

ห้องสมุดงานวิจัย สำนักงานคณะกรรมการการวิจัยแห่งชาติ



E42153



GENETIC ANALYSIS AND PEOPLING OF THAI
POPULATIONS

MS. SATTARA HATTIRAT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
(BIOINFORMATICS AND SYSTEMS BIOLOGY)
SCHOOL OF BIORESOURCES AND TECHNOLOGY AND
SCHOOL OF INFORMATION TECHNOLOGY
KING MONCKUT'S UNIVERSITY OF TECHNOLOGY THONEURI

2010



E42153

Genetic Analysis and Peopling of Thai Populations

Ms. Sattara Hattirat B.Sc (Biology)

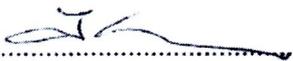
A Thesis Submitted in Partial Fulfillment of the Requirements for
The Degree of Master of Science (Bioinformatics and Systems Biology)
School of Bioresources and Technology and School of Information Technology
King Mongkut's University of Technology Thonburi
2010

Thesis Committee

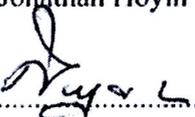



.....
(Asst. Prof. Kanokwan Poonputsa, Ph.D.)

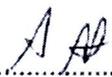
Chairman of Thesis Committee


.....
(Asst. Prof. Jonathan Hoyin Chan, Ph.D.)

Member and Thesis Advisor


.....
(Researcher. Sissades Tongsim, Ph.D.)

Member and Thesis Co-Advisor


.....
(Researcher. Apichart Intarapanich, Ph.D.)

Member


.....
(Researcher, Philip James Shaw, Ph.D.)

Member

Copyright reserved

Thesis Title	Genetic Analysis and Peopling of Thai Population
Thesis Credit	12
Candidate	Miss Sattara Hattirat
Thesis Advisor	Asst. Prof. Dr. Jonathan Hoyin Chan
Thesis Co-advisor	Dr. Sissades Tongsim
Program	Master of Science
Field of Study	Bioinformatics and Systems Biology
Faculty	School of Bioresources and Technology and School of Information Technology
B.E.	2553

Abstract

 E 42153

This thesis aims to analyze the genetic data of recent Thai populations, using Single Nucleotide Polymorphism (SNP) markers. The genetic data are compared to indigenous populations in Thailand and recent populations in neighboring countries.

Approaches used in the analyses include population genetics, statistics, non-parametric clustering, model-based estimation of population genetic admixture and haplotype analysis. The analyses are geared toward a better understanding of the genetic structure and genetic diversity of Thai populations in four geographical regions of Thailand: the North, the Central region, the Northeast and the South. Historical aspects and ethnolinguistic backgrounds of Thai population are investigated to provide comparative insights.

It is found that the genetic patterns of recent Thai people in the four regions are highly admixed and diverse. There are discreet substructures in the Thai populations which require further subsampling. The genetic profiles of Thai people, composing of shared ancestral components with Southern Chinese, Iban and Indian populations, are similar to those of Khmer people. The genetic profiles of Ayutthaya populations are close to Thai Mon. This suggests genetic assimilation of early Austroasiatic speaking populations into the newly arrived Tai populations.

Keywords: Population Structure/ Genetic Admixture/ Population Genetics/
Demographic History/ Ethnicity/ Ethnolinguistics/ Mainland Southeast Asia

หัวข้อ	การวิเคราะห์ทางพันธุศาสตร์ประชากรในคนไทย
หน่วยกิต	12
ผู้เขียน	นางสาว ศรัทธา หัตถิรัตน์
อาจารย์ที่ปรึกษา	ผศ. ดร. โจนathan โสอิน ชาน
อาจารย์ที่ปรึกษา (ร่วม)	ดร. ศิษณุ ทงสิมา
หลักสูตร	วิทยาศาสตร์มหาบัณฑิต
สาขาวิชา	ชีวสารสนเทศและชีววิทยาระบบ
คณะ	ทรัพยากรชีวภาพและเทคโนโลยี และ คณะเทคโนโลยีสารสนเทศ
พ.ศ.	2553

บทคัดย่อ

E 42153

งานวิจัยนี้มุ่งวิเคราะห์ข้อมูลทางพันธุกรรมของประชากรไทยในปัจจุบัน โดยใช้เครื่องหมายทางพันธุกรรมชนิดซิงเกิลนิวคลีโอไทด์โพลิมอร์ฟิซึม (Single Nucleotide Polymorphism หรือ SNP) โดยเปรียบเทียบกับข้อมูลทางพันธุกรรมของคนพื้นถิ่นในประเทศไทย และประชากรปัจจุบันในประเทศไทยและประเทศใกล้เคียง วิธีศึกษาที่ใช้ประกอบด้วยวิธีการทางพันธุศาสตร์ประชากร สถิติ การแบ่งกลุ่มแบบไม่อิงพารามิเตอร์ (non-parametric clustering) การประเมินสัดส่วนการผสมกันทางพันธุกรรมของประชากรซึ่งอิงโมเดลทางชีววิทยา (model-based estimation of population genetic admixture) และการวิเคราะห์แฮปโลไทป์ (haplotype) เพื่อให้เกิดความเข้าใจโครงสร้างและความหลากหลายทางพันธุศาสตร์ของประชากรไทยในสี่ภาคของประเทศ คือ ภาคเหนือ กลาง ตะวันออกเฉียงเหนือ และภาคใต้ โดยนำแง่มุมทางประวัติศาสตร์และภาษาศาสตร์ชาติพันธุ์มาเปรียบเทียบเพื่อการนี้ด้วย ผลการวิจัยพบว่ารูปแบบทางพันธุกรรมของประชากรไทยในปัจจุบันทั้งสี่ภาคมีความหลากหลายทั้งระหว่างภาคและภายในภาคด้วยตนเอง แสดงให้เห็นถึงโครงสร้างย่อยที่ซ่อนอยู่ (discreet substructures) ภายในกลุ่มประชากรไทย ควรเก็บข้อมูลทางพันธุศาสตร์เป็นกลุ่มที่ย่อยลงไปมากขึ้นเมื่อทำการศึกษาทางพันธุศาสตร์การแพทย์หรือทางพันธุศาสตร์ประชากรต่อไป นอกจากนี้ คนไทยมีโครงสร้างทางพันธุกรรมบางส่วนร่วมกับตัวอย่างประชากรจากจีน คนไอบาน (มาเลเซีย) และอินเดีย ซึ่งเป็นลักษณะ (profile) ทางพันธุกรรมที่คล้ายกับคนเขมรมาก ในขณะที่คนพื้นเมืองในเขตภาคกลาง เช่น อุรุษยาและจังหวัดโดยรอบ มีลักษณะทางพันธุกรรมที่ใกล้เคียงกับคนพื้นเมืองมอญ ซึ่งให้เห็นความสำคัญของการผสมผสานทางพันธุกรรมของกลุ่มประชากรที่พูดภาษากลุ่มออสโตรเอเชียติก (Austroasiatic language family) และกลุ่มประชากรที่พูดภาษากลุ่มไต (Tai)

คำสำคัญ: โครงสร้างประชากร/ การผสมทางพันธุกรรม/ พันธุศาสตร์ประชากร/ คนพื้นถิ่น/
ประวัติศาสตร์ประชากร/ ภาษชาติพันธุ์/ เอเชียตะวันออกเฉียงใต้

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my two supervisors, Asst. Prof. Jonathan Hoyin Chan of the School of Information Technology, the Chairperson of Computer Science (B.Sc.) program, King Mongkut's University of Technology Thonburi, and Dr. Sissades Tongsimma, Head of Bioinformatics and Biostatistics Laboratory at Genome Institute, National Center for Genetic Engineering and Biotechnology (BIOTEC), for their detailed and constructive comments. Their logical ways of thinking and insights in computational methods have been of great value to me. Their understanding, encouragement and personal guidance have provided a good basis for the present thesis. I am also grateful to Dr. Apichart Intarapanich for his wise ways around computational methods and mathematics and Dr. Philip J. Shaw for his sound and constructive comments for my work. I always enjoy and benefit from working with Asst. Prof. Kanokwan Poomputsa, a beloved and respected faculty at School of Bioresources and Technology.

During this work I have collaborated with many colleagues for whom I have great regards for. I owe my utmost sincere gratitude to Mr. Chumpol Ngampew for his patient guidance on python programming which is fundamental in my work. Without his praised programming experience and teaching skills, this work could not have become successful. I warmly thank Mr. Pongsakorn Wangkumhang for his MATLAB skills and friendly help on technical issues. His shared insights for discussion are much appreciated. My sincere thanks are due to Dr. Anunchai Assawamakin for his unyielding stimulation for me to keep working. He guided me through the most difficult initial period of my work with his sharp insights and encouragements. I am also thankful to my friends and families whose mental support makes the unavoidable hardship endurable and the working life enjoyable. The financial support of King Mongkut's University of Technology Thonburi and BIOTEC is gratefully acknowledged.

CONTENTS

	PAGE
ENGLISH ABSTRACT	i
THAI ABSTRACT	ii
ACKNOWLEDGMENT	iii
CONTENTS	iv
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER	
1. INTRODUCTION	1
1.1 Rationale	1
1.2 Scope	4
1.3 Goals	4
1.4 Objectives	4
2. BACKGROUND	5
2.1 Human genetic variation	5
2.2 Genetic admixture	10
2.3 Allele frequencies	10
2.4 F_{ST}	10
2.5 Language families in mainland Southeast Asia	11
3. METHODOLOGIES	12
3.1 Samples	12
3.2 Data preparation	13
3.3 Principle Component Analysis (PCA) and iterative pruning PCA (ipPCA)	14
3.4 ADMIXTURE analysis	14
3.5 Distance analysis	15
3.6 F_{ST}	16
3.7 Individual tree reconstruction	16
3.8 Haplotype inference	16
3.9 Haplotype analysis	17
4. RESULTS	18
4.1 Thai genetic structure comparing with global and Eurasian populations	18
4.2 Thai genetic structure comparing with Southeast and East Asian populations	24
4.3 Genetic structure within Thai population	27
4.4 Haplotype sharing	35
5. DISCUSSION	37
5.1 Genetic characteristics of Thai people	37
5.2 Comparative population genetic structures: the Thai with the others	38
5.3 Comparison to mtDNA and NRY studies	40
6. CONCLUSION	47

CONTENTS

	PAGE
REFERENCES	48
APPENDIX	
A PROTOCOLS FOR GENETIC DATA PREPARATION	55
A.1 PED FILES FOR 2 COMBINED GENOTYPE DATA (USED FOR ADMIXTURE ANALYSIS)	55
A.2 ENCODED GENOTYPE FILES FOR PCA	55
A.3 Files for AWClust	55
B PYTHON CODES	56
B.1 Querying and encoding into ped file format	56
B.2 Checking for matching strands with querying	58
B.3 Checking for matching SNPs from 2 files, changing into compatible strands	58
B.4 Combining overlapping SNP data from 2 genotype files	59
CIRRICULUM VITAE	61

LIST OF TABLES

TABLE		PAGE
3.1	Original sets of samples used in this study	13
3.2	Sets of combined samples and overlapping markers used in all analyses	13
4.1	Pairwise analysis of the norms of ancestry estimates (vectors) in each population for the larger set of 992 Thai individuals and denser markers of 553,892 SNPs.	33
4.2	Pairwise analysis of the norms of ancestry estimates (vectors) in each population for a set of 1,024 Thai individuals (the dataset above plus 32 Thai Ayutthaya) and less dense overlapping markers of 39,049 SNPs.	33
4.3	Pairwise analysis of the norms of ancestry estimates (vectors) in each ethnic population for a set of 337 individuals, 12,375 SNPs.	34
4.4	Pairwise FST of Thai populations from four regions of Thailand.	35
4.5	Number of haplotypes in chromosome 21 and haplotype sharing in populations with >20 individual samples.	35
4.6	Number of haplotypes in chromosome 20 and haplotype sharing in populations with >20 individual samples.	36
5.1	Genetic studies on the populations in Thailand	41
5.2	Relevant genetic studies of the peopling of Thailand	44

LIST OF FIGURES

FIGURE	PAGE
2.1 Linkage disequilibrium around an ancestral mutation (triangle).	7
2.2 The erosion of linkage disequilibrium by recombination.	8
2.3 Formation of a hybrid population or genetic admixture	10
2.4 Ethnolinguistic pattern of modern Southeast Asia	11
3.1 Illustration of how ADMIXTURE identify and cluster individuals into groups of populations.	15
4.1 PCA and EM clustering (ADMIXTURE, K=13) of Thai (Ayutthaya and surrounding cities in Central area) and 40 populations from Xing <i>et al.</i> , 2010 which also include a Thai population from the South (Phuket and Moken).	19
4.2 ADMIXTURE analysis of global individuals, including the new Thai samples from Ayutthaya and surrounding cities in Central area, from K = 3-13.	20
4.3 Zoom-in graph illustrating the admixture patterns of indigenous populations including Ayutthaya and surrounding cities in the Central region of Thailand.	21
4.4 ADMIXTURE analysis (K = 2-10) of Eurasian and some Polynesian individuals from Xing <i>et al.</i> (2010) and the new indigenous Thai samples from the central part of Thailand.	22
4.5 Individual phylogenetic tree constructed from allele sharing distance calculated from all individuals with neighbor joining method.	23
4.6 Expected heterozygosity calculated from genotype data from each population reported by ThaiSNPdb including 32 Thai data (THAI) and the data from 11 HapMap populations.	24
4.7 Non-parametric analysis—ipPCA iteration 3 (A) and iteration 4 (B)—of genetic variance in Southeast Asian populations.	25
4.8 Principal component analysis of some genetic variation among populations whose global results show close genetic proximity to Thai populations (A) and phylogenetic Tree for 8 populations (2 Chinese, 2 Thai, Khmer, Vietnamese, Iban and Bhramin Indian) (B).	26
4.9 Averaged (A) and Individual (B) ancestral estimation for Thai and populations which have shared ancestry in the larger set of samples.	27
4.10 PCA featuring a clinal pattern of genetic variation among Thai individuals in accordance to geography of Thailand (A) and PCA of a larger set of individuals and SNP markers featuring discernible patterns of genetic variation consistent with the four geographical regions of Thailand (B).	28
4.11 PCA with different sets of Thai populations: indigenous Thai population in central region (Ayutthaya) and southern region (S Xing) and recent populations in all regions of Thailand (N, NE, C, BKK, S).	29
4.12 Phylogenetic tree reconstructed from individual allele sharing distance from recent Thai individuals from four regions of Thailand	30
4.13 Phylogenetic tree reconstructed from individual allele sharing distance from the Northern, the Northeastern and the Southern Thai populations	30

LIST OF FIGURES

FIGURE		PAGE
4.14	Phylogenetic tree reconstructed from individual allele sharing distance of recent Thai individuals from 4 regions of Thailand and individuals from other closely related populations in East and South Asia.	31
4.15	Ancestry estimation of recent Thai individuals who live in the North, the Northeast, the Central area and the South of Thailand with two subsampling of the people in Bangkok and Ayutthaya and surrounding cities, ADMIXTURE K = 2-5, 455 individuals, 39, 049 SNPs.	32
4.16	Ancestry estimation of recent Thai individuals in different geographical regions, ADMIXTURE K = 3 at 39,049 SNPs, 1,024 individuals.	33
4.17	Averaged ancestry estimation of Thai populations in Bangkok, the North, the Northeast, the South and the Central are of Thailand with pooled standard deviations and variance of the genetic component values among individuals in each region.	32
4.18	Averaged ancestry estimation Thai Ayutthaya, other ethnic populations in Thailand and one ethnic population from Indonesia (IN-MT).	34