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 MONGKUT'S UNIVERSITY OF TECHNOLOGY THONBURI
2010

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 Tongsima, Ph.D.) Member and Thesis Co-Advisor

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

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

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 Tongsima
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

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 ศรัทธารา หัตถีรัตน์
อาจารย์ที่ปรึกษา	ผศ. ดร. โจนาธาน โฮยิน ชาน
อาจารย์ที่ปรึกษา (ร่วม)	ดร. ศิษเฎศ ทองสิมา
หลักสูตร	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	ชีวสารสนเทศและชีววิทยาระบบ
คณะ	ทรัพยากรชีวภาพและเทคโนโลยี และ คณะเทคโนโลยีสารสนเทศ
พ.ศ.	2553

บทคัดย่อ

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

คำสำคัญ : โครงสร้างประชากร/ การผสมทางพันธุกรรม/ พันธุศาสตร์ประชากร/ คนพื้นถิ่น/ ประวัติศาสตร์ประชากร/ ภาษาชาติพันธุ์/ 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 Tongsima, 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

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES