Thesis Title Ethnicity Prediction of Population in Northern Thailand Using Single Nucleotide Polymorphism Data

Author Miss. Massupa Kaewgahya

Degree Master of Science (Bioinformatics)

Thesis Adivisory Committee Assoc.Prof.Dr. Daoroong Kangwanpong Advisor

Assoc.Prof.Dr. Jeerayut Chaijaruwanich Co-advisor

Dr. Jatupol Kampuansai Co-advisor

ABSTRACT

Northern Thailand had a long history, leading to the diverse social structure with mixed culture of people from many different cities. Since the past till the present day, a lot of population migrations from neighboring countries, crossing the northern border area, still continuously occurred. Thus these populations movement cause the diversity of ethnic groups in North Thailand. Each ethnic group has its own culture and tradition, as well as the language, indicating the ethnic uniqueness.

In 2009, there was a published research work from HUGO Pan Asian SNP on using SNPs in human genome of the Southeast Asian populations, including the northern Thai populations, for their genetic relationship study. Since the SNP can be used to determine the different between any two unrelated individuals, it should be assumed that some numbers of SNPs might be specified to each ethnic group. Thus in this thesis, the computational concept was applied to select the specific SNP which can be used as the genetic marker for identifying each population in the upper northern part of Thailand.

In this thesis, the Mutual Information (MI) principle was employed to rank the 58960 SNP loci, genotyping by the Affymetrix SNPArray 50K Xba, in 256 unrelated individuals from 13 ethnic groups, comprising Karen, Hmong, Yao, Lua, H'tin, Mlabri, Mon, Paluang, Plang, Yuan, Yong, Lue and Khuen. The calculated MI value was descending ranked, and the specific SNPs were selected for the ethnic group differentiation, using the decision tree classification model. To test the discrimination efficiency of the selected group of SNPs, the correspondence analysis was used. The genetic distances among populations were analyzed and then the phylogenetic tree was reconstructed.

The result showed that, when the frequencies of SNPs, ranked from calculated MI values, were considered, the top 100 SNP loci of each ethnic group could be selected. The subsets of 10, 20, … 100 were then created and used as the training and testing data in decision tree model. The top 60 SNP loci of each ethnic group gave the highest classification accuracy value of 89.84 %. When the SNP genotype frequency was used in the correspondence analysis and the graph of the result was plotted, almost all ethnic groups could be discriminated by the selected SNPs, except the Karen (KA), Tai Yong (TY) and Lawa (LW) which were very close together. Genetic relationships among populations were consistent within the Tai-Kadai speaking groups only, testing by the genetic distance and the phylogenetic tree reconstruction.

This study demonstrated the computational application with genetic data. All employed methods could be used to select a small number of specific SNPs for genetic relationship of the populations with satisfying result. In addition, these results can also be beneficial for further population genetic studies.

ชื่อเรื่องวิทยานิพนธ์ การทำนายชาติพันธุ์ของประชากรในภาคเหนือของไทย โดยใช้ข้อมูลภาวะพหุสัณฐานของนิวคลีโอไทด์เดี่ยว

ผู้เขียน นางสาวมาสสุภา แก้วกายา

ปริญญา วิทยาศาสตรมหาบัณฑิต (ชีวสารสนเทศศาสตร์)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์

รศ.ดร. ดาวรุ่ง กังวานพงศ์	อาจารย์ที่ปรึกษาหลัก
รศ.ดร. จีรยุทธ ไชยจารุวณิช	อาจารย์ที่ปรึกษาร่วม
ดร. จตุพล คำปวนสาย	อาจารย์ที่ปรึกษาร่วม

บทคัดย่อ

ภาคเหนือของประเทศไทยมีประวัติศาสตร์อันยาวนาน ก่อให้เกิดโครงสร้างของสังคมที่มีความหลากหลาย มีการผสมผสานกันทางวัฒนธรรมของผู้คนที่มาจากหลายบ้านต่างเมือง ตั้งแต่อดีตจนถึงปัจจุบันยังคงมีการเคลื่อนย้ายถิ่นฐานของประชากรจำนวนมากจากประเทศเพื่อนบ้าน ซึ่งข้ามพรมแดนไทยในบริเวณภาคเหนือตอนบน นำไปสู่การเปลี่ยนแปลงด้านประชากรอย่างต่อเนื่อง มีผลให้เกิดความหลากหลายของกลุ่มชาติพันธุ์ในประชากรของภาคเหนือ โดยที่แต่ละชาติพันธุ์ต่างล้วนมีเอกลักษณ์ทางวัฒนธรรม ประเพณี และภาษาพูด ที่บ่งบอกถึงความเป็นกลุ่มชนของตนเองได้เป็นอย่างดี

ในปี ค.ศ. 2009 ได้มีงานวิจัยตีพิมพ์ของ HUGO Pan Asian SNP ซึ่งวิเคราะห์สนิปส์ในจีโนมของประชากรต่างๆ ในภูมิภาคเอเชียตะวันออกเฉียงใต้ รวมถึงประชากรในภาคเหนือของประเทศไทย เพื่อศึกษาความสัมพันธ์ระหว่างประชากร ประกอบกับการที่สนิปส์มีคุณสมบัติซึ่งสามารถระบุความแตกต่างระหว่างสองบุคคลใดๆ ที่ไม่มีความเกี่ยวข้องกันได้ จึงคาดว่าน่าจะมีสนิปส์อยู่จำนวนหนึ่งที่แสดงถึงความจำเพาะสำหรับแต่ละกลุ่มประชากร งานวิจัยนี้จึงประยุกต์ความคิดในเชิงการคำนวณ เพื่อค้นหาสนิปส์ที่มีความจำเพาะและสามารถใช้เป็นเครื่องหมายทางพันธุกรรมในการแยกแยะแต่ละกลุ่มชาติพันธุ์ในภาคเหนือตอนบนของไทยได้

งานวิจัยนี้ได้ใช้หลักการการเกิดขึ้นร่วมกันของข้อมูล (MI) ในการจัดอันดับสนิปส์จำนวน 58,960 ตำแหน่ง ซึ่งได้จากการทำจีโนไทป์ด้วย Affymetrix SNPArray 50K Xba ในประชากรจำนวน 256 คน จาก 13 กลุ่มชาติพันธุ์ ได้แก่ กะเหรี่ยง ม้ง เย้า ลัวะ ถิ่น มลาบรี มอญ ปะหล่องพล่าง ไทยวน ไทยอง ไทลื้อ และไทเขิน จากนั้นจัดเรียงลำดับสนิปส์ตามค่า MI จากสูงไปหาต่ำ ในการคัดเลือกสนิปส์ที่มีความจำเพาะ เพื่อนำไปจำแนกกลุ่มชาติพันธุ์ดังกล่าวด้วยแบบจำลองวิธีการต้นไม้การตัดสินใจ แล้วจึงทดสอบประสิทธิภาพในการแยกแยะกลุ่มประชากรด้วยสนิปส์ที่คัดเลือกนั้น ด้วยการวิเคราะห์การสมนัย พร้อมทั้งศึกษาความสัมพันธ์ทางพันธุกรรมด้วยการคำนวนค่าระยะห่างทางพันธุกรรมและการวิเคราะห์ phylogenetic tree

ผลการศึกษาพบว่า เมื่อเรียงลำดับสนิปส์จากค่า MI ที่คำนวนได้ แล้วพิจารณาความถี่ของสนิปส์ในแต่ละช่วงของค่า MI จะสามารถคัดเลือกสนิปส์ได้ 100 ตำแหน่งแรกของแต่ละกลุ่มประชากร เมื่อแบ่งสนิปส์เป็นชุดย่อยๆ ตามจำนวน 10, 20, , 100 แล้วใช้เป็นข้อมูลสอนให้กับแบบจำลองต้นไม้การตัดสินใจ ซึ่งพบว่าสนิปส์ 60 ตำแหน่งแรกของแต่ละกลุ่มประชากรให้ค่าความถูกต้องในการจำแนกกลุ่มสูงที่สุด คือ ร้อยละ 89.94 เมื่อนำข้อมูลความถี่จีโนไทป์ของสนิปส์ดังกล่าวไปวิเคราะห์การสมนัยและนำผลการวิเคราะห์มาสร้างกราฟ จะเห็นได้ว่าสนิปส์ที่คัดเลือกมานั้น สามารถจำแนกประชากรเกือบทุกกลุ่มออกจากกันได้อย่างชัดเจน โดยมีเพียงกลุ่มชาติพันธุ์ กะเหรี่ยง (KA) ไทยอง (TY) และ ลัวะ (LW) เท่านั้น ที่มีความใกล้เคียงกันมาก และเมื่อดูความสัมพันธ์ระหว่างประชากรโดยใช้การคำนวนค่าระยะห่างทางพันธุกรรมและการวิเคราะห์ phylogenetic tree พบว่ามีความสอดคล้องกับการจัดกลุ่มประชากรตามตระกูลภาษาเฉพาะในกลุ่มตระกูลที่พูดภาษาไท-กะได

การศึกษาครั้งนี้แสดงให้เห็นถึงการนำแนวคิดเชิงการคำนวนมาปรับใช้กับข้อมูลทางพันธุศาสตร์ โดยวิธีการทั้งหมดนี้สามารถเลือกใช้สนิปส์ที่จำเพาะและมีศักยภาพในการแยกแยะเพียงจำนวนน้อยตำแหน่ง ก็สามารถศึกษาความแตกต่างทางพันธุกรรมของประชากรได้ในระดับที่น่าพอใจ จึงคาดว่า ผลการศึกษาครั้งนี้จะเป็นข้อมูลพื้นฐานในการนำไปประยุกต์ใช้กับการศึกษาด้านพันธุศาสตร์ประชากรต่อไป