

**GENETIC ARCHITECTURE OF DUSUNIC,
MURUTIC AND PAITANIC ETHNIC GROUPS
OF SABAH, MALAYSIA AS REVEALED BY
HIGH DENSITY GENOTYPING ARRAYS**



YEW CHEE WEI

UNIVERSITI MALAYSIA SABAH
PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH

**BIOTECHNOLOGY RESEARCH INSTITUTE
UNIVERSITI MALAYSIA SABAH
2016**

**GENETIC ARCHITECTURE OF DUSUNIC,
MURUTIC AND PAITANIC ETHNIC GROUPS
OF SABAH, MALAYSIA AS REVEALED BY
HIGH DENSITY GENOTYPING ARRAYS**

YEW CHEE WEI



UMS
PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH
UNIVERSITI MALAYSIA SABAH

**THESIS SUBMITTED IN FULFILLMENT FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY**

**BIOTECHNOLOGY RESEARCH INSTITUTE
UNIVERSITI MALAYSIA SABAH
2016**

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS

JUDUL: **GENETIC ARCHITECTURE OF DUSUNIC, MURUTIC AND PAITANIC ETHNIC GROUPS OF SABAH, MALAYSIA AS REVEALED BY HIGH DENSITY GENOTYPING ARRAYS**

IJAZAH: **DOCTOR OF PHYLOSOPHY IN BIOTECHNOLOGY (MOLECULAR GENETICS)**

Saya **YEW CHEE WEI**, sesi pengajian **2011-2016**, mengaku membenarkan tesis Doktor Falsafah ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis ini adalah hak milik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan (✓)

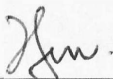
SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA 1972)

TERHAD

(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

TIDAK TERHAD



YEW CHEE WEI

Disahkan oleh,
MURULAIN BINTI ISMAIL
LIBRARIAN



(Tandatangan Pustakawan)



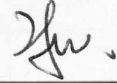
Prof. Madya Dr. Vijay Kumar
(Penyelia)

Tarikh: 22 Ogos 2016

DECLARATION

I hereby declare that the material in this thesis is of my own effort, except for quotations, excerpts, equations, references and summaries which have been duly acknowledged and cited clearly as sources.

22 August 2016



Yew Chee Wei
PB20119025



UMS
UNIVERSITI MALAYSIA SABAH

PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH

CERTIFICATION

NAME : **YEW CHEE WEI**

MATRIC NO. : **PB20119025**

TITLE : **GENETIC ARCHITECTURE OF DUSUNIC, MURUTIC AND PAITANIC ETHNIC GROUPS OF SABAH, MALAYSIA AS REVEALED BY HIGH DENSITY GENOTYPING ARRAYS**

DEGREE : **DOCTOR OF PHYLOSOPHY IN BIOTECHNOLOGY (MOLECULAR GENETICS)**

VIVA DATE : **15 April 2016**

DECLARED BY;

1. SUPERVISOR

Assoc. Prof. Dr. Vijay Kumar



UMS
UNIVERSITI MALAYSIA SABAH

Signature

ACKNOWLEDGEMENT

I would like to express my deepest gratitude and appreciation to my supervisor, Associate Professor Dr. Vijay Kumar, who has been patient and supportive to advise, guide and supervise me throughout the study. His efforts had encouraged me to complete the research. At the same time, he had given opportunities for me in participating a number of activities such as lab attachment, conferences, competitions and application of scholarships, which had enhanced my soft skills in addition to the laboratory skills and knowledge of the research subjects.

Besides, I would like to thank the Director of Biotechnology Research Institute, Professor Dr. Clemente Michael Wong, and the administration of the institute, for giving me the opportunity to further my study in a well-equipped and well-managed laboratory. In addition, I also would like to thank the Ministry of Science, Technology and Innovation Malaysia for supporting this research through a research grant (project no.: 100-RMI/BIOTEK 16/6/2 B (1/2011)). I am very thankful to the Ministry of Science, Technology and Innovation for supporting my expenses through a scholarship under National Science Fellowship Scheme.

Most importantly, I am very grateful to the kindhearted indigenous people of Sabah for being voluntary to contribute their DNA samples, without which this thesis would not be successful. I also wish to express my appreciation to Mr. Alexander Minsong and Mr. Arzan Ranjuban, for their great efforts in assisting sample collection activities. Lastly, I am in my deepest gratitude to my wife, Wee Ching Ching for her moral support. Thank you!

Yew Chee Wei
22 Aug 2016

UNIVERSITI MALAYSIA SABAH

ABSTRACT

The native ethnic groups of Sabah are categorized under the 'North Borneo' stock of the Austronesian linguistic family. It is generally believed that the native groups of Northern Borneo are plausibly descendants of the 'Out-of-Taiwan' Austronesian wave of human migration. While there may be some anthropological evidence support for this, the lack of genetic evidence makes the hypothesis inconclusive. As such, this study aimed to unravel and compare the population metrics, genetic structure and genetic relationships of the Northern Borneo indigenous ethnic groups (North Borneans) against Southern China and Southeast Asian populations, and subsequently provide inference of their migration history. Ethical clearance was obtained and blood samples were collected from healthy individuals. A total of 117 individuals representing five indigenous ethnic groups namely Dusun, Rungus, Sonsogon, Sungai-Lingkabau and Murut-Paluan were genotyped with ~2.4 million genome-wide single nucleotide polymorphism (SNP) markers. The genotype data were then merged with public datasets i.e. HapMap, Human Genome Diversity Project (HGDP), Singapore Genome Variation Project (SGVP), and Pan-Asian SNP Consortium (PASNP) data to form a comprehensive meta-dataset composing of 89 regional and worldwide populations. Population metrics namely decay of linkage disequilibrium, genetic heterozygosity, genetic differentiation (F_{ST}) and phylogeny were analyzed. Next, comparative population genetic structure analysis was performed to determine the genetic gradient among populations, and to assign genetic component and its admixture across the tested populations. Finally, the genetic relationships among populations were inferred by a combinatorial correlation of these outputs. The results showed that the North Borneans were subdivided into three subgroups which were 'Dusun-Rungus', 'Sonsogon-Sungai', and 'Muruts'. The 'Sonsogon-Sungai' grouping, which is made up of Dusunic and Paitanic-speaking group respectively, indicated that the linguistic groupings of the ethnic groups do not necessarily reflect their genetic affinity. Meanwhile, the North Borneans had reduced heterozygosity and were highly differentiated among themselves. Clustering with principal components clearly depicted that each ethnic group is an independent genetic entity. As a whole, they formed a unique genetic ancestry, which was not found in previous reports. Importantly, they were closest to the non-Negrito Filipinos and the Cosmopolitan Malays of Singapore. However, phylogenetic analysis clustered the North Borneans to the Filipinos and Taiwan Natives, but not to other Island Southeast Asians. On the contrary, the Bidayus (West Borneo) was clustered with the Javanese and Temuans. Subsequent estimation of gene flow direction revealed that statistically probable migration event(s) was unidirectional, from North Borneo towards mainland Southeast Asia, but not the reverse. As such, a new hypothesis is postulated that the five ethnic groups descended from Taiwan Natives and Borneo Island served as one of the cross-road for two distinct waves of migration from mainland Southeast Asia and Taiwan, respectively. In conclusion, the findings indicated that Sabah's indigenous population, as a whole, has a unique yet distinctive pool of genetic variants, which are important for anthropological and medical genetic studies.

ABSTRAK

Arkitek Genetik Kumpulan-Kumpulan Etnik Dusunik, Murutik and Paitanik di Sabah, Malaysia yang dicirikan dengan Kaedah Genotip Tersusun Padat

Golongan pribumi Sabah adalah tertakluk kepada kategori 'Borneo Utara' di bawah kumpulan linguistik 'Austronesia'. Asal-usul mereka dipercayai berketurunan dari penghijrahan manusia yang dinamakan 'Keluar dari Taiwan'. Walaupun hipotesis ini disokong dengan bukti antropologi, kekurangan bukti genetik menyebabkan asal-usul mereka masih tidak jelas. Oleh itu, kajian ini bertujuan untuk meneliti metrik populasi dan membandingkan struktur populasi dan hubungan genetik kumpulan etnik Borneo Utara terhadap populasi China Selatan dan Asia Tenggara, lalu memaparkan hipotesis baru tentang sejarah penghijrahan mereka. Kelulusan etika telah didapatkan sebelum sampel darah dikumpulkan dari individu yang sihat. Sejumlah 117 individu yang mewakili lima kumpulan bumiputera iaitu Dusun, Rungus, Sonsogon, Sungai-Lingkabau dan Murut-Paluan, telah diuji genotip dengan ~2.4 juta penanda 'single nucleotide polymorphism' SNP pada seluruh genom. Data genotip ini digabungkan dengan set-set data awam iaitu HapMap, Human Genome Diversity Project (HGDP), Singapore Genome Variation Project (SGVP) dan Pan-Asian SNP Consortium (PASNP) untuk menghasilkan satu set data komprehensif yang mengandungi 89 populasi serantau dan sedunia. Metrik populasi iaitu perpautan ketidakseimbangan, heterozigositi, pembezaan genetik (F_{ST}) dan filogeni turut dikaji. Kemudian, perbandingan struktur genetik populasi dijalankan untuk menentukan kecerunan genetik antara populasi, dan juga menetapkan komponen genetik dan percampurannya antara populasi. Akhirnya, hubungan genetik disimpulkan dengan mengkaitkan semua keputusan yang didapati. Penduduk Borneo Utara dibahagikan kepada tiga kumpulan kecil yang terdiri daripada gabungan 'Dusun-Rungus', 'Sonsogon-Sungai' dan 'Murut'. 'Sonsogon-Sungai' yang masing-masing merupakan penutur Dusunik dan Paitanik, menunjukkan bahawa pengumpulan linguistik tidak semestinya mencerminkan hubungan genetik mereka. Sementara itu, penduduk Borneo Utara memaparkan pengurangan heterozigositi dan adalah berbeza dengan ketara di kalangan mereka. Analisis kelompok komponen utama telah menggambarkan dengan jelas bahawa setiap kumpulan etnik adalah satu entiti genetik yang bebas. Secara amnya, mereka membentuk satu keturunan genetik yang unik yang tidak pernah dilaporkan. Yang pentingnya, profil genetik mereka adalah paling dekat dengan orang Filipin bukan negro and dan Melayu Kosmopolitan dari Singapore. Namun begitu, analisis filogeni mengumpulkan penduduk Borneo Utara dengan Filipino dan Bumiputera Taiwan, tetapi bukan populasi lain di kawasan Kepulauan Asia Tenggara. Sebaliknya, Bidayuh (Borneo Barat) dikelompokkan dengan orang Java dan Temuan. Jangkaan pengaliran gen menunjukkan migrasi sehalu dari Borneo Utara ke Asia Tenggara, dan bukan sebaliknya. Justeru, satu hipotesis baru telah dicadangkan bahawa lima kumpulan etnik tersebut adalah berketurunan dari Bumiputera Taiwan, dan Pulau Borneo pula berfungsi sebagai salah satu simpangan jalan kepada dua migrasi yang masing-masing berasal dari Taiwan dan benua Asia Tenggara. Kesimpulannya, keputusan kajian ini menunjukkan bahawa penduduk pribumi Sabah mengandungi variasi genetik yang unik dan adalah penting untuk kajian-kajian anthropologi and genetik perubatan pada masa kelak.

LIST OF CONTENTS

	Page
TITLE	i
DECLARATION	ii
CERTIFICATION	iii
ACKNOWLEDGEMENT	iv
ABSTRACT	v
ABSTRAK	vi
LIST OF CONTENTS	vii
LIST OF TABLES	xiii
LIST OF FIGURES	xiv
LIST OF ABBREVIATIONS	xvi
LIST OF APPENDIX	xvii
CHAPTER 1: INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problems	2
1.3 Objectives	4
CHAPTER 2: LITERATURE REVIEW	6
2.1 The Background of Sabah	6
2.1.1 Multi-Ethnic Populations in Sabah	7
2.1.2 The Legend of Nunuk Ragang and Seven Brothers	8
2.2 An Overview of the Population Background of Southeast Asia	9
2.3 Pre-Historical Evidence on the Peopling of Southeast Asia	11
2.3.1 Pre-History of Sabah	11
2.3.2 Peopling of Peninsular Malaysia, Borneo and the Philippines	13
2.3.3 The 'Two-Layer' Hypothesis	17
2.4 Linguistic Evidence and the Theory of 'Out of Taiwan'	18
2.4.1 The 'Greater North Borneo' Hypothesis	19
2.5 Genetic Evidence on Human Migration History	20
2.5.1 Inference by Classical Genetic Markers	21

2.5.2	Gender-specific Migration Pattern	23
a.	Maternal Lineages in Southeast Asia	24
b.	Paternal Lineages in Southeast Asia	26
2.5.3	Inference by Genome-wide Autosomal Markers	27
2.6	Genetic-based Inference for the Peopling of Borneo Island	29
2.7	A Summary of Previous Genetic Reports on Sabahans	30
2.8	A Review on Analysis of Population Genetic Structure	32
2.8.1	Principal Component Analysis	32
2.8.2	Genetic Component and Admixture	33
CHAPTER 3: GENERAL METHODOLOGY		35
3.1	Ethical Clearance and Blood Sample Collection	35
3.2	Isolation of White Blood Cells from Whole Blood	40
3.3	Preparation of DNA Samples	41
3.3.1	Genomic DNA Isolation from White Blood Cells	41
3.3.2	Quantification and Quality Assessment of Genomic DNA	42
3.4	Genome-wide SNP Genotyping with High Density Array Beadchip	43
3.4.1	Whole Genome Amplification	43
3.4.2	Fragmentation and Precipitation of Amplified Genomic DNA	44
3.4.3	Resuspension of Fragmented DNA	45
3.4.4	Hybridization of Fragmented DNA on Beadchip	45
3.4.5	Single Base Extension and Fluorescence Labeling	47
3.4.6	Coating Beadchip and Image Scanning	49
3.4.7	Calling of SNP Alleles	50
3.5	Quality Assessment of Samples and SNPs	50
3.5.1	Quality Assessment of Samples	50
a.	Removing Individuals with <98% Call Rate	50
b.	Removing Individuals with Discrepant Reported Gender	50
c.	Removing Individuals in First-Degree Relationships	51
3.5.2	Quality Assessment of SNPs	51
a.	Removing non-Autosomal SNPs	51
b.	Removing SNPs with Missing Call Rate > 5%	51

c.	Removing SNPs Deviated from Hardy-Weinberg Equilibrium	52
3.6	Data Merging with Public Datasets	52
3.7	Analysis of Population Metrics, Genetic Structure and Migration Direction	53
3.7.1	List of Population Metrics Analysis	53
3.7.2	List of Genetic Structure Analysis	53
3.7.3	List of Migration Direction Analysis	53
 CHAPTER 4: GENETIC STRUCTURE OF FIVE INDIGENOUS ETHNIC GROUPS OF SABAH		 54
4.1	Introduction	54
4.2	Objectives	55
4.3	Methodology	58
4.3.1	Sample Genotyping with Genome-wide SNPs	58
4.3.2	Quality Assessment and Merging with Public Datasets	58
4.3.3	Population Metrics and Genetic Structure Analysis within the North Borneo Ethnic Groups	59
a.	Distribution of Allele Frequencies	59
b.	Decay of Linkage Disequilibrium among North Borneans	59
c.	Observed Heterozygosity	60
d.	Pairwise Genetic Differentiation, F_{ST}	60
e.	Phylogenetic Relationships	60
f.	Genetic Structure and Admixture	60
4.3.4	Population Structure Analysis with Worldwide Populations	61
a.	Decay of Linkage Disequilibrium in Comparison to Regional Populations	61
b.	Observed Heterozygosity and Worldwide Phylogenetic Relationships	62
c.	Regional and Worldwide Population Structure	62
d.	Estimating the Direction of Gene Flow	63
4.4	Results	63
4.4.1	Distribution of Allele Frequencies	63

4.4.2	Pattern of Decay of Linkage Disequilibrium	63
4.4.3	Population Metrics within North Borneans	67
	a. Observed Heterozygosities	67
	b. Pairwise Genetic Differentiation, F_{ST}	67
	c. Phylogenetic Relationships among the North Borneans	68
4.4.4	Genetic Structure and Admixture among the North Borneans	69
	a. Principal Component Analysis	69
	b. Genetic Component and Admixture Analysis	69
4.4.5	Analysis against Worldwide and Regional Populations	71
	a. Decay of Linkage Disequilibrium	71
	b. Genetic Heterozygosity	71
	c. Pairwise F_{ST}	71
	d. Worldwide Phylogenetic Relationships	75
	e. Worldwide and Regional Principal Component Analysis	75
	f. Genetic Component and Admixture Analysis	76
	g. Estimation of the Direction of Gene Flow	83
4.5	Discussion	87
4.5.1	Population Metrics and Genetic Structure within North Borneans	87
4.5.2	Comparative Population Metrics and Genetic Structure with Worldwide and Regional Populations	89
4.6	Conclusion	95

CHAPTER 5: INFERRING THE GENETIC RELATIONSHIPS OF NORTH BORNEANS TO MULTI-ETHNIC SOUTHEAST ASIANS

5.1	Introduction	96
5.2	Objectives	98
5.3	Methodology	99
5.3.1	Merging with Public Datasets	99
	a. Quality Assessment before Merging Datasets	101
	b. Sample Quality Assessment	101
	c. SNP Quality Assessment	101
	d. Merging Public Datasets	101
	e. Subdividing the meta-Dataset	102

5.3.2	Analysis of Population Metrics	103
	a. Genetic Heterozygosity	103
	b. Genetic Differentiation, F_{ST}	103
	c. Phylogenetic Analysis	104
5.3.3	Analysis of Genetic Structure	104
	a. Principal Component Analysis	104
	b. Genetic Component and Admixture Ancestry	105
5.4	Results	106
5.4.1	Reduction of Genetic Heterozygosity	106
5.4.2	High Pairwise F_{ST} among North Borneans	106
5.4.3	Phylogenetic Relationships	112
5.4.4	Principal Component Analysis	114
	a. Worldwide Populations	114
	b. Regional Populations without Negritos	114
	c. Austronesians	116
5.4.5	Genetic Component and Admixture Analysis	119
	a. Worldwide Populations	119
	b. Regional Populations without Negritos	121
	c. Austronesians	122
5.5	Discussion	127
5.5.1	Reduced Heterozygosity and High Genetic Differentiation in North Borneans	127
5.5.2	Phylogenetic Relationships of the Southeast Asians	131
5.5.3	Population Genetic Structure of the Southeast Asians	132
5.5.4	Inferring Genetic Relationships of the Southeast Asians	134
5.6	Conclusion	136
CHAPTER 6: GENERAL DISCUSSION		138
6.1	Proposing a New Hypothesis of Human Migration History in Southeast Asia	138
6.2	Proposed New Hypothesis: The North Borneans were Descended from Taiwan Natives, and the Borneo Island Served as one of the Cross-	

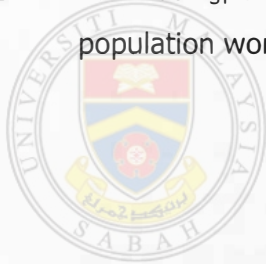
Road for Two Distinct Waves of Migration from Mainland Southeast Asia and Taiwan, Respectively	139
6.2.1 The Origin of North Borneans	139
6.2.2 The Borneo Island Served as the Cross-Road of Migration between Mainland Southeast Asians and Austronesians	143
CHAPTER 7: SUMMARY	146
REFERENCES	152
APPENDIX	173



UMS
UNIVERSITI MALAYSIA SABAH

LIST OF TABLES

	Page
Table 2.1: A summary of archaeological findings in Sabah	14
Table 3.1: Twelve sampling activities in four districts of Sabah	38
Table 3.2: Ethnic groups and number of samples collected	39
Table 4.1: Percentage of monomorphic SNPs genotyped by the ascertained SNP panel	65
Table 4.2: Observed heterozygosity for the five ethnic groups	67
Table 4.3: Pairwise population differentiation index, F_{ST} , among the North Borneo ethnic groups	68
Table 4.4: Pairwise F_{ST} of North Borneans against worldwide populations	74
Table 4.5: The f_3 statistics test proved that there was gene flow from North Borneo and Cambodians towards MAS.	86
Table 5.1: Pairwise F_{ST} of North Borneans in comparison to other population worldwide.	109



UMS
UNIVERSITI MALAYSIA SABAH

LIST OF FIGURES

	Page
Figure 3.1: Sampling sites and sample distribution	37
Figure 4.1: Location of North Borneo (Malaysian Sabah) and the approximate distribution of indigenous ethnic groups by language family	57
Figure 4.2: Distribution of allele frequencies of all autosomal SNPs for The five indigenous ethnic groups of Sabah	65
Figure 4.3: Decay of LD for chromosome 21 of the Sabah five ethnic groups	66
Figure 4.4: Phylogenetic relationships of the five ethnic groups of Sabah	68
Figure 4.5: Principal component analysis of five ethnic groups of Sabah	70
Figure 4.6: Cross validation test for ADMIXTURE analysis within five ethnic groups of Sabah	70
Figure 4.7: Linkage disequilibrium analysis of North Borneo ethnic groups together with Southern China and Southeast Asian populations	72
Figure 4.8: Comparison of heterozygosity between the North Borneans against 21 regional populations	73
Figure 4.9: Phylogenetic relationships (Neighbor-Joining tree) of the North Borneo ethnic groups with other worldwide populations	78
Figure 4.10: Principal component analysis of worldwide and regional populations from North-East Asia down to Southeast Asia	79
Figure 4.11: Cross validation test with the ADMIXTURE program among 63 worldwide populations	80
Figure 4.12: Cross validation test with the ADMIXTURE program among 26 regional (NEA-SC-SEA-NB) populations.	80
Figure 4.13: Ancestry assignment for 63 worldwide populations, and proportions of hypothetical ancestries.	81
Figure 4.14: Ancestry assignment for 26 regional populations, and proportions of hypothetical ancestries.	82

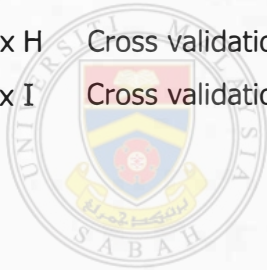
Figure 4.15:	Maximum likelihood tree of 26 regional populations, with assumption of no gene flow by TREEMIX program.	84
Figure 4.16:	Gene flow from North Borneo towards Cosmopolitan Malays of Singapore (MAS) was consistently detected.	85
Figure 5.1:	Map of distribution for 'Regional Population', particularly Southern China Minorities and all Southeast Asians of the meta-dataset.	100
Figure 5.2:	Observed heterozygosity of each individual population of the 89 worldwide populations.	108
Figure 5.3:	Phylogenetic relationship (maximum likelihood tree) of all 89 worldwide populations.	113
Figure 5.4:	Principal component analysis of 89 worldwide populations.	115
Figure 5.5:	Principal component analysis of 57 regional populations.	117
Figure 5.6:	Principal component analysis of 22 Austronesians groups.	118
Figure 5.7:	Assignment of genetic ancestry of 89 worldwide populations.	124
Figure 5.8:	Assignment of genetic ancestry of 57 regional populations.	125
Figure 5.9:	Assignment of genetic ancestry of 22 Austronesians groups.	126
Figure 6.1:	The new hypothesis postulates the origin of the North Borneans, and Borneo Island has served as human migration cross-road between mainland Southeast Asians and Austronesians.	140

LIST OF ABBREVIATIONS

SNP	-	Single nucleotide polymorphisms markers
BP	-	years before present
CE	-	common era
NB	-	North Borneo
SEA	-	Southeast Asia
ISEA	-	Islands/insular Southeast Asia
EA	-	East Asia
SC	-	Southern China
NE	-	Northeast Asia
HGDP	-	Human Genome Diversity Project
PASNP	-	Pan-Asia SNP Consortium
SGVP	-	Singapore Genome Variation Project
S5E	-	Sabah Five Ethnic Groups
DDS	-	Dusun
DRG	-	Rungus
DSO	-	Sonsogon
PSG	-	Sungai-Lingkabau
RPL	-	Murut-Paluan
MAS	-	Cosmopolitan Malays of Singapore
YRI	-	Yoruba in Ibadan, Nigeria
CEU	-	Utah residents with Northern and Western European ancestry
GIH	-	Gujarati Indians in Houston, Texas
CHB	-	Han Chinese in Beijing, China
JPT	-	Japanese in Tokyo, Japan
LD	-	Linkage disequilibrium

LIST OF APPENDIX

	Page	
Appendix A	A screenshot of the Genome Studio software	173
Appendix B	A screenshot of the .ped and .map file for the PLINK program	174
Appendix C	List of 63 populations by merging four datasets	175
Appendix D	Pairwise F_{ST} among selected Southeast Asia, Southern China and Northeast Asia populations, of the 63 populations meta-dataset	177
Appendix E	List of 89 populations by merging five datasets	178
Appendix F	Pairwise F_{ST} for selected populations in close geographical vicinity or similar in culture, for the 89 populations meta-dataset	181
Appendix G	Cross validation test for 89 worldwide populations	182
Appendix H	Cross validation test for 57 regional populations	183
Appendix I	Cross validation test for 22 Austronesians populations	184



UMS
UNIVERSITI MALAYSIA SABAH

CHAPTER 1

INTRODUCTION

1.1 Research Background

Sabah has more than 40 indigenous ethnic groups. Out of these, there are 32 indigenous Austronesian groups (King and King, 1984; Pugh-Kitingan, 2012). Linguistic study by Summer Institute of Language (SIL) International shows that the 'North Borneo' language stock, which is under the great Austronesian super family, extends from as far as the south islands of Philippines into the vast majority of Sabah and towards the interior lands of Sarawak and Kalimantan (Lewis *et al.*, 2015). The 'North Borneo' language stock in Sabah can be divided into three major groups, namely the Dusunic, Paitanic and Murutic families. The Dusunic family, which is the major population in the state, spans from the Northeast, Central and into the West Coast regions; the Paitanic family is concentrated in the interior lands of the East Sabah, and along the Kinabatangan river valley; whereas the Murutic family ranges from the interior lands into the South-West and expands all the way into the heartland of Borneo (Lewis *et al.*, 2015).

North Borneo (NB) is geographically the nearest to the Southern Philippines Islands, possibly serving as a viaduct of the 'Out-of-Taiwan' wave of human migration towards the archipelago of Islands Southeast Asia (Tabbada *et al.*, 2010). The lack of genetic data from Sabah presents a void in obtaining of a better picture on the migration history of the archipelago. As the peopling of Southeast Asia (SEA) and its archipelago by Austroasiatic-speakers and Austronesians is still under debate (Jinam *et al.*, 2012), it is possible that the extant populations of North Borneo may be the descendants of ancient populations that had undergone complex demographic history in the past.

The region of Southeast Asia (SEA) spans from the great landmass at the southeast end of the Eurasia continent, to the staggering islands, which makes up of eleven countries. The human population of this region composes of several hundreds of ethnolinguistic indigenous ethnic groups (Lewis, 2015). This profound diversity prompt debates pertaining to peopling of SEA due to non-conclusiveness and contradicting hypothesis in the previous reports. The commonly known ones are the Southern Route dispersal of the 'Out of Africa' theory (~40 kya) and Austronesian Expansion (~5 kya) (Macaulay *et al.*, 2005; Lipson *et al.*, 2014). Respectively, it postulated the origin of the Negrito natives found in Andaman Islands, Peninsular Malaysia, the Philippines, East Indonesia, Papuan New Guinea, Australia and Pacific Islands; and the widespread Austronesian languages from the most diverse in Taiwan, down to ISEA, to Madagascar in the West and to Pacific Islands in the East (Rasmussen *et al.*, 2011; Cox *et al.*, 2012). However, the contribution of ancestries originated from mainland Southeast Asia, and minorities groups from Southern China, was only hypothesized as 'Early Train' recently (Jinam *et al.*, 2012).

In accordance to multiple aspects of studies based on archaeological, linguistic, social-political and genetics findings, there were putatively four waves of human migration mediated by the anatomically modern human. This migration history was anticipated to be chronologically composed of the pre-historical 'Out of Africa' Negritos, Upper Palaeolithic/Mesolithic cultural diffusion, Austronesian expansion i.e. 'Out of Taiwan', and recent interaction of maritime kingdoms (Lansing *et al.*, 2011). As such, the Island Southeast Asia should be considered as a great melting pot of diversified genetic ancestries.

1.2 Research Problems

The island of Borneo flourishes with multiple ethnic groups of diverse languages, cultures and plausibly genetic entities across three countries i.e. Brunei, Indonesia (Kalimantan) and Malaysia (which includes the states of Sabah, formerly known as North Borneo, and Sarawak). Previous studies on genetic population structure using genome-wide SNPs panels often shows a sparse and limited sampling coverage from Borneo. Apart from the small number of Ibans and Bidayus from

Sarawak, Dayaks and an unnamed population from Kalimantan, there is no representative ethnic groups from Sabah (North Borneo) in these studies (Li *et al.*, 2008; HUGO Pan-Asia SNP Consortium, 2009; Xing *et al.*, 2009; Wollstein *et al.*, 2010).

Meanwhile, many genetic-based studies tend to group the Dusunic-speaking ethnic groups under an umbrella socio-political term 'Kadazandusun' (Teh *et al.*, 2014), which is the largest indigenous ethnic grouping of Sabah. Disregarding the plausible underlying genetic stratification among the linguistic-related ethnic groups may portray a risk in faulty interpretation in genome-wide association study of diseases among the ethnic groups in the future (Price *et al.*, 2008). As such, this study aimed to unravel the genetic population structure of the North Bornean natives and its genetic relatedness to other neighboring populations in the Southern China and Southeast Asian regions.

However, a recent report by the HUGO Pan-Asian SNP Consortium (2009), that performed genetic analysis with genome-wide SNPs, argued that the Southeast Asians should be ancestral to the North-East populations, such as Taiwan Natives and Han Chinese, instead of being descendants, as commonly believed. The inference was based on the findings that the SEAsians were the most genetically diverse than the North-East ones.

In contrary, there were previous reports based on patrilineal and matrilineal genetic analysis had contradictive inference against the consortium's one. They articulated that, the SEAsians, especially the islanders, had not only 20% of their variation derived from the putative 'Out of Taiwan' ancestry (Capelli, *et al.*, 2001; Hill *et al.*, 2007), but also had genetic contribution originated from Southern China or mainland South East Asia (Jinam *et al.*, 2012). Meanwhile, the Negritos from Philippines and Indonesians were not only found to be heavily admixed with Austronesians (Cox *et al.*, 2010; Delphin *et al.*, 2011), but only had their original languages shifted to the Austronesians'. This language replacement corroborated the 'Out of Taiwan' theory. More importantly, non-genetic studies based on archaeology, linguistic, material cultures and maritime trading activities, were all in

agreement with the 'Out of Taiwan' and 'Early Train' theories, that suggest Taiwan-origin and mainland SEA-origin of the Island Southeast Asians' ancestries (Bellwood and Dizon, 2005; Hung *et al.*, 2007; Friedlaender *et al.*, 2008; Gray *et al.*, 2009; Tabbada *et al.*, 2010; Jinam *et al.*, 2012; Ko *et al.*, 2014; Trejaut *et al.*, 2014).

It is impossible to arrive at a consensus standpoint, if the analysis were not based on comprehensive sampling of populations from the islands and landmass of SEA, Southern China and Taiwan. This is a prerequisite as it will provide the highest resolution of phylogeny and population genetic structure with the help of bioinformatic tools. As such, this study aimed to illustrate the genetic relationships of the population residing in this region, by performing analysis of genetic structure, admixture and population differentiation. This is achievable by merging the publicly available datasets from HapMap Project, Human Genome Diversity Project (HGDP), Singapore Genome Variation Project (SGVP), and the genotypes of five ethnic groups of Sabah which were generated in this study. With a combination of clustering-based analysis, phylogeny, population heterozygosity and differentiation, the direction of migration was then postulated. In addition, as there is no reported data from North Borneo before, whilst her geographical location is the closest to the Philippines, it is possible that this area serve as a entry/transition point for human migration.

1.3 Objectives

To address the current paucity of information on the genetic structure of North Borneans and its relevance to regional population genetics, the objectives of this study are:

1. To characterize distribution of allele frequencies, heterozygosity, linkage disequilibrium, genetic differentiation and phylogenetic relationships of five indigenous ethnic groups of Sabah, i.e. Dusun, Rungus, Sonsogon, Sungai-Lingkabau and Murut-Paluan, through array-based genome-wide high density single nucleotide polymorphisms (SNPs) markers.

2. To characterize the genetic structure of five indigenous ethnic groups of Sabah via principal component analysis, and genetic component and admixture analysis.

3. To postulate the genetic relationship of the five indigenous ethnic groups of Sabah to regional and worldwide populations, as inferred by population metrics, genetic structure and estimation of gene flow, which are derived from comprehensive meta-analysis of public datasets.



UMS
UNIVERSITI MALAYSIA SABAH