

ฐานข้อมูลความหลากหลายทางพันธุกรรมในคนไทยแบบสลิปและซีเอ็นวี

อนันต์ชัย อัสวามะสิน¹ ชุมพล งามพิว¹ อรนุช ประดิษฐ์ทรัพย์² ชนินทร์ ลิ้มวงศ์^{2,3} และ ศิษฏา ทองสิมา^{1*}

Anunchai Assawamakin¹, Chumpol Ngamphiw¹, Oranud Praditsap², Chanin Linwongse^{2,3} and Sissades Tongsim^{1*}

¹ห้องปฏิบัติการชีวสถิติและสารสนเทศ สถาบันจีโนม ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ อุทยานวิทยาศาสตร์ประเทศไทย คลองหลวง ปทุมธานี 12120

²หน่วยอนุพันธุศาสตร์ สถานส่งเสริมการวิจัย ³หน่วยเวชพันธุศาสตร์ ภาควิชาอายุรศาสตร์ คณะแพทยศาสตร์ ศิริราชพยาบาล มหาวิทยาลัยมหิดล บางกอกน้อย กรุงเทพฯ 10700

¹Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand

²Division of Molecular Genetics, Department of Research and Development, ³Division of Medical Genetics, Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkoknoi, Bangkok 10700, Thailand

* Corresponding author: sissades@biotec.or.th

บทคัดย่อ

การศึกษาความสัมพันธ์ระหว่างความหลากหลายทางพันธุกรรม (genetic polymorphisms) ของจีโนมมนุษย์ซึ่งเป็นเอกลักษณ์ของแต่ละบุคคล เช่น สลิป (single nucleotide polymorphism, SNP) และ ซีเอ็นวี (copy number variation, CNV) กับการพยากรณ์ความเสี่ยงต่อการเกิดโรค ความต้านทานโรค การตอบสนองต่อยา สามารถนำไปใช้เลือกชนิดและขนาดยาที่เหมาะสม รวมทั้งการค้นหายาและพัฒนาวัคซีนใหม่ที่เหมาะสมเฉพาะกลุ่มประชากร การประยุกต์ใช้ความรู้ทางระบาดวิทยาพันธุศาสตร์ (genetic epidemiology) การวิจัยด้านพันธุศาสตร์ในประเทศไทยมีความจำเป็นในการศึกษาชนิดความถี่ของแอลลีลต่างๆ บนยีน เพื่อใช้เป็นข้อมูลพื้นฐานในการ

ประเมินขนาดตัวอย่าง และหลีกเลี่ยงการทำซ้ำในสลิปที่ไม่มีความหลากหลาย งานวิจัยนี้มีจุดประสงค์ในการจัดสร้างฐานข้อมูลความหลากหลายทางพันธุกรรมในคนไทยทั้งจีโนม เพื่อประโยชน์ในงานวิจัยด้านมนุษยพันธุศาสตร์ และระบาดวิทยาพันธุศาสตร์ โดยใช้เทคโนโลยี Affymetrix SNP genotyping 500K array เพื่อระบุจีโนมไทยและความถี่ของสลิปและซีเอ็นวี และนำมาเปรียบเทียบกับฐานข้อมูลอ้างอิงสาธารณะอื่นๆ เช่น NCBI dbSNP, Pan-Asian SNP Database, International HapMap Project, JSNP และ Database of Genomic Variant โดยข้อมูลของคนไทยและข้อมูลจากฐานข้อมูลสาธารณะเหล่านี้ ได้ถูกจัดเก็บในฐานข้อมูลซึ่งสามารถเข้าใช้ผ่านทางเครือข่ายอินเทอร์เน็ตที่ <http://www.biotec.or.th/thaisnp2>

Abstract

The determination of association between single nucleotide polymorphisms (SNPs) and copy number variations (CNV) and the risks of development of diseases will be useful in development of predictive system based on genetic makeup of individual. Moreover, such polymorphisms can also be used in refining the efficacy and safety of drugs which are different amongst people. To identify genetic risks of common diseases, the information of linkage disequilibrium patterns in the study population is needed for estimation of sample sizes and averting genotyping of non-polymorphic markers in the study population. This information will particularly be very important for genotype-phenotype association. This project aims to determine SNPs and CNVs in Thai population by using 5th generation Affymetrix SNP genotyping arrays, each array contains approximately 500,000 known SNP genotyping positions. Furthermore, this analysis offers genetic comparative analysis of Thais against other NCBI dbSNP, Pan-Asian SNP Database, International HapMap Project, JSNP and Database of Genomic Variant. The genotypic information is made available for public access at <http://www.biotech.or.th/thaisnp2>.

คำสำคัญ: สนิป, ซีเอ็นวี, ไทย, ความหลากหลายทางพันธุกรรม

Keywords: SNPs, CNV, Thais, genetic polymorphism

บทนำ

ในยุคหลังจีโนมิกส์ความก้าวหน้าในด้านอนุพันธุศาสตร์และเวชพันธุศาสตร์ ที่เป็นผลมาจากโครงการศึกษาจีโนมมนุษย์ (Human Genome Project) ทำให้ความรู้ในด้านพยาธิกำเนิดระดับโมเลกุล (molecular pathogenesis) ความรู้เรื่องยีนก่อโรค และ

ยีนก่อความเสี่ยงต่อโรค (predisposing gene) มีมากขึ้น เป็นที่คาดกันว่าความรู้ดังกล่าว จะเป็นประโยชน์โดยตรงต่อการรักษาโรคในเวชปฏิบัติ และจะนำไปสู่การพัฒนาวิธีการรักษาโรคที่จำเพาะ สำหรับพื้นฐานทางพันธุกรรมของแต่ละบุคคล (individualized medicine หรือ personalized medicine) การศึกษาความสัมพันธ์ของความหลากหลายของจีโนมมนุษย์ที่เป็นเอกลักษณ์ของแต่ละบุคคลกับขบวนการต่างๆ ครอบคลุมตั้งแต่การพยากรณ์ความเสี่ยงต่อการเกิดโรค การป้องกันโรค การตอบสนองต่อยา การเลือกใช้ยา และขนาดยาที่เหมาะสม การค้นหายาใหม่ การพัฒนายาใหม่และการค้นหาขบวนการใหม่ที่เหมาะสมเฉพาะกลุ่มประชากร ซึ่งสิ่งเหล่านี้มาจากองค์ความรู้พื้นฐานว่ามนุษย์แต่ละคนมีความแตกต่างของรหัสพันธุกรรมในยีนที่เกี่ยวข้องกับวิถีพยาธิกำเนิดของโรค ยีนที่กำหนดปัจจัยทางเภสัชพลศาสตร์และเภสัชจลนศาสตร์ที่แตกต่างกัน การทราบถึงความแตกต่างเหล่านี้โดยละเอียด ย่อมนำไปสู่การเลือกใช้ยาตามความเหมาะสมกับโรค ระยะและความรุนแรงของโรค สาเหตุของโรค ลักษณะทางชีววิทยาของผู้ที่จะได้รับยา และขนาดยาที่เหมาะสม เพื่อให้เกิดประโยชน์ในการรักษาสูงสุด และลดอาการอันไม่พึงประสงค์จากการใช้ยากับผู้ป่วย ในปัจจุบันการศึกษาทางระบาดวิทยาพันธุศาสตร์เป็นการพิสูจน์แนวคิดที่ผลของความหลากหลายทางพันธุกรรม สัมพันธ์กับลักษณะต่างๆ ทางคลินิก การศึกษาเกี่ยวกับความหลากหลายทางพันธุกรรมของยีน สามารถแบ่งออกเป็น 2 แบบ ได้แก่ การศึกษาเกี่ยวกับผลของความหลากหลายทางพันธุกรรม ต่อความโน้มเอียงต่อการเกิดโรค (disease susceptibility) และการศึกษาเกี่ยวกับผลของความหลากหลายทางพันธุกรรม กับการตอบสนองต่อยา (drug responsiveness) โดยความแตกต่างทางพันธุศาสตร์ทั้งหมด เป็นผลประกอบขึ้นจากยีนหลายยีน หรือยีนกับสิ่งแวดล้อม ที่เรียกว่า โรคพหุปัจจัย

ข้อมูลจากโครงการศึกษาจีโนมมนุษย์ นำมาสู่โครงการค้นหาสนิป (single nucleotide polymorphism,

SNP) ของประชากรเชื้อชาติต่างๆ ในโครงการ International HapMap Project (HapMap) (International HapMap Consortium, 2007) และความสามารถในการทำจีโนมไทป์สไนป์ที่มากขึ้นนำมาสู่การค้นหายีน (copy number variation, CNV) ที่เกิดการขาดหายและแทรกเข้าของชิ้นส่วนดีเอ็นเอขนาดใหญ่ ซึ่งมีการศึกษาถึงผลของยีนดังกล่าวต่อการเกิดโรค (Bora *et al.*, 2007; Estivill *et al.*, 2007; Lachman *et al.*, 2007; Sebat *et al.*, 2007) และเป็นความหลากหลายขั้นพื้นฐานของมนุษย์ ข้อมูลความหลากหลายทั้งสองชนิด เมื่อนำมาใช้ในการศึกษา จะช่วยเพิ่มโอกาสและความรวดเร็ว ในการทำความเข้าใจเกี่ยวกับความสัมพันธ์ของสารพันธุกรรม กับกระบวนการเปลี่ยนแปลงต่างๆ ในร่างกายที่มีผลต่อการเกิดโรคและความรุนแรงของโรค ซึ่งจะนำไปสู่การค้นพบยาใหม่สำหรับการรักษาโรค รวมทั้งการวินิจฉัยเพื่อการทำนายโอกาสที่จะเป็นโรค ทำให้เกิดการวางแผนการป้องกัน และรักษาโรคที่มีความถูกต้องมากขึ้น การประยุกต์ใช้ความรู้ทางระบาดวิทยาพันธุศาสตร์ในประเทศไทยมีความจำเป็นอย่างยิ่งที่จะต้องมีการวิจัยทางด้านพันธุศาสตร์เกี่ยวกับชนิด ความถี่ของแอลลีลต่างๆ บนยีนในคนไทย จากผลสืบเนื่องของโครงการศึกษาจีโนมมนุษย์ ได้มีการจัดเก็บข้อมูลความหลากหลายทางพันธุกรรมของมนุษย์ในโครงการ NCBI dbSNP และได้เริ่มมีการวิจัยทางระบาดวิทยาพันธุศาสตร์ในแต่ละประเทศ ได้แก่ JSNP Project จากประเทศญี่ปุ่น (Haga *et al.*, 2002; Hirakawa *et al.*, 2002) และ International HapMap Project (International HapMap Consortium, 2007) เป็นต้น ส่วนในประเทศไทยได้มีการดำเนินโครงการ Thailand SNP Discovery Project (ThaiSNP1) เมื่อปี ค.ศ. 2004 และได้สิ้นสุดลงเมื่อปี ค.ศ. 2006 โดยได้ค้นหายีนที่อยู่บริเวณส่วนที่ถอดรหัสของยีนจำนวน 194 ยีน จากการหาลำดับเบสจากตัวอย่างดีเอ็นเอของคนไทย 32 คน ที่ได้คัดเลือกตามเกณฑ์การรับเข้าศึกษาในโครงการดังต่อไปนี้

1. อายุอยู่ระหว่าง 50-60 ปี
2. ครอบครัวใน 3 ชั่วรุ่น เป็นคนไทย
3. มีสุขภาพแข็งแรงสมบูรณ์ปราศจากประวัติการเจ็บป่วยเรื้อรัง

โดยมีข้อมูลสไนป์อื่นที่ได้จากโครงการวิจัยที่ศึกษาแบบทั้งจีโนม อาทิ ข้อมูลสไนป์แบบแอลลีโลไทป์ (allelotype) ของผู้ป่วยโรคธาลัสซีเมียและมีข้อมูล tagSNP และ แอลลีโลไทป์ในกลุ่มยีนที่เกี่ยวข้องกับกระบวนการเปลี่ยนแปลงยา (drug metabolizing enzyme gene) จำนวน 228 ยีน จากความร่วมมือของโครงการ Thailand Center of Excellence for Life Science (TCELS) กับศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ ซึ่งมีการจัดเก็บ ไว้ที่ <http://www4a.biotech.or.th/thaimut> อย่างไรก็ตาม ข้อมูลความหลากหลายเหล่านี้ ยังมีข้อจำกัดในการนำมาใช้ประโยชน์ทางระบาดวิทยาพันธุศาสตร์กล่าวคือ

- ข้อมูลจากโครงการ ThaiSNP1 มีข้อมูลที่ไม่ครอบคลุมทั้งจีโนม ทำให้ขาดข้อมูลของยีนส่วนใหญ่ที่นักวิจัยสนใจศึกษา
- ข้อมูลจากโครงการ ThaiSNP1 ไม่สามารถนำมาใช้ในการวิเคราะห์เชิงมนุษย์พันธุศาสตร์เนื่องจากขาดข้อมูลสไนป์ที่อยู่นอกยีน และระเบียบวิธีที่ใช้ในการค้นหายีนเหล่านี้ ไม่เอื้ออำนวยในการเปรียบเทียบกับข้อมูลจากประชากรอื่นๆ
- ฐานข้อมูลที่มีข้อมูลเฉพาะแอลลีโลไทป์ (มีเพียงข้อมูลรูปแบบ/ตำแหน่ง/ความถี่ แต่ไม่มีข้อมูลของแต่ละบุคคล) ไม่สามารถนำมาวิเคราะห์เชิงลึกเพื่อให้ได้มาซึ่งข้อมูลพหุคูณได้
- ไม่มีข้อมูลของยีนในฐานข้อมูลทั้งหมดในประเทศไทย

ดังนั้นงานวิจัยนี้จึงต้องการจัดสร้างฐานข้อมูลความหลากหลายทางพันธุกรรมในคนไทย แบบสไนป์และยีนทั้งจีโนม เพื่อประโยชน์ในงานวิจัยด้านมนุษย์พันธุศาสตร์ และระบาดวิทยาพันธุศาสตร์ โดยใช้เทคโนโลยี 5th generation Affymetrix SNP genotyping 500K array

อุปกรณ์และวิธีการ

เซลล์ตัวอย่างส่งตรวจและการสกัดดีเอ็นเอ

เซลล์ตัวอย่างส่งตรวจ 32 ตัวอย่าง ได้รับความอนุเคราะห์จากสถาบันราชานุกูล กรมสุขภาพจิต การสกัดดีเอ็นเอของตัวอย่างตรวจใช้วิธีมาตรฐาน ฟีนอลคลอโรฟอร์ม ตรวจวัดความเข้มข้นของดีเอ็นเอ และตรวจสอบคุณภาพของดีเอ็นเอโดยวิธีเจลิเล็กโทรโฟริซิส

การหาจีโนไทป์แต่ละสลิป

การระบุตำแหน่งและความถี่ (ข้อมูลปฐมภูมิ) ของความหลากหลายทางพันธุกรรมแบบสลิป ใช้เทคโนโลยี 5th generation Affymetrix SNPs genotyping array ซึ่งทุกขั้นตอนจะมีการตรวจสอบคุณภาพของดีเอ็นเอที่ใช้เป็นสารตั้งต้น โดยวัดปริมาณความเข้มข้นของดีเอ็นเออีกครั้ง จากนั้นจึงตรวจสอบคุณภาพของดีเอ็นเอ ด้วยวิธีเจลิเล็กโทรโฟริซิสที่ความเข้มข้นเจล 2.0% หลังจากนั้นตัดดีเอ็นเอตั้งต้นด้วยเอนไซม์ตัดจำเพาะ *StyI* และ *NspI* แล้วติดฉลากสารเรืองแสง ด้วยวิธีการเพิ่มปริมาณดีเอ็นเอโดยใช้ไพรเมอร์แบบ universal แล้วใช้เป็นโพรบ (probe) ในการจับสายดีเอ็นเอคู่สมที่ถูกติดอยู่บนสไลด์แก้ว จากนั้นล้างโพรบออก และตรวจสอบสัญญาณจากตัวอย่างตรวจแต่ละตัวอย่าง เริ่มจากการกวาดสัญญาณเพื่อหาความเข้มของสารเรืองแสงจากโพรบ ที่สามารถจับดีเอ็นเอเป้าหมายได้ ซึ่งจะบอกถึงสลิปแต่ละตำแหน่งว่าเป็นชนิดใดของแอลลีล ขั้นตอนดังกล่าวใช้อัลกอริทึมแบบ BRLMM-P ในการแปลงข้อมูลสัญญาณความเข้มสารเรืองแสง ให้เป็นข้อมูลจีโนไทป์ในแต่ละสลิป ในรูปแบบ CEL ไฟล์

ผลการทดลอง

การระบุตำแหน่งความหลากหลายทางพันธุกรรมแบบสลิป และซีเอ็นวี

เนื่องจากข้อมูลที่ได้จากการเปรียบเทียบความหลากหลายทางพันธุกรรมแบบสลิปของคนไทย 32 คน

มีการบอกตำแหน่งบนจีโนมด้วยข้อมูลลำดับเบสข้างเคียง (flanking sequence) ที่สามารถเข้าคู่กันได้ ตำแหน่งที่ถูกต้องของสลิปถูกตรวจสอบ โดยการนำข้อมูลหมายเลขและลำดับเบสข้างเคียงของสลิปไปเปรียบเทียบกับฐานข้อมูลสลิปสาธารณะ dbSNP และข้อมูลลำดับเบสจากฐานข้อมูล Refseq ข้อมูลตำแหน่งของสลิปถูกจัดเก็บอยู่ในฐานข้อมูลโดยใช้โปรแกรมบริหารจัดการฐานข้อมูล MySQL ส่วนการระบุตำแหน่งความหลากหลายทางพันธุกรรมแบบซีเอ็นวี ใช้ความเข้มของสัญญาณสารเรืองแสงของแต่ละสลิป มาระบุตำแหน่ง ซึ่งสามารถแสดงให้เห็นเป็นขั้นตอนมาตรฐานดัง Figure 1

เมื่อได้ข้อมูลความเข้มแสงในแต่ละตำแหน่งของสลิปแล้ว การหาตำแหน่งของซีเอ็นวีนั้นจะเป็นไปตามข้อมูลอ้างอิงของตำแหน่งของสลิปที่อยู่ติดกันไปเรื่อยๆ บนโครโมโซม โดยมีขั้นตอน คือ การปรับสัญญาณลงที่เกิดจากไฮบริไดซ์ผิดของโพรบ การหักลบสัญญาณให้เป็นมาตรฐาน (normalization) และการสรุปรวมสัญญาณ (summarization) ของตัวอย่างตรวจทั้งหมด หลังจากนั้นจะกวาดสัญญาณไปที่ละสลิปบนโครโมโซม เพื่อหาการเพิ่มขึ้นหรือลดลงของปริมาณความเข้มแสงที่ไปด้วยกันบนชุดของสลิปที่อยู่ติดกัน จะทำให้ทราบถึงตำแหน่งและปริมาณของซีเอ็นวีทั่วทั้งจีโนม ตำแหน่งซีเอ็นวีทั้งหมดที่ได้จาก 32 ตัวอย่างตรวจ ถูกระบุโดยใช้โปรแกรม Aroma.affymetrix (Bengtsson *et al.*, 2008) ที่ใช้อัลกอริทึม circular binary segmentation (CBS) และเก็บข้อมูลลงในฐานข้อมูล

สร้างฐานข้อมูลของความหลากหลายทางพันธุกรรมแบบสลิป และซีเอ็นวี

การจัดเก็บข้อมูลความหลากหลายทางพันธุกรรม แบบสลิปและซีเอ็นวีของคนไทย ที่ได้จากโครงการนี้ ได้พัฒนาต่อจากฐานข้อมูลสลิปของคนไทยจากโครงการ ThaiSNP1 โดยจัดเก็บข้อมูลของสลิปทั้งหมดได้แก่ข้อมูลหมายเลขกำกับสลิปตำแหน่ง

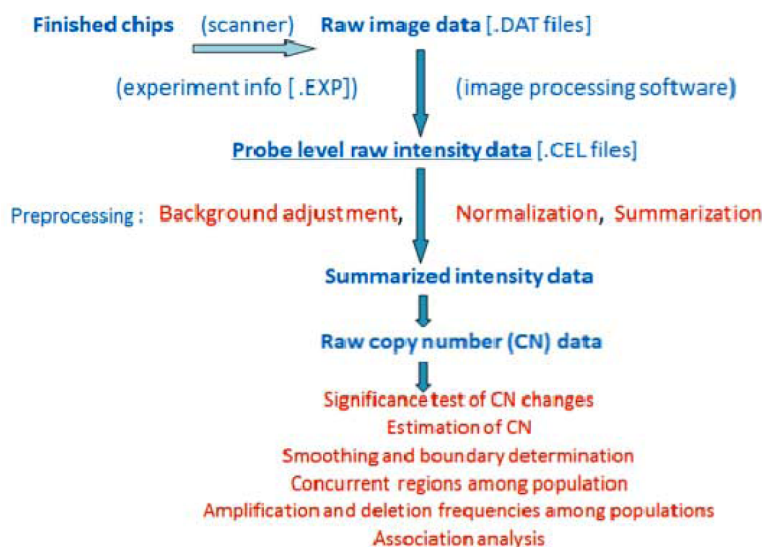


Figure 1 The protocol of CNVs identification from SNP intensity data files (.CEL files).

ของสปี ข้อมูลแอลลีล ข้อมูลจีโนไทป์ ความถี่ในการเกิดสปี ข้อมูลสปีที่เกี่ยวข้องกับยีน โดยได้มีการเปรียบเทียบกับข้อมูลสปีของเชื้อชาติอื่นๆ ได้แก่ ข้อมูลสปีจากฐานข้อมูลสาธารณะ คือ dbSNP, HapMap และ JSNP นอกจากนี้ยังได้รวมข้อมูลสปีจากชนเผ่าต่างๆ ในภูมิภาคเอเชียจากโครงการ Pan-Asian SNP ข้อมูลสปีของคนไทยที่ได้จากโครงการ ThaiSNP1 ซึ่งสปีบางส่วนเป็นข้อมูลสปีใหม่ที่ไม่อยู่ในฐานข้อมูลสาธารณะเข้ามาไว้ด้วย เพื่อใช้เปรียบเทียบและช่วยในการตัดสินใจเลือกเครื่องหมายทางพันธุกรรม นอกจากการจัดเก็บข้อมูลสปีแล้ว ยังมีการออกแบบตารางข้อมูลเพื่อจัดเก็บข้อมูลของยีน โรคที่เกี่ยวข้องกับยีน โดยอ้างอิงจากฐานข้อมูลสาธารณะ Refseq genome build 36.3 เพื่อประโยชน์ในการสืบค้นข้อมูลสปีและซีเอ็นวีทั้งหมดที่ได้ถูกจัดเก็บไว้ในรูปแบบของตาราง โดยสามารถสรุปได้ดังนี้

1. ข้อมูลสปีที่ได้จากโครงการ ThaiSNP2 พร้อมทั้งข้อมูลจีโนไทป์จำนวน 443,816 สปี
2. ข้อมูลสปีจากโครงการ ThaiSNP1 จำนวน 3,860 สปี
3. ข้อมูลสปีจากโครงการ Pan-Asian SNP จำนวน 56,013 สปี จาก 75 กลุ่มประชากร รวม

ทั้งสิ้น 1,928 คน

4. ข้อมูลสปีจาก dbSNP build 129 จำนวน 14,735,067 สปี
5. ข้อมูลสปีจาก HapMap public release 27 จำนวน 4,165,577 สปี จาก 11 กลุ่มประชากร รวมทั้งสิ้น 1,207 คน
6. ข้อมูลสปีจาก JSNP release 35 จำนวน 184,081 สปี
7. ข้อมูลซีเอ็นวีจากข้อมูลของคนไทย 32 คน ได้จำนวนทั้งสิ้น 3,230 ตำแหน่ง โดยเลือกเฉพาะที่มีขนาดความยาวมากกว่า 5,000 เบส และสามารถนับรวมเป็น CNV Region ได้ทั้งหมด 1,889 บริเวณ

ข้อมูลในฐานข้อมูล ThaiSNP2 จะประกอบไปด้วย 2 ส่วน คือ ส่วนข้อมูลที่เป็นรายละเอียดของโครโมโซมยีนและสปีจะถูกเก็บอยู่ในฐานข้อมูลสัมพันธ์ สำหรับข้อมูลที่น่าสนใจในรูปแบบของภาพกราฟฟิค ได้ใช้ฐานข้อมูลในรูปแบบของ generic feature format (GFF) และแสดงผลเป็นภาพกราฟฟิคด้วยโปรแกรม GBrowse (Stein *et al.*, 2002) โดยข้อมูล GFF ได้มาจากโครงการฐานข้อมูล HapMap ซึ่งมีข้อมูลจีโนมที่ประกอบไปด้วยโครโมโซม คอนทิก (contig) ยีน dbSNP HapMap และซีเอ็นวี ฐาน

ข้อมูล Database of Genomic Variants โดยแปลงข้อมูล SNP ของคนไทยทั้งจากโครงการ ThaiSNP1 และ ThaiSNP2 ข้อมูล JSNP ข้อมูลจาก Pan-Asian SNP และข้อมูล CNV Region ของคนไทยเพิ่มเข้ามา เพื่อนำเสนอร่วมกับข้อมูล GFF จาก HapMap ที่มีอยู่ โดยการนำเสนอข้อมูลอยู่ในรูปของเว็บไซต์เพื่อแสดงและแสดงผลข้อมูลในรูปแบบกราฟิกด้วย GBrowse เพื่อให้ผู้ใช้สามารถเข้าถึงและสืบค้นข้อมูลจากฐานข้อมูลได้ง่าย โดยเข้าใช้งานผ่านทางเว็บไซต์ที่ <http://www.biotech.or.th/thaisnp2> ซึ่งมีเมนูในการเข้าใช้งานต่างๆ ประกอบด้วย

- เมนู Login สำหรับให้ผู้ใช้สมัครสมาชิกหรือล็อกอินเข้ามายังระบบ ในกรณีที่ต้องการเข้าถึงข้อมูลที่มีความสำคัญ เช่น การเข้าถึงข้อมูลจีโนม
- เมนู SNPs Search ใช้สำหรับค้นหาข้อมูล SNP จากยีน ตำแหน่งบนโครโมโซม โรคที่สนใจ หรือค้นหาจาก SNP ID
- เมนู CNVRs Search ใช้สำหรับค้นหาข้อมูล CNV Region บนโครโมโซม
- เมนู Genotype & LD Blocks ใช้สำหรับดึงข้อมูลจีโนมไทยของ 32 คนไทย ในบริเวณที่สนใจได้แก่ จากรายชื่อยีน หรือตำแหน่งบนโครโมโซม โดยอยู่ในรูปแบบไฟล์ที่ค้นด้วยจุลภาค (CSV) และในรูปแบบ HapMap ซึ่งสามารถนำไปคำนวณหา LD Blocks ด้วยโปรแกรม Haploview (Barrett *et al.*, 2005) ได้
- เมนู Local Tools ใช้เพื่อเชื่อมต่อไปยังเว็บไซต์ต่างๆ ที่พัฒนาขึ้นภายในศูนย์พันธุฯ ซึ่งเกี่ยวข้องกับโครงการนี้
- เมนู Link-outs ใช้เพื่อเชื่อมต่อไปยังเว็บไซต์ที่เกี่ยวข้องภายนอก เช่น NCBI, HapMap, JSNP, Genomic Variants และ GeneCards

ในการค้นหาข้อมูล SNP ผู้ใช้สามารถค้นหาจากยีน ตำแหน่งบนโครโมโซม โรคที่สนใจ หรือจาก SNP ID เมื่อสืบค้นโดยใช้ชื่อยีน ระบบจะเข้าไปติดต่อกับฐานข้อมูลและแสดงผลการค้นหาออกมาเป็นข้อมูล

ของยีน และข้อมูลของ SNP บนยีน (Figure 2A) โดยจะแสดงข้อมูลของ SNP จากหลายฐานข้อมูล เพื่อให้ง่ายต่อการเปรียบเทียบของผู้ใช้ ได้มีการใช้รูปกำกับที่แสดงถึงแหล่งที่มาของความหลากหลายในแต่ละ SNP ผู้ใช้งานสามารถเชื่อมต่อเพื่อไปดูข้อมูล SNP ทั้งหมดที่อยู่ในบริเวณที่สนใจ และข้อมูลอื่นๆ โดยเลือกรูปแบบขยายบริเวณ Graphical View (Figure 2B) ในการดูข้อมูลรายละเอียดของ SNP ผู้ใช้งานสามารถเลือกที่ SNP ที่ต้องการดูข้อมูล จากหมายเลข SNP (Figure 2A) หรือจากรูป SNP (Figure 2B) ระบบจะสืบค้นข้อมูลต่างๆ ของ SNP ซึ่งประกอบไปด้วยตำแหน่งของ SNP ในจีโนม ลำดับเบสข้างเคียงตำแหน่ง SNP ชื่อของ SNP จากฐานข้อมูลต่างๆ ที่มีตำแหน่งตรงกัน ความถี่ที่ปรากฏในแต่ละเชื้อชาติ ตำแหน่งของ SNP ในยีน ข้อมูลยืนยันของการเป็น SNP และเชื่อมโยงออกไปเพื่อแสดงตำแหน่งของ SNP ด้วย GBrowse (Figure 3)

ฐานข้อมูล ThaiSNP2 มีข้อมูลของ SNP จากหลายฐานข้อมูลดังที่ได้กล่าวไปแล้ว (Figure 3) ได้แก่ ข้อมูล SNP 11 กลุ่มประชากรจาก HapMap ได้แก่ ASW: African ancestry in Southwest USA, CEU: Utah residents with Northern and Western European ancestry from the CEPH collection, CHB: Han Chinese in Beijing, China, CHD: Chinese in Metropolitan Denver, Colorado, GIH: Gujarati Indians in Houston, Texas, JPT: Japanese from Tokyo, Japan, LWK: Luhya in Webuye, Kenya, MEX: Mexican ancestry in Los Angeles, California, MKK: Maasai in Kinyawa, Kenya, TSI: Tuscans in Italy และ YRI: Yoruba from Ibadan, Nigeria ข้อมูล SNP จาก 73 กลุ่มประชากรในแถบทวีปเอเชียทั้งหมด ซึ่งเป็นข้อมูลจากงานวิจัย HUGO Pan-Asian Initiative Consortium ที่คณะผู้วิจัยได้มีส่วนร่วมวิจัยในการศึกษาประวัติศาสตร์ความเป็นมาของชาติพันธุ์ในแถบเอเชียทั้งหมด โดยรวมข้อมูลตามประเทศที่อยู่ของกลุ่มประชากรหรือแหล่งที่มาของข้อมูล จำนวน 11 กลุ่ม

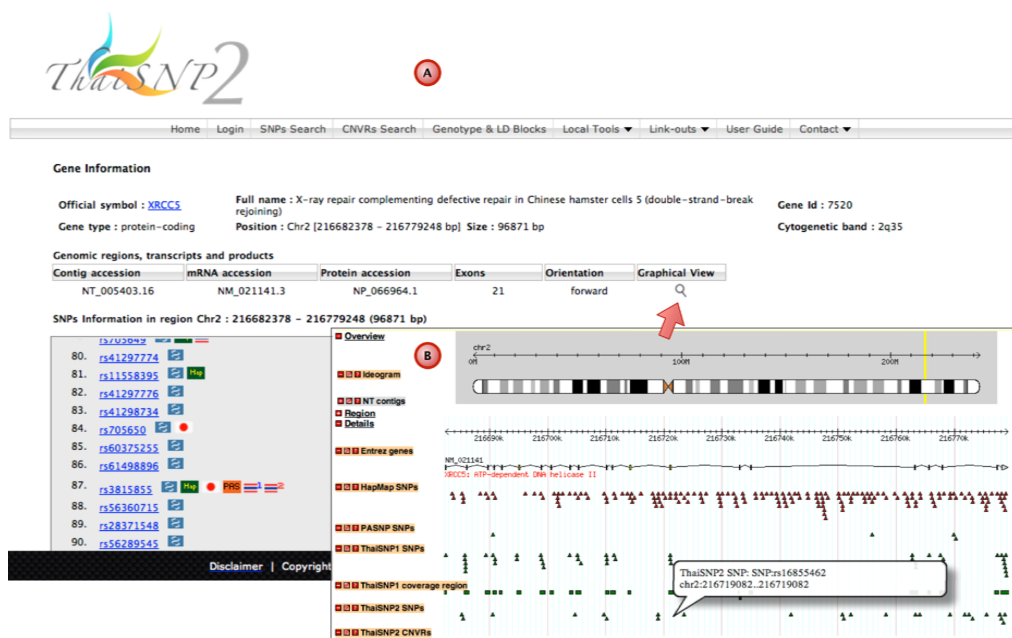


Figure 2 SNP results from the query of the SNP search page. (A) The listing of SNPs in tabular format. (B) GBrowse interface to display SNP locations on gene and chromosome.

Polymorphisms Information :

SNP Id : rs3815855, th1563, IMS-JST176242
 Class : snp
 Sequences : GAAGGGCACTCAGGCAAGTACTTTAAGTCATCACATAGTT
 [C/G]
 AGTGTCCACAATTTCCACACGGTGGACTTCATTGGAAG
 Position : at NT_005403.16 : 67191787 Chr2 : 216690615 Strand : +
 Validation Status : Validated
 Polymorphism in gene : XRCC5 (GeneID:7520) mRNA : NM_021141 Protein : NP_066964 Function : Intron
 Source :
 Graphical view :

Variation Summary :

Source	Population	Sample size	Allele Freq
dbSNP	---	1106	G [0.174] / C [0.826]
JSNP	Japan	1496	C [0.896] / G [0.104]
HapMap	ASW	166	C [0.849] / G [0.151]
HapMap	CEU	328	C [0.732] / G [0.268]
HapMap	CHB	168	C [0.845] / G [0.155]
HapMap	CHD	170	C [0.812] / G [0.188]
HapMap	GIH	176	C [0.892] / G [0.108]
HapMap	JPT	172	C [0.890] / G [0.110]
HapMap	LWK	180	C [0.828] / G [0.172]
HapMap	MEX	154	C [0.838] / G [0.162]
HapMap	MKK	342	C [0.751] / G [0.249]
HapMap	TSI	176	C [0.767] / G [0.233]
HapMap	YRI	334	C [0.868] / G [0.132]
PASNP	AX	50	C [0.900] / G [0.100]
PASNP	CEU	120	C [0.758] / G [0.242]
PASNP	CHB	90	C [0.822] / G [0.178]
PASNP	CN	486	C [0.780] / G [0.220]
PASNP	ID	576	C [0.858] / G [0.142]
PASNP	IN	348	C [0.839] / G [0.161]
PASNP	JP	240	C [0.917] / G [0.083]
PASNP	JPT	88	C [0.875] / G [0.125]
PASNP	KR	180	C [0.883] / G [0.117]
PASNP	MY	432	C [0.852] / G [0.148]
PASNP	PI	286	C [0.902] / G [0.098]
PASNP	SG	180	C [0.794] / G [0.206]
PASNP	TH	488	C [0.820] / G [0.180]
PASNP	TW	160	C [0.744] / G [0.256]
PASNP	YRI	120	C [0.875] / G [0.125]
ThaiSNP1	Thai	60	C [0.833] / G [0.167]
ThaiSNP2	Thai	64	G [0.844] / C [0.156]

Close

Figure 3 SNP summary results categorized by different populations and database sources.

ได้แก่ AX: Populations from Affymetrix, CN: China, ID: Indonesia, IN: India, JP: Japan, KR: Korea, MY: Malaysia, PI: Philippines, SG: Singapore, TH: Hill tribe in Thailand และ TW: Taiwan

สำหรับการสืบค้นข้อมูลซีเอ็นวี ผู้ใช้งานสามารถค้นหาข้อมูลจากโครโมโซม ซึ่งผลการสืบค้นจะแสดงข้อมูลของบริเวณซีเอ็นวี (CNV Region, CNVR) ทั้งหมดที่อยู่บนโครโมโซมที่สนใจซึ่ง CNVRs ที่แสดงจะมีขนาดใหญ่มากกว่า 5,000 เบส โดยแสดงจำนวนของซีเอ็นวีทั้งหมดที่ได้จากแต่ละคน เพื่อให้ผู้ใช้สามารถเข้าไปดูรายละเอียดของซีเอ็นวีได้ (Figure 4A) ซึ่งแสดงรายการของ CNVs ทั้งหมดที่อยู่ใน CNVR เดียวกัน โดยแสดงชื่อตัวอย่าง ตำแหน่งของซีเอ็นวีบนโครโมโซมขนาดของซีเอ็นวีจำนวน สนิปทั้งหมดที่อยู่ในบริเวณ CNVs ชนิดของ CNVs ซึ่งแบ่งได้ดังนี้ คือ 0: homozygous deletion, 1: hemizygous deletion, 2: normal, 3: single copy gain และ 4-6: multiple copy gain นอกจากนี้ผู้ใช้สามารถเข้าไปดูตำแหน่งของ CNVRs บนโครโมโซมในรูปแบบภาพกราฟิกบน GBrowse ได้ โดยเลือกที่รูปแว่นขยายตรงบริเวณ

Graphical View (Figure 4B)

เมื่อมองภาพรวมทั้งหมดโดยใช้ฐานข้อมูล ThaiSNP2 จะสามารถเปรียบเทียบข้อมูลความหลากหลายทางพันธุกรรมในหลายๆ มุมมองตั้งแต่การเปรียบเทียบในแง่ของความสัมพันธ์ของความหลากหลายหลายชนิดอื่นๆ รวมถึงความหลากหลายที่จำเพาะในแต่ละประชากรที่สามารถนำมาเปรียบเทียบกันในงานวิจัยทางพันธุศาสตร์อื่นๆ สำหรับการใช้เทคโนโลยีของ GBrowse นั้น เป็นการลดเวลาในการที่จะนำเอาข้อมูลในหลายๆ มิติเข้ามาซ้อนทับ เพื่อให้ผู้ใช้ได้เห็นภาพมากขึ้น ซึ่งทำให้ผู้ใช้เข้าใจรูปแบบเฉพาะบางอย่าง หรือที่ผู้ใช้ไม่เคยได้เห็นมุมมองแบบนี้ในการเข้าไปที่แหล่งอื่นๆ ที่ให้บริการข้อมูลในลักษณะเดียวกันนี้ ซึ่งฐานข้อมูล ThaiSNP2 เป็นฐานข้อมูลที่มีความครบถ้วนมากที่สุดของข้อมูลความหลากหลายทางพันธุกรรมในประเทศไทย

เตรียมข้อมูลปฐมภูมิสำหรับการวิเคราะห์ข้อมูลทุติยภูมิ

ในขั้นตอนการวิเคราะห์ข้อมูลทุติยภูมิในระดับที่สูงขึ้น เช่น การคำนวณความถี่ของแอลลีลและ

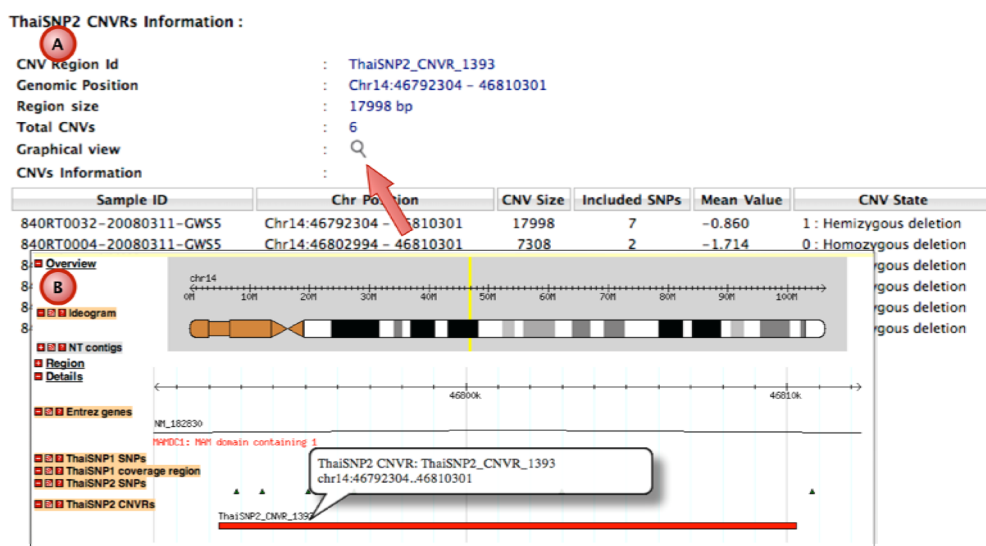


Figure 4 CNV Region. (A) Exhibition of all CNVs that belong to the specific CNVR. (B) Using GBrowse to display CNVRs and other related information.

จีโนไทป์ Hardy-Weinberg equilibrium heterozygosity และ functional SNP prediction ผู้ใช้สามารถที่จะกำหนดปัจจัยที่ต้องการศึกษาได้ เช่น จำนวนกลุ่มของประชากรที่ต้องการศึกษา สนิปที่ต้องการในตำแหน่งบนยีน หรือบนจีโนม อนึ่งเนื่องจากข้อมูลความหลากหลายทางพันธุกรรมในฐานข้อมูล ThaiSNP2 มีเป็นจำนวนมาก การจัดเตรียมข้อมูลที่ใช้ต้องการจะจัดทำแบบตามที่ใช้ร้องขอ(on-demand) โดยผู้ใช้สามารถเลือกที่เมนู Genotype & LD block โดยข้อมูลส่วนนี้ผู้ใช้จะต้องล็อกอินเข้ามาในระบบเมื่อผู้ใช้ใส่ข้อมูลที่น่าสนใจ เช่น รายชื่อยีน หรือส่วนของโครโมโซมที่น่าสนใจ ระบบจะค้นหาสนิปทั้งหมดที่อยู่ในบริเวณที่ผู้ใช้ต้องการและแสดงข้อมูลของสนิปเบื้องต้น เพื่อให้ผู้ใช้เลือกสนิปที่ต้องการอีกครั้ง เมื่อผู้ใช้กดปุ่ม Export ระบบจะสร้างไฟล์ข้อมูลจีโนไทป์ในรูปแบบของไฟล์ที่กั้นด้วยจุลภาค (CSV)

ส่วนการคำนวณหาชุดของแอลลีลที่อยู่เรียงติดกันเป็นชุดหรือที่เรียกว่าแฮพโลไทป์ (haplotype) ที่มีความเกี่ยวข้องกับการเกิดโรคโดยเฉพาะอย่างยิ่งโรคที่มีลักษณะเป็น multifactorial disease เช่น โรคหลอดเลือดหัวใจ โรคเบาหวาน โรคความจำเสื่อม โรคซึมเศร้าและโรคมะเร็ง เป็นต้น ซึ่งรูปแบบแฮพโลไทป์ จะเป็นเครื่องหมายทางพันธุกรรมโดยอ้อม เพื่อนำไปสู่ความเข้าใจกลไกการแสดงออกของยีนที่แท้จริงต่อไป รูปแบบต่างๆ ของแอลลีลที่จะส่งผ่านสู่ลูกหลานเป็น block เดียวกันจะต้องมีลักษณะเป็น “linkage disequilibrium-LD” กล่าวคือ block ของแอลลีลดังกล่าวจะไม่ถูกแยกออกจากกันโดยกระบวนการรีคอมบิเนชันในช่วงไมโอซิส จากความสำคัญของแฮพโลไทป์ จึงมีกลุ่มองค์กรชื่อว่า HapMap ที่รวบรวมเอารูปแบบแฮพโลไทป์ จากกลุ่มประชากรที่มีความหลากหลายทางเชื้อชาติ เพื่อใช้ทำนายการเกิดโรค การตอบสนองต่อการรักษา ซึ่งในช่วงแรกเป็นความร่วมมือระหว่างประเทศ ญี่ปุ่น สหราชอาณาจักร แคนาดา จีน ไนจีเรีย และสหรัฐอเมริกา (3 กลุ่มเชื้อชาติ African, Asian และ European) และในระยะที่

สองของการเก็บรวบรวม HapMap ได้เก็บกลุ่มประชากรมากขึ้น ซึ่งข้อมูลของ HapMap ได้ถูกดึงมาเก็บที่ฐานข้อมูลไว้ทั้งหมดเพื่อความสะดวกรวดเร็วในการที่ผู้ใช้จะเข้ามาในฐานข้อมูลเพียงแห่งเดียวที่เตรียมข้อมูลไว้อย่างครบถ้วน ซึ่งผู้ใช้สามารถดึงข้อมูลจีโนไทป์ไปวิเคราะห์รูปแบบ LD ข้อมูล haplotype phasing ข้อมูล LD block partitioning และ htSNP ได้ โดยการใช้เครื่องมือที่มีชื่อว่า Haploview โดยสามารถคำนวณจากข้อมูลจีโนไทป์ในรูปแบบของ hapmap เมื่อผู้ใช้กดปุ่ม Export for Haploview นอกเหนือจากนี้ฐานข้อมูล ThaiSNP2 ยังได้มีการติดตั้งโปรแกรม Haploview เพื่อคำนวณ LD block โดยอัตโนมัติ เมื่อผู้ใช้กดปุ่ม ThaiSNP2 LD block จะแสดงผลการคำนวณ Haplotype block และภาพ LD Plot (Figure 5)

วิจารณ์และสรุปผลการทดลอง

ฐานข้อมูลความหลากหลายทางพันธุกรรม ThaiSNP2 เป็นการจัดเก็บและรวบรวมข้อมูลความหลากหลายทางพันธุกรรมพื้นฐาน ที่มีประโยชน์ในแง่ของการวิจัยต่อยอดในงานวิจัยทางด้านพันธุศาสตร์ และนำเสนอส่วนของข้อมูลในรูปแบบของภาพกราฟิกทางคณะผู้วิจัยได้เลือกใช้ฐานข้อมูลในรูปแบบของ GFF ซึ่งเป็นรูปแบบการจัดเก็บข้อมูลจีโนม ให้อยู่ในรูปแบบมาตรฐานเดียวกัน สำหรับการแสดงผลเป็นภาพกราฟิกด้วยโปรแกรม GBrowse งานวิจัยนี้รวบรวมความหลากหลายทางพันธุกรรมแบบสนิป และซีเอ็นวีไว้เป็นจำนวนมาก ข้อมูลความหลากหลายทางพันธุกรรมที่มีการระบุอยู่ในฐานข้อมูลนี้ ได้ทำอย่างเป็นมาตรฐานมากที่สุด เพื่อความถูกต้องของข้อมูล โดยเลือกใช้อัลกอริธึมมาตรฐาน ในการระบุตำแหน่งของข้อมูลความหลากหลายทางพันธุกรรมของทั้งสองแบบ คือสนิปและซีเอ็นวี ข้อมูลทั้งหมดจะแสดงในรูปแบบของการเปรียบเทียบในแต่ละเชื้อชาติ จึงทำให้ผู้ใช้สามารถใช้งานได้อย่างสะดวก และทำให้เห็นมุมมองที่ไม่เคยเห็นมาก่อนในการเปรียบเทียบความ

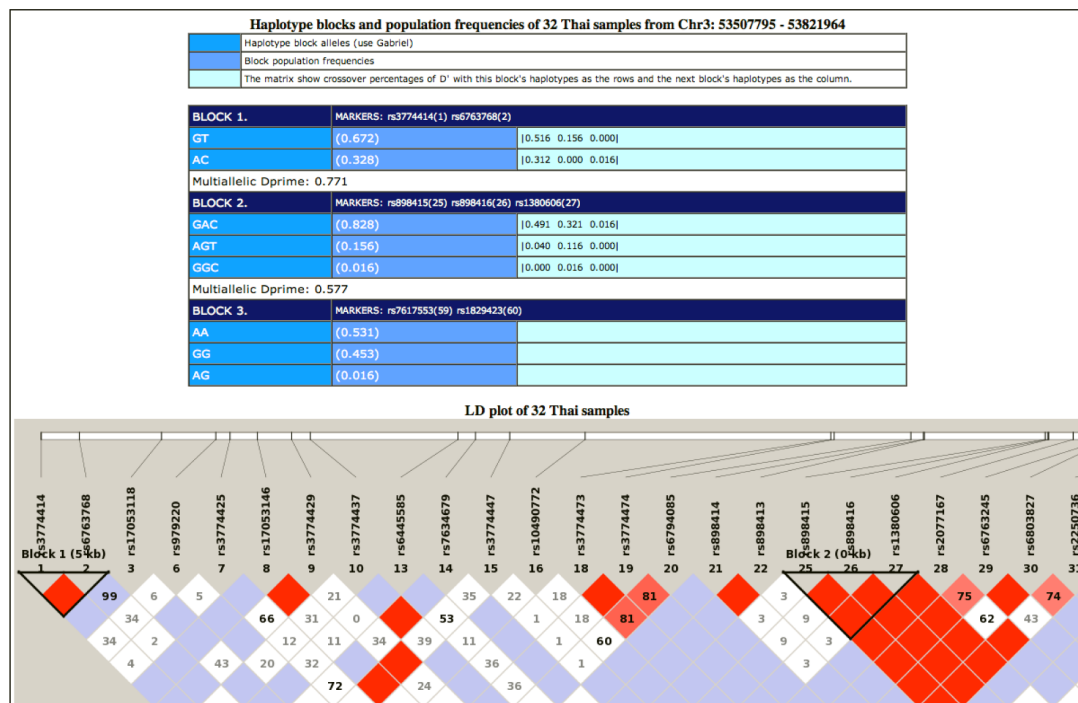


Figure 5 The result of haplotype blocks and LD plot of 32 Thais.

หลากหลายทางพันธุกรรมในระดับจีโนม จะเห็นได้ว่าโครงการนี้ได้เก็บรวบรวมความหลากหลายทางพันธุกรรมที่มีมากที่สุดของคนไทย และจัดเก็บเพื่อให้มีการสืบค้น วิเคราะห์ เปรียบเทียบ การใช้ประโยชน์จากฐานข้อมูล ThaiSNP2 นั้นมีมากมาย เช่น เป็นข้อมูลพื้นฐานที่สำคัญและจำเป็น ในการวิจัยต่อยอดทางระบาดวิทยาพันธุศาสตร์ ที่ช่วยทำให้เกิดความเข้าใจกระบวนการต่างๆ ครอบคลุมตั้งแต่การพยากรณ์ความเสี่ยงต่อการเกิดโรคตัวอย่างของการใช้ประโยชน์จากข้อมูลพื้นฐานที่จัดเก็บอยู่ในฐานข้อมูล ThaiSNP2 ได้แก่

- การศึกษาทางพันธุกรรมแบบยีนเดี่ยว (monogenic disease) ที่ต้องการใช้เครื่องหมายทางพันธุกรรมในการติดตามใน linkage study นั้น ผู้วิจัยสามารถเข้ามาเลือกเครื่องหมายทางพันธุกรรม ที่มีความหลากหลายเพียงพอในการศึกษาได้ในทันที

- การศึกษาทางพันธุกรรมในแง่ของเซลล์พันธุศาสตร์ (cytogenetics) ผู้วิจัยสามารถเข้ามา

เปรียบเทียบ หรือตรวจสอบความผิดปกติที่ตรวจพบว่าเป็นเพียงความหลากหลายทางพันธุกรรมชนิดซีเอ็นวี หรือว่าเป็นความผิดปกติที่เป็นสาเหตุของการเกิดโรคในครอบครัวที่ศึกษาได้

- การใช้ข้อมูลของซีเอ็นวี ในการวิจัยเภสัชพันธุศาสตร์เกี่ยวกับการตอบสนองต่อยา ซึ่งมีความเชื่อมโยงกับการมีความหลากหลายทางพันธุกรรมแบบซีเอ็นวี ซึ่งจะมีจำนวนชุดซ้ำ (copy) ไม่เท่ากันในกลุ่มที่มีอาการไม่พึงประสงค์จากการใช้ยานี้

- การศึกษาความใกล้ชิดทางชาติพันธุ์ และการวิเคราะห์หาโครงสร้างประชากร 4 แบบ คือ 1) การตรวจหาโครงสร้างประชากร 2) การกำหนดกลุ่มประชากรให้กับตัวอย่าง 3) หาจำนวนกลุ่มของประชากรย่อยและ 4) สามารถหาสัดส่วนของประชากรในแต่ละกลุ่มประชากร เป็นการศึกษาความแตกต่างของความถี่ของแอลลีลในแต่ละประชากร เนื่องจากกลไกของวิวัฒนาการ

จะเห็นได้ว่า ข้อมูลพื้นฐานที่ได้มาจากการงาน

วิจัยการสร้างฐานข้อมูล ThaiSNP2 นั้นมีประโยชน์อย่างมากในงานวิจัยทางพันธุกรรมในหลายๆ ด้าน ทั้งงานวิจัยทางด้านมนุษยพันธุศาสตร์ ระบาดวิทยา พันธุศาสตร์ โรคพันธุศาสตร์ พันธุศาสตร์ประชากร นิติวิทยาศาสตร์พันธุกรรม ซึ่งไม่ใช่เป็นเพียงการใช้ข้อมูลเท่านั้น ทั้งยังเป็นข้อมูลที่ใช้ในการพัฒนาเครื่องมือช่วยในการวิจัยด้านชีวสารสนเทศ (bioinformatics) อื่นๆ ต่อไป

กิตติกรรมประกาศ

คณะผู้วิจัยขอขอบคุณสำนักบริหารจัดการคลังเตอร์และโปรแกรมวิจัย สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ และหน่วยอนุพันธุศาสตร์ สถานส่งเสริมการวิจัย คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล ที่ให้ทุนร่วมในการทำวิจัยโครงการนี้ และขอขอบคุณ ดร.นพ.วีรยุทธ ประพันธ์พจน์ จากสถาบันราชานุกูล กรมสุขภาพจิตที่ได้ให้ความร่วมมือและความช่วยเหลือให้เซลล์ตัวอย่างของคนไทย สุดท้ายนี้คณะผู้วิจัยขอขอบคุณผู้ที่เข้าร่วมโครงการค้นหา SNP ของคนไทยที่ได้อนุญาตให้นำดีเอ็นเอไปเพื่อศึกษาวิจัยในครั้งนี้

เอกสารอ้างอิง

- Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263-265.
- Bengtsson, H., Simpson, K., Bullard, J. and Hansen K. 2008. *Aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory*. Tech Report #745, Department of Statistics, University of California, Berkeley.
- Bora, N.S., Kaliappan, S., Jha, P., Xu, Q., Sivasankar, B., Harris, C.L., Morgan, B.P. and Bora, P.S. 2007. CD59, a complement regulatory protein, controls choroidal neovascularization in a mouse model of wet-type age-related macular degeneration. *J Immunol* 178:1783-1790.
- Estivill, X. and Armengol, L. 2007. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* 3: 1787-1799.
- Haga, H., Yamada, R., Ohnishi, Y., Nakamura, Y. and Tanaka, T. 2002. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. *J Hum Genet* 47: 605-610.
- Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T. and Nakamura, Y. 2002. JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res* 30:158-162.
- International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.
- Lachman, H.M., Pedrosa, E., Petruolo, O.A., Cockerham, M., Papolos, A., Novak, T., Papolos, D.F. and Stopkova, P. 2007. Increase in GSK3beta gene copy number variation in bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet* 144: 259-265.
- Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.H., Hicks, J., Spence, S.J., Lee, A.T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, P.K., Bregman, J., Sutcliffe, J.S., Jobanputra, V., Chung, W., Warburton, D., King, M.C., Skuse, D., Geschwind, D.H., Gilliam, T.C., Ye, K. and Wigler, M. 2007.

- Strong association of de novo copy number mutations with autism. *Science* 316: 445-449.
- Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. and Lewis, S. 2002. The generic genome browser: a building block for a model organism system database. *Genome Res* 12:1599-1610.