Genet*, 73(Pt 3), 314-334.
- Dulik, M. C., Zhadanov, S. I., Osipova, L. P., Askapuli, A., Gau, L., Gokcumen, O., et al. (2012). Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians. *Am J Hum Genet*, 90(2), 229-246.
-

-
- Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour*, 1-3.
- Edelmann, J., Deichsel, D., Hering, S., Plate, I., & Szibor, R. (2002). Sequence variation and allele nomenclature for the X-linked STRs DXS9895, DXS8378, DXS7132, DXS6800, DXS7133, GATA172D05, DXS7423 and DXS8377. *Forensic Sci Int*, 129(2), 99-103.
- Elia, J., Sackett, J., Turner, T., Schardt, M., Tang, S. C., Kurtz, N., et al. (2012). Attention-deficit/hyperactivity disorder genomics: update for clinicians. *Curr Psychiatry Rep*, 14(5), 579-589.
- Endicott, P., Metspalu, M., Stringer, C., Macaulay, V., Cooper, A., & Sanchez, J. J. (2006). Multiplexed SNP typing of ancient DNA clarifies the origin of Andaman mtDNA haplogroups amongst South Asian tribal populations. *PLoS One*, 1, e81.
- Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*, 14(8), 2611-2620.
- Excoffier, L., Laval, G., & Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online*, 1, 47-50.
- Flint, J., Rochette, J., Craddock, C., Dode, C., Vignes, B., Horsley, S., et al. (1996). Chromosomal stabilisation by a subtelomeric rearrangement involving two closely related Alu elements. *Hum Mol Genet*, 5(8), 1163-1169.
- Forster, P. (2004). Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Philos Trans R Soc Lond B Biol Sci*, 359(1442), 255-264; discussion 264.
- Forster, P., & Romano, V. (2007). Timing of a back-migration into Africa. *Science*, 316(5821), 50-53.
- Foster, A., & Laurin, N. (2012). Development of a fast PCR protocol enabling rapid generation of AmpFI STR(R) Identifiler(R) profiles for genotyping of human DNA. *Investig Genet*, 3, 6.
-

-
- Frayer, D. W., Wolpoff, M. H., Thorne, A. G., Smith, F. H., & Pope, G. G. (1993). Theories of Modern Human Origins - the Paleontological Test. *Am Anthropol*, 95(1), 14-50.
- Friedlaender, J. S., Friedlaender, F. R., Hodgson, J. A., Stoltz, M., Koki, G., Horvat, G., et al. (2007). Melanesian mtDNA complexity. *PLoS One*, 2(2), e248.
- Fucharoen, G., Fucharoen, S., & Horai, S. (2001). Mitochondrial DNA polymorphisms in Thailand. *J Hum Genet*, 46(3), 115-125.
- Gamble, C., Davies, W., Pettitt, P., & Richards, M. (2004). Climate change and evolving human diversity in Europe during the last glacial. *Philos Trans R Soc Lond B Biol Sci*, 359(1442), 243-253; discussion 253-244.
- Gill, P., Brown, R. M., Fairley, M., Lee, L., Smyth, M., Simpson, N., et al. (2008). National recommendations of the Technical UK DNA working group on mixture interpretation for the NDNAD and for court going purposes. *Forensic Sci Int Genet*, 2(1), 76-82.
- Gill, P., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., et al. (1994). Identification of the remains of the Romanov family by DNA analysis. *Nat Genet*, 6(2), 130-135.
- Giros, B., el Mestikawy, S., Godinot, N., Zheng, K., Han, H., Yang-Feng, T., et al. (1992). Cloning, pharmacological characterization, and chromosome assignment of the human dopamine transporter. *Mol Pharmacol*, 42(3), 383-390.
- Gomez, M. V., Reyes, M. E., Cardenas, H., & Garcia, O. (2003). Genetic variation for 12 STRs loci in a Colombian population (Department of Valle del Cauca). *Forensic Sci Int*, 137(2-3), 235-237.
- Goudet, J. (1995). FSTAT (Version 1.2): A computer program to calculate F-statistics. *J Hered*, 86(6), 485-486.
- Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A Draft Sequence of the Neandertal Genome. *Science*, 328(5979), 710-722.
- Greenberg, J. H., Turner, C. G., & Zegura, S. L. (1986). The Settlement of the America - a Comparison of the Linguistic, Dental, and Genetic-Evidence. *Curr Anthropol*, 27(5), 477-497.
-

-
- Guindalini, C., Howard, M., Haddley, K., Laranjeira, R., Collier, D., Ammar, N., et al. (2006). A dopamine transporter gene functional variant associated with cocaine abuse in a Brazilian sample. *Proc Natl Acad Sci U S A*, 103(12), 4552-4557.
- Gunnarsdottir, E. D., Li, M., Bauchet, M., Finstermeier, K., & Stoneking, M. (2011). High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res*, 21(1), 1-11.
- Gunnarsdottir, E. D., Nandineni, M. R., Li, M., Myles, S., Gil, D., Pakendorf, B., et al. (2011). Larger mitochondrial DNA than Y-chromosome differences between matrilineal and patrilineal groups from Sumatra. *Nat Commun*, 2, 228.
- Hagelberg, E., Goldman, N., Lio, P., Whelan, S., Schiefenhover, W., Clegg, J. B., et al. (1999). Evidence for mitochondrial DNA recombination in a human population of island Melanesia. *Proc Biol Sci*, 266(1418), 485-492.
- Haile-Selassie, Y. (2001). Late Miocene hominids from the Middle Awash, Ethiopia. *Nature*, 412(6843), 178-181.
- Hall, T. A. (1999). *BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT*. Paper presented at the Nucleic Acids Symp. Ser.
- Harihara, S., Hirai, M., Suutou, Y., Shimizu, K., & Omoto, K. (1992). Frequency of a 9-bp deletion in the mitochondrial DNA among Asian populations. *Hum Biol*, 64(2), 161-166.
- Hartmann, A., Thieme, M., Nanduri, L. K., Stempf, T., Moehle, C., Kivisild, T., et al. (2009). Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Hum Mutat*, 30(1), 115-122.
- He, J. D., Peng, M. S., Quang, H. H., Dang, K. P., Trieu, A. V., Wu, S. F., et al. (2012). Patrilineal perspective on the Austronesian diffusion in Mainland Southeast Asia. *PLoS One*, 7(5), e36437.
- Hering, S., Augustin, C., Edelmann, J., Heide, M., Dressler, J., Rodig, H., et al. (2006). DXS10079, DXS10074 and DXS10075 are STRs located within a 280-kb region of Xq12 and provide stable haplotypes useful for complex kinship cases. *Int J Legal Med*, 120(6), 337-345.
-

-
- Hertzberg, M., Mickleson, K. N., Serjeantson, S. W., Prior, J. F., & Trent, R. J. (1989). An Asian-specific 9-bp deletion of mitochondrial DNA is frequently found in Polynesians. *Am J Hum Genet*, 44(4), 504-510.
- Hill, C., Soares, P., Mormina, M., Macaulay, V., Clarke, D., Blumbach, P. B., et al. (2007). A mitochondrial stratigraphy for island southeast Asia. *Am J Hum Genet*, 80(1), 29-43.
- Hill, C., Soares, P., Mormina, M., Macaulay, V., Meehan, W., Blackburn, J., et al. (2006). Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol Biol Evol*, 23(12), 2480-2491.
- Horai, S., Murayama, K., Hayasaka, K., Matsubayashi, S., Hattori, Y., Fucharoen, G., et al. (1996). mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. *Am J Hum Genet*, 59(3), 579-590.
- Houck, C. M., Rinehart, F. P., & Schmid, C. W. (1979). A ubiquitous family of repeated DNA sequences in the human genome. *J Mol Biol*, 132(3), 289-306.
- Hudjashov, G., Kivisild, T., Underhill, P. A., Endicott, P., Sanchez, J. J., Lin, A. A., et al. (2007). Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc Natl Acad Sci U S A*, 104(21), 8726-8730.
- Hurles, M. E., Sykes, B. C., Jobling, M. A., & Forster, P. (2005). The dual origin of the Malagasy in Island Southeast Asia and East Africa: evidence from maternal and paternal lineages. *Am J Hum Genet*, 76(5), 894-901.
- Inglehearn, C. F., & Cooke, H. J. (1990). A VNTR immediately adjacent to the human pseudoautosomal telomere. *Nucleic Acids Res*, 18(3), 471-476.
- Ingman, M., & Gyllensten, U. (2003). Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. *Genome Res*, 13(7), 1600-1606.
- Ivanova, R., Astrinidis, A., Lepage, V., Djoulah, S., Wijnen, E., Vu-Trieu, A. N., et al. (1999). Mitochondrial DNA polymorphism in the Vietnamese population. *Eur J Immunogenet*, 26(6), 417-422.
- Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14), 1801-1806.
-

-
- Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature*, 314(6006), 67-73.
- Jin, L., & Su, B. (2000). Natives or immigrants: modern human origin in east Asia. *Nat Rev Genet*, 1(2), 126-133.
- Jinam, T. A., Hong, L. C., Phipps, M. E., Stoneking, M., Ameen, M., Edo, J., et al. (2012). Evolutionary history of continental southeast Asians: "early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. *Mol Biol Evol*, 29(11), 3513-3527.
- Jones, C. W., & Kafatos, F. C. (1982). Accepted mutations in a gene family: evolutionary diversification of duplicated DNA. *J Mol Evol*, 19(1), 87-103.
- Jorde, L. B., & Wooding, S. P. (2004). Genetic variation, classification and 'race'. *Nat Genet*, 36(11 Suppl), S28-33.
- Kang, A. M., Palmatier, M. A., & Kidd, K. K. (1999). Global variation of a 40-bp VNTR in the 3'-untranslated region of the dopamine transporter gene (SLC6A3). *Biol Psychiatry*, 46(2), 151-160.
- Kapitonov, V., & Jurka, J. (1996). The age of Alu subfamilies. *J Mol Evol*, 42(1), 59-65.
- Karwautz, A., Campos de Sousa, S., Konrad, A., Zesch, H. E., Wagner, G., Zormann, A., et al. (2008). Family-based association analysis of functional VNTR polymorphisms in the dopamine transporter gene in migraine with and without aura. *J Neural Transm*, 115(1), 91-95.
- Kasamatsu, H., & Vinograd, J. (1974). Replication of circular DNA in eukaryotic cells. *Annu Rev Biochem*, 43(0), 695-719.
- Kayser, M., Brauer, S., Weiss, G., Schiefenhover, W., Underhill, P., Shen, P., et al. (2003). Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am J Hum Genet*, 72(2), 281-302.
- Kee, B. P., Lian, L. H., Lee, P. C., Lai, T. X., & Chua, K. H. (2011). Genetic data for 15 STR loci in a Kadazan-Dusun population from East Malaysia. *Genet Mol Res*, 10(2), 739-743.

-
- Kong, Q. P., Sun, C., Wang, H. W., Zhao, M., Wang, W. Z., Zhong, L., et al. (2011). Large-scale mtDNA screening reveals a surprising matrilineal complexity in east Asia and its implications to the peopling of the region. *Mol Biol Evol*, 28(1), 513-522.
- Kong, Q. P., Yao, Y. G., Sun, C., Bandelt, H. J., Zhu, C. L., & Zhang, Y. P. (2003). Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. *Am J Hum Genet*, 73(3), 671-676.
- Kramerov, D. A., & Vassetzky, N. S. (2005). Short retroposons in eukaryotic genomes. *Int Rev Cytol*, 247, 165-221.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860-921.
- Le Couteur, D. G., Leighton, P. W., McCann, S. J., & Pond, S. (1997). Association of a polymorphism in the dopamine-transporter gene with Parkinson's disease. *Mov Disord*, 12(5), 760-763.
- Leakey, M., & Walker, A. (1997). Early hominid fossils from Africa. *Sci Am*, 276(6), 74-79.
- Leeflang, E. P., Liu, W. M., Hashimoto, C., Choudary, P. V., & Schmid, C. W. (1992). Phylogenetic evidence for multiple Alu source genes. *J Mol Evol*, 35(1), 7-16.
- Lell, J. T., Sukernik, R. I., Starikovskaya, Y. B., Su, B., Jin, L., Schurr, T. G., et al. (2002). The dual origin and Siberian affinities of Native American Y chromosomes. *Am J Hum Genet*, 70(1), 192-206.
- Levinson, G., & Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol*, 4(3), 203-221.
- Li, H., Wen, B., Chen, S. J., Su, B., Pramoongago, P., Liu, Y., et al. (2008). Paternal genetic affinity between Western Austronesians and Daic populations. *BMC Evol Biol*, 8, 146.
- Lim, L. S., Ang, K. C., Mahani, M. C., Shahrom, A. W., & Md-Zain, B. M. (2010). Mitochondrial DNA polymorphism and phylogenetic relationships of Proto Malays in Peninsular Malaysia. *J Biol Sci*, 10, 71-83.
-

-
- Lipska, B. K., Peters, T., Hyde, T. M., Halim, N., Horowitz, C., Mitkus, S., et al. (2006). Expression of DISC1 binding partners is reduced in schizophrenia and associated with DISC1 SNPs. *Hum Mol Genet*, 15(8), 1245-1258.
- Liu, P., Scherer, J. R., Greenspoon, S. A., Chiesl, T. N., & Mathies, R. A. (2011). Integrated sample cleanup and capillary array electrophoresis microchip for forensic short tandem repeat analysis. *Forensic Sci Int Genet*, 5(5), 484-492.
- Loo, J. H., Trejaut, J. A., Yen, J. C., Chen, Z. S., Lee, C. L., & Lin, M. (2011). Genetic affinities between the Yami tribe people of Orchid Island and the Philippine Islanders of the Batanes archipelago. *BMC Genet*, 12.
- Lum, J. K., & Cann, R. L. (2000). mtDNA lineage analyses: origins and migrations of Micronesians and Polynesians. *Am J Phys Anthropol*, 113(2), 151-168.
- Lum, J. K., Rickards, O., Ching, C., & Cann, R. L. (1994). Polynesian mitochondrial DNAs reveal three deep maternal lineage clusters. *Hum Biol*, 66(4), 567-590.
- Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., et al. (2005). Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*, 308(5724), 1034-1036.
- Maity, B., Nunga, S. C., & Kashyap, V. K. (2003). Genetic polymorphism revealed by 13 tetrameric and 2 pentameric STR loci in four Mongoloid tribal population. *Forensic Sci Int*, 132(3), 216-222.
- Malyarchuk, B., Grzybowski, T., Derenko, M., Perkova, M., Vanacek, T., Lazur, J., et al. (2008). Mitochondrial DNA phylogeny in Eastern and Western Slavs. *Mol Biol Evol*, 25(8), 1651-1658.
- Mamedov, I. Z., Shagina, I. A., Kurnikova, M. A., Novozhilov, S. N., Shagin, D. A., & Lebedev, Y. B. (2010). A new set of markers for human identification based on 32 polymorphic Alu insertions. *Eur J Hum Genet*, 18(7), 808-814.
- Mansur, K., Kogid, M., & Madais, S. J. (2010). Human capital investment: Literature review analysis and a study case among Kadazan-Dusun of Pulutan village, Menggatal, Kota Kinabalu, Sabah, Malaysia. *Soc Sci*, 5(3), 264-269.
- Matera, A. G., Hellmann, U., Hintz, M. F., & Schmid, C. W. (1990). Recently transposed Alu repeats result from multiple source genes. *Nucleic Acids Res*, 18(20), 6019-6023.
-

-
- McCallum, L. K., Fernandez, F., Quinlan, S., Macartney, D. P., Lea, R. A., & Griffiths, L. R. (2007). Association study of a functional variant in intron 8 of the dopamine transporter gene and migraine susceptibility. *Eur J Neurol*, 14(6), 706-707.
- Meacham, W. (1988). On the improbability of Austronesian origins in South China. *Asian Perspect*, 26, 89-106.
- Melton, T., Peterson, R., Redd, A. J., Saha, N., Sofro, A. S., Martinson, J., et al. (1995). Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. *Am J Hum Genet*, 57(2), 403-414.
- Merriwether, D. A., Hall, W. W., Vahlne, A., & Ferrell, R. E. (1996). mtDNA variation indicates Mongolia may have been the source for the founding population for the New World. *Am J Hum Genet*, 59(1), 204-212.
- Metspalu, M., Romero, I. G., Yunusbayev, B., Chaubey, G., Mallick, C. B., Hudjashov, G., et al. (2011). Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet*, 89(6), 731-744.
- Mighell, A. J., Markham, A. F., & Robinson, P. A. (1997). Alu sequences. *FEBS Lett*, 417(1), 1-5.
- Mignini, F., Napolioni, V., Codazzo, C., Carpi, F. M., Vitali, M., Romeo, M., et al. (2012). DRD2/ANKK1 TaqIA and SLC6A3 VNTR polymorphisms in alcohol dependence: association and gene-gene interaction study in a population of Central Italy. *Neurosci Lett*, 522(2), 103-107.
- Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., et al. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*, 16(9), 1182-1190.
- Mills, R. E., Pittard, W. S., Mullaney, J. M., Farooq, U., Creasy, T. H., Mahurkar, A. A., et al. (2011). Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res*, 21(6), 830-839.
- Mitchell, R. J., & Hammer, M. F. (1996). Human evolution and the Y chromosome. *Curr Opin Genet Dev*, 6(6), 737-742.
-

-
- Mona, S., Grunz, K. E., Brauer, S., Pakendorf, B., Castri, L., Sudoyo, H., et al. (2009). Genetic admixture history of Eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. *Mol Biol Evol*, 26(8), 1865-1877.
- Morin, P. A., Luikart, G., & Wayne, R. K. (2004). SNPs in ecology, evolution and conservation. *Trends Ecol Evol*, 19(4), 208-216.
- Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet*, 19(R2), R131-136.
- Nei, M., Tajima, F., & Tatenno, Y. (1983). Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J Mol Evol*, 19(2), 153-170.
- Niemi, M., Blauer, A., Iso-Touru, T., Nystrom, V., Harjula, J., Taavitsainen, J. P., et al. (2013). Mitochondrial DNA and Y-chromosomal diversity in ancient populations of domestic sheep (*Ovis aries*) in Finland: comparison with contemporary sheep breeds. *Genet Sel Evol*, 45(1), 2.
- Nimmo, H. A. (1968). Reflections on Bajau History. *Philippine Studies: Historical and Ethnographic Viewpoints*, 16(1), 32-59.
- Nishimura, Y., Yoshinari, T., Naruse, K., Yamada, T., Sumi, K., Mitani, H., et al. (2006). Active digestion of sperm mitochondrial DNA in single living sperm revealed by optical tweezers. *Proc Natl Acad Sci U S A*, 103(5), 1382-1387.
- O'Gara, C., Stapleton, J., Sutherland, G., Guindalini, C., Neale, B., Breen, G., et al. (2007). Dopamine transporter polymorphisms are associated with short-term response to smoking cessation treatment. *Pharmacogenet Genomics*, 17(1), 61-67.
- Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., et al. (2006). The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science*, 314(5806), 1767-1770.
- Olley, J. M., Roberts, R. G., Yoshida, H., & Bowler, J. M. (2006). Single-grain optical dating of grave-infill associated with human burials at Lake Mungo, Australia. *Quat Sci Rev*, 25(19-20), 2469-2474.
- Oppenheimer, S. (2009). The great arc of dispersal of modern humans: Africa to Australia. *Quatern Int*, 202, 2-13.
-

-
- Oppenheimer, S., & Richards, M. (2001a). Fast trains, slow boats, and the ancestry of the Polynesian islanders. *Sci Prog*, 84(Pt 3), 157-181.
- Oppenheimer, S., & Richards, M. (2001b). Polynesian origins. Slow boat to Melanesia? *Nature*, 410(6825), 166-167.
- Pakendorf, B., & Stoneking, M. (2005). Mitochondrial DNA and human evolution. *Annu Rev Genomics Hum Genet*, 6, 165-183.
- Palanichamy, M. G., Sun, C., Agrawal, S., Bandelt, H. J., Kong, Q. P., Khan, F., et al. (2004). Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am J Hum Genet*, 75(6), 966-978.
- Passarino, G., Semino, O., Modiano, G., & Santachiara-Benerecetti, A. S. (1993). COII/tRNA(Lys) intergenic 9-bp deletion and other mtDNA markers clearly reveal that the Tharus (southern Nepal) have Oriental affinities. *Am J Hum Genet*, 53(3), 609-618.
- Peakall, R., & Smouse, P. (2012). GenAEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*.
- Peng, M. S., Palanichamy, M. G., Yao, Y. G., Mitra, B., Cheng, Y. T., Zhao, M., et al. (2011). Inland post-glacial dispersal in East Asia revealed by mitochondrial haplogroup M9a'b. *BMC Biol*, 9, 2.
- Peng, M. S., Quang, H. H., Dang, K. P., Trieu, A. V., Wang, H. W., Yao, Y. G., et al. (2010). Tracing the Austronesian footprint in Mainland Southeast Asia: a perspective from mitochondrial DNA. *Mol Biol Evol*, 27(10), 2417-2430.
- Pepinski, W., Niemcunowicz-Janica, A., Janica, J., Skawronska, M., Koc-Zorawska, E., Rydzewska, M., et al. (2001). Population data for the STR systems D8S1132, CD4, VWA and TH01 in the region of Podlasie (northeastern Poland). *Med Sci Monit*, 7(1), 130-133.
- Persico, A. M., Bird, G., Gabbay, F. H., & Uhl, G. R. (1996). D2 dopamine receptor gene TaqI A1 and B1 restriction fragment length polymorphisms: enhanced frequencies in psychostimulant-preferring polysubstance abusers. *Biol Psychiatry*, 40(8), 776-784.

-
- Persico, A. M., & Macciardi, F. (1997). Genotypic association between dopamine transporter gene polymorphisms and schizophrenia. *Am J Med Genet*, 74(1), 53-57.
- Pierson, M. J., Martinez-Arias, R., Holland, B. R., Gemmell, N. J., Hurles, M. E., & Penny, D. (2006). Deciphering past human population movements in Oceania: provably optimal trees of 127 mtDNA genomes. *Mol Biol Evol*, 23(10), 1966-1975.
- Plummer, T. (2004). Flaked stones and old bones: biological and cultural evolution at the dawn of technology. *Am J Phys Anthropol, Suppl* 39, 118-164.
- Prasad, B. V., Ricker, C. E., Watkins, W. S., Dixon, M. E., Rao, B. B., Naidu, J. M., et al. (2001). Mitochondrial DNA variation in Nicobarese Islanders. *Hum Biol*, 73(5), 715-725.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945-959.
- Puyok, A., & Bagang, T. P. (2011). Ethnicity, culture and indigenous leadership in modern politics: the case of the KadazanDusun in Sabah, East Malaysia. *Kajian Malaysia*, 29(Supp. 1), 177-197.
- Rajeevan, H., Cheung, K. H., Gadagkar, R., Stein, S., Soundararajan, U., Kidd, J. R., et al. (2005). ALFRED: an allele frequency database for microevolutionary studies. *Evol Bioinform Online*, 1, 1-10.
- Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K. E., Rasmussen, S., Albrechtsen, A., et al. (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science*, 334(6052), 94-98.
- Ray, D. A., Xing, J., Hedges, D. J., Hall, M. A., Laborde, M. E., Anders, B. A., et al. (2005). Alu insertion loci and platyrrhine primate phylogeny. *Mol Phylogenet Evol*, 35(1), 117-126.
- Razafindrazaka, H., Ricaut, F. X., Cox, M. P., Mormina, M., Dugoujon, J. M., Randriamarolaza, L. P., et al. (2010). Complete mitochondrial DNA sequences provide new insights into the Polynesian motif and the peopling of Madagascar. *Eur J Hum Genet*, 18(5), 575-581.
-

-
- Redd, A. J., Takezaki, N., Sherry, S. T., McGarvey, S. T., Sofro, A. S., & Stoneking, M. (1995). Evolutionary history of the COII/tRNA^{Lys} intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Mol Biol Evol*, 12(4), 604-615.
- Reed, D. L., Smith, V. S., Hammond, S. L., Rogers, A. R., & Clayton, D. H. (2004). Genetic analysis of lice supports direct contact between modern and archaic humans. *PLoS Biol*, 2(11), e340.
- Regueiro, M., Carril, J. C., Caeiro, B., Pontes, M. L., Abrantes, D., Lima, G., et al. (2004). A study of STR loci (D18S51, FGA, TH01, TPOX) in sub-Saharan populations. *Int Congr Ser*, 1261(0), 130-132.
- Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327), 1053-1060.
- Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M. R., Pugach, I., et al. (2011). Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet*, 89(4), 516-528.
- Richards, M., Bandelt, H.-J., Kivisild, T., & Oppenheimer, S. (2006). A model for the dispersal of modern humans out of Africa *Human mitochondrial DNA and the evolution of Homo sapiens* (pp. 225-265): Springer.
- Rosenberg, N. A. (2003). DISTRUCT: a program for the graphical display of population structure. *Mol Ecol Notes*, 4(1), 137-138.
- Rosenberg, N. A., Burke, T., Elo, K., Feldman, M. W., Freidlin, P. J., Groenen, M. A. M., et al. (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics*, 159(2), 699-713.
- Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822), 928-933.
- Santos, F. R., Pandya, A., Tyler-Smith, C., Pena, S. D., Schanfield, M., Leonard, W. R., et al. (1999). The central Siberian origin for native American Y chromosomes. *Am J Hum Genet*, 64(2), 619-628.
-

-
- Santovito, A., Cervella, P., Selvaggi, A., Caviglia, G. P., Burgarello, C., Sella, G., et al. (2008). DAT1 VNTR polymorphisms in a European and an African population: identification of a new allele. *Hum biol*, 80(2), 191-198.
- Schurr, T. G., Sukernik, R. I., Starikovskaya, Y. B., & Wallace, D. C. (1999). Mitochondrial DNA variation in Koryaks and Itel'men: population replacement in the Okhotsk Sea-Bering Sea region during the Neolithic. *Am J Phys Anthropol*, 108(1), 1-39.
- Schwerdtner Manez, K., & Ferse, S. C. (2010). The history of Makassan trepang fishing and trade. *PLoS One*, 5(6), e11346.
- Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316(5829), 1341-1345.
- Seah, L. H., Jeevan, N. H., Othman, M. I., Jaya, P., Ooi, Y. S., Wong, P. C., et al. (2003). STR Data for the AmpFISTR Identifier loci in three ethnic groups (Malay, Chinese, Indian) of the Malaysian population. *Forensic Sci Int*, 138(1-3), 134-137.
- Senut, B., Pickford, M., Gommery, D., Mein, P., Cheboi, K., & Coppens, Y. (2001). First hominid from the Miocene (Lukeino Formation, Kenya). *Comptes Rendus De L Academie Des Sciences Serie Ii Fascicule a-Sciences De La Terre Et Des Planetes*, 332(2), 137-144.
- Settin, A., Abdel-Hady, H., El-Baz, R., & Saber, I. (2007). Gene Polymorphisms of TNF- α - 308, IL-10- 1082, IL-6- 174, and IL-1Ra VNTR Related to Susceptibility and Severity of Rheumatic Heart Disease. *Pediatr Cardiol*, 28(5), 363-371.
- Shang, H., Tong, H., Zhang, S., Chen, F., & Trinkaus, E. (2007). An early modern human from Tianyuan Cave, Zhoukoudian, China. *Proc Natl Acad Sci U S A*, 104(16), 6573-6578.
- Shedlock, A. M., & Okada, N. (2000). SINE insertions: powerful tools for molecular systematics. *Bioessays*, 22(2), 148-160.
- Shedlock, A. M., Takahashi, K., & Okada, N. (2004). SINEs of speciation: tracking lineages with retroposons. *Trends Ecol Evol*, 19(10), 545-553.
-

-
- Shen, M. R., Batzer, M. A., & Deininger, P. L. (1991). Evolution of the master Alu gene(s). *J Mol Evol*, 33(4), 311-320.
- Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1), 308-311.
- Simonson, T. S., Xing, J., Barrett, R., Jerah, E., Loa, P., Zhang, Y., et al. (2011). Ancestry of the Iban is predominantly Southeast Asian: genetic evidence from autosomal, mitochondrial, and Y chromosomes. *PLoS One*, 6(1), e16338.
- Simsek, M., Al-Sharbati, M., Al-Adawi, S., Ganguly, S. S., & Lawatia, K. (2005). Association of the risk allele of dopamine transporter gene (DAT1*10) in Omani male children with attention-deficit hyperactivity disorder. *Clin Biochem*, 38(8), 739-742.
- Smith, C. (2005). Genomics: SNPs and human disease. *Nature*, 435(7044), 993.
- Soares, P., Alshamali, F., Pereira, J. B., Fernandes, V., Silva, N. M., Afonso, C., et al. (2012). The Expansion of mtDNA Haplogroup L3 within and out of Africa. *Mol Biol Evol*, 29(3), 915-927.
- Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Rohl, A., et al. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet*, 84(6), 740-759.
- Soares, P., Rito, T., Trejaut, J., Mormina, M., Hill, C., Tinkler-Hundal, E., et al. (2011). Ancient voyaging and Polynesian origins. *Am J Hum Genet*, 88(2), 239-247.
- Soares, P., Trejaut, J. A., Loo, J. H., Hill, C., Mormina, M., Lee, C. L., et al. (2008). Climate change and postglacial human dispersals in southeast Asia. *Mol Biol Evol*, 25(6), 1209-1218.
- Sobrinho, B., Brion, M., & Carracedo, A. (2005). SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Sci Int*, 154(2-3), 181-194.
- Soodyall, H., Vigilant, L., Hill, A. V., Stoneking, M., & Jenkins, T. (1996). mtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in sub-Saharan Africans. *Am J Hum Genet*, 58(3), 595-608.
-

-
- Steinlechner, M., Berger, B., Scheithauer, R., & Parson, W. (2001). Population genetics of ten STR loci (AmpFI STR SGM plus) in Austria. *Int J Legal Med*, 114(4-5), 288-290.
- Stoneking, M. (2000). Hypervariable sites in the mtDNA control region are mutational hotspots. *Am J Hum Genet*, 67(4), 1029-1032.
- Stoneking, M. (2008). Human origins. The molecular perspective. *EMBO Rep*, 9 Suppl 1, S46-50.
- Stoneking, M., & Delfin, F. (2010). The human genetic history of East Asia: weaving a complex tapestry. *Curr Biol*, 20(4), R188-193.
- Stringer, C. B., & Andrews, P. (1988). Genetic and fossil evidence for the origin of modern humans. *Science*, 239(4845), 1263-1268.
- Su, B., Xiao, C., Deka, R., Seielstad, M. T., Kangwanpong, D., Xiao, J., et al. (2000). Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum Genet*, 107(6), 582-590.
- Su, Z., Gay, L. J., Strange, A., Palles, C., Band, G., Whiteman, D. C., et al. (2012). Common variants at the MHC locus and at chromosome 16q24.1 predispose to Barrett's esophagus. *Nat Genet*, 44(10), 1131-1136.
- Suadi, Z., Siew, L. C., Tie, R., Hui, W. B., Asam, A., Thiew, S. H., et al. (2007). STR data for the AmpFI STR Identifier loci from the three main ethnic indigenous population groups (Iban, Bidayuh, and Melanau) in Sarawak, Malaysia. *J Forensic Sci*, 52(1), 231-234.
- Sutovsky, P., Navara, C. S., & Schatten, G. (1996). Fate of the sperm mitochondria, and the incorporation, conversion, and disassembly of the sperm tail structures during bovine fertilization. *Biol Reprod*, 55(6), 1195-1205.
- Sykes, B., Leibo, A., Low-Beer, J., Tetzner, S., & Richards, M. (1995). The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. *Am J Hum Genet*, 57(6), 1463-1475.
- Taanman, J. W. (1999). The mitochondrial genome: structure, transcription, translation and replication. *Biochim Biophys Acta*, 1410(2), 103-123.
-

-
- Tabbada, K. A., Trejaut, J., Loo, J. H., Chen, Y. M., Lin, M., Mirazon-Lahr, M., et al. (2010). Philippine mitochondrial DNA diversity: a populated viaduct between Taiwan and Indonesia? *Mol Biol Evol*, 27(1), 21-31.
- Tajima, A., Hayami, M., Tokunaga, K., Juji, T., Matsuo, M., Marzuki, S., et al. (2004). Genetic origins of the Ainu inferred from combined DNA analyses of maternal and paternal lineages. *J Hum Genet*, 49(4), 187-193.
- Takezaki, N., Nei, M., & Tamura, K. (2010). POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. *Mol Biol Evol*, 27(4), 747-752.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*, 28(10), 2731-2739.
- Tanaka, M., Cabrera, V. M., Gonzalez, A. M., Larruga, J. M., Takeyasu, T., Fuku, N., et al. (2004). Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res*, 14(10A), 1832-1850.
- Tchinda, J., & Lee, C. (2006). Detecting copy number variation in the human genome using comparative genomic hybridization. *Biotechniques*, 41(4), 385, 387, 389 passim.
- Templeton, A. R. (1993). The “Eve” hypotheses: a genetic critique and reanalysis. *Am Anthropol*, 95(1), 51-72.
- Thangaraj, K., Chaubey, G., Kivisild, T., Reddy, A. G., Singh, V. K., Rasalkar, A. A., et al. (2005). Reconstructing the origin of Andaman Islanders. *Science*, 308(5724), 996.
- Thurgood, G. (1999). *From Ancient Cham to modern dialects: Two thousand years of language contact and change* (Vol. 28): University of Hawaii Press.
- Tiret, L. (2002). Gene-environment interaction: a central concept in multifactorial diseases. *Proc Nutr Soc*, 61(4), 457-463.
- Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., et al. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, 271(5254), 1380-1387.
-

-
- Torroni, A., Achilli, A., Macaulay, V., Richards, M., & Bandelt, H. J. (2006). Harvesting the fruit of the human mtDNA tree. *Trends Genet*, 22(6), 339-345.
- Torroni, A., Petrozzi, M., Santolamazza, P., Sellitto, D., Cruciani, F., & Scozzari, R. (1995). About the "Asian"-specific 9-bp deletion of mtDNA. *Am J Hum Genet*, 57(2), 507-508.
- Trejaut, J. A., Kivisild, T., Loo, J. H., Lee, C. L., He, C. L., Hsu, C. J., et al. (2005). Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. *PLoS Biol*, 3(8), e247.
- Trejaut, J. A., Yen, J. C., Loo, J. H., & Lin, M. (2011). Modern human migrations in insular Asia according to mitochondrial DNA and non-recombining Y chromosome. *ISBT Sci Series*, 6(2), 361-365.
- Trinkaus, E., Moldovan, O., Milota, S., Bilgar, A., Sarcina, L., Athreya, S., et al. (2003). An early modern human from the Pesteră cu Oase, Romania. *Proc Natl Acad Sci U S A*, 100(20), 11231-11236.
- Trivedi, R., Sitalaximi, T., Banerjee, J., Singh, A., Sircar, P. K., & Kashyap, V. K. (2006). Molecular insights into the origins of the Shompen, a declining population of the Nicobar archipelago. *J Hum Genet*, 51(3), 217-226.
- Tumonggor, M. K., Karafet, T. M., Hallmark, B., Lansing, J. S., Sudoyo, H., Hammer, M. F., et al. (2013). The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific. *J Hum Genet*, 58(3), 165-173.
- Turner, C. G., 2nd. (1987). Late Pleistocene and Holocene population history of East Asia based on dental variation. *Am J Phys Anthropol*, 73(3), 305-321.
- Ulluwishewa, R., Roskrige, N., Harmsworth, G., & Antaran, B. (2008). Indigenous knowledge for natural resource management: a comparative study of Māori in New Zealand and Dusun in Brunei Darussalam. *GeoJournal*, 73(4), 271-284.
- Underhill, P. A., & Kivisild, T. (2007). Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu Rev Genet*, 41, 539-564.
- Vali, U., Brandstrom, M., Johansson, M., & Ellegren, H. (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genet*, 9, 8.
-

-
- Vandenbergh, D. J., Persico, A. M., Hawkins, A. L., Griffin, C. A., Li, X., Jabs, E. W., et al. (1992). Human dopamine transporter gene (DAT1) maps to chromosome 5p15.3 and displays a VNTR. *Genomics*, 14(4), 1104-1106.
- Vandenbergh, D. J., Thompson, M. D., Cook, E. H., Bendahhou, E., Nguyen, T., Krasowski, M. D., et al. (2000). Human dopamine transporter gene: coding region conservation among normal, Tourette's disorder, alcohol dependence and attention-deficit hyperactivity disorder populations. *Mol Psychiatry*, 5(3), 283-292.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science*, 291(5507), 1304-1351.
- Wahls, W. P., Wallace, L. J., & Moore, P. D. (1990). Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells. *Cell*, 60(1), 95-103.
- Ward, C., Leakey, M., & Walker, A. (1999). The new hominid species *Australopithecus anamensis*. *Evol Anthropol*, 7(6), 197-205.
- Washburn, S. L. (1964). The Origin of Races: Weidenreich's Opinion. *Am Anthropol*, 66(5), 1165-1167.
- Watkins, W. S., Rogers, A. R., Ostler, C. T., Wooding, S., Bamshad, M. J., Brassington, A. M., et al. (2003). Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res*, 13(7), 1607-1618.
- Watson, E., Forster, P., Richards, M., & Bandelt, H. J. (1997). Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet*, 61(3), 691-704.
- Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., & Marth, G. (2002). Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet*, 71(4), 854-862.
- Wen, B., Li, H., Gao, S., Mao, X., Gao, Y., Li, F., et al. (2005). Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol Biol Evol*, 22(3), 725-734.
- Wendorf, F., Schild, R., Close, A. E., Donahue, D. J., Jull, A. J., Zabel, T. H., et al. (1984). New radiocarbon dates on the cereals from wadi kubbaniya. *Science*, 225(4662), 645-646.
-

-
- Whittle, M. R., Romano, N. L., & Negreiros, V. A. (2004). Updated Brazilian genetic data, together with mutation rates, on 19 STR loci, including D10S1237. *Forensic Sci Int*, 139(2-3), 207-210.
- Wiesner, R. J., Ruegg, J. C., & Morano, I. (1992). Counting target molecules by exponential polymerase chain reaction: copy number of mitochondrial DNA in rat tissues. *Biochem Biophys Res Commun*, 183(2), 553-559.
- Wolpoff, M. H., Hawks, J., & Caspari, R. (2000). Multiregional, not multiple origins. *Am J Phys Anthropol*, 112(1), 129-136.
- Wolpoff, M. H., Spuhler, J. N., Smith, F. H., Radovic, J., Pope, G., Frayer, D. W., et al. (1988). Modern human origins. *Science*, 241(4867), 772-774.
- Wood, B. (2002). Palaeoanthropology: Hominid revelations from Chad. *Nature*, 418(6894), 133-135.
- Wood, B., & Strait, D. (2004). Patterns of resource use in early Homo and Paranthropus. *J Hum Evol*, 46(2), 119-162.
- Xi, T., Jones, I. M., & Mohrenweiser, H. W. (2004). Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics*, 83(6), 970-979.
- Xing, J., Watkins, W. S., Shlien, A., Walker, E., Huff, C. D., Witherspoon, D. J., et al. (2010). Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics*, 96(4), 199-210.
- Xu, S., Pugach, I., Stoneking, M., Kayser, M., & Jin, L. (2012). Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *Proc Natl Acad Sci U S A*, 109(12), 4574-4579.
- Yao, Y. G., Watkins, W. S., & Zhang, Y. P. (2000). Evolutionary history of the mtDNA 9-bp deletion in Chinese populations and its relevance to the peopling of east and southeast Asia. *Hum Genet*, 107(5), 504-512.
- Yong, R. Y., Aw, L. T., & Yap, E. P. (2004). Allele frequencies of 15 STR loci of three main ethnic populations in Singapore using an in-house marker panel. *Forensic Sci Int*, 141(2-3), 175-183.
-

-
- Zhivotovsky, L. A., Rosenberg, N. A., & Feldman, M. W. (2003). Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am J Hum Genet*, 72(5), 1171-1186.
- Zimmermann, B., Rock, A., Huber, G., Kramer, T., Schneider, P. M., & Parson, W. (2011). Application of a west Eurasian-specific filter for quasi-median network analysis: Sharpening the blade for mtDNA error detection. *Forensic Sci Int Genet*, 5(2), 133-137.
- Zuckerkindl, E., & Cavalli, G. (2007). Combinatorial epigenetics, "junk DNA", and the evolution of complex organisms. *Gene*, 390(1-2), 232-242.

APPENDICES

Appendix 1 : Sample of informed consent form for volunteers.

VOLUNTEER PERSONAL INFORMATION FORM

1. Name: 2. Age:
3. Nationality: 4. Ethnic:

5. Family Information:

Father's Side		Mother's Side	
Subject	Ethnic/tribe *	Subject	Ethnic/tribe *
Father		Mother	
Grandfather		Grandfather	
Grandmother		Grandmother	

*Please specify which ethnics

7. Do you have any health problem?
(a) No. I am healthy. (b) Yes (Please specify

I, Identity Card No.
(Name of subject)
of hereby agree to take part in the
study titled:
Genomic and proteomic approaches for the identification of biomarkers of colorectal cancer in a Malaysian-based population.

The nature and purpose of which has been explained in detail to me by Mr/Ms.
I have been informed about the nature of the study in terms of methodology, possible drawbacks, and expected outcomes (as per subject information sheet) and I allow the sample to be analysed in any equipped laboratory. After knowing and understanding all the information provided, I voluntarily consent on my own free will to participate in the study specified above.
I understand that I can withdraw from this study at any time without assigning any reason.



Signature:
(Subject)

IN THE PRESENCE OF
Name:
I.C. No.:
Signature:
(Witness)


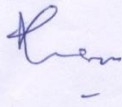
I confirm that I have explained to the subject the nature and purpose of the above-mentioned study.
Signature:
(Attending researcher)

THANK YOU FOR YOUR COOPERATION.

Appendix 2 : Approval letter for ethical clearance (reference number: 770.21).

 UNIVERSITI MALAYA KUALA LUMPUR PUSAT PERUBATAN UM		JAWATANKUASA ETIKA PERUBATAN PUSAT PERUBATAN UNIVERSITI MALAYA ALAMAT: LEMBAH PANTAI, 59100 KUALA LUMPUR, MALAYSIA TELEFON: 03-79494422 samb. 3209 FAKSIMILI: 03-79494638	
NAME OF ETHICS COMMITTEE/IRB: Medical Ethics Committee, University Malaya Medical Centre ADDRESS: LEMBAH PANTAI 59100 KUALA LUMPUR		ETHICS COMMITTEE/IRB REFERENCE NUMBER: 770.21	
PROTOCOL NO: TITLE: Assessment 7 Analysis Of The Genomic Diversity In The HV3, DAT-1, CEL, DYS287, REGION V, D7S820, D13S317 & D16S39 Markers Of East Malaysian Indigenous Population Samples			
PRINCIPAL INVESTIGATOR: Prof. Madya Chua Kek Heng TELEPHONE: KOMTEL:		SPONSOR:	
The following item <input checked="" type="checkbox"/> have been received and reviewed in connection with the above study to be conducted by the above investigator.			
<input checked="" type="checkbox"/> Borang Permohonan Pindaan Penyelidikan <input type="checkbox"/> Study Protocol <input type="checkbox"/> Investigator Brochure <input checked="" type="checkbox"/> Amendment Subject Information Sheet <input type="checkbox"/> Consent Form <input type="checkbox"/> Questionnaire <input type="checkbox"/> Investigator(s) CV's (if applicable)		Ver date: 14 Jan 10 Ver date: Ver date: Ver date: Ver date:	
and have been <input checked="" type="checkbox"/>			
<input checked="" type="checkbox"/> Approved <input type="checkbox"/> Conditionally approved (identify item and specify modification below or in accompanying letter) <input type="checkbox"/> Rejected (identify item and specify reasons below or in accompanying letter)			
Comments:			
Date of approval: 24 th FEBRUARY 2010			
		 PROF. LOOI LAI MENG Chairman Medical Ethics Committee	

Appendix 3 : Approval letter for ethical clearance (reference number: 612.16).

 JAWATANKUASA ETIKA PERUBATAN PUSAT PERUBATAN UNIVERSITI MALAYA ALAMAT: LEMBAH PANTAI, 59100 KUALA LUMPUR, MALAYSIA TELEFON: 03-79494422 FAKSIMILI: 03-79545682															
NAME OF ETHICS COMMITTEE/IRB: Medical Ethics Committee, University Malaysia Medical Centre ADDRESS: LEMBAH PANTAI 59100 KUALA LUMPUR	ETHICS COMMITTEE/IRB REFERENCE NUMBER: 612.16														
PROTOCOL NO: TITLE: Assessment and Analysis of the Genomic Diversity in the HV3, DAT-1, CEL, DYS287, region V, D7S820, D13S317, and D16S39 markers of East Malaysian indigenous population samples															
PRINCIPAL INVESTIGATOR: Dr Chua Kek Heng TELEPHONE:	SPONSOR: KOMTEL:														
<p>The following item <input checked="" type="checkbox"/> have been received and reviewed in connection with the above study to be conducted by the above investigator.</p> <table border="0"> <tr> <td><input checked="" type="checkbox"/> Borang Permohonan Penyelidikan</td> <td>Ver date: 11 Oct 2007</td> </tr> <tr> <td><input type="checkbox"/> Study Protocol</td> <td>Ver date:</td> </tr> <tr> <td><input type="checkbox"/> Investigator Brochure</td> <td>Ver date:</td> </tr> <tr> <td><input checked="" type="checkbox"/> Patient Information Sheet</td> <td>Ver date:</td> </tr> <tr> <td><input checked="" type="checkbox"/> Consent Form</td> <td>Ver date:</td> </tr> <tr> <td><input type="checkbox"/> Questionnaire</td> <td>Ver date:</td> </tr> <tr> <td><input checked="" type="checkbox"/> Investigator(s) CV's (Dr Chua Kek Heng)</td> <td>Ver date:</td> </tr> </table> <p>and have been <input checked="" type="checkbox"/></p> <p><input checked="" type="checkbox"/> Approved <input type="checkbox"/> Conditionally approved (identify item and specify modification below or in accompanying letter) <input type="checkbox"/> Rejected (identify item and specify reasons below or in accompanying letter)</p> <p>Comments:</p> <p>i. Investigator is required to follow instructions, guidelines and requirements of the Medical Ethics Committee.</p> <p>ii. Investigator is required to report any protocol deviations/violations through the Clinical Investigation Centre and provide annual/closure reports to the Medical Ethics Committee.</p> <p>Date of approval: 31st October 2007</p> <p>s.k Ketua Jabatan Perubatan Molekul Timbalan Dekan (Penyelidikan) Fakulti Perubatan, Universiti Malaya Setiausaha Jawatankuasa Penyelidikan Pusat Perubatan Fakulti Perubatan, Universiti Malaya</p> <div style="text-align: right;">  PROF. LOOI LAI MENG Chairman Medical Ethics Committee </div>		<input checked="" type="checkbox"/> Borang Permohonan Penyelidikan	Ver date: 11 Oct 2007	<input type="checkbox"/> Study Protocol	Ver date:	<input type="checkbox"/> Investigator Brochure	Ver date:	<input checked="" type="checkbox"/> Patient Information Sheet	Ver date:	<input checked="" type="checkbox"/> Consent Form	Ver date:	<input type="checkbox"/> Questionnaire	Ver date:	<input checked="" type="checkbox"/> Investigator(s) CV's (Dr Chua Kek Heng)	Ver date:
<input checked="" type="checkbox"/> Borang Permohonan Penyelidikan	Ver date: 11 Oct 2007														
<input type="checkbox"/> Study Protocol	Ver date:														
<input type="checkbox"/> Investigator Brochure	Ver date:														
<input checked="" type="checkbox"/> Patient Information Sheet	Ver date:														
<input checked="" type="checkbox"/> Consent Form	Ver date:														
<input type="checkbox"/> Questionnaire	Ver date:														
<input checked="" type="checkbox"/> Investigator(s) CV's (Dr Chua Kek Heng)	Ver date:														

Appendix 4 : Frequencies of *Alu* insertions in various populations used for the construction of PCA and NJ phylogenetic trees.

Population	<i>Alu</i> insertion frequency				
	PV92	TPA25	APO	B65	HS4.32
Africa					
Alur	0.1250	0.2500	0.7083	0.7500	0.5000
Biaka Pygmy	0.3000	0.0000	0.7500	0.5000	0.2000
Hema	0.3333	0.3056	0.7778	0.5000	0.6471
Mbuti Pygmy	0.6000	0.0000	0.8000	0.8000	0.2000
Nande	0.2667	0.2353	0.7500	0.6765	0.3667
Nguni	0.2500	0.0833	0.5714	0.5417	0.2273
San	0.3000	0.2000	0.8214	0.6538	0.3214
Sotho/Tswana	0.3235	0.2857	0.7750	0.2647	0.1333
Tsonga	0.3077	0.2083	0.6818	0.5000	0.2308
Zaire Pygmy	0.3750	0.1212	0.9394	0.6094	0.3871
Overall	0.3178	0.1931	0.7823	0.5679	0.3504
Asia					
Cambodian	1.0000	0.5417	0.7917	0.4167	0.5455
Chinese	0.8529	0.4412	0.8824	0.4706	0.4375
Japanese	0.8571	0.5000	0.8438	0.4118	0.4375
Malay	0.5000	0.2500	0.8333	0.5000	0.5000
Vietnamese	0.8750	0.2778	0.9444	0.4444	0.5556
Overall	0.8571	0.3973	0.8562	0.4333	0.4722
Europe					
Finnish	0.1563	0.4444	0.9737	0.4211	0.5789
French	0.2750	0.7250	0.9500	0.6250	0.7000
Northern European	0.2537	0.5522	0.9697	0.6186	0.5859
Polish	0.1250	0.7500	0.9500	0.3500	0.7000
Overall	0.2342	0.5826	0.9652	0.5602	0.6150
India					
Brahmin	0.4083	0.5185	0.8167	0.4831	0.5690
Irulaa	0.4853	0.7879	0.8088	0.4394	0.7647
Kapu	0.4828	0.4818	0.8103	0.4052	0.5439
Kondadoraa	0.4038	0.7115	0.8704	0.5370	0.4630
Kshatriya	0.1364	0.4545	0.7778	0.3182	0.7273
Madiga	0.4483	0.6552	0.8571	0.3103	0.4655
Mala	0.5000	0.6154	0.8077	0.4231	0.5000
Maria Gonda	0.3864	0.5714	0.6591	0.6364	0.5238
Relli	0.6316	0.5263	0.7895	0.5789	0.4211
Santala	0.6250	0.6538	0.7308	0.3750	0.5417
Vysya	0.4444	0.6500	0.8000	0.6000	0.5000
Yadava	0.5472	0.6132	0.8774	0.4717	0.5000
Overall	0.4662	0.5957	0.8148	0.4596	0.5435
Average (all populations)	0.4398	0.4868	0.8372	0.4949	0.5088

Appendix 5 : Genetic distance (D_A) of 48 populations generated from allelic frequencies of 13 STR loci.

[illegible]

Appendix 6 : Genetic distance (D_A) of 40 populations generated from allelic frequencies of 12 STR loci.

[illegible]

Appendix 7 : Mt polymorphisms of 150 Kadazan-Dusun individuals and their determined haplogroups, presented in the EMPOP database format.

[illegible]

Appendix 7 : continuation.

[illegible]

Appendix 8 : Mt polymorphisms of 150 Bajau individuals and their determined haplogroups, presented in the EMPOP database format.

Mitochondrial DNA data of 150 Bafau individuals; Kee Boon Pin (bpkee99@yahoo.com)
150 individuals from Bafau tribe (indigenous population), state of Sabah, East Malaysia.
#16024-16365 73-308 439-576 8291-8289

mtHV-01	B4c1b2a2	1	16104C	16181C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	315 C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL		
mtHV-02	R9b2	1	16104C	16182C	73G	263G	309 C	315 C	151C	523DEL	524DEL																
mtHV-03	B4c2	1	16147T	16183C	16184A	16189C	16217C	16235G	16293G	73G	263G	315 C	151C	6281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL					
mtHV-04	B4c1b2a2	1	16104C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	315 C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL			
mtHV-05	M43b	1	16092C	16218T	16311C	16319A	16357C	73G	152C	199C	263G	309 C	315 C	489C													
mtHV-06	R9c1a	1	16157C	16256T	16304C	16335G	73G	240DEL	263G	309 C	309 C	315 C															
mtHV-07	M2D	1	16093C	16212A	16209C	16221T	16273G	73G	152C	225A	240DEL	263G	309 C	315 C	316A	489C	523DEL	524DEL									
mtHV-08	F1a	1	16157C	16261T	16291T	16304C	16335G	73G	240DEL	263G	309 C	315 C															
mtHV-09	B4a1a	1	16182C	16183C	16189C	16217C	16261T	73G	146C	263G	315 C	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL					
mtHV-10	R9c1a	1	16157C	16256T	16304C	16335G	73G	240DEL	263G	309 C	309 C	315 C															
mtHV-11	F1a	1	16157C	16261T	16291T	16304C	16335G	73G	240DEL	263G	309 C	315 C															
mtHV-12	B4b1a+207	1	16129A	16190C	16182C	16183C	16189C	16217C	16292T	73G	207A	263G	309 C	315 C	499A	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL	
mtHV-13	B4a1a	1	16129A	16190C	16182C	16183C	16189C	16217C	16292T	73G	146C	263G	309 C	315 C	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL		
mtHV-14	R9c1a	1	16157C	16256T	16304C	16335G	73G	240DEL	263G	309 C	309 C	315 C															
mtHV-15	M7c2b	1	16182C	16183C	16189C	16217C	16261T	73G	146C	263G	309 C	315 C	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL				
mtHV-16	B4a1a	1	16129A	16190C	16182C	16183C	16189C	16217C	16292T	73G	146C	263G	309 C	315 C	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL		
mtHV-17	N5	1	16111T	16168T	16172C	16183C	16189C	16223T	16242T	16263C	16311C	16362G	73G	152C	263G	315 C											
mtHV-18	R9c1a	1	16157C	16256T	16304C	16335G	73G	240DEL	263G	309 C	309 C	315 C															
mtHV-19	B4c1b2a2	1	16104C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	196C	263G	309 C	309 C	315 C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL
mtHV-20	M9	1	16223T	16362G	73G	263G	309 C	315 C	151C	489C																	
mtHV-21	B4c2	1	16147T	16183C	16184A	16189C	16217C	16235G	73G	263G	309 C	315 C	6281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL						
mtHV-22	M2D	1	16093C	16212A	16209C	16221T	16273G	73G	152C	225A	240DEL	263G	309 C	315 C	316A	489C	523DEL	524DEL									
mtHV-23	B4b1a+207	1	16129A	16190C	16182C	16183C	16189C	16217C	16292T	73G	207A	263G	315 C	499A	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL		
mtHV-24	N8	1	16124C	16223T	16263C	16274A	16311C	16343G	16357C	73G	152C	263G	309 C	309 C	315 C												
mtHV-25	B4b	1	16129A	16182C	16183C	16189C	16217C	16261T	73G	152C	309 C	309 C	315 C	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL			
mtHV-26	R9c2b	1	16126C	16223T	16261T	16362G	73G	146C	199C	263G	309 C	315 C	489C	263G	315 C	291DEL	292DEL	293DEL	294DEL	315 C	489C						
mtHV-27	M7c3c	1	16223T	16295T	16362G	73G	146C	199C	263G	309 C	315 C	489C	523DEL	524DEL													
mtHV-28	F3b1	1	16220C	16295C	16298T	16311C	16356T	16362G	73G	150T	152C	207A	240DEL	263G	315 C												
mtHV-29	M7c3c	1	16223T	16278T	16295T	16362G	73G	146C	199C	263G	309 C	315 C	489C	523DEL	524DEL												
mtHV-30	R9b2	1	16104C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	309 C	309 C	315 C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL	
mtHV-31	B4c1b2a2	1	16104C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	309 C	309 C	315 C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL	
mtHV-32	F3b1	1	16220C	16295C	16298T	16311C	16356T	16362G	73G	150T	152C	204C	274A	240DEL	263G	315 C											
mtHV-33	R9b2	1	16104C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	309 C	309 C	315 C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL	
mtHV-34	N5	1	16111T	16168T	16172C	16183C	16189C	16223T	16242T	16263C	16311C	16362G	73G	152C	263G	315 C											
mtHV-35	M7b1245678+16192	1	16126C	1618A	16192T	16223T	16297C	16362G	73G	150T	199C	207A	263G	309 C	315 C	318C	489C										
mtHV-36	V2	1	16126C	16129A	16231C	16311C	73G	146C	263G	309 C	315 C	489C															
mtHV-37	F1a1	1	16104C	16182C	16183C	16189C	16217C	16261T	73G	146C	263G	309 C	315 C	523DEL	524DEL												
mtHV-38	E2	1	16051G	16223T	16362G	73G	146C	195C	263G	309 C	315 C	489C															
mtHV-39	R9b2	1	16104C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	309 C	309 C	315 C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL	
mtHV-40	F4a1	1	16172C	16129A	16231C	16311C	16356T	16362G	73G	150T	152C	204C	274A	240DEL	263G	315 C											
mtHV-41	N8	1	16129A	16223T	16263C	16274A	16311C	16343G	16357C	73G	234G	263G	309 C	309 C	315 C												
mtHV-42	R9c1a	1	16157C	16256T	16304C	16335G	73G	240DEL	263G	309 C	309 C	315 C															
mtHV-43	R22	1	16249C	16268C	16304C	73G	152C	263G	315 C	329A	523DEL	524DEL															
mtHV-44	F1a	1	16104C	16182C	16183C	16189C	16217C	16261T	73G	146C	263G	309 C	315 C	523DEL	524DEL												
mtHV-45	F1a3a	1	16129A	16172C	16234T	16304C	16311C	73G	240DEL	263G	309 C	315 C	523DEL	524DEL													
mtHV-46	B4c1b2a2	1	16104C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	196C	263G	315 C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL		
mtHV-47	M2c1c	1	16181T	16311C	16319A	16357C	73G	152C	199C	263G	309 C	315 C	489C														
mtHV-48	F4c3c	1	16104C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	309 C	315 C	489C										
mtHV-49	R14	1	16181T	16218C	16288C	16304G	16360T	73G	207A	234G	263G	309 C	309 C	315 C	573 C	489C	523DEL	524DEL	573 C	573 C	573 C	573 C					
mtHV-50	F1a	1	16129A	16172C	16304C	73G	240DEL	263G	315 C	523DEL	524DEL																
mtHV-51	M9	1	16181T	16223T	16362G	73G	263G	309 C	315 C	489C																	
mtHV-52	R9c	1	16051G	16304C	73G	199C	209C	234G	240DEL	263G	309 C	315 C	523DEL	524DEL													
mtHV-53	F1a	1	16129A	16172C	16304C	73G	240DEL	263G	315 C	523DEL	524DEL																
mtHV-54	B4a1a	1	16182C	16183C	16189C	16217C	16261T	73G	146C	263G	309 C	309 C	315 C	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL			
mtHV-55	F1a	1	16157C	16261T	16291T	16304C	16335G	73G	240DEL	263G	309 C	309 C	315 C	523DEL	524DEL												
mtHV-56	F1a	1	16129A	16172C	16304C	73G	240DEL	263G	315 C	523DEL	524DEL																
mtHV-57	F1a	1	16129A	16172C	16304C	73G	240DEL	263G	315 C	523DEL	524DEL																
mtHV-58	F3b1	1	16129A	16172C	16304C	73G	240DEL	263G	315 C	523DEL	524DEL																
mtHV-59	F3b1	1	16181T	16220C	16295C	16298T	16362G	73G	150T	152C	240DEL	263G	315 C	489C													

Appendix 8 : continuation.

mtHV-B86	B4c1b2a2	1	16140C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	315.1C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL		
mtHV-B87	B4c1b2a2	1	16140C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	309.1C	309.2C	315.1C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL
mtHV-B88	M20	1	16093C	16129A	16209C	16223T	16272G	73G	152C	225A	249DEL	263G	309.1C	315.1C	316A	489C	523DEL	524DEL								
mtHV-B89	M7c3c	1	16223T	16278T	16295T	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL											
mtHV-B90	M7c3c	1	16223T	16278T	16295T	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL											
mtHV-B91	B4c1b2a2	1	16140C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	315.1C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL		
mtHV-B92	B4a1a	1	16182C	16183C	16189C	16217C	16281T	73G	146C	263G	309.1C	315.1C	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL			
mtHV-B93	M7c3c	1	16223T	16278T	16295T	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL											
mtHV-B94	B4c1b2a2	1	16140C	16182C	16183C	16189C	16217C	16274A	16335G	73G	146C	150T	195C	263G	315.1C	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL		
mtHV-B95	M7c3c	1	16223T	16295T	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL												
mtHV-B96	E1a1a	1	16185T	16223T	16278T	16291T	16362C	73G	263G	309.1C	315.1C	489C														
mtHV-B97	F3b1	1	16188T	16220C	16265G	16298C	16362C	73G	150T	152C	249DEL	263G	315.1C													
mtHV-B98	F1a	1	16129A	16172C	16304C	73G	249DEL	263G	315.1C	523DEL																
mtHV-B99	E1a1a	1	16223T	16291T	16362C	73G	263G	309.1C	315.1C	489C																
mtHV-B100	E2	1	16051G	16223T	16362C	73G	195C	263G	315.1C	489C																
mtHV-B101	B4b1a+207	1	16129A	16136C	16182C	16183C	16189C	16217C	16292T	73G	207A	263G	315.1C	499A	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL	
mtHV-B102	E1a1a	1	16185T	16223T	16291T	16362C	73G	263G	309.1C	309.2C	315.1C	489C	524.1A	524.2C												
mtHV-B103	R14	1	16187T	16271C	16288C	16304G	73G	207A	234G	263G	310C	573.1C	573.2C	573.3C	573.4C	573.5A	573.6C									
mtHV-B104	B4b1a+207	1	16129A	16136C	16182C	16183C	16189C	16217C	16292T	73G	207A	263G	315.1C	499A	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL	
mtHV-B105	B4a2b	1	16182C	16183C	16189C	16217C	16281T	73G	146C	152C	263G	309.1C	315.1C	489C	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL	
mtHV-B106	E1a1a	1	16183C	16189C	16223T	16291T	16362C	73G	263G	309.1C	315.1C	489C														
mtHV-B107	M7c3c	1	16223T	16295T	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL												
mtHV-B108	M9	1	16186T	16223T	16362C	73G	263G	309.1C	315.1C	489C																
mtHV-B109	M74b1	1	16223T	16246T	16311C	16362C	73G	195C	263G	309.1C	315.1C	489C														
mtHV-B110	B4b1a+207	1	16136C	16183C	16189C	16217C	73G	204C	207A	263G	309.1C	309.2C	315.1C	499A	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL			
mtHV-B111	B5b	1	16140C	16183C	16189C	16243C	73G	103A	152C	263G	315.1C	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL				
mtHV-B112	B4a1a	1	16182C	16183C	16189C	16217C	16281T	73G	146C	152C	263G	309.1C	309.2C	315.1C	524.1A	524.2C										
mtHV-B113	Y2	1	16128C	16231C	16311C	73G	263G	309.1C	309.2C	315.1C	482C															
mtHV-B114	M7b3	1	16086C	16129A	16297C	16324C	73G	199C	263G	315.1C	489C															
mtHV-S1	E1b+16261	1	16223T	16261T	16362C	73G	152C	263G	309.1C	309.2C	315.1C	489C														
mtHV-S2	M44	1	16223T	16301T	16362C	73G	152C	263G	309.1C	309.2C	315.1C	489C														
mtHV-S3	R9c1a	1	16157C	16256T	16304C	16311C	16335G	73G	249DEL	263G	309.1C	315.1C														
mtHV-S4	M51	1	16223T	16278T	73G	150T	152C	263G	315.1C	489C	523DEL	524DEL														
mtHV-S5	M20	1	16093C	16129A	16209C	16223T	16272G	73G	152C	225A	249DEL	263G	309.1C	315.1C	316A	489C	523DEL	524DEL								
mtHV-S78	M43b	1	16092C	16218T	16311C	16319A	16357C	73G	152C	199C	263G	309.1C	315.1C	489C												
mtHV-S84	D6	1	16189C	16223T	16311C	16362C	73G	263G	315.1C	489C																
mtHV-S96	M7c3c	1	16223T	16295T	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL												
mtHV-S97	D5b	1	16189C	16223T	16362C	73G	150T	248G	309.1C	309.2C	315.1C	456T	489C													
mtHV-S98	Y2	1	16126C	16231C	16311C	73G	263G	309.1C	315.1C	482C																
mtHV-S99	F3b1	1	16220C	16258C	16265G	16298C	16362C	73G	150T	152C	249DEL	263G	315.1C													
mtHV-S100	E1a1a	1	16185T	16223T	16291T	16362C	73G	263G	309.1C	315.1C	489C															
mtHV-S101	F3b1	1	16220C	16258C	16265G	16298C	16362C	73G	150T	152C	249DEL	263G	315.1C													
mtHV-S102	M7b1'2'4'5'6'7'8+(16192)	1	16126C	16129A	16192T	16223T	16297C	73G	150T	199C	263G	309.1C	309.2C	315.1C	489C											
mtHV-S103	B4c2	1	16147T	16172C	16183C	16184A	16189C	16217C	16235G	73G	263G	309.1C	315.1C													
mtHV-S104	F3b1	1	16220C	16265G	16298C	16362C	73G	150T	152C	249DEL	263G	315.1C														
mtHV-S105	B4c2	1	16147T	16183C	16184A	16189C	16217C	16235G	73G	263G	309.1C	315.1C														
mtHV-S106	F3b	1	16093C	16220C	16298C	16311C	16362C	73G	150T	152C	249DEL	263G	309.1C	315.1C												
mtHV-S107	R9c1a	1	16157C	16256T	16304C	16311C	16335G	73G	249DEL	263G	309.1C	315.1C														
mtHV-S108	R9c1a	1	16157C	16256T	16291T	16304C	16335G	73G	249DEL	263G	315.1C															
mtHV-S109	Q1	1	16129A	16144C	16148T	16223T	16241G	16265C	16311C	16343G	73G	89C	146C	263G	309.1C	315.1C	489C									
mtHV-S110	B4a1a	1	16182C	16183C	16189C	16217C	16261T	73G	146C	263G	309.1C	315.1C	523DEL	524DEL	8281DEL	8282DEL	8283DEL	8284DEL	8285DEL	8286DEL	8287DEL	8288DEL	8289DEL			
mtHV-S111	E1a1a	1	16223T	16291T	16362C	73G	263G	309.1C	315.1C	489C																
mtHV-S112	N8	1	16223T	16263C	16274A	16311C	16343G	16357C	73G	152C	263G	309.1C	315.1C	450C												
mtHV-S114	E1a1a	1	16185T	16223T	16291T	16362C	73G	263G	315.1C	489C																
mtHV-S115	JT	1	16126C	16129A	16183C	16189C	16278T	73G	263G	291.1A	309.1C	309.2C	315.1C	489C	560T											
mtHV-S116	M7c3c	1	16223T	16278T	16295T	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL											
mtHV-S117	R9b1a1a	1	16192T	16234T	16288C	16304C	16309G	73G	143A	163G	263G	309.1C	315.1C	523DEL	524DEL	573.1C										
mtHV-S118	R9c1a	1	16157C	16256T	16304C	16335G	73G	249DEL	263G	309.1C	309.2C	315.1C														
mtHV-S119	R9c1a	1	16157C	16256T	16304C	16335G	73G	195C	249DEL	263G	309.1C	309.2C	315.1C													
mtHV-S121	F3b1	1	16220C	16265G	16298C	16362C	73G	150T	152C	249DEL	263G	309.1C	315.1C													
mtHV-S122	M7c3c	1	16223T	16278T	16295T	16311C	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL										
mtHV-S123	E1a1a	1	16223T	16291T	16362C																					

Appendix 9 : Mt polymorphisms of 150 Rungus individuals and their determined haplogroups, presented in the EMPOP database format.

# Mitochondrial DNA data of 150 Rungus individuals; Kee Boon Pin (bpkee99@yahoo.com)		# 150 individuals from Rungus tribe (indigenous population), state of Sabah, East Malaysia.	
# 150 individuals from Rungus tribe (indigenous population), state of Sabah, East Malaysia.		# 150 individuals from Rungus tribe (indigenous population), state of Sabah, East Malaysia.	
mHV-R1	M7c3c	1	16223T 16295T 16346C 16362C 73G 146C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R2	D4b1	1	16223T 16319A 16362C 73G 263G 309.1C 315.1C 489C
mHV-R3	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 315.1C 456T 489C 523DEL 524DEL
mHV-R4	M7c3c	1	16093C 16223T 16295T 16311C 16337T 16362C 73G 146C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R5	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 309.2C 315.1C 456T 489C 523DEL 524DEL
mHV-R6	E1b	1	16126C 16223T 16255A 16257T 16259T 16278T 16362C 73G 152C 178G 263G 291DEL 292DEL 293DEL 294DEL 315.1C 489C 523DEL 524DEL
mHV-R7	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16209C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 315.1C 456T 489C
mHV-R8	E1b+16261	1	16187T 16223T 16261T 16319A 16362C 73G 152C 263G 309.1C 315.1C 489C
mHV-R10	M7c3c	1	16223T 16295T 16346C 16362C 73G 146C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R11	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 315.1C 489C
mHV-R12	M7a4	1	16223T 16297C 73G 150T 199C 204C 263G 271T 309.1C 315.1C 489C
mHV-R13	E1b+16261	1	16223T 16261T 16362C 73G 152C 263G 309.1C 315.1C 489C
mHV-R14	E1a1a	1	16185T 16223T 16291T 16362C 73G 263G 309.1C 309.2C 315.1C 489C
mHV-R15	M33a1b	1	16144C 16223T 73G 146C 199C 235G 263G 309.1C 315.1C 334C 489C 523DEL 524DEL
mHV-R16	E2	1	16051G 16223T 16362C 73G 146C 195C 200G 263G 309.1C 315.1C 489C
mHV-R17	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 309.2C 315.1C 456T 489C 523DEL 524DEL
mHV-R18	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16209C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 309.2C 315.1C 456T 489C 523DEL 524DEL
mHV-R19	M7b1245678+16192	1	16126C 16129A 16192T 16223T 16297C 16362C 73G 150T 199C 263G 309.1C 315.1C 489C
mHV-R20	E1a1a	1	16223T 16291T 16311C 16362C 73G 146C 263G 309.1C 309.2C 315.1C 489C
mHV-R21	E1b+16261	1	16187T 16223T 16261T 16319A 16362C 73G 152C 263G 309.1C 315.1C 489C
mHV-R22	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 309.2C 315.1C 489C
mHV-R23	N5	1	16111T 16188T 16172C 16183C 16189C 16223T 16263C 16268T 16311C 16362C 73G 152C 263G 309.1C 309.2C 315.1C
mHV-R24	E1b+16261	1	16223T 16261T 16278T 16319A 16362C 73G 152C 263G 309.1C 315.1C 489C
mHV-R25	M7c3c	1	16223T 16295T 16346C 16362C 73G 146C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R26	E2	1	16051G 16223T 16362C 73G 146C 195C 200G 263G 309.1C 309.2C 315.1C 489C
mHV-R27	N5	1	16111T 16188T 16172C 16183C 16189C 16223T 16263C 16268T 16311C 16362C 73G 152C 263G 309.1C 309.2C 315.1C
mHV-R28	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 309.2C 315.1C 489C
mHV-R29	JT	1	16126C 16129A 16183C 16189C 16278T 73G 263G 291.1A 309.1C 309.2C 315.1C 489C 560T
mHV-R30	B4a1a	1	16182C 16183C 16189C 16217C 16223T 16261T 73G 146C 263G 309.1C 309.2C 315.1C 523DEL 524DEL 8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL
mHV-R31	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 309.2C 315.1C 456T 489C 523DEL 524DEL
mHV-R32	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 309.2C 315.1C 456T 489C 523DEL 524DEL
mHV-R33	M7c3c	1	16223T 16295T 16346C 16362C 73G 146C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R34	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 315.1C 489C
mHV-R35	R9c1a	1	16157C 16296T 16304C 16335G 73G 249DEL 263G 309.1C 315.1C
mHV-R36	M7c3c	1	16093C 16223T 16295T 16337T 16362C 73G 146C 199C 263G 309.1C 315.1C 489C
mHV-R37	N5	1	16111T 16188T 16172C 16183C 16189C 16223T 16263C 16268T 16311C 16362C 73G 152C 235G 263G 309.1C 315.1C 489C
mHV-R38	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 315.1C 456T 489C 523DEL 524DEL
mHV-R39	B4a1a	1	16179T 16182C 16183C 16189C 16217C 16261T 73G 146C 263G 315.1C 523DEL 524DEL 8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL
mHV-R40	M7c3c	1	16223T 16295T 16346C 16362C 73G 146C 199C 263G 309.1C 315.1C 489C
mHV-R41	R9c1a	1	16157C 16296T 16304C 16335G 73G 249DEL 263G 309.1C 315.1C
mHV-R42	E1a1a	1	16185T 16223T 16291T 16362C 73G 236C 263G 309.1C 315.1C 489C
mHV-R43	B4a1a	1	16182C 16183C 16189C 16217C 16223T 16261T 73G 146C 263G 315.1C 523DEL 524DEL 8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL
mHV-R44	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 315.1C 456T 489C 523DEL 524DEL
mHV-R45	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 309.2C 315.1C 489C
mHV-R46	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 309.2C 315.1C 489C
mHV-R47	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 309.2C 315.1C 489C
mHV-R48	M7b1245678+16192	1	16126C 16129A 16192T 16223T 16297C 16362C 73G 150T 199C 263G 309.1C 315.1C 489C
mHV-R49	B4a1a	1	16182C 16183C 16189C 16217C 16223T 16261T 73G 146C 263G 309.1C 309.2C 315.1C 523DEL 524DEL 8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL
mHV-R50	M7c3c	1	16093C 16223T 16295T 16337T 16362C 73G 146C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R51	M7c3c	1	16093C 16223T 16295T 16337T 16362C 73G 146C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R52	N21+195	1	16193T 16223T 73G 150T 195C 263G 309.1C 315.1C 337DEL
mHV-R53	M7c3c	1	16223T 16295T 16346C 16362C 73G 146C 199C 263G 309.1C 315.1C 489C
mHV-R54	E1a1a	1	16189C 16223T 16291T 16362C 73G 263G 309.1C 315.1C 489C
mHV-R55	E1b	1	16126C 16223T 16255A 16257T 16259T 16278T 16362C 73G 152C 178G 263G 291DEL 292DEL 293DEL 294DEL 309.1C 315.1C 489C
mHV-R56	E1b+16261	1	16187T 16223T 16261T 16319A 16362C 73G 152C 263G 309.1C 315.1C 489C
mHV-R57	M7c3c	1	16223T 16295T 16346C 16362C 73G 146C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R58	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 315.1C 489C
mHV-R59	B4a1a	1	16182C 16183C 16189C 16217C 16223T 16261T 73G 146C 263G 315.1C 523DEL 524DEL 8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL
mHV-R60	E1a1a	1	16223T 16291T 16311C 16362C 73G 146C 263G 309.1C 315.1C 489C
mHV-R61	E1a1a	1	16223T 16291T 16311C 16362C 73G 263G 309.1C 315.1C 489C
mHV-R62	M8b	1	16182C 16183C 16189C 16223T 16336C 73G 150T 152C 185A 263G 310C 313DEL 314DEL 315DEL 489C 523DEL 524DEL
mHV-R63	E1b+16261	1	16187T 16223T 16261T 16319A 16362C 73G 152C 263G 315.1C 489C
mHV-R64	R9c1a	1	16157C 16296T 16304C 16335G 73G 249DEL 263G 315.1C
mHV-R65	R9c1a	1	16051G 16304C 73G 186C 16304C 234G 249DEL 263G 309.1C 315.1C
mHV-R66	R9c1a	1	16157C 16296T 16304C 16335G 73G 249DEL 263G 309.1C 315.1C
mHV-R67	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 309.2C 315.1C 456T 489C 523DEL 524DEL
mHV-R68	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 315.1C 489C
mHV-R69	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 315.1C 489C
mHV-R70	M7c3c	1	16093C 16223T 16295T 16337T 16362C 73G 146C 152C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R71	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 309.2C 315.1C 456T 489C 523DEL 524DEL
mHV-R72	JT	1	16126C 16129A 16183C 16189C 16278T 73G 263G 291.1A 309.1C 309.2C 315.1C 489C 560T
mHV-R73	M7c3c	1	16093C 16223T 16295T 16337T 16362C 73G 146C 152C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R74	B4a1a	1	16182C 16183C 16189C 16217C 16223T 16261T 73G 146C 263G 315.1C 523DEL 524DEL 8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL
mHV-R75	R9c1a	1	16157C 16296T 16291T 16304C 16335G 73G 249DEL 263G 315.1C
mHV-R76	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 309.2C 315.1C 456T 489C 523DEL 524DEL
mHV-R77	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 315.1C 456T 489C 523DEL 524DEL
mHV-R78	R9c1a	1	16157C 16296T 16304C 16335G 73G 249DEL 263G 309.1C 315.1C
mHV-R79	R9c1a	1	16157C 16296T 16304C 16335G 73G 249DEL 263G 309.1C 315.1C
mHV-R80	R9c1a	1	16157C 16296T 16304C 16335G 73G 249DEL 263G 309.1C 315.1C
mHV-R81	E1b+16261	1	16187T 16223T 16261T 16319A 16362C 73G 152C 263G 315.1C 489C
mHV-R82	D5b1c1	1	16092C 16148T 16182C 16183C 16189C 16223T 16362C 73G 150T 152C 185A 263G 309.1C 315.1C 456T 489C 523DEL 524DEL
mHV-R83	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 315.1C 489C
mHV-R84	M7c3c	1	16093C 16223T 16295T 16337T 16362C 73G 146C 152C 199C 263G 309.1C 315.1C 489C 523DEL 524DEL
mHV-R85	R9c1a	1	16157C 16296T 16304C 16335G 73G 249DEL 263G 315.1C 548T
mHV-R86	E1a1a	1	16223T 16291T 16362C 73G 263G 309.1C 315.1C 489C

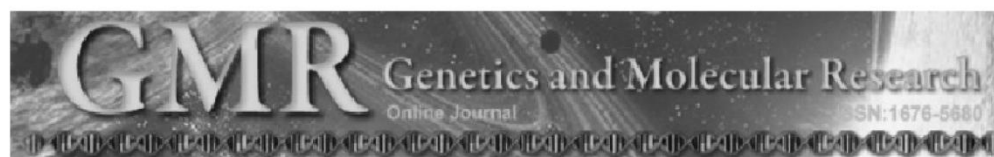
Appendix 9 : continuation.

mthV-R87	R9c1a	1	16157C	16256T	16304C	16311C	16335G	73G	249DEL	263G	309.1C	315.1C	573.1C						
mthV-R88	E2	1	16051G	16223T	16382C	73G	146C	195C	200G	263G	309.1C	309.2C	315.1C	489C					
mthV-R89	M7c3c	1	16167G	16223T	16295T	16346C	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL			
mthV-R90	M7c3c	1	16223T	16295T	16346C	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL				
mthV-R91	D5b1c1	1	16092C	16148T	16182C	16183C	16189C	16223T	16362C	73G	150T	152C	185A	263G	309.1C	315.1C	456T	489C	523DEL 524DEL
mthV-R92	E1a1a	1	16223T	16291T	16362C	73G	263G	309.1C	309.2C	315.1C	489C								
mthV-R93	F3b1	1	16093C	16220C	16265G	16298C	16362C	73G	150T	152C	249DEL	263G	315.1C						
mthV-R94	JT	1	16126C	16129A	16183C	16189C	16278T	73G	263G	291.1A	309.1C	309.2C	309.3C	315.1C	489C	560T			
mthV-R95	M7c3c	1	16223T	16295T	16346C	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL				
mthV-R96	E1a1a	1	16223T	16291T	16362C	73G	263G	309.1C	315.1C	489C									
mthV-R97	F3b1	1	16093C	16220C	16265G	16298C	16362C	73G	150T	152C	249DEL	263G	315.1C						
mthV-R98	M9	1	16223T	16362C	73G	263G	309.1C	315.1C	489C										
mthV-R99	M7c3c	1	16223T	16295T	16346C	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL				
mthV-R100	M5a1	1	16093C	16129A	16223T	16291T	16362C	73G	263G	309.1C	309.2C	315.1C	489C						
mthV-R101	D5b1c1	1	16092C	16148T	16182C	16183C	16189C	16223T	16362C	73G	150T	152C	185A	263G	309.1C	309.2C	315.1C	456T	489C 523DEL 524DEL
mthV-R102	R9c1a	1	16157C	16256T	16304C	16311C	16335G	73G	249DEL	263G	309.1C	315.1C	489C	523DEL	524DEL				
mthV-R103	M7c3c	1	16223T	16295T	16346C	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL				
mthV-R104	R9c1a	1	16157C	16256T	16304C	16335G	73G	249DEL	263G	315.1C									
mthV-R105	E1a1a	1	16145A	16185T	16223T	16291T	16362C	73G	236C	263G	309.1C	315.1C	489C						
mthV-R106	R9c1a	1	16157C	16256T	16304C	16335G	73G	249DEL	263G	315.1C									
mthV-R107	JT	1	16126C	16129A	16183C	16189C	16278T	73G	263G	291.1A	309.1C	309.2C	315.1C	489C	560T				
mthV-R108	B4a1a	1	16182C	16183C	16189C	16217C	16223T	16261T	73G	146C	260A	263G	309DEL	315.1C	523DEL	524DEL	524DEL	8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL	
mthV-R109	M7c3c	1	16223T	16295T	16346C	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL				
mthV-R110	E1b+16261	1	16093C	1618T	16223T	16261T	16318A	16362C	73G	152C	195C	263G	315.1C	489C					
mthV-R111	E1a1a	1	16223T	16291T	16362C	73G	263G	309.1C	315.1C	489C									
mthV-R112	R9c1a	1	16157C	16256T	16304C	16335G	73G	249DEL	263G	309.1C	315.1C								
mthV-R113	E1b+16261	1	16223T	16261T	16319A	16362C	73G	152C	263G	315.1C	489C								
mthV-R114	E1a1a	1	16223T	16291T	16362C	73G	263G	309.1C	309.2C	315.1C	489C								
mthV-R115	B4a1a	1	16182C	16183C	16189C	16217C	16223T	16261T	73G	146C	195C	263G	315.1C	523DEL	524DEL	8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL			
mthV-R116	M7c3c	1	16093C	16223T	16295T	16337T	16362C	73G	146C	152C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL		
mthV-R117	E1a1a	1	16185T	16223T	16291T	16362C	73G	263G	309.1C	309.2C	315.1C	489C							
mthV-R118	R9c	1	16193T	16223T	16304C	16335G	73G	249DEL	263G	309.1C	315.1C	573.1C							
mthV-R119	R9c1a	1	16157C	16256T	16304C	16311C	16335G	73G	150T	195C	263G	309.1C	315.1C	337DEL					
mthV-R120	D5b1c1	1	16092C	16148T	16182C	16183C	16189C	16223T	16362C	73G	150T	152C	185A	263G	309.1C	309.2C	309.3C	315.1C	456T 489C 523DEL 524DEL
mthV-R121	E1a1a	1	16111T	16168T	16172C	16189C	16223T	16263C	73G	16189C	16362C	263G	315.1C	235G	263G	315.1C			
mthV-R122	F3b1	1	16093C	16220C	16265G	16298C	16362C	73G	150T	152C	249DEL	263G	315.1C						
mthV-R123	R9c1a	1	16157C	16256T	16304C	16335G	73G	249DEL	263G	309.1C	315.1C								
mthV-R124	B4a1a	1	16182C	16183C	16189C	16217C	16223T	16261T	73G	146C	263G	315.1C	523DEL	524DEL	8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL				
mthV-R125	E1a1a	1	16223T	16291T	16362C	73G	263G	309.1C	309.2C	315.1C	489C								
mthV-R126	R9c1a	1	16157C	16256T	16304C	16335G	73G	249DEL	263G	309.1C	315.1C	523DEL	524DEL						
mthV-R127	B4a1a	1	16182C	16183C	16189C	16217C	16223T	16261T	73G	146C	263G	315.1C	523DEL	524DEL	8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL				
mthV-R128	D5b1c1	1	16092C	16148T	16182C	16183C	16189C	16209C	16223T	16362C	73G	150T	152C	185A	263G	309.1C	309.2C	315.1C	456T 489C 523DEL 524DEL
mthV-R129	E1a1a	1	16223T	16291T	16311C	16362C	73G	146C	263G	309.1C	315.1C	489C							
mthV-R130	E2	1	16051G	16223T	16362C	73G	146C	195C	200G	263G	309.1C	309.2C	315.1C	489C					
mthV-R131	B4a1a	1	16182C	16183C	16189C	16217C	16223T	16261T	73G	146C	263G	309.1C	309.2C	315.1C	523DEL	524DEL	8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL		
mthV-R132	D5b1c1	1	16092C	16148T	16182C	16183C	16189C	16209C	16223T	16362C	73G	150T	152C	185A	263G	309.1C	309.2C	315.1C	456T 489C 523DEL 524DEL
mthV-R133	R9c1a	1	16157C	16256T	16304C	16311C	16335G	73G	249DEL	263G	309.1C	315.1C							
mthV-R134	E2	1	16051G	16223T	16362C	73G	146C	195C	200G	263G	309.1C	309.2C	315.1C	489C					
mthV-R135	D5b1c1	1	16092C	16148T	16182C	16183C	16189C	16223T	16362C	73G	150T	152C	185A	263G	309.1C	315.1C	456T	489C	523DEL 524DEL
mthV-R136	D5b1c1	1	16092C	16148T	16182C	16183C	16189C	16223T	16362C	73G	150T	152C	185A	263G	309.1C	315.1C	456T	489C	523DEL 524DEL
mthV-R137	E1a1a	1	16185T	16223T	16291T	16362C	73G	263G	309.1C	315.1C	489C								
mthV-R138	B4a1a	1	16182C	16183C	16189C	16217C	16261T	16325C	73G	146C	263G	309.1C	315.1C	523DEL	524DEL	8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL			
mthV-R139	D5b1c1	1	16092C	16148T	16182C	16183C	16189C	16223T	16362C	73G	150T	152C	185A	263G	309.1C	309.2C	315.1C	456T	489C 523DEL 524DEL
mthV-R140	R9c1a	1	16157C	16256T	16304C	16311C	16335G	73G	249DEL	263G	309.1C	315.1C							
mthV-R141	B5a	1	16140C	16183C	16189C	16266A	16293G	73G	210G	263G	315.1C	523DEL	524DEL	8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL					
mthV-R142	M7c3c	1	16223T	16295T	16346C	16362C	73G	146C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL				
mthV-R143	N21+195	1	16193T	16223T	73G	150T	195C	263G	309.1C	315.1C	337DEL								
mthV-R144	R9c1a	1	16157C	16256T	16304C	16311C	16335G	73G	249DEL	263G	309.1C	315.1C							
mthV-R145	B5a	1	16140C	16183DEL	16189C	16266A	16293G	73G	210G	263G	315.1C	523DEL	524DEL	8281DEL 8282DEL 8283DEL 8284DEL 8285DEL 8286DEL 8287DEL 8288DEL 8289DEL					
mthV-R146	R9c1a	1	16157C	16256T	16304C	16311C	16335G	73G	249DEL	263G	309.1C	315.1C							
mthV-R148	N21+195	1	16193T	16223T	73G	150T	195C	263G	309.1C	315.1C	337DEL								
mthV-R149	D5b1c1	1	16092C	16148T	16182C	16183C	16189C	16223T	16362C	73G	150T	152C	185A	263G	309.1C	315.1C	456T	489C	523DEL 524DEL
mthV-R150	M7b1245678+16192	1	16126C	16129A	16192T	16223T	16297C	16362C	73G	150T	199C	263G	309.1C	315.1C	489C				
mthV-R151	E1a1a	1	16124C	16185T	16223T	16291T	16362C	73G	263G	309.1C	315.1C	489C	524.1A	524.2C					
mthV-R152	M7c3c	1	16093C	16223T	16295T	16337T	16362C	73G	146C	152C	199C	263G	309.1C	315.1C	489C	523DEL	524DEL		

Appendix 10 : Frequencies of mt haplogroup of various populations used in the construction of PCA.

Population	Mitochondrial haplogroup frequency																																											
	B4a*	B4a1a1	B4a2	B4b1	B4c1b	B4c2	B5a	B5b	C	D*	D4	D5	E1*	E1a	E1b	F1a*	F1a1*	F1a1a	F1a3	F1a4	F1a5	F3b	F4	M*	M7*	M7b*	M7b1	M7b3	M7c3c	M17	M21a	M73	N*	N9a6	N21	Q	R9b	R9b1	R9c	R9c1a	R21	R22	Y2	Z
Cham	4.2	0.0	0.0	0.6	2.4	10.7	16.1	1.8	0.6	1.8	0.0	0.0	0.6	0.0	0.0	1.8	1.2	6.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	3.0	0.0	1.2	4.2	0.0	2.4	0.0	3.0	1.8	0.0	3.0	0.0	1.8	0.0	0.0	4.2	0.0	0.0
Kinh	3.6	0.0	0.0	0.7	2.2	2.9	12.9	0.7	4.3	0.7	0.0	2.2	0.0	0.0	0.0	6.5	4.3	5.8	0.0	0.0	0.0	0.0	0.0	0.0	0.7	4.3	6.5	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	4.3	0.0	2.2	0.0	0.0	0.0	3.6	
Vietnam	2.4	0.0	0.0	1.0	1.7	1.0	7.9	0.7	3.4	4.1	0.0	2.4	0.0	0.0	0.0	9.2	5.1	6.2	0.0	0.0	0.0	0.0	0.0	0.0	3.4	0.0	5.5	5.8	0.3	0.0	0.3	0.0	0.0	0.0	1.7	0.0	0.0	3.4	0.0	1.4	0.0	0.0	0.0	1.0
Cambodia	0.0	0.0	0.0	0.0	0.0	16.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.8	0.0	0.0	0.0	0.0	0.0	22.6	0.0	0.0	0.0	0.0	0.0	6.5	0.0	3.2	0.0	0.0	6.5	0.0	3.2	0.0	0.0	0.0	6.5	0.0	0.0	
Hainan	8.8	0.0	0.0	3.1	3.8	0.6	6.3	0.6	5.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9	0.6	0.0	0.0	0.0	3.8	3.8	5.0	2.5	6.9	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	5.7	0.0	1.3	0.0	0.0	0.0	0.6	
Taiwanese Aborigine	1.0	9.2	5.7	10.3	0.4	0.0	5.4	0.0	0.0	1.1	0.0	4.2	12.0	0.0	0.0	1.9	2.9	0.0	0.0	0.0	0.0	7.9	12.0	0.7	0.0	0.1	0.7	10.9	4.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	0.0	0.0	1.4	0.0	
Orang Asli	0.7	0.0	0.0	0.0	0.0	0.0	0.7	5.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0	2.9	0.0	25.4	0.0	0.0	6.1	8.6	0.0	8.6	0.0	0.0	0.0	19.3	0.0	0.0	
Kadazan-Dusun	3.3	0.0	0.0	4.7	0.0	0.7	3.3	0.0	0.0	0.0	1.3	4.7	0.0	8.0	4.7	0.0	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23.3	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.3	0.0	0.0	0.0	0.0
Bajau	6.7	0.0	2.0	5.3	7.3	2.7	0.7	2.0	0.7	0.0	0.0	0.7	0.0	7.3	1.3	4.7	0.0	0.7	0.7	0.7	0.0	6.7	0.0	0.0	0.0	0.0	2.0	1.3	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.7	9.3	0.0	1.3	2.7	0.0		
Rungus	8.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.7	15.3	0.0	18.7	7.3	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	2.0	0.0	15.3	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	1.3	14.0	0.0	0.0	0.0	0.0
Iban	13.6	0.0	0.0	0.0	3.7	4.9	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	3.7	0.0	1.2	0.0	0.0	0.0	16.1	0.0	0.0	13.6	1.2	0.0	0.0	1.2	0.0	9.9	0.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0	2.5	12.4	11.1
NW China	1.0	0.0	0.0	2.0	0.2	0.0	2.2	1.0	4.7	9.9	3.2	3.0	0.2	0.0	0.0	1.0	2.5	0.7	0.0	0.0	0.0	0.0	0.2	11.1	0.7	1.0	1.7	0.2	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.7	0.0	0.2	0.0	0.0	0.0	3.2
NE China	3.4	0.0	0.2	2.7	2.1	0.0	3.4	1.4	4.3	12.6	3.4	7.1	0.0	0.0	0.0	2.1	2.5	0.2	0.2	0.2	0.0	0.0	0.0	4.1	0.2	1.1	3.2	0.0	0.5	0.0	0.0	0.0	1.8	0.0	0.0	0.0	0.2	0.0	0.2	0.0	0.0	0.7	4.6	
SW China	6.5	0.0	0.3	1.8	0.3	0.8	6.8	0.6	5.7	4.9	4.6	2.6	0.1	0.0	0.0	4.9	1.9	3.7	0.0	0.2	0.0	0.0	0.2	8.0	0.9	3.4	4.6	0.0	0.2	0.5	0.0	0.0	0.4	0.3	0.0	0.0	0.0	1.7	0.0	0.3	0.0	0.0	0.0	1.3
SE China	6.1	0.0	0.2	3.1	1.6	1.0	4.9	2.6	5.2	6.3	4.7	5.6	0.2	0.0	0.0	2.9	3.3	2.7	0.0	0.0	0.0	0.0	0.0	7.0	1.2	3.0	2.1	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.9	0.0	0.4	0.0	0.0	0.0	1.2	
Thailand	5.3	0.0	0.0	0.0	0.0	2.3	10.5	0.0	3.6	4.0	0.4	1.6	0.0	0.0	0.0	4.7	1.2	8.9	0.0	0.5	0.0	0.2	0.8	4.7	0.8	2.0	2.8	0.0	0.4	0.5	6.9	0.0	0.0	0.8	0.4	0.0	0.0	1.2	0.0	0.9	0.0	0.8	0.0	0.4
Malayu Malay	0.9	0.0	0.0	0.9	2.7	2.8	9.2	0.9	0.9	0.0	0.0	0.9	0.0	0.9	3.7	3.8	3.7	8.3	1.9	0.0	0.0	0.0	14.2	0.0	0.0	3.7	0.0	4.6	1.9	4.6	3.8	1.8	2.8	1.8	1.8	0.0	0.9	0.0	0.0	0.0	1.8	0.0		
Taiwanese Han	9.2	0.0	11.5	6.6	3.3	0.0	5.1	0.0	0.0	0.2	0.9	4.1	2.4	6.2	0.0	0.2	2.5	0.0	0.7	0.7	0.0	7.1	11.8	0.7	0.0	0.1	0.7	10.2	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8	0.0	0.0	1.8	0.0	1.2	0.0
Philippines	11.3	0.0	0.0	1.6	0.0	0.0	0.0	9.7	0.0	0.0	0.0	0.0	0.0	8.1	0.0	1.6	0.0	0.0	8.2	0.0	0.0	19.0	0.0	2.4	0.0	0.0	0.0	0.0	11.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3	0.0	0.0	12.9	0.0
Sumatra	6.1	0.0	0.0	0.6	4.4	2.2	3.9	2.2	0.0	0.0	0.0	0.0	0.6	3.3	2.2	4.4	0.0	5.0	1.1	1.1	1.7	0.0	1.7	10.1	2.8	0.6	1.7	0.0	8.9	3.9	0.0	5.0	2.8	3.3	0.6	0.0	0.0	1.7	0.0	0.0	0.0	0.6	6.7	0.6
Borneo	7.0	1.3	0.0	1.9	1.9	5.7	4.5	1.3	1.3	0.6	0.0	4.5	0.6	8.9	1.9	0.6	1.3	0.6	1.9	1.9	0.0	5.7	0.0	8.3	0.6	1.3	0.6	0.0	7.0	2.5	0.6	6.4	0.6	0.6	0.0	1.3	0.0	0.6	0.0	1.9	0.0	1.9	1.9	0.3
Java	2.2	0.0	0.0	0.0	2.2	0.0	2.2	0.0	2.2	0.0	0.0	0.0	0.0	0.0	2.2	2.8	2.2	4.4	5.6	0.0	13.0	0.0	0.0	13.0	0.0	0.0	2.2	0.0	10.9	2.2	0.0	4.3	15.2	2.2	0.0	0.0	0.0	4.4	0.0	0.0	2.2	2.2	0.0	
Bali	2.4	0.0	0.0	0.0	6.0	7.2	4.8	0.0	0.0	0.0	0.0	1.2	1.2	3.6	1.2	6.0	4.8	2.4	0.0	0.0	0.0	0.0	12.0	1.2	1.2	4.8	0.0	6.0	4.8	0.0	4.8	0.0	0.0	1.2	1.2	0.0	0.0	0.0	0.0	0.0	7.2	1.2	0.0	
Lombok	4.5	2.3	0.0	0.0	0.0	0.0	6.8	2.3	2.3	0.0	0.0	2.3	0.0	2.3	6.8	9.1	0.0	9.1	2.2	2.3	0.0	0.0	11.4	0.0	0.0	6.8	0.0	2.3	2.3	0.0	9.1	0.0	0.0	0.0	0.0	0.0	2.3	0.0	0.0	11.4	0.0	0.0		
Sumba	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	12.0	2.0	2.0	4.0	4.0	6.0	0.0	2.0	0.0	8.0	2.0	0.0	0.0	8.0	12.0	2.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	2.0	0.0	8.0	0.0	0.0		
Sulawesi	5.1	3.8	0.0	0.8	3.8	0.8	2.5	2.5	0.4	2.5	0.0	8.4	2.1	17.7	3.8	1.7	0.4	1.3	1.7	5.1	0.4	0.0	0.0	3.0	1.3	0.0	0.4	1.3	11.0	1.3	0.0	3.0	0.0	0.8	0.4	2.1	0.0	0.8	0.0	0.8	0.4	1.7	0.0	
Alor	0.0	2.2	0.0	6.6	0.0	0.0	4.4	0.0	0.0	0.0	2.2	0.0	0.0	6.6	4.4	2.2	0.0	2.2	0.0	6.7	0.0	0.0	0.0	2.2	0.0	0.0	0.0	4.4	0.0	0.0	4.4	0.0	0.0	4.4	28.9	0.0	0.0	0.0	11.1	0.0	2.2	0.0	0.0	
Ambon	9.3	14.0	0.0	4.7	0.0	0.0	7.0	0.0	0.0	4.7	0.0	2.3	7.0	4.7	0.0	0.0	0.0	2.3	7.0	4.7	0.0	0.0	0.0	2.3	2.3	2.3	2.3	0.0	2.3	2.3	0.0	2.3	0.0	0.0	11.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Genetics and Molecular Research (2011)



Short Communication

Genetic data for 15 STR loci in a Kadazan-Dusun population from East Malaysia

B.P. Kee¹, L.H. Lian¹, P.C. Lee², T.X. Lai² and K.H. Chua¹

¹Department of Molecular Medicine, Faculty of Medicine,
University of Malaya, Kuala Lumpur, Malaysia

²Biotechnology Program, School of Science and Technology,
Universiti Sabah Malaysia, Kota Kinabalu, Sabah, Malaysia

Corresponding author: L.H. Lian

E-mail: lhlian@um.edu.my

Genet. Mol. Res. 10 (2): 739-743 (2011)

Received September 15, 2010

Accepted November 25, 2010

Published April 26, 2011

DOI 10.4238/vol10-2gmr1064

ABSTRACT. Allele frequencies of 15 short tandem repeat (STR) loci, namely D5S818, D7S820, D13S317, D16S539, TH01, TPOX, Penta D, Penta E, D3S1358, D8S1179, D18S51, D21S11, CSF1PO, vWA, and FGA, were determined for 154 individuals from the Kadazan-Dusun tribe, an indigenous population of East Malaysia. All loci were amplified by polymerase chain reaction, using the Powerplex 16 system. Alleles were typed using a gene analyzer and the Genemapper ID software. Various statistical parameters were calculated and the combined power of discrimination for the 15 loci in the population was calculated as 0.9999999999999999. These loci are thus, informative and can be used effectively in forensic and genetic studies of this indigenous population.

Key words: Short tandem repeats; Population data; East Malaysia; Powerplex 16 system

Annals of Human Biology, 2012; Early Online: 1–6
Copyright © Informa UK, Ltd.
ISSN 0301-4460 print/ISSN 1464-5033 online
DOI: 10.3109/03014460.2012.719548

informa
healthcare

RESEARCH PAPER

Population data of six *Alu* insertions in indigenous groups from Sabah, Malaysia

B. P. Kee¹, K. H. Chua¹, P. C. Lee² & L. H. Lian¹

¹Department of Molecular Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia, and ²Biotechnology Program, School of Science and Technology, Universiti Sabah Malaysia, Kota Kinabalu, Sabah, Malaysia

Background and aim: The present study is the first to report the genetic relatedness of indigenous populations of Sabah, Malaysia, using a set of *Indel* markers (HS4.32, TPA25, APO, PV92, B65 and HS3.23). The primary aim was to assess the genetic relationships among these populations and with populations from other parts of the world by examining the distribution of these markers.

Subjects and methods: A total of 504 volunteers from the three largest indigenous groups, i.e. Kadazan-Dusun, Bajau and Rungus, were recruited for the study. Six *Alu* insertions were typed by PCR with specific primer sets.

Results: All insertions were found to present at different frequencies, ranging from 0.170–0.970. The heterozygosity of most of the markers was high (>0.4), with the exception of HS3.23 and APO. A genetic differentiation study revealed that these populations are closely related to each other ($G_{ST} = 0.006$). A principle component plot showed that these populations have higher affinity to Mainland South East Asia/East Asia populations, rather than Island Southeast Asia (ISEA) populations.

Conclusion: In summary, these indigenous groups were closely associated in terms of their genetic composition. This finding also supports the colonization model of ISEA, which suggests that the inhabitants of this region were mostly descendants from Southern China.

Keywords: *Alu* insertion, Kadazan-Dusun, Bajau, Rungus

INTRODUCTION

The *Alu* element, one of the most prominent examples of repetitive DNA, is a member of the short interspersed element (SINE), which is estimated to exist in more than 10% of the entire human genome (Houck et al. 1979; Smit 1996). Each *Alu* element is ~300 nucleotides in length and is believed to derive from 7SL RNA (Weiner et al. 1986). The *Alu* element is mobilized within the genome by

a gene jumping mechanism known as retroposition—a RNA-mediated transcription process. This mechanism has contributed to the random yet wide distribution of *Alu* elements, with varied density, throughout the genome. The insertion of *Alu* elements began as early as 65 million years ago (Shen et al. 1991). The current amplification rate of *Alu* insertions, at a rate of one new insertion in 200 newborns, has been found to be much lower than in the past, being merely 1% of previous peak rates (Deininger and Batzer 1999). There are more than one million accumulated copies of *Alu* elements being reported in the human genome (Lander et al. 2001). Most *Alu* insertions are constant in all human beings, regardless of the origin of population. However, a small number of them (~5%) are polymorphic. These insertions arose recently during global colonization by modern humans and have great potential to reveal information about modern human expansion and migration events. Most *Alu* insertions are shared by both human and primate genomes, but ~7000 of these insertions are unique to humans (The Chimpanzee Sequencing and Analysis Consortium 2005).

Over decades of extensive studies, *Alu* insertions have been found useful in different aspects of scientific research, especially in cancer and evolutionary studies (Flint et al. 1996; Deininger and Batzer 1999; Batzer and Deininger 2002). Scientists have postulated that *Alu* elements are the major contributor to the evolutionary process throughout primate history, including that of humans (Batzer and Deininger 2002). Massive insertions of *Alu* elements have caused genomic instability that has facilitated the process of speciation (Challem and Taylor 1998). Recently, the study of *Alu* elements has focused on the search for population-specific markers. Researchers have demonstrated the possibility of inferring the origin of a population using a combination of polymorphic *Alu* markers (Ray et al. 2005).

Correspondence: Dr L. H. Lian, Department of Molecular Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia.
Tel: (+60)3 7967 5740. Fax: (+60)3 7967 4957. E-mail: lhlian@um.edu.my
(Received 16 January 2012; accepted 31 July 2012)

Poster presentation, The 14th Biological Sciences Graduate Congress (2009)

189

14th BSGC

CBB-PO 30

Genetic data for 9 STR loci (CSF1PO, TPOX, TH01, FESFPS, F13A01, vWA, D7S820, D13S317, D16S539) in the Kadazan-dusun population of East Malaysia

Kee, B. P.¹, Lian, L. H.¹, Lee, P. C.², and Chua, K. H.¹

¹Department of Molecular Medicine, Faculty of Medicine, University Malaya, 50603, Kuala Lumpur, Malaysia

²Biotechnology Program, School of Science and Technology, University Malaysia Sabah, 88999, Kota Kinabalu, Sabah, Malaysia

Short Tandem Repeats (STRs) have gained popularity in modern human identity testing due to its high power of discrimination and ability to be amplified via multiplex-polymerase chain reaction (PCR). STRs are repetitive sequences found scattered throughout the human genome and are characterized by the occurrence of repeating units ranging from one to four base pairs in length. These sequences are located adjacent to each other and can duplicate up to 100 times. The amount of STR loci present in the human genome is expected to be more than one million and accounts for up to 3% of the entire genome. The objective of our study was to access the genomic diversity of the East Malaysian indigenous population. Kadazan-dusun is the largest indigenous group in the state of Sabah, East Malaysia and make up 17.8% of the local population. The data generated will be of use to in establishing the genetic structure of the population and facilitating future research in various aspects, such as forensic investigations, anthropology, and evolutionary studies. A total of nine STR loci were amplified via the Promega GenePrint STR System. The PCR products were resolved using denaturing polyacrylamide gel electrophoresis (PAGE) and visualized with a silver staining method. Basic population statistical parameters were calculated to characterize the structure of the population studied.

Abstract of the 14th Biological Sciences Graduate Congress

10th -12th December, 2009, Chulalongkorn University, Bangkok, Thailand