รายงานวิจัยฉบับสมบูรณ์

การพัฒนาระเบียบวิธีการเรียนรู้ด้วยเครื่องเพื่อใช้ในการเลือกเครื่องหมาย
ทางพันธุกรรมซึ่งบ่งชี้ลักษณะจำเพาะที่สนใจจากข้อมูลสนิปทั่วจีโนม

Development of Machine Learning Metaheuristics for SNP Biomarker Selection in Genome Wide Studies

ดร.ศิษเฎศ ทองสิมา
ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ

15 มิถุนายน 2557

สัญญาเลขที่ RSA5480026

รายงานวิจัยฉบับสมบูรณ์

การพัฒนาระเบียบวิธีการเรียนรู้ด้วยเครื่องเพื่อใช้ในการเลือกเครื่องหมาย
ทางพันธุกรรมซึ่งบ่งชี้ลักษณะจำเพาะที่สนใจจากข้อมูลสนิปทั่วจีโนม
Development of Machine Learning Metaheuristics for SNP Biomarker
Selection in Genome Wide Studies

ดร.ศิษเฎศ ทองสิมา
ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย
และศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ

กิตติกรรมประกาศ

คณะผู้วิจัยขอขอบพระคุณสำนักงานกองทุนสนับสนุนการวิจัย ที่ให้การสนับสนุนทุนในการทำงานวิจัยครั้งนี้ ขอขอบพระคุณศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติที่ให้การสนับสนุนนักวิจัย สถานที่และอุปกรณ์ในการทำวิจัย ขอขอบคุณ ศ.นพ. สุทัศน์ ฟู่เจริญ และคณะนักวิจัยจากศูนย์วิจัยธาลัสซีเมีย สถาบันชีววิทยาศาสตร์โมเลกุล มหาวิทยาลัยมหิดล นพ. วีรยุทธ ประพันธ์พจน์ ศูนย์วิจัยพันธุศาสตร์การแพทย์ สถาบันราชนุกูล กรมสุขภาพจิต และ The Wellcome Trust Case Control Consortium (WTCCC) ประเทศอังกฤษที่เอื้อเฟื้อในส่วนของข้อมูล และผู้เกี่ยวข้องทุกท่านที่มีส่วนช่วยให้ได้มาซึ่งข้อมูล งานวิจัยนี้จะไม่สามารถประสบความสำเร็จได้หากไม่ได้รับความร่วมมือและความช่วยเหลือจากบุคคลและหน่วยงานที่กล่าวมาข้างต้น

นายศิษเฏศ ทองสิมา

บทคัดย่อ

รหัสโครงการ: RSA5480026

ชื่อโครงการ: การพัฒนาระเบียบวิธีการเรียนรู้ด้วยเครื่องเพื่อใช้ในการเลือกเครื่องหมายทางพันธุกรรมซึ่งบ่งชี้ลักษณะจำเพาะที่สนใจจากข้อมูลสนิปทั่วจีโนม

ชื่อนักวิจัย และสถาบัน: นายศิษเฏศ ทองสิมา ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ

E-mail Address: sissades@biotec.or.th

ระยะเวลาโครงการ: 3 ปี

เครื่องหมายทางพันธุกรรมแบบสนิปส์ คือ ความหลายหลายทางพันธุกรรมที่เป็นตัวกำหนดความแตกต่างกันในสิ่งมีชีวิตแต่ละชนิด งานวิจัยมากมายมีการใช้สนิปส์เป็นเครื่องมือในการตรวจสอบการสัมพันธ์ไปกับการเกิดโรคที่ซับซ้อน การใช้วิธีการทางสถิติเพื่อค้นหาสนิปส์ที่มีความสัมพันธ์กับการเกิดโรคเหล่านั้นไม่สามารถบรรลุข้อประสงค์ได้กับทุกๆกลุ่มประชากร เนื่องจากปัญหาของการมีอยู่ของโครงสร้างประชากร(population structure) ในข้อมูลที่จะเป็นตัวลดสัญญาณของนัยสำคัญทางสถิติได้

คณะผู้วิจัยได้พัฒนาอัลกอริธึมที่ชื่อ i2pPCA ที่ใช้กลไกการหยุดชื่อว่า “EigenDev” เพื่อควบคุมการทำซ้ำของอัลกอริธึม ซึ่งมีความสามารถในการตรวจพบและกำจัดโครงสร้างประชากรในข้อมูลสนิปส์ได้ เราได้ประยุกต์ใช้อัลกอริธึม i2pPCA กับข้อมูลสนิปส์ของคนไทยที่นำไปรวมกับข้อมูลสนิปส์ของประชากรอื่นจากทั่วโลกอีกจำนวน 40 ประชากร อัลกอริธึมนี้สามารถตรวจหาโครงสร้างประชากร และแยกกลุ่มตามโครงสร้าง ท้ายสุดสามารถบอกได้ว่าคนไทยประกอบไปด้วย 4 กลุ่มย่อย ซึ่งเป็นข้อบ่งชี้ว่าในการศึกษาความสัมพันธ์ไปกับการเกิดโรคในอนาคต จำเป็นจะต้องคำนึงถึงการมีอยู่ของโครงสร้างประชากรในคนไทยด้วย นอกจากนี้คณะผู้วิจัยยังได้พัฒนาอัลกอริธึมที่ชื่อ iLOCi สำหรับใช้ค้นหาปฏิสัมพันธ์ระหว่างสนิปส์ที่สัมพันธ์กับการเกิดโรคที่ซับซ้อน ซึ่ง iLOCi สามารถค้นหาสนิปส์เหล่านั้นเจอเมื่อประยุกต์ใช้กับข้อมูล WTCCC ซึ่งประกอบไปด้วยโรคต่างๆ คือ โรค Bipolar disorder(BD), โรค Coronary artery disease (CAD), โรค Crohn’s disease (CD), โรค Hypertension (HT), โรค Rheumatoid arthritis (RA), โรค Type1 diabetes (T1D) และโรค Type2 diabetes (T2D)

ผลการวิเคราะห์ที่ประชากรไทยสามารถแบ่งได้เป็น 4 กลุ่มประชากรย่อยนี้ จะเป็นประโยชน์สำคัญอย่างยิ่งต่อการศึกษาความสัมพันธ์ของการเกิดโรคในคนไทย เพราะต้องคำนึงถึงการรบกวนของโครงสร้างประชากรของคนไทยที่มีอยู่

คำหลัก : สนิปส์, โครงสร้างประชากร, การศึกษาความสัมพันธ์กับการเกิดโรค, ประชากรไทย

Abstract

Project Code: RSA5480026

Project Title: Development of Machine Learning Metaheuristics for SNP Biomarker Selection in Genome Wide Studies

Investigator: Mr. Sissades Tongsima

E-mail Address: sissades@biotec.or.th

Project Period: 3 years

Single nucleotide polymorphisms (SNPs) are the most common genetic variant that differentiate human individuals and many diploid organisms. Hence, SNPs can be used to detect certain abnormalies if they are associated with a trait of interest. Much research has been conducted to identify those associated SNPs whose impact will be on the development of personalized medicine, therapeutic and preventive interventions of many complex diseases. To date, most published predictive SNP markers chosen by using statistical techniques cannot replicate the success due to confounding effects from population stratification.

We developed an algorithm called i2pPCA with a novel 'EigenDev' stopping criterion to control the iterative pruning steps of the algorithm. The algorithm can efficiently detect and correct population stratification that confounds the genetic disease susceptibility signal. We applied i2pPCA to the recent Thai population SNP genotype data combined with 40 world-wide populations. As a result, the recent Thai genetic structures were resolved to 4 main subpopulations. This indicates that Thai population hidden structure should be taken into account when performing disease association studies in the future. We also developed an algorithm, iLOCi, that can efficiently and correctly identify putative SNP markers for predicting the risks of common complex diseases. We successfully identified predictive markers complex diseases including Bipolar Disorder, Crohn's Disease, Coronary Artery Disease, Type 1 and 2 Diabetes Mellitus, Hypertension and Rheumatoid Arthritis from the Wellcome Trust Case Control Consortium (WTCCC).

Our contribution provides an elucidation of 4 Thai population structures using our novel algorithm. This finding would be important to genetic association studies to account for population-structure confounding effects.

Keywords: Single nucleotide polymorphisms (SNPs), population structure, genetic association studies, Thai population

บทนำ

Genetic association studies เป็นการศึกษาที่มุ่งค้นหายีนที่เกี่ยวข้องกับการเกิดโรคที่ซับซ้อน เช่น โรคเบาหวานและโรคมะเร็ง อย่างไรก็ตามด้วยความซับซ้อนของกลไกของการเกิดโรค ทำให้การค้นหายีนเหล่านั้นไม่ได้ทำได้โดยง่ายและตรงไปตรงมาเหมือนโรคทางเมนเดลเลียนที่มีเพียงยีนจำนวนไม่กี่ยีนที่เป็นสาเหตุของการเกิดโรค โดยทั่วไปแล้วโรคที่ซับซ้อนมักจะมียีนที่เกี่ยวข้องจำนวนมาก รวมถึงปัจจัยภายนอกหรือสิ่งแวดล้อมล้วนแล้วแต่มีผลต่อการเกิดโรคทั้งสิ้น เทคนิค Genome Wide Association Studies (GWAS) เป็นวิธีทางสถิติที่ใช้กันอย่างแพร่หลาย โดยจะใช้เครื่องหมายทางพันธุกรรมแบบสนิปส์ที่กระจายอยู่ทั่วทั้งจีโนมในการหาความสัมพันธ์กับยีนที่เกี่ยวข้องกับการเกิดโรค ผลที่ได้จากการศึกษานี้สามารถประยุกต์ใช้ในการรักษาเฉพาะบุคคลในโรคที่ซับซ้อนหรือ personalized medicine โดยทั่วไปการวิเคราะห์สนิปส์ทั่วทั้งจีโนมนั้นต้องอาศัยวิธีการทางสถิติที่มีประสิทธิภาพสูง เนื่องจากเป็นโรคที่ซับซ้อนและสนิปส์มีจำนวนมากถึงหลักล้านตำแหน่ง อย่างไรก็ตามวิธีที่มีอยู่ในปัจจุบันยังไม่สามารถค้นหาสนิปส์ที่สัมพันธ์ไปกับการเกิดโรคที่ซับซ้อนได้อย่างมีประสิทธิภาพ เป็นไปได้ว่าอาจเป็นเพราะยังไม่เคยมีการพิจารณา epistasis หรือปฏิสัมพันธ์ระหว่างยีน ดังนั้นได้มีการตีพิมพ์ผลงานด้านนี้อย่างต่อเนื่อง [1-3] และด้วยเทคโนโลยีจีโนไทป์แบบทั่วทั้งจีโนม จึงได้มีการจีโนไทป์ข้อมูลสนิปส์ของคนไทยขึ้นโดยเป็นการรวมกันของข้อมูลกลุ่ม GWAS ธาลัสซีเมีย [4] และ GWAS โรคซึมเศร้า จำนวน 992 คน และสนิปส์จำนวน 500,000 ตำแหน่ง ร่วมกับข้อมูลจาก Wellcome Trust Case Control Consortium (WTCCC) ซึ่งได้ตีพิมพ์ข้อมูลสนิปส์ขนาดใหญ่ ประกอบไปด้วยตัวอย่างจำนวน 17,000 คน และสนิปส์จำนวน 500,000 ตำแหน่ง [5] โดยการใช้ข้อมูลทั้งสองชุดนี้สร้างความเป็นไปได้ในการพัฒนาวิธีการทางสถิติใหม่ๆ ที่มีประสิทธิภาพมากกว่าเดิมในการค้นหาสนิปส์ที่สัมพันธ์กับการเกิดโรคโดยเฉพาะโรคที่เกิดขึ้นกับคนไทย

โดยทั่วไปแล้ว GWAS ประกอบไปด้วย 4 ขั้นตอน คือ 1) การเก็บตัวอย่างสำหรับการทดลองคือผู้ที่เป็นโรค (case) และผู้ที่สุขภาพปกติ (control) 2) ทำการจีโนไทป์สนิปส์ทั่วทั้งจีโนมโดยใช้สนิปอาเรย์ 3) ทำการคัดกรองสนิปส์ที่ไม่ผ่านการควบคุมคุณภาพออกเพื่อการวิเคราะห์ทางสถิติที่ดีขึ้น 4) วิเคราะห์เปรียบเทียบความแตกต่างของความถี่ของอัลลีลระหว่าง case และ control เพื่อหาสนิปส์ที่มีความแตกต่างกันมากที่สุด อย่างไรก็ตามสำหรับโรคที่ซับซ้อนไม่เป็นเรื่องง่ายที่จะพบสนิปส์เหล่านั้น เนื่องด้วย epistasis และ heterogeneity ยิ่งไปกว่านั้นหากข้อมูลนั้นมีขนาดใหญ่ หมายถึงมีการใช้ตัวอย่างจำนวนมากด้วยเจตนาที่จะสามารถมั่นใจได้มากขึ้นกับค่าทางสถิติ แต่ก็อาจเกิดผลเสียจากการเกิดโครงสร้างประชากรในข้อมูล ทำให้ผลการทดสอบทางสถิติที่ได้เป็นสนิปส์ที่สัมพันธ์กับโครงสร้างประชากรแทนที่จะสัมพันธ์กับการเกิดโรค ด้วยเหตุผลนี้จึงทำให้ต้องตรวจหาโครงสร้างประชากรก่อนทำการวิเคราะห์ทางสถิติเสมอ

ปัจจุบันมีนักวิจัยได้เสนอวิธีในการตรวจพบและกำจัดโครงสร้างประชากรในข้อมูล GWAS หลายวิธี เช่น การปรับค่า χ^2 ให้มีค่าแบบ non-central χ^2 distribution [6] โดย Pritchard และคณะ [7] ก็ได้นำเสนอวิธีกำจัดโครงสร้างประชากรในข้อมูลโดยใช้โมเดลทางสถิติและมีผู้นำมาประยุกต์ใช้งาน

อย่างแพร่หลาย [8-12] แต่อย่างไรก็ตามวิธีนี้ยังไม่มีประสิทธิภาพเมื่อจำนวนสนิปส์มีขนาดเกิน 100,000 ตำแหน่ง และเมื่อไม่นานมานี้ Price [8] และคณะได้นำเสนอวิธีการใหม่ในการตรวจหาโครงสร้างประชากรแต่ก็เป็นวิธีที่ใช้งานได้ยากในการทำ GWAS

ในการหาสนิปส์ที่สัมพันธ์ไปกับการเกิดโรคที่ซับซ้อนนั้นได้มีการนำเสนอการหาปฏิสัมพันธ์ของยีนโดย Musani [13] และตามด้วย [14-17] โดยแบ่งงานได้เป็นสองแนวทาง คือ งานทางด้านสถิติและการเรียนรู้ด้วยเครื่องหรือ Machine Learning (ML) ทางด้านสถิตินั้นจะเป็นการเน้นการทดสอบหาสนิปส์ที่ส่งผลต่อความเสี่ยงต่อการเกิดโรค อย่างไรก็ตามด้วยความซับซ้อนของโรคเองจึงทำให้ไม่ประสบผลสำเร็จมากนัก [17-19] ส่วนทางด้านการเรียนรู้ของเครื่องนั้นจะเป็นการอาศัยโมเดลในการจัดกลุ่ม ที่สามารถทำนายตัวอย่างที่ยังไม่เคยได้รับการตรวจสอบได้ว่าเป็นผู้ป่วยโรคนั้นหรือไม่ อย่างไรก็ตามผลทำนายจะแม่นยำหรือไม่ขึ้นอยู่กับจำนวนของตัวอย่างที่ใช้สร้างโมเดล (trained sample) ต้องมีความเหมาะสมไม่ over-fit จนเกินไป

เพื่อเป็นการแก้ไขปัญหาของ GWAS ที่ได้กล่าวมาข้างต้น เราเสนอระเบียบวิธีการทำ GWAS อย่างมีประสิทธิภาพโดยเลือกสนิปส์ที่สามารถทำนายโอกาสการเกิดโรคที่ซับซ้อน โดยประกอบด้วยสามขั้นตอนคือ 1) ระเบียบวิธีในการค้นหาโครงสร้างประชากร 2) ระเบียบวิธีในการหาปฏิสัมพันธ์ของสนิปส์ที่เกี่ยวข้องกับการเกิดโรคที่ซับซ้อน และ 3) สร้างโมเดลทางการเรียนรู้ของเครื่องเพื่อใช้สนิปส์จากข้อ 2) ทำนายความเสี่ยงในการเกิดโรค

วิธีการทดลอง

1. พัฒนาอัลกอริธึมเพื่อใช้ในการแยกกลุ่มประชากรจากข้อมูลระดับจีโนมที่มีขนาดใหญ่และมีความซับซ้อนค่อนข้างสูง โดยพัฒนาต่อยอดจากระเบียบวิธีจำแนกพีซีเอแบบทำซ้ำ (Iterative pruning PCA) เดิม โดยปรับปรุงเงื่อนไขการหยุดโดยใช้อัลกอริธึมไอเก็นเดฟ (EigenDev) โดยสร้างเป็นซอฟแวร์ชื่อ i2pPCA

2. ประยุกต์ใช้อัลกอริธึม i2pPCA ที่ได้พัฒนาขึ้นเอง ร่วมกับอัลกอริธึมประชากรเชิงพันธุศาสตร์อื่นๆ ได้แก่ Neighbor Joining Tree และ ADMIXTURE วิเคราะห์ข้อมูลประชากรไทยโดยเปรียบเทียบกับประชากรจากทั่วโลกจำนวน 40 กลุ่มจากข้อมูลของ Xing และคณะ [20]

3. ประยุกต์ใช้อัลกอริธึมที่ได้พัฒนาขึ้นเพื่อใช้ในการตรวจหาปัจจัยรบกวนและผลกระทบที่เกิดจากโครงสร้างของกลุ่มประชากร เพื่อทำให้การคัดเลือกกลุ่มของสนิปส์ที่มีความสัมพันธ์กับโรคทางพันธุกรรมที่ซับซ้อนจากการศึกษาด้วย GWAS มีความถูกต้องและมีประสิทธิภาพมากขึ้น โดยใช้ข้อมูล 1) ข้อมูลผู้ป่วยโรคเบต้าธาลัสซีเมีย และ 2) ข้อมูลจากโครงการศึกษาพันธุกรรมของผู้ป่วยโรคซึมเศร้า

4. ค้นหาชุดของสนิปส์ที่ใช้จำแนกประชากรไทยออกจากกันตามโครงสร้างของประชากรด้วยค่าสถิติ Fst เพื่อการประยุกต์ใช้ก่อนการวิเคราะห์ GWAS

5. ออกแบบและพัฒนากระบวนการการเรียนรู้ด้วยเครื่องคอมพิวเตอร์โดยใช้หลักการทางสถิติเพื่อสร้างโมเดลที่ใช้ในการแยกแยะหรือทำนายความเสี่ยงในการเกิดโรคทางพันธุกรรมจากรูปแบบความหลากหลายทางพันธุกรรมของแต่ละบุคคลได้

6. พัฒนาอัลกอริธึมเป็นซอฟท์แวร์ที่มีความสามารถในการประมวลผลแบบขนาน ซึ่งทำให้สามารถเพิ่มประสิทธิภาพและลดเวลาที่ใช้ในการประมวลผลข้อมูลระดับจีโนมที่มีขนาดใหญ่และมีความต้องการการคำนวณที่มีประสิทธิภาพสูงและซับซ้อนได้

7. สามารถค้นหาชุดของสนิปส์ที่สามารถใช้ในการทำนายหรือแบ่งกลุ่มผู้ป่วยออกจากคนปรกติได้ โดยใช้ข้อมูลจากโรคทางพันธุกรรมที่มีความซับซ้อนจำนวน 7 โรคจาก WTCCC ได้แก่ โรค Bipolar disorder(BD), โรค Coronary artery disease (CAD), โรค Crohn’s disease (CD), โรค Hypertension (HT), โรค Rheumatoid arthritis (RA), โรค Type1 diabetes (T1D) และโรค Type2 diabetes (T2D)

ผลการทดลอง

การพัฒนาอัลกอริธึมเพื่อใช้ในการแยกกลุ่มประชากรจากข้อมูลระดับจีโนมที่มีขนาดใหญ่และมีความซับซ้อนค่อนข้างสูง คณะผู้วิจัยได้ทำการพัฒนาต่อยอดจากระเบียบวิธีการจำแนกพีซีเอแบบทำซ้ำ (Iterative pruning PCA) เดิมที่คณะผู้วิจัยได้พัฒนาขึ้น โดยพัฒนาปรับปรุงเงื่อนไขที่ใช้ในการหยุดกระบวนการทำซ้ำใหม่โดยใช้อัลกอริธึมไอเก้นเดฟ (EigenDev) ซึ่งจากการทดสอบกับข้อมูลจำลองและข้อมูลจริงที่มีขนาดใหญ่และมีความซับซ้อนของกลุ่มประชากร พบว่าอัลกอริธึมที่พัฒนาใหม่นี้มีความสามารถในการอนุมานจำนวนกลุ่มประชากรย่อยและจำแนกคนเข้าในกลุ่มประชากรในระดับแยกย่อยได้อย่างถูกต้องแม่นยำและมีประสิทธิภาพมากขึ้นเมื่อเปรียบเทียบกับระเบียบวิธี ipPCA เดิม รวมทั้งอัล-กอริธึมใหม่ยังถูกออกแบบให้รองรับข้อมูลอื่นนอกจากสนิปส์ ได้แก่ ข้อมูล Short Tandem Repeat (STR), microsatellite, minisatellite และ CNVs และทำการประยุกต์ใช้ระเบียบวิธีนี้ใน 2 ด้าน คือ 1) การศึกษาทางด้านประชากรเชิงพันธุศาสตร์ของประชากรไทยเมื่อเปรียบเทียบกับประชากรกลุ่มอื่นๆ ในภาคพื้นเดียวกัน และ 2) การค้นหาสัญญาณรบกวนของโครงสร้างประชากรที่มีในข้อมูลจีโนไทป์ก่อนการวิเคราะห์ GWAS

สำหรับการศึกษาแรก คณะผู้วิจัยได้ทำการประยุกต์ใช้ระเบียบวิธีการจำแนกพีซีเอแบบทำซ้ำ (Iterative pruning PCA) รวมถึงกระบวนการทำซ้ำใหม่โดยใช้อัลกอริธึมไอเก้นเดฟ (EigenDev) เพื่อใช้ในการศึกษาประชากรเชิงพันธุศาสตร์กับข้อมูลที่ได้จากการรวมกันของข้อมูลจีโนไทป์ของผู้ป่วยคนไทยที่เป็นโรคเบต้าธาลัสซีเมียรวมกับข้อมูลของผู้ป่วยโรคซึมเศร้า เพื่อนำมาเป็นข้อมูลตัวแทนของกลุ่มประชากรไทยที่เป็นข้อมูลตัวแทนจากประชากรไทย รวมกับข้อมูลของ Xing และคณะ ซึ่งประกอบไปด้วยข้อมูลจีโนไทป์ของประชากรมนุษย์จากทุกภาคพื้นทวีปอีก 40 กลุ่ม ได้แก่ Alur, Hema, Pygmy, AP Brahmin, N. European, Khmer Cambodian, Chinese, Samoan, Tongan, Slovenian, Totonac, Bambaran, Dogon, Kyrgyzstani, Kurd, Bolivian, Thai Moken, Pakistani, Buryat, Nepalese, TN Dalit, Irula, Japanese, AP Madiga, AP Mala, CEU, YRI, CHB, JPT, Luhya, Tuscan, !Kung, Pedi, Sotho/Tswana, Stalskoe, Iban, TN Brahmin, Urkarah, Vietnamese, Nguni เมื่อรวมกันแล้วข้อมูลมีจำนวน 1842 คน 41,789 สนิปส์ ผลที่ได้จากการวิเคราะห์ i2pPCA คือ การแบ่งกลุ่มตามพันธุกรรมได้จำนวน 24 กลุ่มย่อยตามโครงสร้างประชากร ซึ่งกลุ่มที่ได้นี้แสดงถึง ความสัมพันธ์ในระดับพันธุกรรมของเชื้อชาติต่างๆ รวมถึงกลุ่มประชากรไทยซึ่งสามารถแบ่งได้เป็น 4 กลุ่มย่อย โดยที่ตัวอย่างส่วนใหญ่ของแต่ละกลุ่มประกอบด้วย ประชากรจาก 1) ภาคเหนือ-กลาง 2) ตะวันออกเฉียงเหนือ-กลาง 3) ใต้-กลาง และ 4) กลาง เป็นที่น่าสังเกตว่าในกลุ่มที่ 4 นั้นมีตัวอย่างที่มาจาก CHB และ Vietnamese เป็นการบ่งชี้ว่ากลุ่มนี้มีความไกล้เคียงกันและสามารถอธิบายได้ถึง recent migration ที่เกิดขึ้นไม่นานมานี้ได้

คณะผู้วิจัยได้ทำการประยุกต์ใช้ระเบียบวิธีที่พัฒนาขึ้นกับข้อมูลของคนไทยเพียงกลุ่มเดียว เพื่อตรวจหาการรบกวนของโครงสร้างของกลุ่มประชากรก่อนการศึกษาด้วย GWAS จากการรวมกันของข้อมูลธาลัสซีเมียและโรคซึมเศร้าจะได้จำนวนของตัวอย่างเท่ากับ 992 คน และ 477,704 ตำแหน่ง ผลจากการวิเคราะห์จากข้อ 1 เราทราบว่าประชากรไทยควรมีอย่างน้อย 4 กลุ่มย่อย เนื่องจากจำนวนสนิปส์ในการทดลองที่ 2 นี้มีขนาดมากกว่าเกือบ 10 เท่า จึงเป็นการวิเคราะห์ที่มีความแม่นยำและละเอียดมากกว่า พบว่ากลุ่มย่อยของประชากรยังมีค่าเท่ากับ 4 กลุ่มเหมือนเดิม แต่ตัวอย่างส่วนใหญ่ในกลุ่มย่อยมีการผสมกันระหว่างภาคน้อยลง หรือกล่าวคือมี homogeniety มากขึ้น

ประชากรไทยมีรูปแบบของความหลากหลายทางพันธุกรรมเป็นของตัวเอง และแตกต่างจากชนชาติอื่นๆในแถบนี้ โดยเมื่อพิจารณาภายในกลุ่มประชากรไทย สามารถแบ่งออกได้เป็น 4 กลุ่มประชากรย่อย ซึ่งการค้นพบนี้แสดงให้เห็นว่าอัลกอริธึม i2pPCA สามารถนำไปประยุกต์ใช้ในการศึกษาพันธุศาสตร์เชิงประชากรและกำจัดการรบกวนของสัญญาณโครงสร้างประชากรในการวิเคราะห์ GWAS ได้

ผลที่ได้ทำให้เราทราบว่าการศึกษา GWAS ของประชากรไทยในอนาคต จำเป็นจะต้องคำนึงถึงการมีอยู่ของโครงสร้างประชากรที่มีอยู่ในประชากรไทยเพื่อเป็นการป้องกันการรบกวนของโครงสร้างประชากรต่อ Association test เราจำเป็นที่จะต้องทำการจัดกลุ่มประชากรไทยที่ต้องการทำ GWAS เสียก่อน คณะผู้วิจัยจึงได้ประยุกต์ใช้การวิเคราะห์ Fst เพื่อเป็นการค้นหาชุดของสนิปส์ที่เกี่ยวข้องกับโครงสร้างของประชากร โดยมีการจับคู่เพื่อคำนวณหาค่า Fst สำหรับสนิปส์แต่ละตำแหน่งได้ทั้งหมด 6 คู่ และจากนั้นเรียงลำดับจากมากไปน้อยแล้วเลือก top rank ของแต่ละคู่ออกมาได้จำนวน 400 สนิปส์ ในสนิปส์กลุ่มนี้หลายๆ สนิปส์เคยถูกรายงานแล้วว่ามีความเกี่ยวข้องกับลักษณะการแสดงออกหรือฟีโนไทป์และ susceptability ของโรคต่างๆ เช่น Skin pigmentation, Pediatric Asthma, Adaptation in Asian populations, Skeleton growth, Lung cancer, Alcohol dependency, Upper aerodigestive tract cancer, metabolic effect of alcohol, metabolic syndrome, type II diabetes และ Brugada disease

ส่วนที่สองคณะผู้วิจัยได้พัฒนาอัลกอริธึมเพื่อใช้สำหรับค้นหาปฏิสัมพันธ์ระหว่างสนิปส์ในระดับจีโนมที่เกี่ยวข้องกับการเกิดโรคทางพันธุกรรมที่ซับซ้อนและพบบ่อยโดยอาศัยหลักการความแตกต่างของปฏิสัมพันธ์ระหว่างสนิปส์ของกลุ่มคนปรกติและกลุ่มคนที่เป็นโรค โดยทำการค้นหาความแตกต่างของปฏิสัมพันธ์จากสนิปส์ที่เป็นไปได้ทุกคู่ และทำการเรียงค่าความแตกต่างของปฏิสัมพันธ์ที่ได้จากสูงสุดไปจนถึงต่ำสุด โดยคู่สนิปส์ที่ให้ค่าความแตกต่างของปฏิสัมพันธ์ระหว่างกลุ่มคนปรกติและกลุ่มคนที่เป็นโรคสูงกว่าจะมีความสัมพันธ์กับการเกิดโรคมากกว่า สำหรับหลักการคำนวณที่ใช้ในการค้นหาความแตกต่างของปฏิสัมพันธ์ระหว่างสนิปส์สามารถดูรายละเอียดเพิ่มเติมได้จากเอกสารในภาคผนวก

คณะผู้วิจัยได้พัฒนาอัลกอริธึมดังกล่าวอยู่ในรูปซอฟท์แวร์ชื่อ iLOCi โดยได้พัฒนาให้มีลักษณะเป็นการประมวลผลแบบขนาน สามารถใช้หน่วยประมวลผลหลายตัวร่วมกันคำนวณเพื่อเพิ่มประสิทธิภาพในการประมวลผลได้ ข้อมูลอินพุตจะเป็นข้อมูลจีโนทัยป์ของสนิปส์จากกลุ่มคนปรกติและกลุ่มคนที่เป็นโรค โดยได้ทำการทดสอบซอฟท์แวร์ที่พัฒนาขึ้นกับข้อมูลจำลองเพื่อทดสอบความถูกต้องของอัลกอริธึมที่พัฒนาขึ้น และทดสอบกับข้อมูลจริงซึ่งเป็นข้อมูลจีโนไทป์ของสนิปส์ในระดับจีโนมจำนวน

ประมาณเกือบ 500,000 สนิปส์โดยใช้ข้อมูลโรคพันธุกรรมที่ซับซ้อนจำนวน 7 โรค ได้แก่ โรค Bipolar disorder(BD), โรค Coronary artery disease (CAD), โรค Crohn's disease (CD), โรค Hypertension (HT), โรค Rheumatoid arthritis (RA), โรค Type1 diabetes (T1D) และโรค Type2 diabetes (T2D) จาก WTCCC คณะผู้วิจัยได้ทำการศึกษาปฏิสัมพันธ์ของยีนที่เกี่ยวข้องหรือมีสนิปส์ที่มีความแตกต่างของค่าปฏิสัมพันธ์สูงในแต่ละโรค โดยพบว่าผลที่ได้จากอัลกอริธึม iLOCi ที่พัฒนาขึ้นมีทั้งยีนที่มีการรายงานมาแล้วจากวารสารวิชาการต่าง ๆ ว่าเกี่ยวข้องกับโรคโดยอ้างอิงจากฐานข้อมูล HuGE Navigator (http://www.hugenavigator.net) และยีนที่ไม่เคยมีการรายงานมาก่อนว่าเกี่ยวข้องกับโรค โดยกลุ่มยีนเหล่านี้คณะผู้วิจัยได้ใช้เครื่องมือชื่อ ToppGene (http://toppgene.cchmc.org/) เพื่อใช้ในการสร้างความสัมพันธ์ของกลุ่มยีนที่ไม่เคยมีการรายงานมาก่อนว่าเกี่ยวข้องกับโรคกับกลุ่มยีนที่มีการรายงานมาแล้วว่าเกี่ยวข้องกับโรค ซึ่งวิธีการดังกล่าวทำให้สามารถค้นพบยีนอื่นๆ ที่เกี่ยวข้องกับโรคได้ คณะผู้วิจัยได้พัฒนาอัลกอริธึมและผลการทดสอบรวมทั้งข้อมูลจำลอง พร้อมทั้งได้เขียนบทความเรื่อง iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies ซึ่งได้รับการตีพิมพ์ในวารสาร BMC Genomics

อัลกอริธึมที่ได้พัฒนาและตีพิมพ์ในวารสารวิชาการนี้เป็นแนวความคิดจากการที่นักวิจัยเชื่อว่าการค้นหาสนิปที่ทำนายความเสี่ยงของโรคที่ซับซ้อนและพบบ่อยไม่ควรใช้การทดสอบว่า สนิปแต่ละตัวมีปัจจัยที่ทำให้เกิดการเบี่ยงเบนจาก Hardy-Weinburg Equilibrium (HWE) เพราะบางครั้งความเบี่ยงเบนจาก HWE อาจจะไม่สามารถดูได้เพียงการใช้สนิปส์เพียงตัวเดียว หากแต่เป็นผลที่มาจากยีนหลายตัวทำงานร่วมกัน (epistasis) ซึ่งแนวความคิด epistasis หรือ gene-gene interaction นี้ก็ได้รับความนิยมในช่วงที่ผ่านมา หากแต่ว่าการค้นหาสนิปส์หลายตัวที่สัมพันธ์ไปกับโรคพร้อม ๆ กันเป็นการคำนวณที่ซับซ้อนและใช้เวลานานมาก อาทิการหาคู่สนิปส์ที่เป็นไปได้ทั้งหมดจากสนิปอาร์เรย์ 500,000 ตำแหน่ง มีมากถึง ~125,000 ล้านคู่ ทางคณะผู้วิจัยจึงได้สร้างซอฟท์แวร์ที่สามารถทำงานบนเครื่องคอมพิวเตอร์สมรรถนะสูงได้อย่างมีประสิทธิภาพ เพื่อช่วยจัดการคัดกรองคู่สนิปส์ที่เป็นไปได้ทั้งหมดเพื่อให้เหลือจำนวนสนิปส์จำนวนหนึ่งที่สามารถทำการศึกษาทางฟังก์ชั่นต่อได้

หลักการการหาความสัมพันธ์ทำโดยใช้หลักการที่ว่า ความสัมพันธ์ของคู่สนิปส์ (two-SNP correlation) ในกลุ่มของคนที่เป็นโรค (case group) ถ้ามีความแตกต่างกับกลุ่มคนปกติ (control group) คู่สนิปส์นี้น่าจะเกี่ยวข้อง (associate) ไปกับการเกิดโรคนั้นๆ หลักการ correlation ดังกล่าวนี้คณะผู้วิจัยได้ทำการพิสูจน์ทางคณิตศาสตร์และสถิติว่ามีความสัมพันธ์กับหลักการการหา Linkage Disequilibrium (LD) คู่สนิปส์ที่ผ่านกระบวนการคัดกรอง SNP-correlation difference นี้จึงเป็นคู่สนิปส์ที่อาจจะนำมาใช้เป็นเครื่องหมายพันธุกรรมในการทำนายความเสี่ยงของการเกิดโรคที่ซับซ้อนและพบบ่อยได้ในอนาคต

บทวิจารณ์

เป้าหมายของงานวิจัยนี้มุ่งพัฒนาระเบียบวิธีทางคอมพิวเตอร์เพื่อใช้ในการค้นหาเครื่องหมายพันธุกรรมแบบตำแหน่งเดียวที่เรียกว่าสนิปส์ (SNPs) เพื่อประโยชน์ในการระบุลักษณะจำเพาะที่สนใจ (Trait) ซึ่งอาจจะเป็นโรคที่ซับซ้อนหรือลักษณะเด่นทางพันธุกรรมของพืชหรือสัตว์ต่างๆ โดยได้จำแนกปัญหาของงานวิจัยออกเป็นสามส่วนหลัก กล่าวคือ 1) การพัฒนาระเบียบวิธีเพื่อแก้ปัญหาสัญญาณทางโครงสร้างประชากร (Population structure/stratification) ที่เข้ามารบกวนการวิเคราะห์วิจัย 2) การพัฒนาระเบียบวิธีวิจัยเพื่อค้นหาสนิปส์ที่สัมพันธ์ไปกับลักษณะจำเพาะ และ 3) การพัฒนาฐานข้อมูลสนิปส์ของคนไทยเพื่อคัดกรองสนิปส์ที่เกี่ยวข้องกับประชากรไทยเพื่อประโยชน์ในการค้นหาสนิปส์ที่สัมพันธ์ไปกับลักษณะจำเพาะนั้น การดำเนินงานทั้งสามปีที่ผ่านมาได้ผลิตผลงานวิจัยเพื่อแก้ไขปัญหาทั้งสามประเด็นดังกล่าวโดยในข้อที่ 1 ผู้วิจัยได้พัฒนาอัลกอริธึมใหม่ในการคัดแยกกลุ่มประชากรจากข้อมูลสนิปส์ชื่อว่า iterative prunning Principal Component Analysis (i2pPCA) โดยได้พัฒนากลไกที่ช่วยในการหยุดการทำงานแบบ “วนซ้ำ” และสร้างกลุ่มประชากรย่อยที่มีความคล้ายคลึงกันในรูปแบบของสนิปส์ออกมา [21] อัลกอริธึมดังกล่าวสามารถนำไปใช้งานแทนอัลกอริธึมการแยกคลัสเตอร์ (clustering algorithm) ได้โดยมีข้อดีในประเด็นที่ i2pPCA ไม่ต้องใช้ค่ากำหนดจำนวนกลุ่มเพื่อใช้สร้างกลุ่มประชากรย่อยในขณะที่โปรแกรมการแยกคลัสเตอร์อื่น ๆ จำเป็นต้องระบุ ผู้วิจัยได้นำเอา i2pPCA ที่ได้พัฒนาและตีพิมพ์ไปใช้งานเพื่อตรวจสอบว่าข้อมูลของประชากรไทยมีโครงสร้างประชากรที่ก่อให้เกิดปัญหาของการวิเคราะห์หาสนิปส์ที่สัมพันธ์กับโรคหรือไม่ โดยได้ใช้ข้อมูลของผู้ป่วยโรคธาลัสซีเมียและผู้ป่วยโรคซึมเศร้ามาทดสอบ ผลที่ได้ คือ เราสามารถเห็นว่าคนไทยจำนวนเกือบ 1000 คน สามารถแบ่งแยกตามความเหมือนของข้อมูลทางพันธุกรรมได้ทั้งหมด 4 กลุ่มหลักๆ [22] ผู้วิจัยเล็งเห็นว่ากลุ่มประชากรย่อยที่ค้นพบนี้น่าจะส่งผลต่อการวิเคราะห์วิจัยหาความสัมพันธ์กับโรค (disease genetic association studies) ผลการวิเคราะห์ด้านพันธุศาสตร์ประชากร (population genetics) ที่ได้ทำขึ้นมีประโยชน์ในการศึกษาชาติพันธุ์ของมนุษย์ที่อาศัยอยู่ในบริเวณเอเชียตะวันออกเฉียงใต้เป็นอย่างยิ่ง ผู้วิจัยได้พัฒนาฐานข้อมูลของสนิปส์สำหรับคนไทยและประชากรในเขตแพนเอเชีย (PanAsian population) [23] ซึ่งในขณะเดียวกันฐานข้อมูลดังกล่าวสามารถนำไปตอบโจทย์ในข้อที่ 3 ที่ได้กล่าวเบื้องต้นอีกด้วย ในด้านการใช้งานเชิงพันธุศาสตร์ประชากร การวิเคราะห์ควรนำเอาข้อมูลประชากรใกล้เคียงอาทิ มาเลเซีย เวียดนาม กัมพูชา ลาว และจีนตอนใต้เข้ามาวิเคราะห์ร่วมกัน ทำให้เราสามารถเห็นความต่อเนื่องของสัญญาณทางพันธุกรรมของประชากรในภูมิภาคนี้ชัดเจนมากขึ้นด้วย

หลังจากที่ได้พัฒนาโปรแกรมตรวจสอบและแก้ไขสัญญาณรบกวน population stratification ผู้วิจัยได้พัฒนาอัลกอริธึมเพื่อค้นหาสนิปส์ที่สัมพันธ์ไปกับโรคที่ซับซ้อนและพบบ่อย (common complex

diseases) งานวิจัยแรกได้พัฒนาซอฟท์แวร์ชื่อว่า iLOCi ซึ่งเป็นซอฟท์แวร์ที่ทำการวิเคราะห์ปฏิสัมพันธ์ชองคู่สนิปส์ (SNP-pair interaction) ร่วมไปกับความเป็นไปได้ที่คู่สนิปส์นี้เกี่ยวข้องไปกับความเสี่ยงของการเกิดโรค (disease association) งานวิจัยนี้เปิดโอกาสให้นักวิจัยค้นหาสนิปส์ในมิติของยีนที่อาจจะทำงานร่วมกันแล้วส่งผลกับการเกิดโรค (epistasis) หลักการที่ใช้เพื่อค้นหาคู่สนิปส์เหล่านี้เป็นหลักการที่เรียบง่ายตามสมมติฐาน (null hypothesis) ที่ว่าความสัมพันธ์ของสนิปส์แต่ละคู่ (correlation of SNPs pair) ควรที่จะเหมือนกันในประชากรกลุ่มควบคุม (control group) และประชากรกลุ่มที่เป็นโรค (case group) ถ้าไม่เป็นไปตามสมมติฐานแสดงว่าคู่สนิปส์ดังกล่าวน่าจะเกี่ยวข้องไปกับความเสี่ยงของการเกิดโรคได้ เนื่องจากการค้นหาคู่สนิปส์จำเป็นที่จะต้องทดสอบคู่สนิปส์ที่เป็นไปได้ทั้งหมด ซึ่งมีจำนวนเท่ากับ O(N*N) ซึ่งปัจจุบันจำนวนของสนิปส์ (N) มีอยู่ในระดับหนึ่งล้านตำแหน่งหรือมากกว่านั้น ผู้วิจัยจึงได้พัฒนาอัลกอริธึมที่สามารถใช้ประสิทธิภาพของเครื่องคอมพิวเตอร์สมรรถนะสูงเพื่อมาจัดการคำนวณให้มีประสิทธิภาพ ปัญหาที่ตามมาอีกอย่าง คือ เรื่องของที่จัดเก็บข้อมูลของ correlation of SNP pairs ทั้งหมด ในปัจจุบันอัลกอริธึม iLOCi จะไม่ได้เก็บข้อมูลทั้งหมดไว้แต่จะทำการจัดเก็บข้อมูลที่มีค่าความแตกต่างของ correlation of SNP pairs จากกลุ่ม case เทียบกับ control สูงที่สุดจำนวนหนึ่งไว้ ทำให้ความสัมพันธ์ในมิติที่มากกว่าสองสนิปส์อาจจะหายไปจากข้อมูล (higher order gene interaction) ในอนาคตคาดว่าจะทำการวิเคราะห์ higher order interaction ในระหว่างการคำนวณหาความแตกต่างของ correlation of SNP pairs เพื่อลดความเสี่ยงของ false negative ลง ในส่วนของการทำนายสนิปส์ที่สามารถใช้ในการทำนายความเสี่ยงของโรคผู้วิจัยได้นำเอาเทคนิคที่ชื่อว่า hidden naïve bayes (HNB) เข้ามาเพื่อทำการค้นหา SNPs (feature selection) จากจำนวนของคู่สนิปส์ที่เป็นไปได้ [24] เทคนิคนี้มีความแม่นยำเทียบเท่าหรือดีกว่าในบางข้อมูลเมื่อเปรียบเทียบกับเทคนิค recursive feature elimination support vector machine (RFE-SVM) เนื่องจากปัญหาของ computational complexity ทำให้ผู้ใช้ไม่สามารถค้นหาคำตอบโดยดูจากข้อมูลทั้งหมดได้ ดังนั้นคำตอบที่ได้อาจจะไม่ใช่คำตอบที่ดีที่สุด

หนังสืออ้างอิง

1. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5(8): 618-25.
2. Culverhouse R, Suarez BK, Lin J, Reich T (2002) A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 70(2):461-71.
3. Heidema AG, Boer JM, Nagelkerke N, Mariman EC, van der A DL, Feskens EJ (2006) The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC Genet 7:23.
4. Nuinoon M, Makarasara W, Mushiroda T, Setianingsih I, Wahidiyat PA, Sripichai O, Kumasaka N, Takahashi A, Svasti S, Munkongdee T, Mahasirimongkol S, Peerapittayamongkol C, Viprakasit V, Kamatani N, Winichagoon P, Kubo M, Nakamura Y, Fucharoen S (2010) A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E. Hum Genet 127(3):303-14.
5. Consortium WT (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.
6. Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA (2006) Centralizing the non-central chi-square: a new method to correct for population stratification in genetic case-control association studies. Genet Epidemiol 30(4):277–89.
7. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170-81.
8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–9.
9. Epstein MP, Allen AS, Satten GA (2007) A simple and improved correction for population stratification in case-control studies. Am J Hum Genet 80:921-30.
10. Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM (2007) A randomization test for controlling population stratification in whole-genome association studies. Am J Hum Genet 81:895–905.
11. Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M (2008) On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82:453-63.

12. Guan W, Liang L, Boehnke M, Abecasis GR (2009) Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. Genet Epidemiol 33(6):508-17.
13. Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB (2007) Detection of gene x gene interactions in genome-wide association studies of human population data. Hum Hered 63(2):67-84.
14. Hoh J, Ott J (2003) Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 4(9):701-9.
15. Moore JH (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human Hered 56(1-3):73-82.
16. Moore JH, Ritchie MD (2004) STUDENTJAMA. The challenges of whole-genome approaches to common diseases. JAMA 291(13):1642-3.
17. Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S (2006) Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet 79(6):1002-16.
18. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37(4):413-7.
19. Millstein J, Siegmund KD, Conti DV, Gauderman WJ (2005) Identifying susceptibility genes by using joint tests of association and linkage and accounting for epistasis. BMC Genet 6 Suppl 1:S147.
20. Xing J, Watkins WS, Shlien A, Walker E, Huff CD, Witherspoon DJ, Zhang Y, Simonson TS, Weiss RB, Schiffman JD, Malkin D, Woodward SR, Jorde LB (2010) Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. Genomics 96(4):199-210.
21. Limpiti T, Intarapanich A, Assawamakin A, Shaw PJ, Wangkumhang P, Piriyapongsa J, Ngamphiw C, Tongsima (2010) Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and STRUCTURE. BMC Bioinformatics 12:255.
22. Wangkumhang P, Shaw PJ, Chaichoompu K, Ngamphiw C, Assawamakin A, Nuinoon M, Sripichai O, Svasti S, Fucharoen S, Praphanphoj V, Tongsima S (2013) Insight into the peopling of mainland Southeast Asia from Thai Population Genetic Structure. PLoS One 8(11):e79522

23. Ngamphiw C, Assawamakin A, Xu S, Shaw PJ, Yang JO, Ghang H, Bhak J, Liu E, Tongsima S (2011) PanSNPdb: The Pan-Asian SNP Genotyping Database. PLoS ONE 6(6):e21451.
24. Assawamakin A, Prueksaaroon S, Kulawonganunchai S, Shaw PJ, Varavithya V, Ruangrajitpakorn T, Tongsima S (2013) Biomarker Selection and Classification of "-Omics" Data Using a Two-Step Bayes Classification Framework. Biomed Res Int 2013:148014.

ผลงานจากโครงการวิจัยที่ได้รับทุนจาก สกว.

ผลงานวิจัยที่ตีพิมพ์ในวารสารวิชาการระดับนานาชาติ

1. Chumpol Ngamphiw, Anunchai Assawamakin, Shuhua Xu, Philip James Shaw, Jin Ok Yang, Ho Ghang, Jong Bhak, Edison Liu, and **Sissades Tongsima**. PanSNPdb: The Pan-Asian SNP Genotyping Database. PLoS ONE 2011; 6(6):e21451. doi:10.1371/journal.pone.0021451.
   บทความนี้ได้รับการคัดเลือกให้เป็น Highlight ของ A-IMBN Research เมื่อวันที่ 5 ต.ค. 2554
2. Tulaya Limpiti, Apichart Intarapanich, Anunchai Assawamakin, Philip James Shaw, Pongsakorn Wangkumhang, Jittima Piriyapongsa, Chumpol Ngamphiw and **Sissades Tongsima**. Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and STRUCTURE. BMC Bioinformatics 2011 Jun 23; 12:255 doi:10.1186/1471-2105-12-255.
3. Jittima Piriyapongsa, Chumpol Ngamphiw, Apichart Intarapanich, Supasak Kulawonganunchai, Anunchai Assawamakin, Chaiwat Bootchai, Philip James Shaw and **Sissades Tongsima**. iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies. BMC Genomics 2012; 13 Suppl 7:S2. doi:10.1186/1471-2164-13-S7-S2.
4. Pongsakorn Wangkumhang, Philip James Shaw, Kridsadakorn Chaichoompu, Chumpol Ngamphiw, Anunchai Assawamakin, Manit Nuinoon, Orapan Sripichai, Saovaros Svasti, Suthat Fucharoen, Verayuth Praphanphoj and **Sissades Tongsima**. Insight into the peopling of mainland Southeast Asia from Thai Population Genetic Structure. PLoS One. 2013 Nov 4;8(11):e79522. doi: 10.1371/journal.pone.0079522.
5. Anunchai Assawamakin, Supakit Prueksaaroon, Supasak Kulawonganunchai, Philip James Shaw, Vara Varavithya, Taneth Ruangrajitpakorn and **Sissades Tongsima**. Biomarker Selection and Classification of “-Omics” Data Using a Two-Step Bayes Classification Framework. Biomed Res Int. 2013;2013:148014. doi: 10.1155/2013/148014.

การนำผลงานวิจัยไปใช้ประโยชน์

คณะผู้วิจัยได้พัฒนาซอฟท์แวร์เพื่อใช้ในการแยกกลุ่มประชากรจากข้อมูลระดับจีโนมที่มีขนาดใหญ่และมีความซับซ้อนค่อนข้างสูงในชื่อ i2pPCA โดยสามารถดาวน์โหลดใช้งานผ่านทางเวบไซต์ http://www4a.biotec.or.th/GI/tools/ippca

ซอฟท์แวร์สำหรับค้นหาปฏิสัมพันธ์ระหว่างสนิปส์ในระดับจีโนมที่เกี่ยวข้องกับการเกิดโรคทางพันธุกรรมที่ซับซ้อนชื่อ iLOCi สามารถดาวน์โหลดใช้งานผ่านทางเวบไซต์ http://www4a.biotec.or.th/GI/tools/iloci

กิจกรรมอื่น ๆ ที่เกี่ยวข้อง ได้แก่

1. หัวหน้าโครงการได้รับเชิญเป็นวิทยากรบรรยายในงานประชุมวิชาการ FAOBMB Meeting ครั้งที่ 22 ในหัวข้อเรื่อง “PASNP work on the Thailand data” เมื่อวันที่ 5-7 ต.ค. 2554 ณ ประเทศสิงคโปร์
2. หัวหน้าโครงการได้รับเชิญเป็นวิทยากรบรรยายในงาน A-IMBN Annual Conference ครั้งที่ 14 ในหัวข้อเรื่อง “Large-scale study of Thai population genetic” เมื่อวันที่ 2 ธ.ค. 2554 จัดที่บ้านวิทยาศาสตร์สิริธร อุทยานวิทยาศาสตร์ประเทศไทย ปทุมธานี
3. หัวหน้าโครงการได้รับเชิญเป็นวิทยากรบรรยายในงาน Workshop on e-Science and High Performance Computing (eHPC 2012) ในหัวข้อเรื่อง “High Performance Computing in Computational Biology: BIOTEC Perspectives” เมื่อวันที่ 1 มิ.ย.2555 ณ มหาวิทยาลัยหอการค้า กรุงเทพฯ
4. จัดอบรมเชิงปฏิบัติการให้แก่นักวิจัยที่สนใจ ในหัวข้อเรื่อง “ชีวสารสนเทศกับการวิเคราะห์ทางด้านพันธุศาสตร์ประชากร” ณ บ้านวิทยาศาสตร์สิรินธร อุทยานวิทยาศาสตร์ประเทศไทย ปทุมธานี ระหว่างวันที่ 30-31 ตุลาคม 2556 โดยมีผู้เข้าร่วมอบรมจำนวน 35 คน
5. จัดอบรมเชิงปฏิบัติการให้แก่นักวิจัยทางด้านการปรับปรุงพันธ์พืช ในหัวข้อเรื่อง “IRRDB Training Workshop on Bioinformatics” ณ สถาบันวิจัยยาง กรมวิชาการเกษตร กรุงเทพฯ ระหว่างวันที่ 18-26 พฤศจิกายน 2556 โดยมีผู้เข้าร่วมอบรม 20 คน

การเชื่อมโยงกับต่างประเทศหรือรางวัลที่ได้รับ

1. หัวหน้าโครงการได้มีความร่วมมือกับคณะนักวิจัยในภูมิภาคเอเชียแปซิฟิก ภายใต้ชื่อ Pan Asian Population Genomics Initiative (PAPGI) เพื่อร่วมกันวางแผนในการดำเนินการวิจัยในโครงการศึกษาเรื่องพันธุศาสตร์ของประชากรในภูมิภาคเอเชียแปซิฟิกในเฟสที่ 2 หัวหน้าโครงการได้รับเลือกเป็น Steering Committee ของโครงการดังกล่าว โดยทำหน้าที่เป็นตัวแทนประเทศไทยเพื่อเจรจาทิศทางการดำเนินงานวิจัย
2. หัวหน้าโครงการได้รับรางวัลThe Meritorious Service Award เมื่อเดือน ธ.ค. 2554 จาก Asia-Pacific Bioinformatics Network (APBioNet) ในฐานะบุคคลที่ได้ทำงานใน

APBioNet สนับสนุนงานวิจัยและการศึกษาทางด้าน Bioinformatics ผ่านกิจกรรมของ APBioNet อาทิงานประชุมวิชาการ the International Conference on Bioinformatics (InCoB)

ภาคผนวก

OPEN ACCESS Freely available online

PLoS one

PanSNPdb: The Pan-Asian SNP Genotyping Database

Chumpol Ngamphiw[1,2], Anunchai Assawamakin[1], Shuhua Xu[3], Philip J. Shaw[1], Jin Ok Yang[4], Ho Ghang[5], Jong Bhak[5,6], Edison Liu[7], Sissades Tongsima[1]*, and the HUGO Pan-Asian SNP Consortium[¶]

1 National Center for Genetic Engineering and Biotechnology (BIOTEC), Klong Luang, Pathumthani, Thailand, **2** Inter-Department Program of BioMedical Sciences, Faculty of Graduate School, Chulalongkorn University, Bangkok, Thailand, **3** Chinese Academy of Sciences and Max Planck Society (CAS-MPG) Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, **4** Korean BioInformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), Yuseong-gu, Deajeon, South Korea, **5** Personal Genomics Institute, Genome Research Foundation, Suwon, South Korea, **6** Theragen BiO Institute, TheragenEtex, Suwon, South Korea, **7** Genome Institute of Singapore, Singapore, Singapore

Abstract

The HUGO Pan-Asian SNP consortium conducted the largest survey to date of human genetic diversity among Asians by sampling 1,719 unrelated individuals among 71 populations from China, India, Indonesia, Japan, Malaysia, the Philippines, Singapore, South Korea, Taiwan, and Thailand. We have constructed a database (PanSNPdb), which contains these data and various new analyses of them. PanSNPdb is a research resource in the analysis of the population structure of Asian peoples, including linkage disequilibrium patterns, haplotype distributions, and copy number variations. Furthermore, PanSNPdb provides an interactive comparison with other SNP and CNV databases, including HapMap3, JSNP, dbSNP and DGV and thus provides a comprehensive resource of human genetic diversity. The information is accessible via a widely accepted graphical interface used in many genetic variation databases. Unrestricted access to PanSNPdb and any associated files is available at: http://www4a.biotec.or.th/PASNP.

Citation: Ngamphiw C, Assawamakin A, Xu S, Shaw PJ, Yang JO, et al. (2011) PanSNPdb: The Pan-Asian SNP Genotyping Database. PLoS ONE 6(6): e21451. doi:10.1371/journal.pone.0021451

Editor: Amanda Ewart Toland, Ohio State University Medical Center, United States of America

Received April 7, 2011; **Accepted** May 27, 2011; **Published** June 23, 2011

Funding: The authors wish to thank the National Center for Genetic Engineering and Biotechnology for financial support. Anunchai Assawamakin was supported by the National Science and Technology Development Agency (NSTDA) Postdoctoral Fellowship offered through the National Center for Genetic Engineering and Biotechnology (BIOTEC). Shuhua Xu was supported by the National Science Foundation of China (30971577) and the Science and Technology Commission of Shanghai Municipality (09ZR1436400, 11QA1407600). Philip J. Shaw is supported by a grant from the Bill and Melinda Gates Foundation under the Grand Challenges Explorations Initiative. Jin Ok Yang was supported by KRIBB Research Initiative Program and the Korean Ministry of Education, Science and Technology (MEST) under grant number (2010-0029345). Ho Ghang and Jong Bhak were supported by MOST and KRIBB internal fund of South Korea. Sissades Tongsima was supported by the Thailand Research Fund (TRF), and also in part by the Center of Excellence in Molecular Genetics of Cancer and Human Diseases, Chulalongkorn University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Jong Bhak is an employee of Theragen BiO Institute and Edison Liu is an employee of the Genome Institute of Singapore. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials, as detailed online in the guide for authors.

* E-mail: sissades@biotec.or.th

[¶] Membership of the HUGO Pan-Asian SNP Consortium is provided in Text S1.

Introduction

In recent years, genome-wide single nucleotide polymorphism (SNP) data from high density array platforms and next generation whole-genome sequencing data have been gathered from various human populations. These data embody the transition from single-locus based studies to genomics analyses of human population structure and disease gene mapping [1–5]. Until recently, Asian populations have been largely underrepresented in genome-wide studies in comparison to other peoples of the world. For example, both the International HapMap project and 1000 Genome project lack population samples from Southeast Asia, which is known to contain the most ethno-linguistically diverse populations in Asia. To address this type of shortcoming, the Human Genome Organization (HUGO) Pan-Asian SNP consortium was established to sample genetic diversity in Asia. This effort culminated in a survey of 1,719 unrelated individuals from 71 populations from China (including Taiwan), India, Indonesia, Japan, Malaysia, the Philippines, Singapore, South Korea and Thailand [6]. These 71 populations represent most of the major linguistic groups in Asia and the Pacific, i.e. Altaic, Austro-Asiatic, Austronesian, Dravidian, Hmong-Mien, Indo-European, Papuan, Sino-Tibetan and Thai-Kadai. Considering the general concordance between linguistic and genetic affiliations of human populations, genome-wide data from these samples also captured the majority of the human genetic diversity in Asia. A distinct north - south cline with increasing genetic diversity was observed and contrary to the two-wave migration hypothesis, our study showed substantial genetic proximity of Southeast Asian and East Asian populations [6]. This suggested that the entry of humans into the Asian continent occurred as a single primary wave, populating the south and then expanding northward.

Beside population genetics, there are many other uses of this information include pharmacogenomics, forensics, and genetic epidemiology. The complexity of this dataset poses difficulties for analysis, since only the genotypic transformations of the data are available from the SNP database from National Center for Biotechnology Information (dbSNP), and are thus accessible only to researchers with advanced bioinformatic capabilities. Hence, a database of various analyses accompanying the data would be of benefit to researchers in different disciplines who may not have the bioinformatic capabilities to obtain the information they require.

The goals of the Pan-Asian SNP database are 1) present the data in different formats to facilitate analysis with different tools by providing a graphical viewing interface; 2) comparison of the Pan-Asian dataset with other genetic variation databases including HapMap3 [7], dbSNP [8], and Japan SNP database (JSNP) [9]; 3) incorporate the results of different analyses, including the previously published patterns of population genetic structure and new analyses (linkage disequilibrium patterns, haplotype blocks inferred from the linkage disequilibrium (LD) patterns, tagSNPs as markers of LD blocks, copy number variations (CNVs) inferred from the SNP raw data); and 4) provide an infrastructure for future deposition of data and analysis pertaining to Asia.

Results and Discussion

Genotyping and allele frequencies

Genotyping of Affymetrix GeneChip Human Mapping 50K Xba arrays was performed at eight different genotyping centers (China, India, Japan, Korea, Malaysia, Singapore, Taiwan and USA), according to the manufacturer's protocols. More information regarding SNP calling can be found in the Supplements of [6]. In addition to these HUGO Pan-Asian SNP consortium data, the data for the matching SNPs from 209 HapMap samples (CEU, CHB, JPT and YRI) were included into PanSNP. The final dataset contained the genotypes of 54,794 and 1,204 SNPs

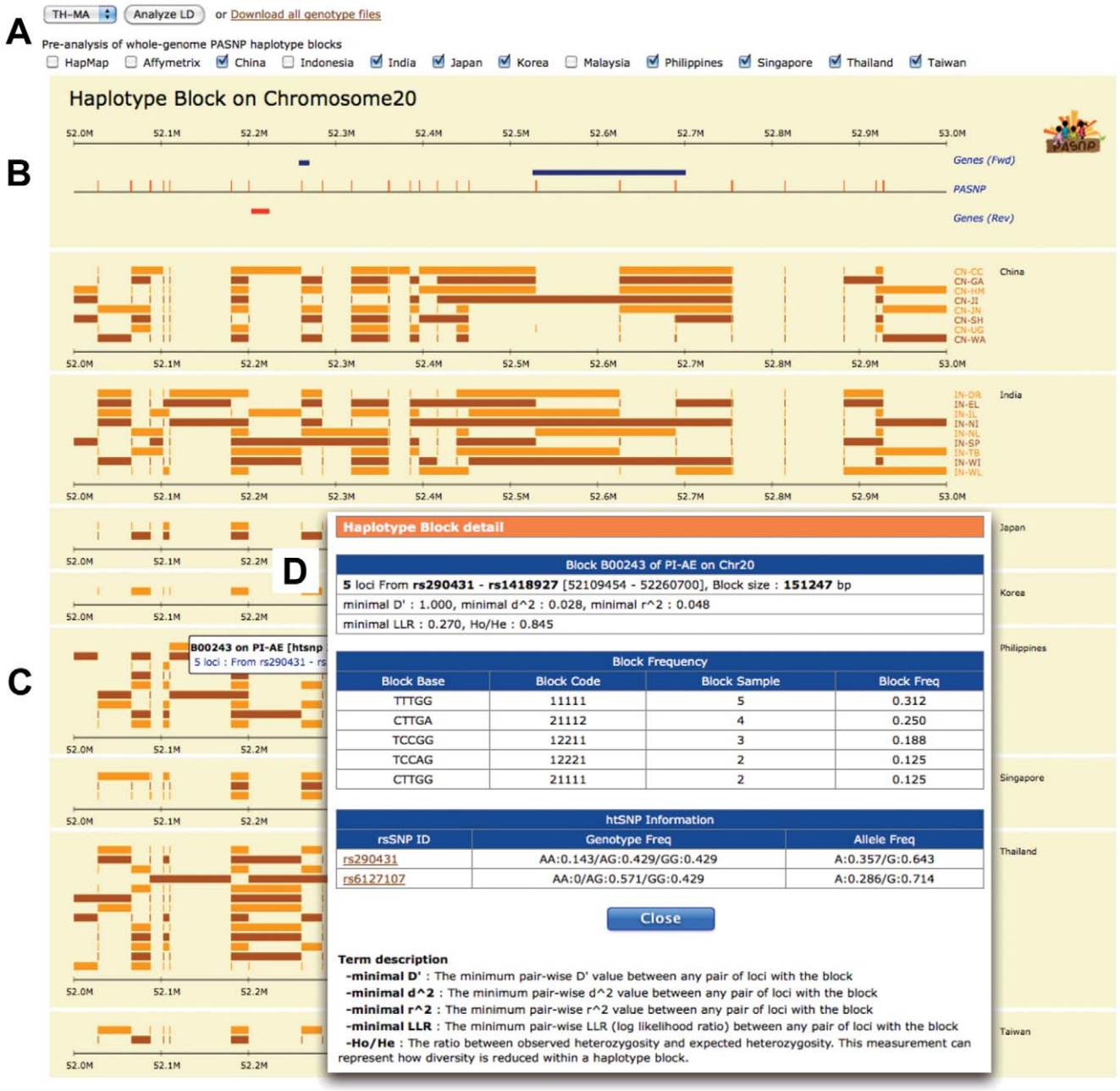
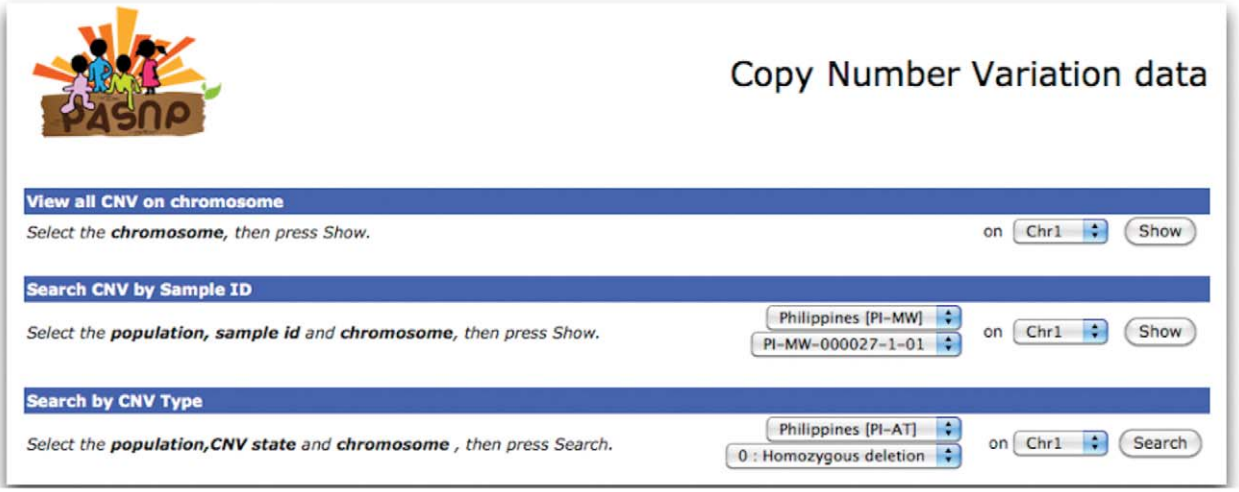
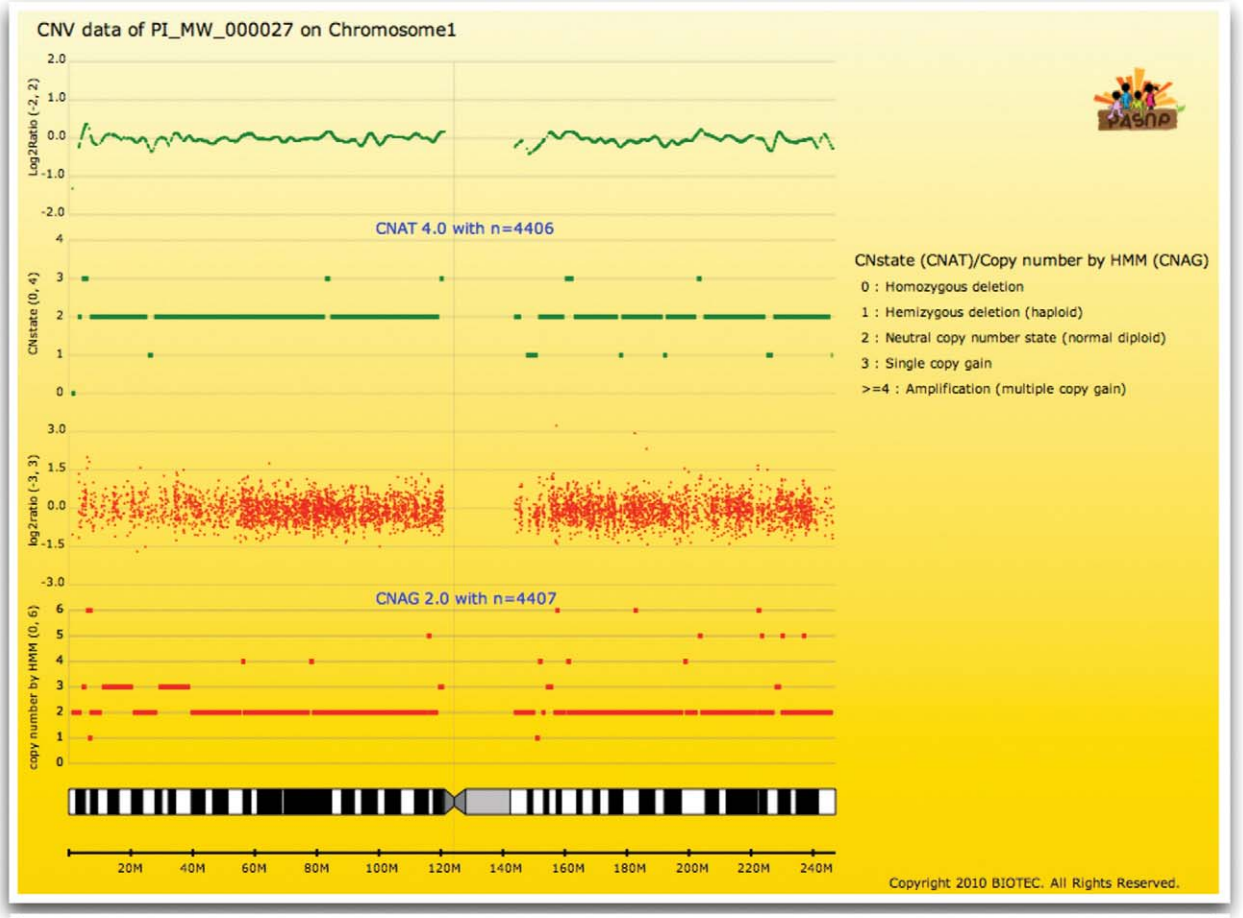


Figure 1. Representation of Haplotype blocks A) haplotype blocks calculation and population selection panel B) SNPs and genes located on chromosome 20 between 52–53 Mb displayed in SVG C) haplotype blocks of the selected populations and D) detailed information (block frequency, tag SNPs) of haplotype blocks displayed by clicking on the SVG view.
doi:10.1371/journal.pone.0021451.g001

A



B



C

CNV Data of PI-AT Population on chromosome 1 — Back

No.	Sample ID	CNV Type	Analyzed Program	Analyzed State	Position	Length (bp)	Number of SNP
1.	PI-AT_000017	Homozygous deletion	CNAG	0	Chr1 : 81778439 - 81785472	7034	3
2.	PI-AT_000017	Homozygous deletion	CNAG	0	Chr1 : 234361448 - 234362983	1536	8
3.	PI-AT_000018	Homozygous deletion	CNAG	0	Chr1 : 217155402 - 217184590	29189	6
4.	PI-AT_000018	Homozygous deletion	CNAG	0	Chr1 : 233084260 - 233084829	570	3
5.	PI-AT_000024	Homozygous deletion	CNAG	0	Chr1 : 94303912 - 94304194	283	3
6.	PI-AT_000043	Homozygous deletion	CNAG	0	Chr1 : 198268749 - 198413314	144566	5
7.	PI-AT_000058	Homozygous deletion	CNAG	0	Chr1 : 85398443 - 85499806	101364	3

Analyzed State : **0**:Homozygous deletion, **1**:Hemizygous deletion, **2**:Neutral copy number state, **3**:Single copy gain, **4-6**:Multiple copy gain

Export CNVs

Figure 2. Copy Number Variation view A) interface to view CNV information B) CNV data of each individuals in SVG, showing log2ratio of signal intensity plots and called states from CNAT 4.0 and CNAG 2.0 programs C) individual CNV results on each chromosome corresponding with CNV type selected in panel A.
doi:10.1371/journal.pone.0021451.g002

mapping to autosomal and sex chromosomes respectively for each individual.

Haplotype inference and block partitioning

Haplotype blocks were predicted exclusively on autosomal chromosomes using HaploBlockFinder [10] using 1928 individuals from 75 populations (excluding AX–AI) based on the four gamete test (FGT) assumption with parameters:

–A3 –D0.8 –B0.01 –M1 –T1 –P0.8 –Q0.2

The haplotypes of each block were inferred using fastPHASE [11] with parameters:

–T20 –C50 –Km1000 –Kp.05

The blocks and their haplotypes are stored in the database and can be graphically displayed through the web interface shown in Figure 1. Detail on SNP distribution of each chromosome is listed in Table S1.

Copy number variation analysis

Copy number analyses were done using Copy Number Analysis Tools version 4.0 (CNAT4.0) [12] and Copy number analyzer for GeneChip(CNAG 2.0) [13]. Since the focus is on the population level, un-paired sample analysis with 1 Mbps genomic smoothing was used in these analyses. Male and female data were analyzed separately for chromosome X. The CNV graphical interface shown in Figure 2 displays the $\log_2$ratio of the probe intensities and CNstate/N_AB results from CNAT4.0 and CNAG2.0 respectively. More information on CNV analysis can be found in Text S2.

Conclusion

Following the publication of the HUGO Pan-Asian SNP consortium study of human genetic diversity in Asia, it became apparent that there was a need for an information resource which integrates the Asian data with other worldwide populations and presents this data is a user friendly format. Similar to the HapMap initiative, PanSNPdb offers genome structural information pertaining to Asian populations in a familiar graphical comparative view based on GBrowse where SNP genotyping from multiple populations can be visualized on the same page. This database also offers pre-computed information of LD blocks and their haplotypes on each chromosome; such information for each population can be visualized both in table and SVG formats and can be exported for future use. Furthermore, users can adjust the number of SNPs for haplotype inferencing and calculate this using Haploview, which is performed by our server. In terms of genome structure, we calculated the CNV information using un-paired sample analysis whose information, e.g., log2ratio and CNV state for individual visualization (SVG) and CNV state at the population level (GBrowse) comparing with CNV information from the database of genomic variations. The database is available for public access at: http://www4a.biotec.or.th/PASNP.

This database offers a comprehensive catalog of Asian population genotypes, which is compatible with the HapMap project. It also serves as the main genotyping repository of the Pan Asian SNP consortium which will contribute further Asian specific genetic information in the future. We anticipate that newer Asian populations with denser genotyping platform along with their analyses from the consortium will be deposited into PanSNPdb. With the advent of more cost effective whole genome sequencing technology, other structural genomic variations among Asian populations will also be explored.

Methods

System Design and Implementation

PanSNPdb manages the genetic variation data, reference information and precomputed haplotypes using the open source database management MySQL version 5.5.1. The web interface was constructed using the content management system (CMS) Plone version 3.3.5. Python scripting language was used to connect to MySQL and draw scalable vector graphic (SVG) images of precomputed haplotype blocks and CNV log2ratio signals. PanSNPdb adopts GBrowse to display population-level comparison of SNP and CNV locations on genes and chromosomes. The database system is hosted by a dedicated computer server equipped with 2xAMD 6-core with a clock speed of 2.8 GHz using 64 Gigabytes of memory and 2 Terabytes of hard disk space.

The PanSNPdb database was constructed using the genotyping information described in [6] consisting of 1,928 unrelated individuals representing 71 Asian populations and 4 populations from HapMap. Information related to each population, such as geographical, ethnic and linguistic data were added to the database; this information is provided in Table S2 and can be visualized through the PanSNPdb web interface. The database was designed and implemented so as to facilitate comparison with genotyping information from other public data sources including HapMap, dbSNP and JSNP. To locate SNPs, the Reference Sequence of Human Genome build 36.3 is used as the template. Since these SNPs may be useful for medical genetic studies, the gene-disease information published by GeneCards was incorporated into the database. These reference data were downloaded, and will be periodically updated when newer versions are announced. Furthermore, copy number variations from the PanAsian SNP dataset were inferred using CNAT and CNAG for future CNV referencing of Asian populations. CNV data from the database of genomic variants (DGV) [14] were incorporated into PanSNPdb so that the comparative view of CNVs across different populations can be rendered. Figure 3A presents the main data sources of the PanSNPdb. Consequently, the comprehensive information in this PanSNPdb can be considered as worldwide data collection, but with special emphasis on Asian populations.

Graphical interface of the data

Figure 3B shows how the graphical interface of PanSNPdb was constructed. In PanSNPdb, SNPs and their corresponding information can be located graphically on the reference sequence along with SNPs from other populations in different tracks. This visualization is made possible using the GBrowse visualization engine [15]. SNPs can be searched via four main entry points: 1) chromosomal location 2) gene name/gene id 3) SNP id or rs number and for medical purpose 4) disease name from GeneCards that are associated with disease-related genes. Similarly, the CNV region information can also be visualized using GBrowse along with other CNVs from DGV.

Haplotype blocks were also inferred at the chromosome level (autosomes) with overlapping regions (see Table S1 for

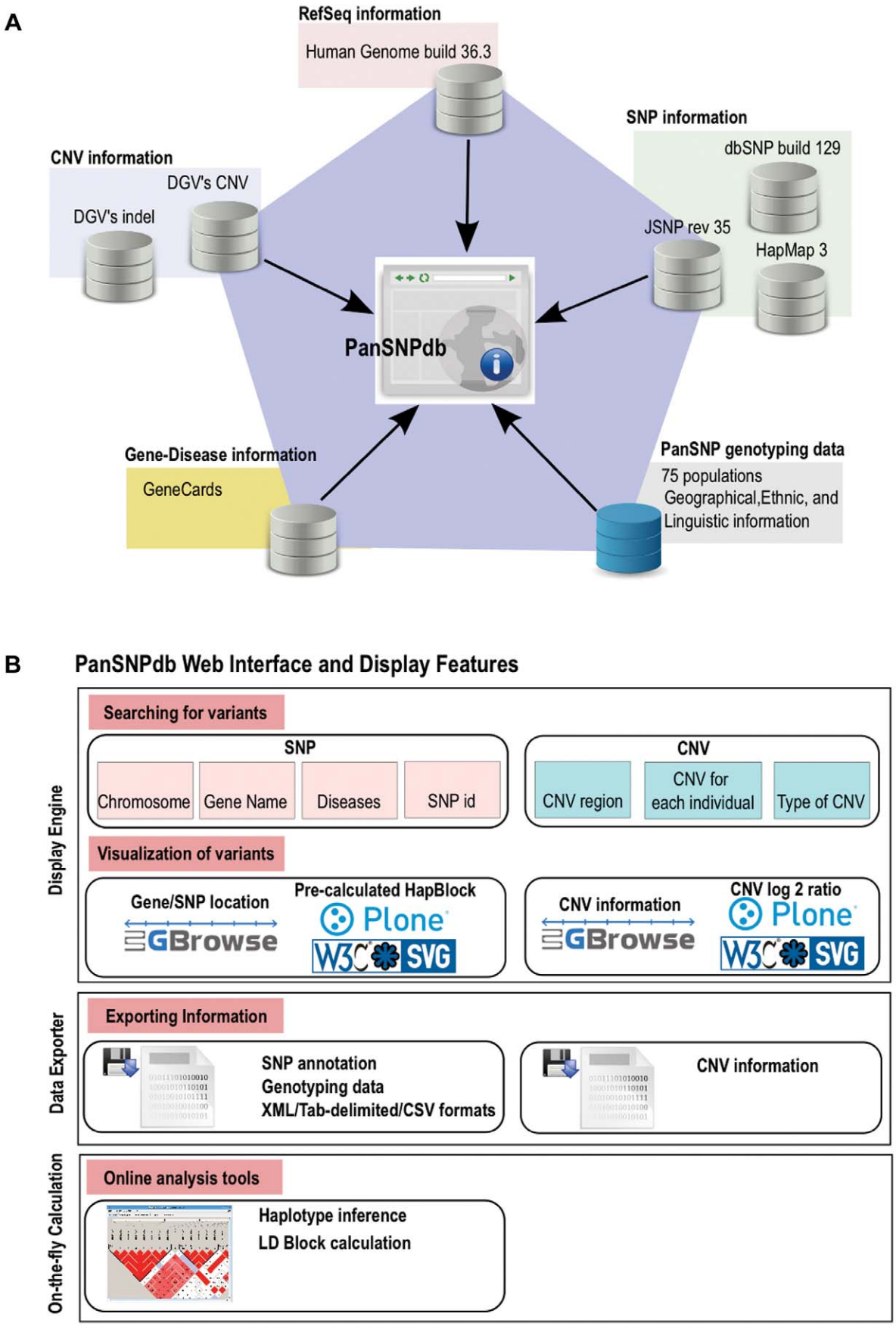


Figure 3. Structure of PanSNPdb A) Architecture of PanSNPdb showing integration of different data sources B) PanSNPdb Web interface and display features.
doi:10.1371/journal.pone.0021451.g003

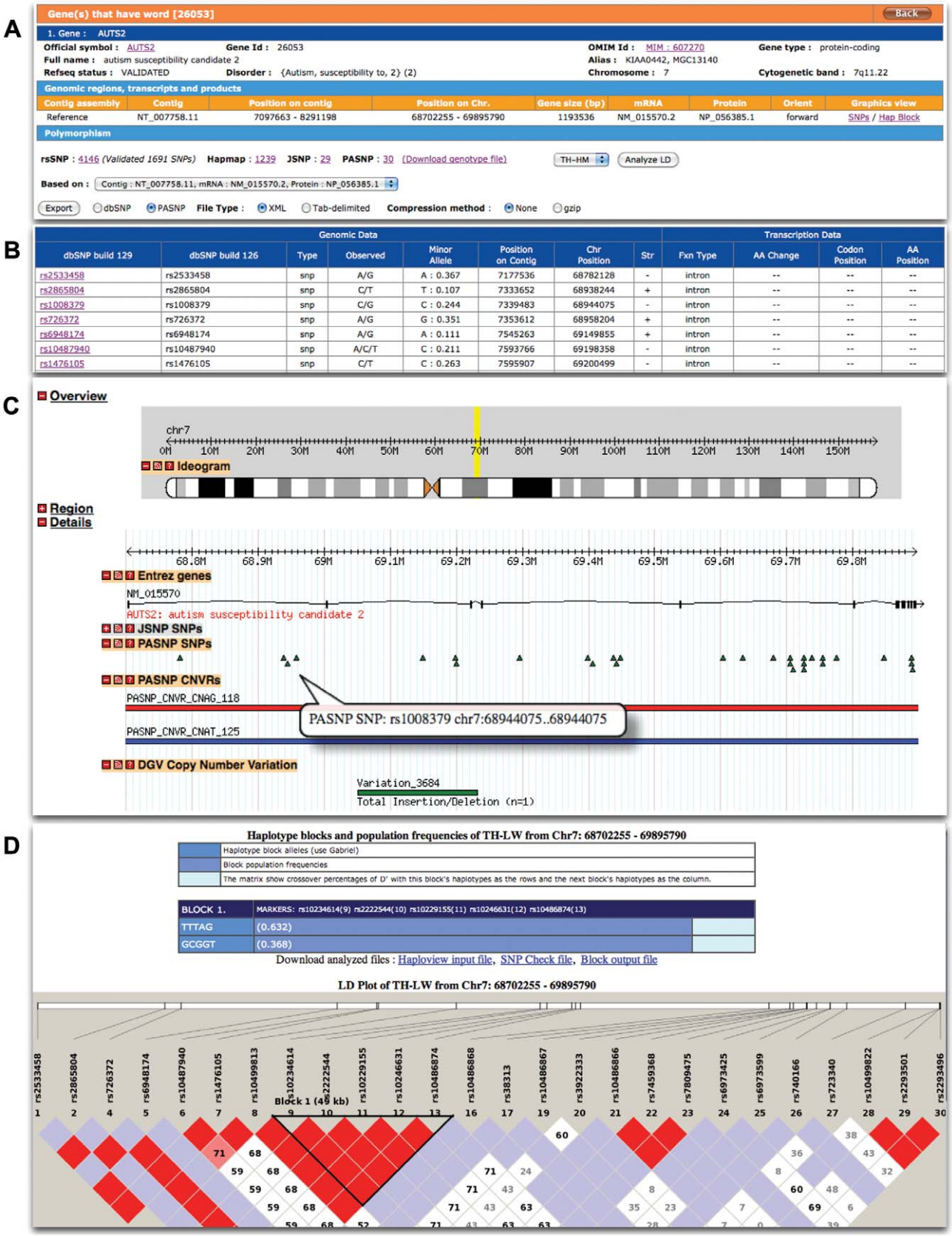


Figure 4. PanSNP results in rich text and graphical formats A) gene and SNPs information that provide export and analyze features B) SNPs and associate informations C) SNPs and genes display with GBrowse D) haplotype blocks calculation with built-in haploview.

doi:10.1371/journal.pone.0021451.g004

distribution of SNPs on each chromosomes). The results can be displayed graphically in any web browser with scalable vector graphic (SVG) supported. The Haploview tool [16] is also integrated into the PanSNPdb website; users can adjust the haplotype inferencing parameters in order to recalculate haplotype blocks "on-the-fly". Lastly, PanSNPdb allows users to export SNP and CNV data, such as location of SNPs, genotyping and CNV data of each individual (in comma separated value (CSV) and/or tab delimited formats). Figure 4 show representative SNP data with beautified text format and a user-interactive graphical view.

Supporting Information

Text S1 The participants of the HUGO Pan-Asian SNP Consortium are arranged by surname alphabetically.
(DOC)

Text S2 PanSNPdb CNV analysis.
(PDF)

Table S1 Total number of SNPs and CNVs (map on RefSeq Genome Build 36.3).
(XLS)

Table S2 Information of 71 Pan-Asian and 4 HapMap populations.
(XLS)

Acknowledgments

The authors wish to thank all the anonymous subjects who contributed their information to PanSNPdb. The authors also acknowledge the National Center for Genetic Engineering and Biotechnology (BIOTEC), and the National Science and Technology Development Agency (NSTDA) for allowing us to host this database on the web/database server, and open for public access.

Author Contributions

Conceived and designed the experiments: ST. Performed the experiments: CN JOY HG JB SX. Analyzed the data: AA PJS SX ST. Contributed reagents/materials/analysis tools: EL ST The HUGO Pan-Asian SNP Consortium. Wrote the paper: ST SX PJS AA.

References

1. Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, et al. (2008) The genetic structure of Pacific Islanders. PLoS Genet 4: e19.
2. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998–1003.
3. Kayser M, Lao O, Saar K, Brauer S, Wang X, et al. (2008) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. Am J Hum Genet 82: 194–198.
4. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.
5. Pennisi E (2010) 1000 Genomes Project gives new map of genetic diversity. Science 330: 574–575.
6. HUGO Pan-Asian SNP Consortium (2009) Mapping human genetic diversity in Asia. Science 326: 1541–1545.
7. International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
8. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2011) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 39: D38–51.
9. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, et al. (2002) JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30: 158–162.
10. Zhang K, Jin L (2003) HaploBlockFinder: haplotype block analyses. Bioinformatics 19: 1300–1301.
11. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–644.
12. Jacobs S, Thompson ER, Nannya Y, Yamamoto G, Pillai R, et al. (2007) Genome-wide, high-resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using microarrays. Cancer Res 67: 2544–2551.
13. Komura D, Shen F, Ishikawa S, Fitch KR, Chen W, et al. (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res 16: 1575–1584.
14. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al. (2004) Detection of large-scale variation in the human genome. Nat Genet 36: 949–951.
15. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, et al. (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12: 1599–1610.
16. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and structure

Tulaya Limpiti[1], Apichart Intarapanich[2], Anunchai Assawamakin[3], Philip J Shaw[3], Pongsakorn Wangkumhang[3], Jittima Piriyapongsa[3], Chumpol Ngamphiw[3] and Sissades Tongsima[3*]

Abstract

Background: The ever increasing sizes of population genetic datasets pose great challenges for population structure analysis. The Tracy-Widom (TW) statistical test is widely used for detecting structure. However, it has not been adequately investigated whether the TW statistic is susceptible to type I error, especially in large, complex datasets. Non-parametric, Principal Component Analysis (PCA) based methods for resolving structure have been developed which rely on the TW test. Although PCA-based methods can resolve structure, they cannot infer ancestry. Model-based methods are still needed for ancestry analysis, but they are not suitable for large datasets. We propose a new structure analysis framework for large datasets. This includes a new heuristic for detecting structure and incorporation of the structure patterns inferred by a PCA method to complement STRUCTURE analysis.

Results: A new heuristic called EigenDev for detecting population structure is presented. When tested on simulated data, this heuristic is robust to sample size. In contrast, the TW statistic was found to be susceptible to type I error, especially for large population samples. EigenDev is thus better-suited for analysis of large datasets containing many individuals, in which spurious patterns are likely to exist and could be incorrectly interpreted as population stratification. EigenDev was applied to the iterative pruning PCA (ipPCA) method, which resolves the underlying subpopulations. This subpopulation information was used to supervise STRUCTURE analysis to infer patterns of ancestry at an unprecedented level of resolution. To validate the new approach, a bovine and a large human genetic dataset (3945 individuals) were analyzed. We found new ancestry patterns consistent with the subpopulations resolved by ipPCA.

Conclusions: The EigenDev heuristic is robust to sampling and is thus superior for detecting structure in large datasets. The application of EigenDev to the ipPCA algorithm improves the estimation of the number of subpopulations and the individual assignment accuracy, especially for very large and complex datasets. Furthermore, we have demonstrated that the structure resolved by this approach complements parametric analysis, allowing a much more comprehensive account of population structure. The new version of the ipPCA software with EigenDev incorporated can be downloaded from http://www4a.biotec.or.th/GI/tools/ippca.

Background

As genotyping platforms incorporate more markers, and the costs for genotyping keep falling, ever larger and more complex datasets are being analyzed. The computationally efficient non-parametric methods for analysis of genotypic datasets are thus increasingly being used to reveal population structure. Resolution of population structure reveals evolutionary relationships between groups of individuals. Furthermore, population structure must be accounted for in genome-wide association studies to reduce spurious associations resulting from ancestral differences between cases and controls [1].

Principal component analysis (PCA) is a widely used non-parametric method for population structure analysis, which uses a covariance matrix for eigenanalysis. The amount and axes of variation among individuals are

* Correspondence: sissades@biotec.or.th
[3]National Center for Genetic Engineering and Biotechnology, Thailand Science Park, Pathumthani 12120, Thailand
Full list of author information is available at the end of the article

captured in the eigenvalues and eigenvectors, respectively. Previously, we developed a PCA framework for population structure analysis which extended the use of PCA beyond its usual application for visualizing the population structure trend by employing an iterative process to simplify the pattern of population structure. The iterative methods used by others, e.g. [2,3] rely on the available ethno-geographical population labels for subjectively grouping individuals, unlike our objective approach.

Our framework, which we dubbed iterative pruning PCA (ipPCA) uses a clustering algorithm to assign individuals into subpopulations without imposing any prior assumptions [4]. ipPCA resolves all subpopulations in a population dataset, and thus reports the total number of primal subpopulations K in addition to assigning individuals contained within them. The term "population" is synonymous with dataset for ipPCA, which is the entire collection of individuals available for analysis. The term "subpopulation" defines a group of individuals assigned by ipPCA in which no further significant substructure is present. ipPCA operates by systematically separating individuals into two clusters using a clustering algorithm based on the Euclidean distances between projected data points and the cluster centroids. The decision to separate individuals requires testing of whether significant structure is present within the dataset (or nested dataset for subsequent iterations of the algorithm). To test for homogeneity among groups of individuals, we previously proposed using the test statistic as implemented in the EIGENSTRAT/SmartPCA algorithm, which reports the probability of structure according to Tracy-Widom (TW) distribution [5]. If no significant structure exists, then the individuals under testing belong to a subpopulation, thus terminating the iterative clustering process. The ipPCA framework is summarized in Figure 1. Using datasets of simulated and real data, we showed how

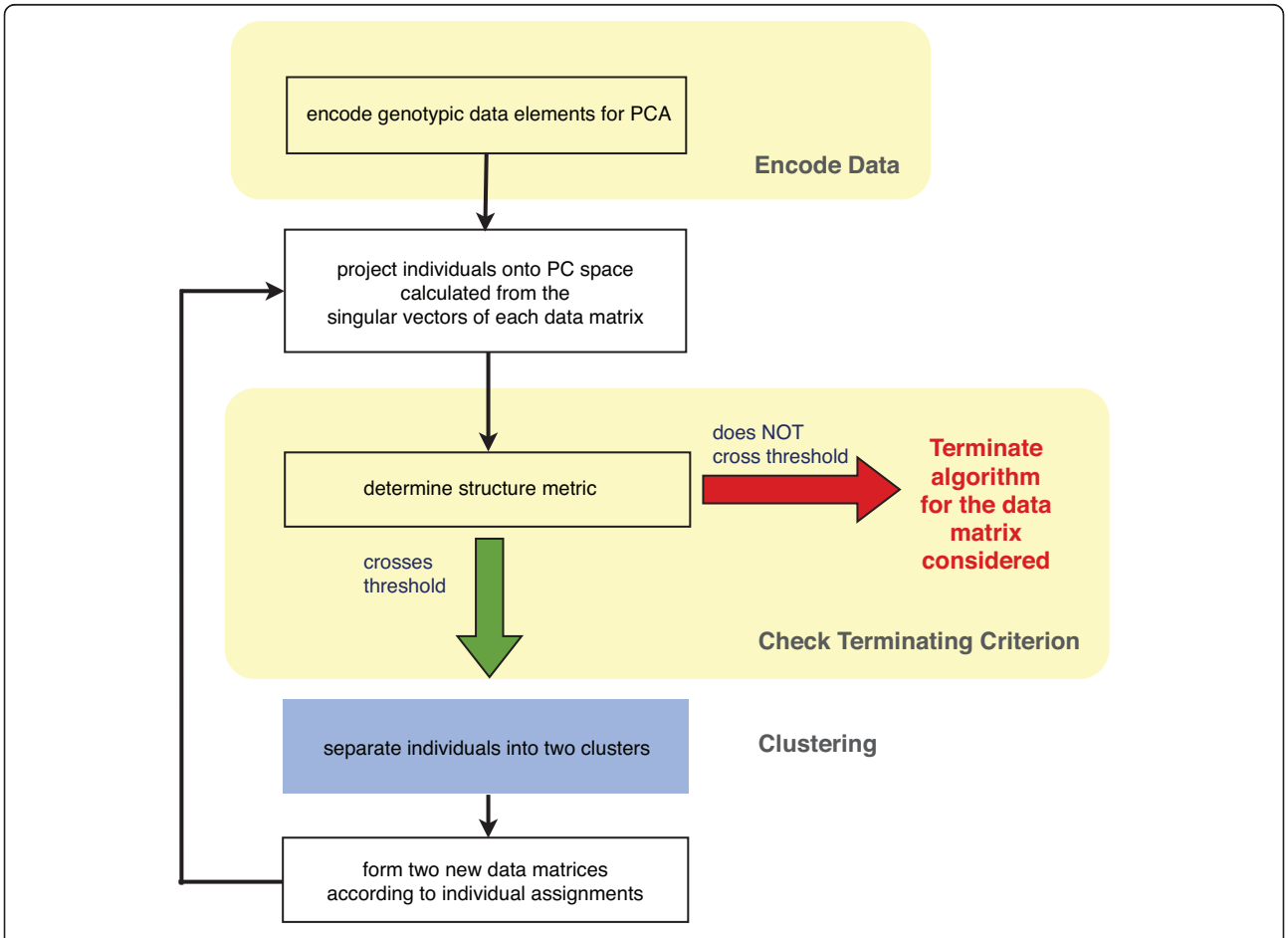


Figure 1 Outline of the ipPCA framework. The framework consists of three main components. First, the genetic data are encoded, zero-means centered and normalized. Then, individuals are projected onto a space spanned by the principal components of the input data matrix. Next, a structure metric is calculated to decide whether to advance to the clustering step or to terminate the algorithm. When the metric does not cross the threshold, a homogenous subpopulation is resolved and subsequently the algorithm terminates. Otherwise, the individuals are bisected. The algorithm iterates until all individuals have been assigned into terminal subpopulations.

ipPCA can correctly assign individuals to subpopulations and infer K. However, the accuracy of ipPCA may be affected by the stopping criterion. An inappropriate termination criterion leads to under- or over-estimation of the number of subpopulations. Moreover, individual assignment errors in early iterations will be compounded and carried forward to later iterations.

Parametric algorithms for clustering individuals into subpopulations, e.g., STRUCTURE, *frappe*, ADMIXTURE, and BAPS, differ from ipPCA in one crucial aspect, namely the method of assigning individuals into subpopulation clusters. The aforementioned parametric algorithms infer ancestral proportions for each individual separately, and group individuals with similar patterns of inferred ancestry. ipPCA and other non-parametric approaches cannot infer ancestry. These techniques attempt to group individuals with similar genetic profiles together. Hence, parametric approaches still offer important information not seen by non-parametric analyses. Large and highly structured population datasets are however intractable for parametric analysis because the number of K ancestral clusters is limited. This is due to the limited number of available samples used to estimate subpopulation allele frequencies. In order to better observe the inherent population structure, a "supervised" structure analysis, with re-sampled individuals, should be performed. The choice of individuals for such supervised analysis is arbitrary and typically guided by available ethno/geographical labels. Nonetheless, careful selection is needed to ensure that individuals being compared have similar ancestries, otherwise the signals of ancestries important for differentiating some groups of individuals may be too weak.

In this paper, we propose a modification to ipPCA by introducing a new stopping criterion called EigenDev for the iterative clustering process which is more robust to spurious patterns in large datasets. The new algorithm is termed EigenDev-ipPCA. To distinguish between the two algorithms in the ipPCA framework, we refer to the previously proposed algorithm which uses the TW statistic as the termination criterion as TW-ipPCA in the subsequent sections. Furthermore, we suggest a new protocol which uses the information from EigenDev-ipPCA to guide parametric analysis. Using real datasets, we demonstrate how this approach can reveal new and structure-informative patterns of ancestry not detectable with unsupervised STRUCTURE analysis.

Methods

New ipPCA terminating criterion

The Tracy-Widom (TW) test statistic, which is implemented in the EIGENSTRAT/SmartPCA algorithm [5], is used as a stopping criterion for the TW-ipPCA algorithm. Although this stopping criterion has been found to work well for some datasets, we found that when much larger datasets containing roughly >1000 individuals were analyzed, the TW-ipPCA resolved far more subpopulations than were expected. We therefore suspected that in some cases when sampling is large, the subpopulations resolved may be spurious, i.e., type I error. Indeed, as pointed out in [5], the relative sample sizes of the underlying subpopulations affect the TW test statistic.

Besides the type I error we found when using the TW statistical test for structure, there are other drawbacks which motivated us to develop an alternative terminating criterion. The first issue is computational difficulty. To obtain the final value of the TW test statistic, too many unknown parameters need to be estimated. No best estimators for these parameters are available, so choices of estimators affect the result. Instead of using the p-values of TW test statistics as thresholds, we propose a new terminating criterion for determining whether the data are structured. The new criterion is based on the eigenvalues of the data matrix and is termed the *EigenDev* heuristic. The EigenDev heuristic follows the same assumption as the TW theory, namely, if the first eigenvalue of the data matrix is significantly larger than the remaining eigenvalues, then substructure exists. However, we extend this observation beyond merely testing the significance of the first eigenvalue to take into account the remaining variance of the data. This allows us to observe structure in higher dimensions. We were inspired to develop EigenDev from the Eigenvalue Grads heuristic, which is applied in the signal processing domain [6]. This work showed that if the data contain only noise and no signal, i.e., non-structured, then there is an excellent linear fit for the eigenvalues ranked in descending order. In population genetic data, the noise represents the natural genetic variation within a (sub)population.

To test for population structure, the EigenDev statistic is calculated from the genotypic data. This calculation first requires that a data matrix is constructed from encoded, zero-means and normalized genotypic data, as described in [5]. This matrix contains rows corresponding to individuals and columns corresponding to alleles. Thus, biallelic SNP markers are encoded by entries in two columns, one for each allele, and STRs by the total number of alleles for that marker locus in the dataset. The presence of an allele is encoded as 1 and its absence as 0. For missing data, i.e., markers with no genotypic call, they are encoded as all 0's.

Given the zero-means, normalized genotype data matrix $\mathbf{X}$ (according to [5]) containing m samples with n allele columns per sample, we construct the sample covariance matrix

$$\mathbf{C} = \frac{1}{m}\mathbf{X}\mathbf{X}^T.$$

The EigenDev value can then be computed from

$$\text{EigenDev} = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(\log(\hat{\sigma}_i^2) - \log(\sigma_i^2))} \quad (1)$$

where

$$\log(\hat{\sigma}_i^2) = \log(\sigma_1^2) + (i-1)c \quad (2)$$

and

$$c = \frac{(\log(\sigma_p^2) - \log(\sigma_1^2)}{(p-1)} \quad (3)$$

where $\sigma_i^2, i = 1, ..., p$, are the first p eigenvalues of $\mathbf{C}$ ranked in descending order. The quantity in Eq. (1) could be negative in some cases. To militate against this possibility, the encoded entries are normalized to have zero mean. This step is important to remove the signal from the common elements, leaving only the differences (genetic variance) between individuals for eigenanalysis. In all empirical studies on both simulated and real data, we found that 90% of the variance in the data always results in a positive value and the convexity constraint in question has never been violated. To account for the rare cases when negative values are encountered, we have included a checking step in the algorithm to detect and report negative values. If negative values are found, the parameter p can be adjusted to ensure a positive quantity in the square root. Recall that $p < \min\{m, n\}$ is the number of eigenvalues used to compute the EigenDev statistic. We also stabilize the variance using log transformation. If the EigenDev value is large, the group of individuals being analyzed would comprise more than one subpopulation and ipPCA progresses to bisect the group; otherwise, the EigenDev-ipPCA algorithm terminates when the EigenDev value falls below a threshold.

Results

Testing

To test the EigenDev concept, several datasets were analyzed:

1. A simulated dataset composed of 10,000 individuals from the same population, each containing 10,000 SNP markers was used for testing the fit of TW distribution. It was generated using the GENOME tool [7] with the following parameters and the following tree file:

```
-pop 1 10000 -N tree.txt -C 20 -S 500
tree.txt:
0 10000
1-1
1 10000
```

Starting at 10,000 founder population individuals, GENOME generates the first generation with the same size as the founder. Each individual has 20 chromosomes and each chromosome contains 500 SNPs.

2. The second dataset was simulated using the same GENOME parameters as the first dataset but with different tree file:

```
tree.txt:
0 5000 5000
1-1 2-1
40 5000
1-1 1-2
80 5000 5000
1-1 2-1
100 10000
```

to generate two subpopulations of size 5,000 individuals each.

3. The third dataset is the Bovine HapMap Project collection of 497 individuals obtained from 19 different breeds, genotyped for 27203 SNPs. It is publicly available from [8].

4. The fourth dataset is publicly available from [9]. It contains 3945 individuals comprising 185 different ethno/geographical labels, typed for 1327 markers (consisting of 848 microsatellites, 476 indels, and 3 SNPs) from [10].

The ipPCA encoded input matrices from the simulated and real complex datasets are also available for download from http://www4a.biotec.or.th/GI/tools/ippca.

Testing metrics for population structure

To test how TW is affected by sampling, a simulated dataset with no substructure was sampled randomly at 20 different sample sizes from 10 to 200 individuals. The corresponding probability-probability (p-p) plots for testing the fit of the TW distribution are shown in Figure 2. It is observed that the TW distribution is violated for most of the sample sizes; good fit is observed only for the sample of 70 individuals. Therefore, the deviation from TW distribution will give a false detection (type I error), particularly for large sample sizes. On the other hand, the TW test is very sensitive for detecting structure, since it is based on a non-linear phase change. It is not susceptible to type II error provided sufficient data are available [5]. However, the non-linearity of the phase change means that an all-or-nothing situation exists where the likelihood of type I cannot be controlled, even across a wide range of p-value thresholds.

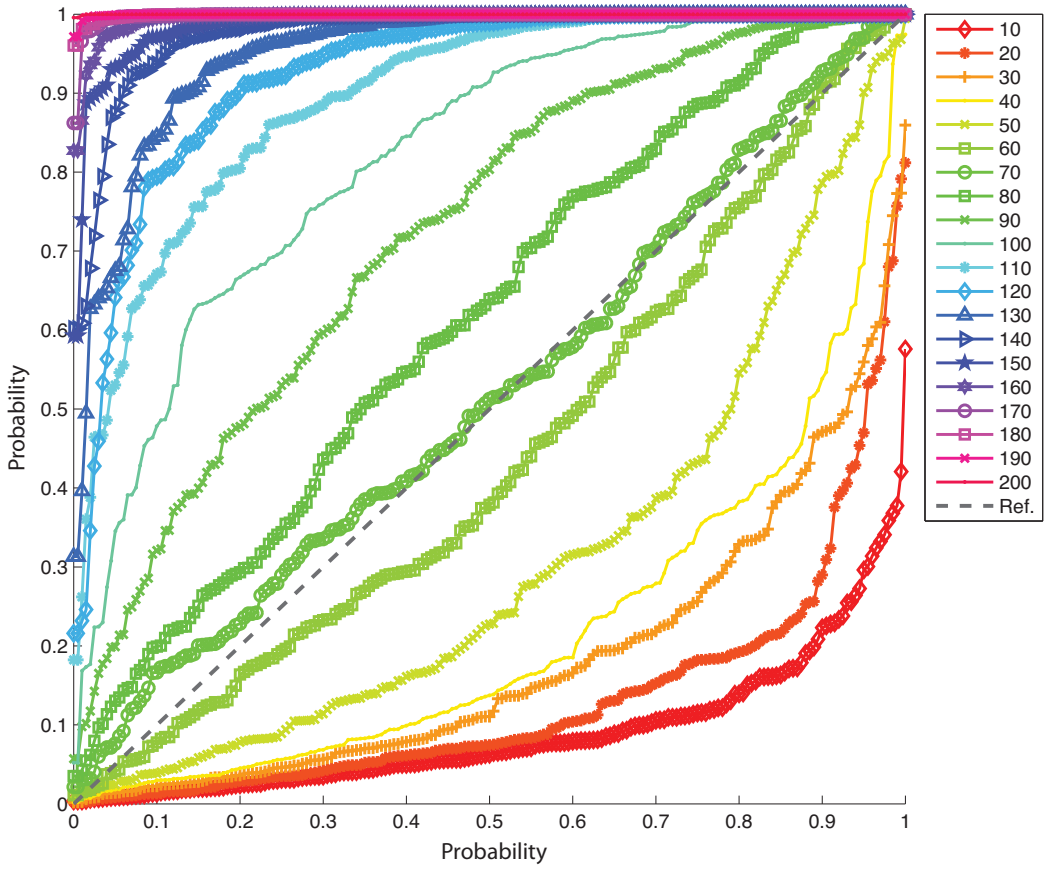


Figure 2 Testing the fit of the TW distribution. A population of size 10,000 individuals with 5,000 markers was simulated using the coalescent model. The p-p plots were generated for sample sizes of 10 to 200 individuals.

To test the performance of the EigenDev heuristic, we simulated the receiver operating characteristic (ROC) curves for three different sample sizes of 100, 200, and 500 individuals from the second simulated dataset, as shown in Figure 3. To obtain the curves, the EigenDev threshold was varied between 0.077 and 0.387. It is observed that the threshold value increases with samples size, and that EigenDev performs better when the sample size is large. An EigenDev threshold of 0.21 was used for analysis of real datasets. This value is an average of the thresholds needed to achieve a 10% false positive rate for the three sample sizes. This value is a good compromise between detecting and resolving all structure present, with minimal spurious structure at typical sample sizes in real datasets.

Guiding parametric analysis with ipPCA

STRUCTURE [11] can be used to perform unsupervised clustering using ancestral components information. However, the high computational complexity of STRUCTURE, especially in finding the maximum posterior probabilities for the number of K ancestral clusters limits practically to K = 20 or fewer. Therefore, highly complex datasets must be divided into sub-datasets, which are then analyzed separately by STRUCTURE. Conventionally, this is done in an arbitrary fashion using prior information, e.g., ethno-geographical population labels. However, the prior information could bias the clustering results. To address this issue, we propose using the unsupervised clustering feature of ipPCA to assist in narrowing the search space for STRUCTURE in a more efficient fashion. In practice, subpopulations assigned by ipPCA can be selected for subsequent STRUCTURE analysis. We call this approach ipPCA-guided STRUCTURE. We applied this method to the Bovine HapMap dataset [8], which is the expanded dataset from the one previously analyzed by us [4]. The result was similar to that reported earlier, i.e. EigenDev-ipPCA resolved 18 subpopulations, each of which are

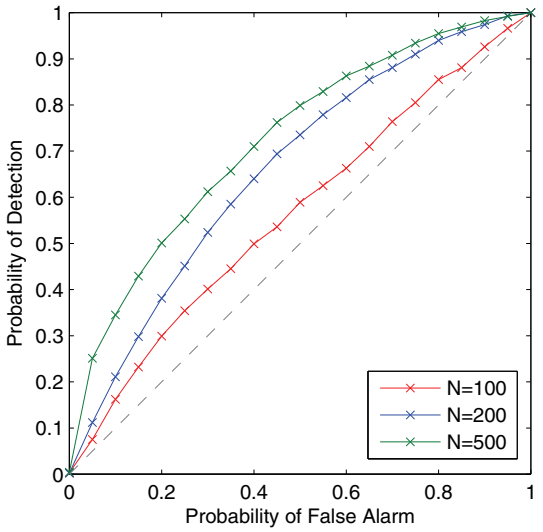


Figure 3 The empirical receiver operating characteristic curve of the EigenDev heuristic. A structured population of 10,000 individuals of 5,000 markers containing two subpopulations (5,000 each) was simulated using the coalescent model. The ROCs were generated for sample sizes of 100, 200, and 500 individuals.

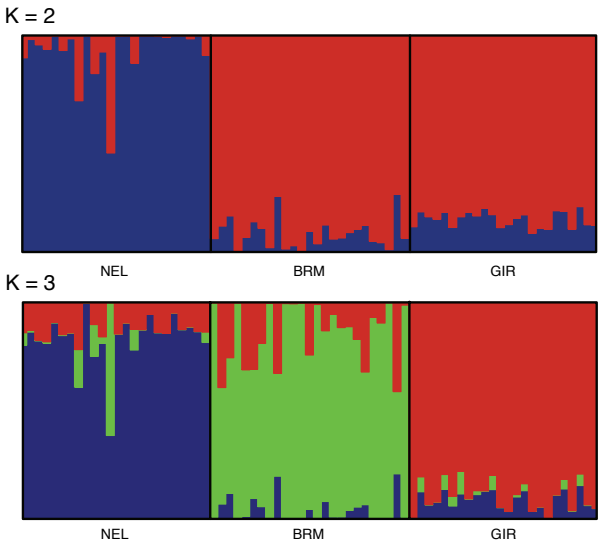


Figure 4 ipPCA-guided STRUCTURE analysis on selected individuals from Bovine HapMap dataset. STRUCTURE analyses were performed on individuals from *B. indicus* breeds (GIR, BRM, and NEL). Results with $K = 2$ and $K = 3$ are shown.

largely composed of individuals of the same breed, except for one subpopulation containing Angus (ANG) and Red Angus (RGU) individuals (the EigenDev-ipPCA results can be viewed from the ipPCA download webpage).

STRUCTURE was used with the default parameters and 10,000 burn in and 10,000 run iterations. Individuals from the Gir (GIR), Brahman (BRM), and Nelore (NEL) breeds resolved as three separate subpopulations by EigenDev-ipPCA) were selected for STRUCTURE analysis to determine whether differences in inferred ancestry exist between these breeds. Furthermore, these three breeds were chosen because they are *B.indicus* breeds, and thus more closely related to each other than the other *B.taurus* breeds in the dataset. STRUCTURE analysis at $K = 3$ on these selected individuals, as shown in Figure 4, revealed breed-distinctive patterns of ancestry not previously reported.

Analysis of a large human dataset by ipPCA

The dataset from Tishko et.al. [10] contains a large number of individuals (3945). Furthermore, these individuals comprise 185 ethno-linguistic distinguishing labels suggesting a large number of genetically distinct groups. The dataset was analyzed by EigenDev-ipPCA, which assigned 49 subpopulations (Figure 5). The assigned subpopulations were largely consistent with the patterns reported earlier [10], in which geographically disparate groups of individuals are genetically distinct, and within Africa, major cultural and linguistic groups are also genetically distinct (see Additional file 1 for more information). In contrast, ipPCA using the TW stopping criterion (TW-ipPCA) assigned 109 subpopulations. Comparison of the subpopulations which differed between the two methods showed that on the whole, subpopulations assigned by TW-ipPCA were sub-clusters of larger subpopulations assigned by EigenDev-ipPCA. For instance, all Indian individuals (15 ethnic labels) were assigned to two subpopulations (SP2 and SP7) by EigenDev-ipPCA, whereas Indians were assigned to 11 subpopulations by TW-ipPCA (see Additional file 1).

ipPCA-guided STRUCTURE analysis

African American is a term used to describe US nationals with self-identified African ancestry, the majority of whom are descended from West African individuals who came to the US via the slave trade. The term African American though is very broad, as it encompasses individuals descended from African ancestors from a broad geographical range, and some also have recent non-African ancestry. African American individuals were assigned into four subpopulations by EigenDev-ipPCA, namely SP4, SP5, SP15 and SP16. Subpopulations SP4 and SP5 contain the majority of African Americans together with predominantly West and Central African Niger-Khordofanian speaking ethnic groups. Five African Americans were assigned to SP15, which

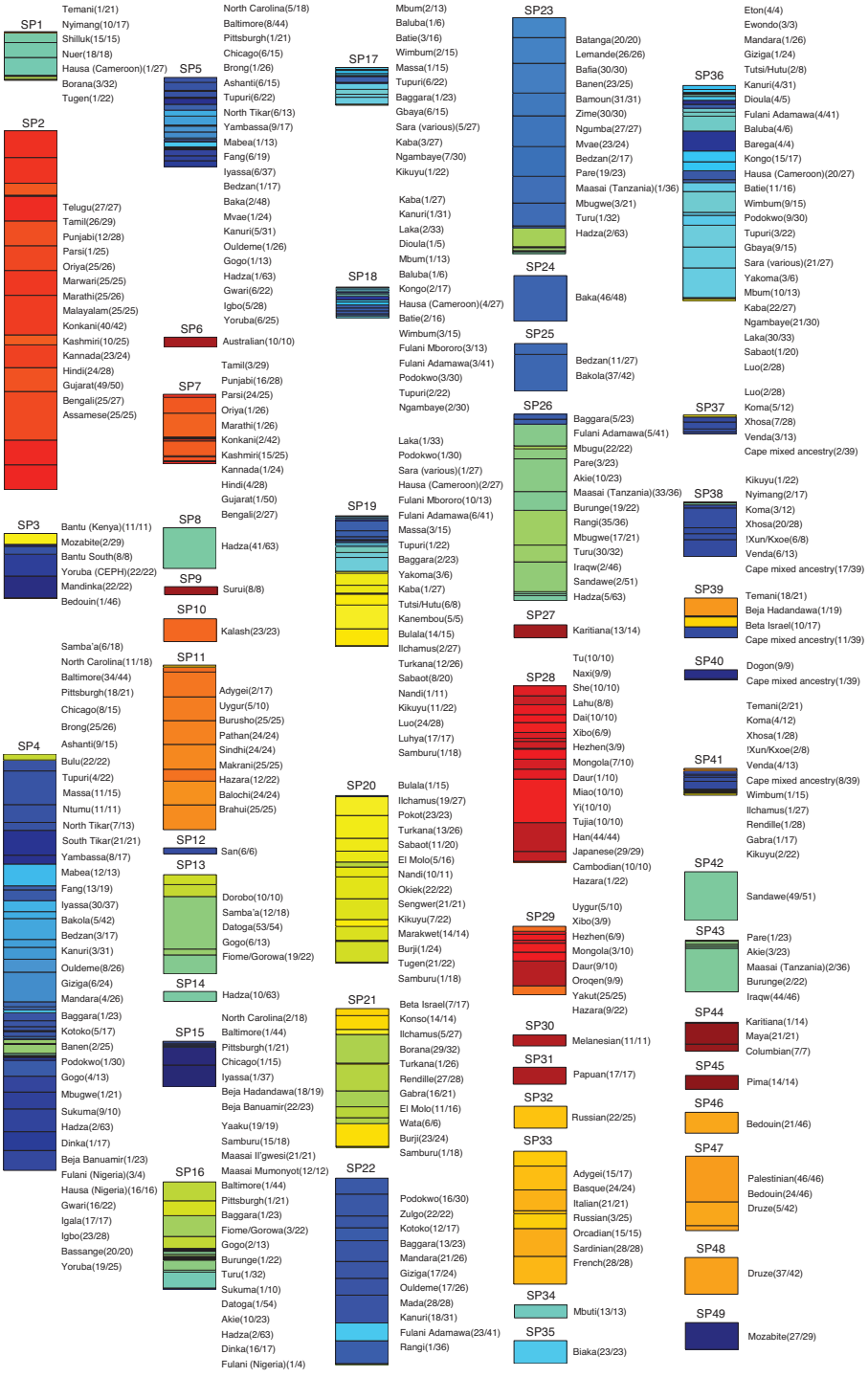


Figure 5 Population assignments of the Tishkoff et al dataset using the EigenDev-ipPCA method. 49 assigned subpopulations are labeled SP1 to SP49. The height of the bars are proportional to the number of assigned individuals in each subpopulation. The population labels of the assigned individuals are shown to the right of each bar with the number of individuals with the same label in parentheses. To aid visualization of the individual assignment, the 185 population labels were grouped into 14 color groups reflecting geographical regions. Color gradients within the color group denote different population labels. For the complete color scheme, see Figure s3 in the Additional file 1.

contains predominantly Afroasiatic Cushitic speaking Bejans from Sudan. Two African Americans were assigned to SP16, which contains predominantly East Africans of mixed Nilo-Saharan Sudanic and Afroasiatic Cushitic speaking ethnic groups.

We then used the information from EigenDev-ipPCA to guide STRUCTURE. All the individuals assigned to SP4, SP5, SP15 and SP16, which included all African-American individuals, were analyzed by STRUCTURE from $K = 2$ to $K = 5$ (see part A in Figure 6). At $K = 3$ or greater, each of the four subpopulations assigned by EigenDev-ipPCA showed distinctive patterns of ancestry, although there appeared to be some overlap between SP15 and SP16 individuals. When focusing on the African-American individuals, distinctive ancestry patterns can also be observed, in particular when comparing SP4 and SP5 assigned individuals (see part B in Figure 6).

Discussion

TW and EigenDev stopping criteria

Analysis of population genetic structure requires first a method for detecting whether significant structure exists

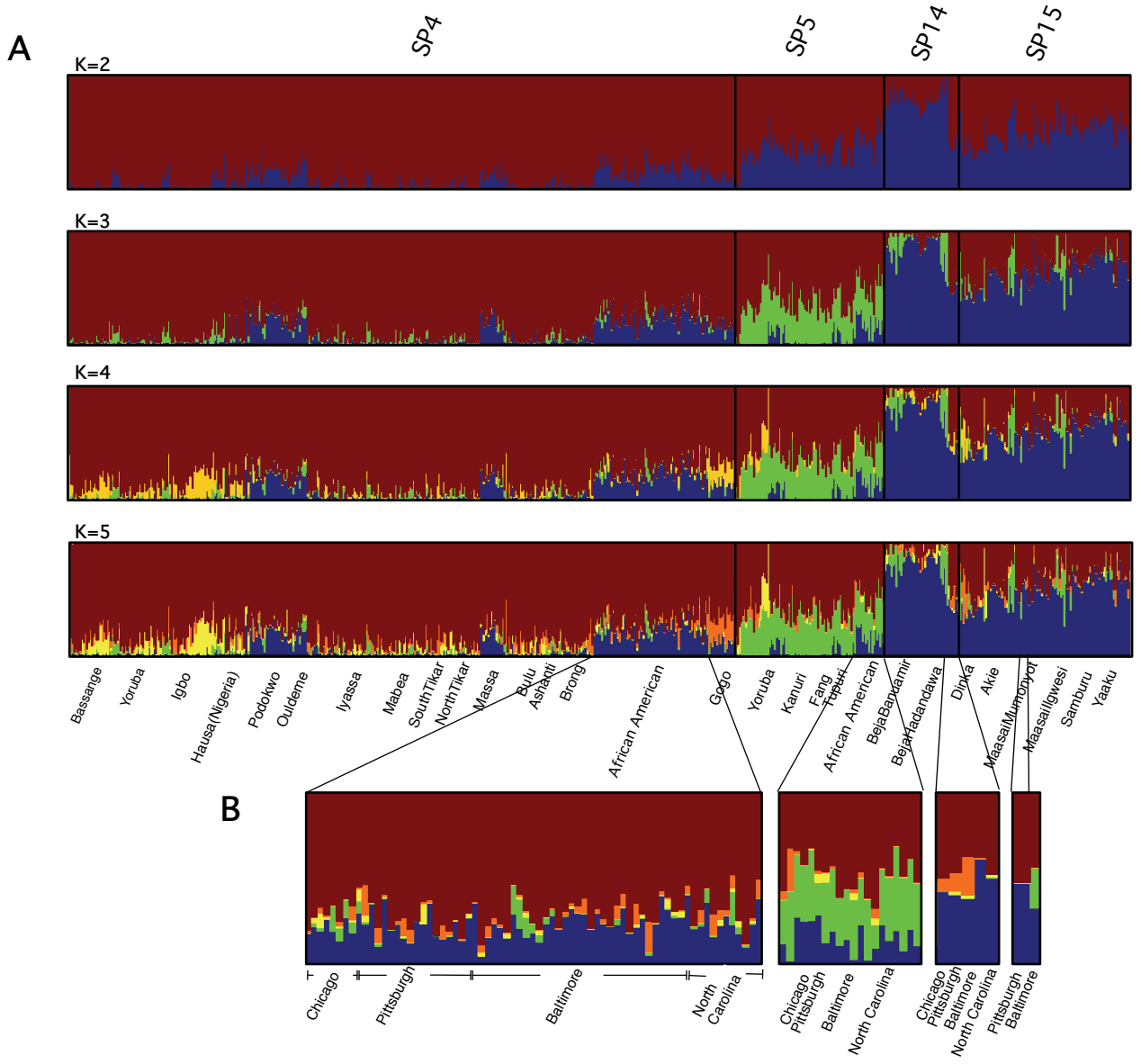


Figure 6 ipPCA-guided STRUCTURE analysis on selected individuals from the Tishkoff et.al. dataset. A) All individuals assigned to SP4, SP5, SP15 and SP16 (see Figure 5), which included all African-American individuals, were analyzed by STRUCTURE from $K = 2$ to $K = 5$. Individuals were sorted according to the ipPCA assignments. Major ethno-linguistic labels for individuals within each subpopulation are also shown (see Figure 5 for complete listing). B) Expanded view of African-American individuals from A).

in the dataset (or nested dataset for further iterations of ipPCA). The current method to obtain this information is to test for deviation from the Tracy-Widom distribution of the largest eigenvalue computed from PCA. A p-value lower than 10^{-12} is considered an acceptable threshold for significance in rejecting the null hypothesis that the data belong to a homogenous (sub)population, and thus are structured [5]. The first experiment with a simulated dataset with no structure revealed that significant deviation from the expected distribution is found, particularly with large sampling (>70 individuals). We infer from this result that when the sample size is large, the TW method suffers from type I error because of this deviation from the TW distribution. Simply using lower p-value thresholds may not give better results, since there is a very small range of p-value that is practical [5]. When applied to real datasets, homogenous (sub)populations sampled at high density may be incorrectly construed as possessing structure. In TW-ipPCA, this would lead to a group of individuals being assigned into separate subpopulations, when they should actually be considered belonging to a single (sub)population.

To alleviate the drawbacks of the TW test statistic, we propose a new termination criterion called EigenDev statistic that is simpler to compute, has no hidden parameters and is shown to be more robust to type I error. For simplicity, one could choose a single EigenDev value to be applied as a universal stopping criterion for ipPCA, which needs to be determined empirically. We determined a threshold of 0.21 from data simulation, which was also appropriate for the real datasets analyzed in this paper.

Analyses of Bovine HapMap dataset

The subpopulation assignment by EigenDev-ipPCA supports the accepted notion that cattle breeds have distinctive genetic profiles. The finding that ANG and RGU were assigned together in the same subpopulation suggests that these breeds are genetically indistinguishable for the markers available, which was also reported by other methods [12]. However, the finding that GIR, BRM, and NEL breeds are resolved as separate subpopulations by EigenDev-ipPCA is novel, since the earlier unsupervised STRUCTURE analysis in [12] on the entire dataset could not distinguish these breeds. ipPCA-guided STRUCTURE analysis on the Bovine HapMap dataset demonstrated differences in ancestries among these breeds, consistent with the assignments by EigenDev-ipPCA. Among these indicine breeds, there is evidence (high heterozygosity and unique SNPs) to suggest that BRM is genetically distinct from others, including GIR and NEL [12]. These results beg the question, why STRUCTURE analysis, when done in a EigenDev-ipPCA guided manner, can reveal differences among these breeds which is not apparent in the unsupervised STRUCTURE analysis? The likeliest explanation is that the overall number of informative markers is low among these indicine breeds in comparison with the others (only 19% of the loci having minor allele frequencies greater than 0.3) [12]. In other words, the allele frequencies among the indicine breeds are highly correlated in comparison with the taurine breeds. Groups of individuals with highly correlated allele frequencies in comparison with other groups tend to be merged by STRUCTURE [11].

Analyses of a large human dataset

The 49 subpopulations assigned by EigenDev-ipPCA each contain individuals largely sharing the same ethnolinguistic label/affiliation, in accordance with [10,13]. Of note, the 426 Indian individuals were assigned to two subpopulations by EigenDev-ipPCA. This grouping is consistent with the parametric analysis of these individuals in [13], which showed weak evidence of structure. Hence, the greater degree of stratification resolved by TW-ipPCA compared with EigenDev-ipPCA is likely to be spurious. The spurious structure resolved by TW-ipPCA is thus attributable to the large sample size (426), which is well above the threshold encountered for type I error from the analysis of simulated data.

Among the African individuals, subpopulations were assigned by EigenDev-ipPCA revealing stratification patterns not described previously. For instance, Niger-Khordofanian speaking non-Pygmy individuals from West and Central Africa could not be distinguished genetically in [10], but were assigned to SP3, SP4 and SP5 subpopulations by EigenDev-ipPCA. The assignment of the majority African Americans to SP4 and SP5 by EigenDev-ipPCA (Figure 5) suggests they have West and Central African Niger-Khordofanian ancestors, in agreement with [10]. On the other hand, the assignment of African Americans to different subpopulations by EigenDev-ipPCA is suggestive of significant structure among these individuals. Supervised STRUCTURE runs performed in [10] to elucidate African American ancestry could only reveal a subtle clinal pattern of variation among the African Americans. The EigenDev-ipPCA guided STRUCTURE analysis, however, shows clear differences in ancestry between SP4 and SP5 African Americans. The SP15 and SP16 assigned African Americans also show ancestry distinct from the SP4 and SP5 assigned individuals, although given the small number of individuals assigned to SP15 and SP16, it is not possible to observe significant ancestry differences between these two groups.

The EigenDev-ipPCA assignment of some African Americans to SP15 and SP16 was unexpected. The contemporary African individuals in these subpopulations

are predominantly from Saharan and East Africa. A recent study of African American ancestry concluded that some individuals have a major ancestral component which is neither West African Niger-Khordofanian, nor European [14]. The possibility that this anomalous ancestry is of Saharan or East African may also be reflected in mtDNA haplotypes, since some African Americans have anomalous haplotypes of unknown African origin [15,16]. The discrepancy between Eigen-Dev-ipPCA guided STRUCTURE and supervised STRUCTURE performed in [10] is due to the choice of individuals in the analysis. When individuals with inappropriately diverged allele frequencies from others are used, key ancestral differences will be missed, the same as was shown in analysis of Bovine data.

Conclusion

We describe EigenDev-ipPCA for analyzing population structure. This approach assigns individuals to subpopulations and determines the total number of subpopulations present. This algorithm incorporates a novel heuristic called EigenDev for detecting substructure, which is applied to the iterative clustering process. EigenDev is robust to population sampling, allowing us to analyze large complex datasets with higher accuracy. The subpopulations assigned by EigenDev-ipPCA reveals overall genetic relatedness among groups of individuals, which can then be used to guide STRUCTURE. Other parametric algorithms such as Admixture and *frappe* could also be used in the same way. Therefore, the combination of EigenDev-ipPCA and STRUCTURE are complementary and can be used together to perform a powerful population stratification analysis. The software both in Matlab source code (m- file) and executable versions on Windows and Linux (64 bit) are available for download at http://www4a.biotec.or.th/GI/tools/ippca.

Additional material

Additional file 1: The detailed analysis and further discussion of the EigenDev-ipPCA results for the Tishkoff et al. dataset.

Acknowledgements and Funding

The authors thank King Mongkut's Institute of Technology Ladkrabang (KMITL), the National Electronics and Computer Technology Center (NECTEC) and the National Center for Genetic Engineering and Biotechnology (BIOTEC) for financial support. In particular, TL and ST acknowledge the funding supported by the National Science and Technology Development Agency (NSTDA). AA was supported by BIOTEC postdoctoral fellowship. JP was supported by BIOTEC platform technology grant and the Thailand Research Fund (TRF) new researcher grant (TRG5380028). PJS was funded by the Bill & Melinda Gates Foundation through the Grand Challenges Explorations Initiative. ST received the support from BIOTEC platform technology and TRF Career Development grant (RSA-54). We thank the reviewers and editors for their constructive comments, which improve the quality and presentation of the manuscript. Finally, we would like to thank the authors and the donors who made the real datasets used in this paper available.

Author details

[1]Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand. [2]National Electronics and Computer Technology Center, Thailand Science Park, Pathumthani 12120, Thailand. [3]National Center for Genetic Engineering and Biotechnology, Thailand Science Park, Pathumthani 12120, Thailand.

Authors' contributions

TL, AI, PJS and ST wrote the manuscript. TL, AI and ST constructed the computational improvement scheme of the new algorithm. AA, PJS and JP conceived the ideas to reanalyze the mixed complex datasets. TL, AI, PW and CN conducted all the experiments presented in this work. TL, AA, PJS, JP and ST analyzed the results. AI, PW and CN wrote the EigenDev-ipPCA program and made it available in executable formats using a Matlab compiler. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 8 October 2010 Accepted: 23 June 2011
Published: 23 June 2011

References

1. Marchini J, Cardon LR, Phillips MS, Donnelly P: **The effects of human population structure on large genetic association studies.** *Nat Genet* 2004, **36(5)**:512-7.
2. Tian C, Plenge R, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver A, Qi L, Gregersen P, Seldin M: **Analysis and application of European genetic substructure using 300 K SNP information.** *PLoS Genet* 2008, **4**:e4.
3. Paschou P, Lewis J, Javed A, Drineas P: **Ancestry informative markers for fine-scale individual assignment to worldwide populations.** *J Med Genet* 2010, **47(12)**:835-47.
4. Intarapanich A, Shaw PJ, Assawamakin A, Wangkumhang P, Ngamphiw C, Chaichoompu K, Piriyapongsa J, Tongsima S: **Iterative pruning PCA improves resolution of highly structured populations.** *BMC Bioinformatics* 2009, **10**:382.
5. Patterson N, Price A, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2(12)**:e190.
6. Luo J, Zhang Z: **Using Eigenvalue Grads Method to Estimate the Number of Signal Source.** In *Proceedings of the 5th International Conference on Signal Processing (WCCC-ICSP 2000). Volume 1*. Beijing, China; 2000:223-225.
7. Liang L, Zollner S, Abecasis GR: **GENOME: a rapid coalescent-based whole genome simulator.** *Bioinformatics* 2007, **23(12)**:1565-7.
8. **The BovineHapMap dataset.** [http://bfgl.anri.barc.usda.gov/cgi-bin/hapmap/affy2/BulkDownloads].
9. **The Tishkoff et. al. dataset.** [http://www.sciencemag.org/content/vol0/issue2009/images/data/1172257/DC1/1172257_dataset.zip].
10. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM: **The genetic structure and history of Africans and African Americans.** *Science* 2009, **324(5930)**:1035-44.
11. Pritchard JK, Stephens M, Donnelly P: **Inference of Population Structure Using Multilocus Genotype Data.** *Genetics* 2000, **155**:945-59.
12. Consortium TBH: **Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds.** *Science* 2009, **324(5926)**:528-32.
13. Rosenberg N, Mahajan S, Gonzalez-Quevedo C, Blum M, Nino-Rosales L, Ninis V, Das P, Hegde M, Molinari L, Zapata G, Weber J, Belmont J, Patel P: **Low levels of genetic divergence across geographically and linguistically diverse populations from India.** *PLoS Genet* 2006, **2(12)**:e215.
14. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA, Bustamante CD: **Genome-wide patterns of population structure and admixture in West Africans and African Americans.** *Proc Natl Acad Sci USA* 2010, **107(2)**:786-91.

15. Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo A: **The African diaspora: mitochondrial DNA and the Atlantic slave trade.** *Am J Hum Genet* 2004, **74(3)**:454-65.
16. Ely B, Wilson JL, Jackson F, Jackson BA: **African-American mitochondrial DNAs often match mtDNAs found in multiple African ethnic groups.** *BMC Biol* 2006, **4**:34.

doi:10.1186/1471-2105-12-255
Cite this article as: Limpiti *et al.*: Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and structure. *BMC Bioinformatics* 2011 **12**:255.

BMC Genomics

PROCEEDINGS Open Access

iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies

Jittima Piriyapongsa[1], Chumpol Ngamphiw[1], Apichart Intarapanich[2], Supasak Kulawonganunchai[1], Anunchai Assawamakin[1], Chaiwat Bootchai[1], Philip J Shaw[1], Sissades Tongsima[1*]

From Asia Pacific Bioinformatics Network (APBioNet) Eleventh International Conference on Bioinformatics (InCoB2012)
Bangkok, Thailand. 3-5 October 2012

Abstract

Background: Genome-wide association studies (GWAS) do not provide a full account of the heritability of genetic diseases since gene-gene interactions, also known as epistasis are not considered in single locus GWAS. To address this problem, a considerable number of methods have been developed for identifying disease-associated gene-gene interactions. However, these methods typically fail to identify interacting markers explaining more of the disease heritability over single locus GWAS, since many of the interactions significant for disease are obscured by uninformative marker interactions e.g., linkage disequilibrium (LD).

Results: In this study, we present a novel SNP interaction prioritization algorithm, named iLOCi (Interacting Loci). This algorithm accounts for marker dependencies separately in case and control groups. Disease-associated interactions are then prioritized according to a novel ranking score calculated from the difference in marker dependencies for every possible pair between case and control groups. The analysis of a typical GWAS dataset can be completed in less than a day on a standard workstation with parallel processing capability. The proposed framework was validated using simulated data and applied to real GWAS datasets using the Wellcome Trust Case Control Consortium (WTCCC) data. The results from simulated data showed the ability of iLOCi to identify various types of gene-gene interactions, especially for high-order interaction. From the WTCCC data, we found that among the top ranked interacting SNP pairs, several mapped to genes previously known to be associated with disease, and interestingly, other previously unreported genes with biologically related roles.

Conclusion: iLOCi is a powerful tool for uncovering true disease interacting markers and thus can provide a more complete understanding of the genetic basis underlying complex disease. The program is available for download at http://www4a.biotec.or.th/GI/tools/iloci.

Background

A major challenge for human genetics is identifying susceptibility genes for complex heritable diseases. Advanced single nucleotide polymorphism (SNP) genotyping technology and genome-wide association study (GWAS) are at the forefront of research in this area. In conventional single locus analysis, each variant is tested individually for disease association. Systematic analysis of GWAS data in this manner can typically uncover multiple SNPs associated with complex diseases [1-3]. These analyses have provided valuable insights into the genetics of complex diseases; however, they typically detect only common, low-risk variants each with small effect and explain only a tiny proportion of disease heritability [4].

The existence of interactions among genes (epistasis) has been proposed to constitute a major proportion of

* Correspondence: sissades@biotec.or.th
[1]National Center for Genetic Engineering and Biotechnology, Pathumthani, 12120, Thailand
Full list of author information is available at the end of the article

disease heritability, which is not captured by single-locus GWAS [5]. The genetical nature of epistasis can be described by several different models as shown in a variety of interaction schema discussed in [6]. Note that genetic factors primarily function through a complex mechanism; thus, epistatic interactions are not limited to independent gene pairs. Multiple genes interacting through a biological network (i.e. indirect interactions) exist which can modify disease penetrance and expressivity.

A number of methods for detecting epistatic interactions among genotypic data have been proposed. Most methods employ a statistical approach to identify interacting marker pairs based on deviation from a null distribution and estimation of type I error. These statistical approaches have been shown to work well in theory, e.g., regression methods [7,8], partitioning chi-square [9], Focused Interaction Testing Framework (FITF) [10], Bayesian model selection [11], and other recent approaches [12,13]. However, the need for control of type I error reduces power to detect interactions in real data, which is exacerbated by the huge number of statistical tests performed in this analysis [14].

Given the challenges for statistical approaches, non-statistical methods such as machine-learning and data-mining methods have been proposed for the study of genetic interactions [15,16]. Instead of model fitting, these methods attempt to explain all of the heritability in terms of marker interactions. Multifactor dimensionality reduction (MDR) is an brute-force method for identifying the most plausible interactions which fit the data [17]. However, MDR and other recently published exhaustive non-parametric approaches [18] are computationally complex and thus impractical for analysis of GWAS data. To overcome the computational burden of non-parametric analysis, several techniques have been developed that employ statistics to assist the non-parametric search for epistasis, including SNPHarvester [19], SNPRuler [20], and BOOST [21]. In these methods, the search space is reduced by a filtering step, usually employing a statistical threshold. The filtered dataset is then used for non-parametric search for epistasis. Although these methods can be applied for analysis of GWAS data, the interactions found rarely offer any new insights since the majority of interacting markers map to the same genomic regions. For example, the analysis of WTCCC (Wellcome Trust Case Control Consortium) data by BOOST revealed that after removal of linked pairs, no interactions were found for five of the seven diseases. Using another approach for exhaustive search of interactions, the most recent paper by Ueki and Tamiya [22] also reported very few interactions in the WTCCC data.

The possible reason for the disappointingly modest improvement of the current hybrid approaches is that they do not adequately account for marker dependencies not related to disease. A well known marker dependency which can confound the identification of genomic regions associated with disease is linkage disequilibrium (LD). LD is non-random association of genotypes at two or more loci that can be on the same or different chromosomes. LD is caused by a number of factors, including genetic linkage and the rate of recombination [23]. Earlier reports [24,25] showed that LD contrast, i.e., differences in LD patterns between case and control groups can reveal the disease signal above the noise of background LD in candidate disease regions. However, to our knowledge, LD contrast has not been employed for comprehensive genetic epistasis study, owing to the high computational complexity.

Clearly, a computationally efficient and comprehensive prioritization technique is required which accounts for marker dependencies unrelated to disease. Moreover, instead of trying to control type I error, a prioritization procedure may be more effective in revealing more of the true disease markers which may have modest individual effects and interact in complex higher-order networks.

In this paper, we propose a novel tool for prioritizing gene-gene interactions called iLOCi (interacting Loci). The iLOCi algorithm ranks all SNP pair combinations according to a novel heuristic that we call ρ_{diff}. The iLOCi program is specifically designed to handle large-scale GWAS data partly through the application of data parallelization. The tests with WTCCC datasets show that the top ranked pairs by our algorithm reveal novel disease genes, several of which are consistent with biological networks underpining disease etiology.

Methods

iLOCi algorithm

The proposed iLOCi algorithm performs genome-wide analysis for identifying SNP pairs that are plausibly associated with a disease. No prior genetical assumptions are employed in the algorithm, which allows the exploration of different dimensions of the association results. The framework can be characterized into two main modules: 1) calculating SNP pair dependencies separately in case and control groups and 2) disease SNP pair prioritization as shown in Figure 1.

Calculation of SNP pair dependencies

iLOCi explores all possible combinations of SNP pairs. Given N SNPs from a SNP array with the SNP index starting from 1 to N, there are a total of $\binom{N}{2} = \frac{N(N-1)}{2}$ possible pairs. Each SNP pair is assigned a unique index (i,j), where $i \neq j$.

From the large number of SNP pairs, it is necessary to identify the dependency unrelated to disease. This dependency includes linkage disequilibrium (LD),

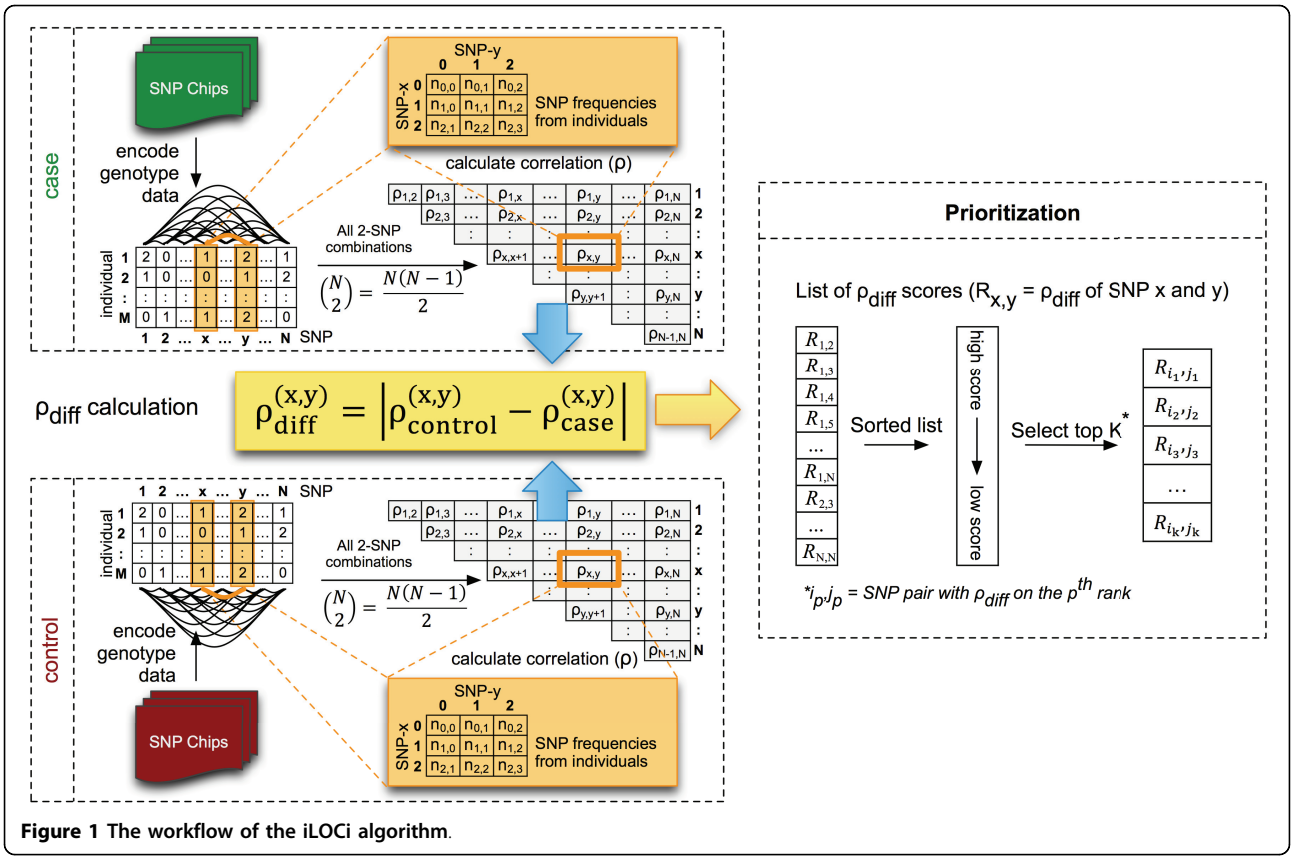


Figure 1 The workflow of the iLOCi algorithm.

population structure, genotype calling artifacts, etc. and is performed separately between the case and control groups. This step of the algorithm is called *dependence test*. Therefore, for each indexed SNP pair, the algorithm calculates two scores, ρ_{case} and $\rho_{control}$. The calculated ρ values using genotypic information were proven to be concordant with LD values (see Additional file 1). LD values are calculated using allelic deviation from the Hardy-Weinberg Equilibrium (HWE) model, which assumes that, without the introduction of specific disturbing factors, the frequencies of alleles and genotypes in a population remain constant from one generation to the next. However, it should be noted that the only information captured by ρ values is the correlation between markers, which is needed for identifying interactions. For LD calculation, the haplotypic phase is also considered, which is computationally very demanding for datasets of this size.

To compute marker ρ values, each SNP locus is considered as a discrete random variable and the numeric values of -1, 0 and 1 are assigned to homozygous wild (w), heterozygous (h), and homozygous variant (v) types respectively. This encoding ensures zero-means, which obviates a normalization step. Let x and y be two discrete random variables of SNPx and SNPy, respectively. Let $P_{(x,y)}$ represents a genotypic joint probability mass function, whose entries are the probability of genotype combinations from both SNPs. Hence, there are nine possible genotypic combinations that are represented by the following matrix:

$$P_{(x,y)} = \begin{bmatrix} P_{ww} & P_{wh} & P_{wv} \\ P_{hw} & P_{hh} & P_{hv} \\ P_{vw} & P_{vh} & P_{vv} \end{bmatrix}$$

For example, P_{ww} is a probability that (x,y) are both homozygous wild type. Each of these probabilities can be calculated by dividing the number of the joint genotypic outcomes with the total number of individuals for either case (N_{case}) or control (N_{control}) groups. For example, $P_{ww}^{\text{ctrl}} = P_{(x=w,y=w)}^{\text{ctrl}} = \frac{N_{(x=w,y=w)}^{\text{ctrl}}}{N_{\text{ctrl}}}$. The dependence test must be performed for all possible SNP pairs. The correlation value ρ_{control} for each SNP pair is calculated as:

$$[\{x]_w y_w P_{ww}^{\text{ctrl}} + x_w y_h P_{wh}^{\text{ctrl}}\} + x_w y_v P_{wv}^{\text{ctrl}}\} + \{x_h y_w P_{hw}^{\text{ctrl}} + x_h y_h P_{hh}^{\text{ctrl}} + x_h y_v P_{hv}^{\text{ctrl}}\} + \{x_v y_w P_{vw}^{\text{ctrl}} + x_v y_h P_{vh}^{\text{ctrl}} + x_v y_v$$

Note that $P_{x=w}^{\text{ctrl}}$, $P_{x=v}^{\text{ctrl}}$, $P_{y=w}^{\text{ctrl}}$, and $P_{y=v}^{\text{ctrl}}$ are the estimated probability of SNPx wild type, SNPx variant type, SNPy wild type and SNPy variant type respectively.

By the same reasoning, ρ_{case} is calculated as:

$$= \frac{P_{ww}^{case} - P_{wv}^{case} - P_{vw}^{case} + P_{vv}^{case}}{\left[\left(P\right]_{x=w}^{case} + P_{x=v}^{case}\right)\left(P_{y=w}^{case} + P_{y=v}^{case}\right)}$$

Disease SNP pair prioritization

The next step is to identify whether the same SNP pair (*x*,*y*) from case and control groups have contrasting patterns of ρ values. A *difference test* is performed by differentiating the ρ values between the case and control groups using a simple subtraction operation, namely $\rho_{diff}=|\rho_{control}-\rho_{case}|$.

To select the highly associated SNP pairs, all SNP pairs are ranked according to the ρ_{diff} values. The ranking of top SNP pairs was chosen, rather than a *P*-value cutoff in order to avoid too many false positive pairs due to the heavy-tailed distribution phenomenon, where the Gaussian distribution decreases faster than the distribution of disease associated SNP pairs [26].

Parallel computing algorithm implemented in iLOCi

The iLOCi algorithm is designed for genome-scale analysis which requires the computation of a huge number of SNP interaction pairs, e.g.≈$1.25x10^{11}$ pairs for a 500,000 SNP dataset. Data parallelization is applied to accelerate this computationally intensive and time-consuming process. The SNP interaction matrix is divided into submatrices of 100,000 or fewer SNPs each. Each SNP interaction submatrix is computed in parallel using a MacPro workstation with 2×2.4 GHz quad-core Intel Xeon processors with 8GB RAM. With this configuration, the complete WTCCC dataset can be analyzed in 19 hours. Details for implemention of the code and data parallelization are available upon request.

Testing iLOCi algorithm performance using simulated data

The performance of iLOCi for detecting disease-associated gene interactions was evaluated and compared with FastEpistasis [27]. The evaluation was made using simulated datasets, which were generated using the GenomeSIM program [28]. The algorithm performance was determined for detection of four different epistatic interaction scenarios:

1) Single pair interaction without marginal effects: Eighteen epistatic models in [29] with heritability (h^2) of 0.2, 0.3, and 0.4 were used for performance comparison (see Additional file 2: Table S1). These heritability levels were chosen to represent those typically found in common complex diseases. The minor allele frequency (MAF), which is the frequency of the less common allele, was assigned to be two levels, 0.2 and 0.4. In total, there are six model groups comprising three models with the same heritability and MAF for each group. 100 independent datasets containing 1600 samples (800 cases and 800 controls) with 100 SNPs were generated for each model group.

2) Single pair interaction with marginal effects: Six epistatic models in [30] with MAF of 0.5 were tested (see Additional file 2: Table S2). 100 independent datasets containing 800 samples (400 cases and 400 controls) and 100 independent datasets containing 1600 samples (800 cases and 800 controls) with 100 SNPs each were generated for each model group.

3) Multiple independent interacting pairs without marginal effects: Eight models of multiple interactions described in supplementary material of [19] were tested. Each of these models were generated from five epistatic models described in [29]. Each model used the same heritability and MAF. 100 independent datasets containing 1600 samples (800 cases and 800 controls) and 100 SNPs were generated for each model group.

4) Higher-order interactions: Data were simulated for the eight interaction network models based on pairwise interaction described in [31] for three-, four-, and five-loci interating networks (see Additional file 2: Table S3). 100 independent datasets containing 800 samples (400 cases and 400 controls) were generated. The number of SNPs varies from model to model.

The algorithm performance was demonstrated by the percentage of accuracy, which is determined by the proportion of 100 independent datasets in which the algorithm correctly identified the interacting SNP pairs. For situations 1 and 2, the identification of disease SNP pair is defined as correct if the disease SNP pair is the top ranked pair with the highest ρ_{diff} score (for iLOCi) or the lowest *P*-value (for FastEpistasis). For multiple independent interacting pairs (case 3), the identification is taken as correct when all five disease SNPs fall in the top five ranked pairs with highest ρ_{diff} score (for iLOCi) or lowest *P*-value (for FastEpistasis). The prediction of higher-order interactions is defined as correct when all disease SNPs are found within all top ranked pairs. The top ranked pairs are defined as all consecutive pairs comprising at least one disease SNP in each pair.

Testing algorithm performance using the WTCCC dataset

In addition to the simulated data, our algorithm was applied to the real genotypic data of WTCCC (Wellcome Trust Case Control Consortium) [3]. This dataset encompasses ~500,000 SNP genotypic data of ~17,000 British samples which are divided into 3000 shared control samples and ~2000 case samples for each of seven complex diseases: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT),

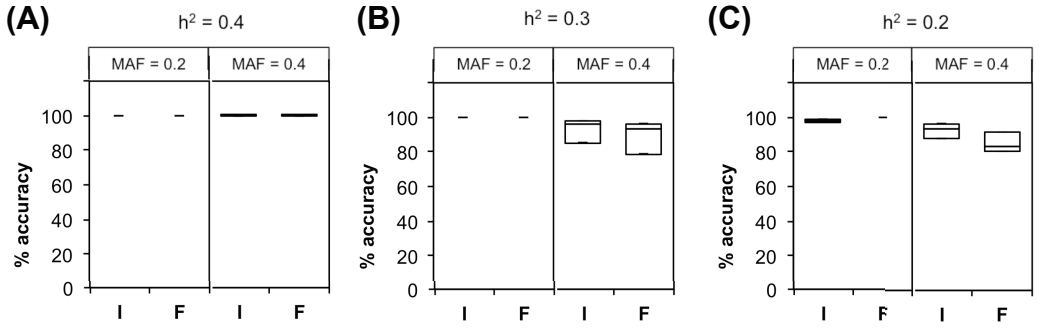


Figure 2 The performance comparison between iLOCi (I) and FastEpistasis (F) on epistatic models without marginal effects. The algorithm performance is shown as the percentage of accuracy, which is the number of simulated datasets (out of 100) in which the correct SNP pair is identified. The accuracy was tested for two different MAF (0.2, 0.4) and three different levels of heritability (A) 0.4, (B) 0.3, and (C) 0.2.

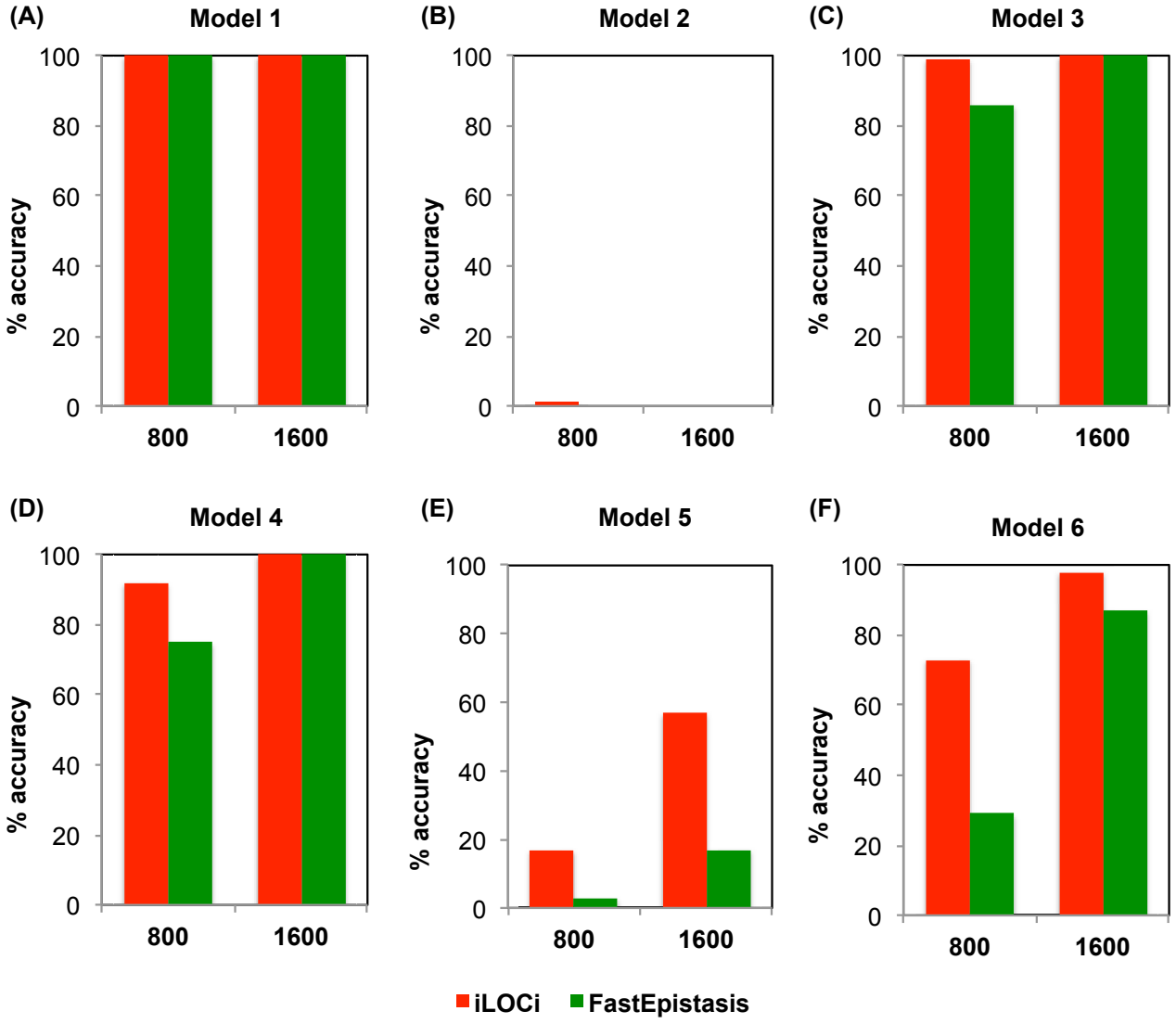


Figure 3 The performance comparison between iLOCi and FastEpistasis on epistatic models with marginal effects. The percentage of accuracy is shown for two different sample sizes (800 and 1600) for six different pairwise interaction models (A-F).

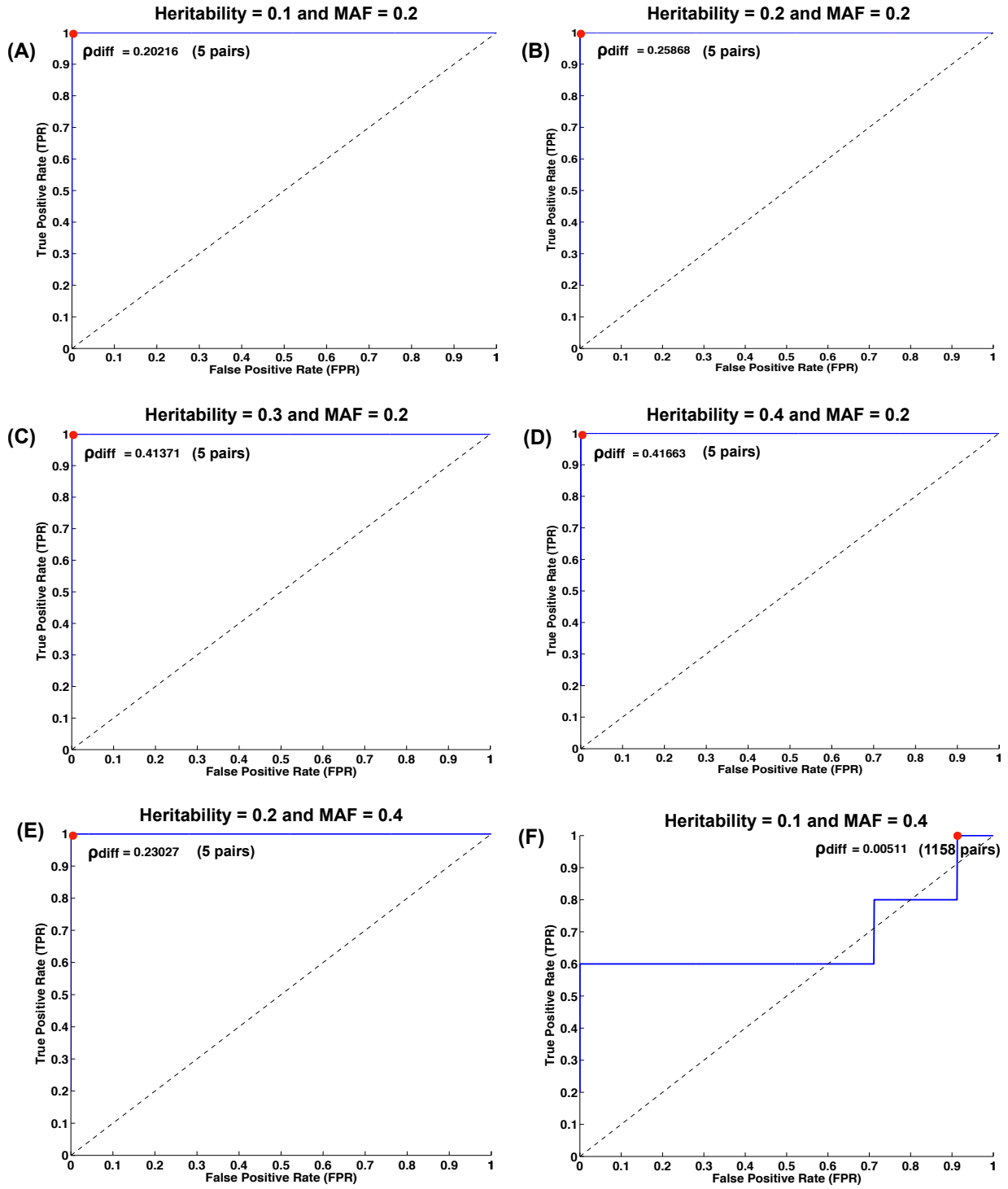


Figure 4 The Receiver Operating Characteristic (ROC) curves for simulation datasets of hybrid models. The ROC curves are displayed for five independent interacting SNP pairs.The MAF and heritability parameters were varied (A) h^2=0.1, MAF = 0.2, (B) h^2=0.2, MAF = 0.2, (C) h^2=0.3, MAF = 0.2, (D) h^2=0.4, MAF = 0.2, (E) h^2=0.2, MAF = 0.4, (F) h^2=0.1, MAF = 0.4. The ρ_{diff}values are shown that give the maximum true positive rate with the lowest false positive rate (red dots).

rheumatoid arthritis (RA), type1 (T1D) and type2 (T2D) diabetes.

For these real datasets, data cleaning was required prior to the analysis. We considered only SNPs and individuals passing WTCCC data quality control [3]. We further filtered the SNP set using MAF>0.05 leaving 355,882 SNPs (complete set) for all diseases. We also generated a SNP marker gene-only subset of 176,148 present in genes (defined as within 10Kb flanking an annotated gene model reported in RefSeq version 36.3).

First, ρ_{diff} values for the seven WTCCC diseases were calculated for all possible ($\approx 63\times10^9$ for complete and $\approx 15\times10^9$ for the gene-only subset) pairs. Next, the empirical ρ_{diff} distributions for each disease were graphed using kernel density plot. For the gene-only SNP subset analysis, the top ranked 1000 SNP pairs were chosen for functional analysis to uncover biological significance. From these pairs, a list of genes was extracted based upon RefSeq (version 36.3) physical locations of SNPs in the genome. To understand the biological significance of the novel genes reported by our algorithm, we also used the candidate gene prioritization feature of ToppGene [32] using the cutoff of *P*-value = 0.01 with Bonferroni correction. The training sets for the ToppGene candidate gene prioritization were the lists of all genes reported in the HuGE Navigator database [33] for the seven diseases. The test sets for the ToppGene analysis were the lists of novel (not reported in HuGE Navigator database) genes represented among the top ranked 1000 SNP pairs obtained from iLOCi.

Results

iLOCi algorithm validation

We used simulated datasets to validate the iLOCi algorithm for identifying various disease-associated epistatic interactions. We chose FastEpistasis for performance comparison with iLOCi due to the fact that the data were simulated according to an interaction model;

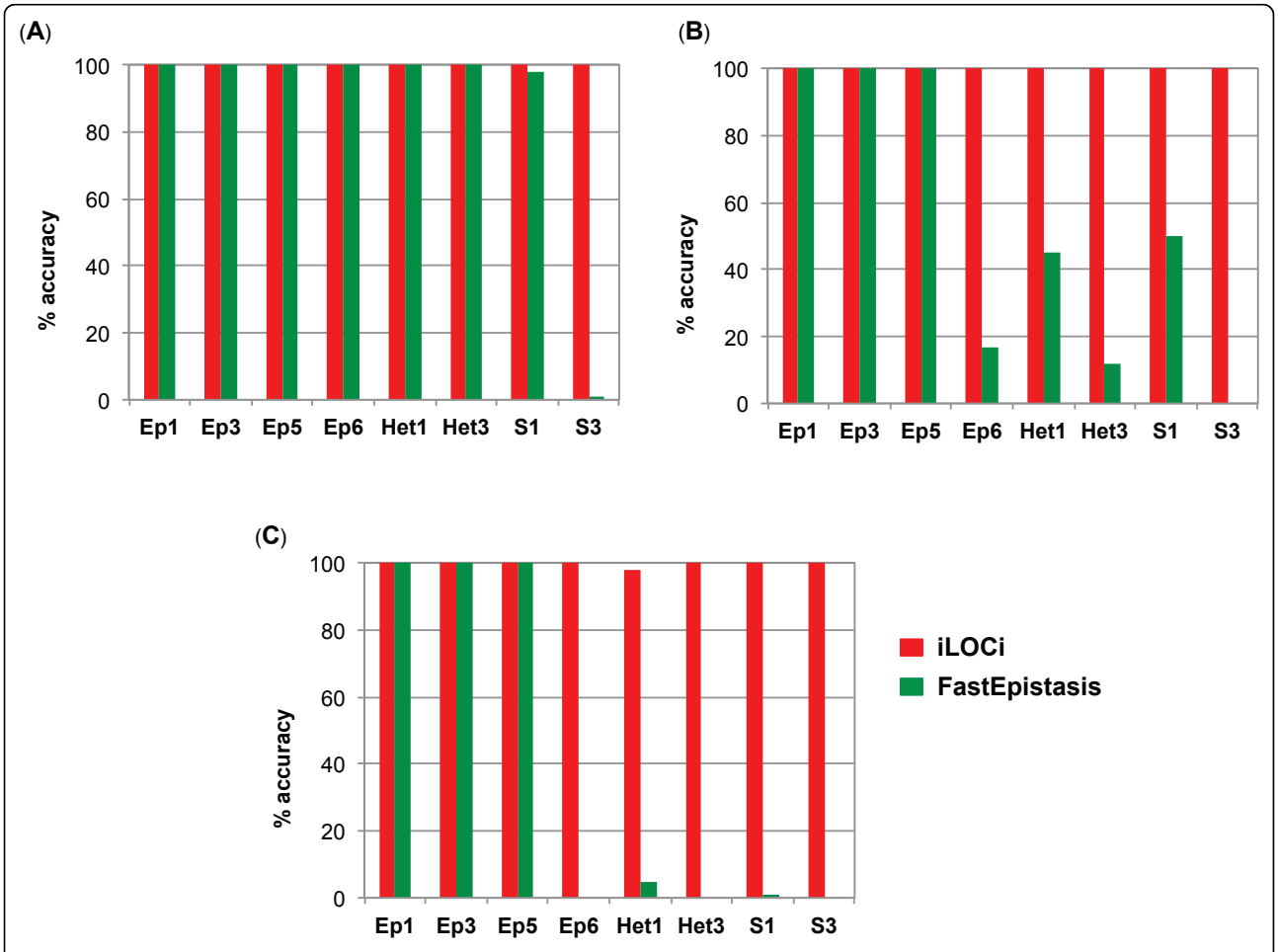


Figure 5 The performance comparisons between iLOCi and FastEpistasis on high-order interaction models. The percentage of accuracy is shown for different models (Ep1, Ep3, Ep5, Ep6, Het1, Het3, S1, S3) of high-order interactions among (A) three-loci, (B) four-loci, and (C) five-loci.

hence this tool would be most suitable for testing. Moreover, the theoretical basis for FastEpistasis is widely accepted for genome-wide analysis.

The first result testing for a single interacting pair demonstrated that the top ranked iLOCi pair was the disease interacting pair in 18 different inheritance models without the presence of marginal effects. Overall, its performance was approximately the same as FastEpistasis for most of the model groups and slightly better in some cases (h^2=0.2, MAF = 0.4; h^2=0.3, MAF = 0.4) as shown in Figure 2. For epistatic interactions with marginal effects, iLOCi outperformed FastEpistasis in most models, except in model 2 for which both methods failed to detect the interacting disease marker pair (Figure 3). Furthermore, we want to demonstrate the specificity as well as sensitivity of iLOCi for detecting multiple interacting disease marker pairs as would be present in a real dataset. Therefore, the receiver operating characteristics (ROC) were plotted for different thresholds of ranked marker pairs, and for different models of heritability and MAF (Figure 4). Generally, iLOCi has high sensitivity and specificity, although the performance tends to be worse with lower degrees of heritability. Moreover, it should be noted that the minimum ρ_{diff} scores that give 100% sensitivity vary greatly from 0.00511 to 0.41663.

In addition to independent interacting pairs, we examined the ability of iLOCi and FastEpistasis to detect higher-order interactions of 3, 4, and 5 loci disease interaction networks for eight models at each level (Figure 5). iLOCi can detect all eight models for all levels of interactions; however, FastEpistasis failed to identify all S3 model interactions. Furthermore, FastEpistasis could detect, with higher than 50% accuracy, in fewer than 50% of the 4-loci network models and only Ep1, Ep3 and Ep5 of the 5-loci network models.

In conclusion, these experiments with simulated data validated the iLOCi algorithm for identifying all four types of higher-order gene interaction. iLOCi performance was comparable to FastEpistasis for a variety of

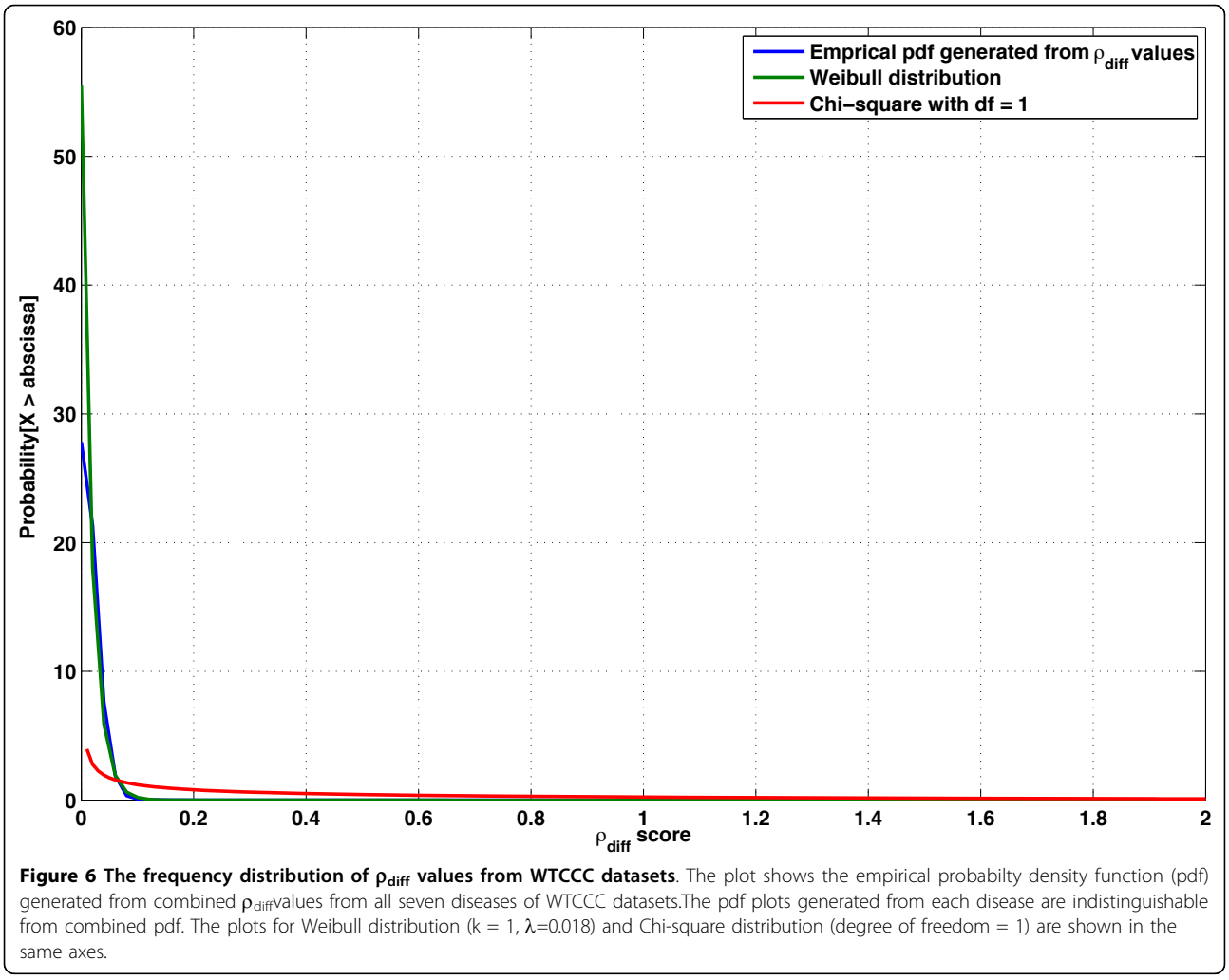


Figure 6 The frequency distribution of ρ_{diff} values from WTCCC datasets. The plot shows the empirical probabilty density function (pdf) generated from combined ρ_{diff}values from all seven diseases of WTCCC datasets.The pdf plots generated from each disease are indistinguishable from combined pdf. The plots for Weibull distribution (k = 1, λ=0.018) and Chi-square distribution (degree of freedom = 1) are shown in the same axes.

Table 1 The lookup table of *P*-values for the associated ρ_{diff} scores

ρ_{diff}score	*P*-value
0.05	6.2177e-2
0.10	3.8659e-3
0.15	2.4037e-4
0.20	1.4945e-5
0.25	9.2925e-7
0.30	5.7777e-8
0.35	3.5924e-9
0.40	2.2336e-10
0.45	1.3888e-11
0.50	8.6353e-13
0.55	5.3735e-14
0.60	3.3307e-15
0.65	2.2204e-16
0.70	<2.2204e-16
0.75	<2.2204e-16
0.80	<2.2204e-16
0.85	<2.2204e-16
0.90	<2.2204e-16
0.95	<2.2204e-16
1.00	<2.2204e-16

The *P*-values were calculated based on the fitted Weibull distribution with k = 1 and λ=0.018.

two-locus interaction models; however, iLOCi was markedly superior for detecting high-order interactions. This would be a major advantage of iLOCi for analysis of real data since high-order interaction is the type of interaction likely to be found in real data of complex diseases and may account for current missing heritability.

iLOCi analyses of WTCCC data

The iLOCi algorithm was tested against real data obtained from WTCCC. The distribution of ρ_{diff} values follows a Weibull distribution pattern for all seven diseases (Figure 6). From the Weibull distribution with k = 1 and λ=0.018, we calculated *P*-values for ρ_{diff} scores ranging from 0.05 to 1.0 (see Table 1). For the seven diseases, we selected the top 1000 pairs for which the calculated minimum *P*-values vary from <2.22e-16 to 1.14e-7 in complete SNP set analysis, and from <2.22e-16 to 4.72e-5 in gene-only SNP analysis (see Table 2).

From iLOCi analysis using the complete SNP marker set, it was found that the great majority of the SNPs have not been previously reported to be associated with the diseases [3]. Furthermore, the majority of these SNPs also do not map to annotated genes. The list of top 1000 SNP pairs is available in Additional File 3. For each disease, iLOCi identified 'hub' SNPs, i.e. SNPs that pair with many other SNPs, e.g., rs1553460 pairs with 1000 other SNPs in BD (Table 3).

Owing to the fact that the majority of interacting SNPs do not map to annotated genes, we re-analyzed the data using the gene-only SNP subset. 'Hub' SNPs were also observed at the gene level (Table 3). From this analysis, it was noted that the top ranked 1000 SNP pairs of all seven diseases map to 321 disease-gene associations that have been annotated on the HuGE Navigator database

Table 2 The ρ_{diff} scores of the 1st and 1000th ranked SNP pairs and their associated *P*-values

Complete set of SNPs (355882 SNPs)					
Disease	**1stρ_{diff}**	**1st*P*-value**	**1000thρ_{diff}**	**1000th*P*-value**	**Avg. ρ_{diff}± SD**
BD	0.2878	1.1410e-7	0.2680	3.4206e-7	0.2718±0.0035
CAD	0.9317	<2.2204e-16	0.9132	<2.2204e-16	0.9171±0.0031
CD	0.3085	3.6109e-8	0.2849	1.3351e-7	0.2887±0.0034
HT	0.2834	1.4510e-7	0.2626	4.6022e-7	0.2667±0.0037
RA	0.9042	<2.2204e-16	0.8866	<2.2204e-16	0.8903±0.0031
T1D	1.0731	<2.2204e-16	0.9996	<2.2204e-16	1.0040±0.0056
T2D	0.3338	8.8226e-9	0.2159	6.1867e-6	0.2198±0.0052
Gene-only SNPs (176148 SNPs)					
Disease	**1stρ_{diff}**	**1st*P*-value**	**1000thρ_{diff}**	**1000th*P*-value**	**Avg. ρ_{diff}± SD**
BD	0.2447	1.2445e-6	0.2224	4.2957e-6	0.2259±0.0032
CAD	0.9294	<2.2204e-16	0.9102	<2.2204e-16	0.9143±0.0035
CD	0.2653	3.9790e-7	0.2248	3.7769e-6	0.2280±0.0033
HT	0.1793	4.7229e-5	0.1561	1.7142e-4	0.1605±0.0043
RA	0.9040	<2.2204e-16	0.8832	<2.2204e-16	0.8875±0.0036
T1D	1.0731	<2.2204e-16	0.9957	<2.2204e-16	1.0007±0.0061
T2D	0.3338	8.8226e-9	0.2127	7.3731e-6	0.2168±0.0052

The highest and the lowest ρ_{diff} scores including their associated *P*-values are displayed with the average scores of top 1000 SNP pairs from the analyses of WTCCC.

Table 3 The hub SNPs/genes identified in the top-ranked 1000 SNP pairs

Hub SNPs from analyses of complete SNP set		
Disease	**Hub SNPs (Genomic position)**	**# Interacting SNPs**
BD	rs1553460 (Chr4:17804959)	1000
CAD	rs3785579 (Chr17:62472963)	1000
CD	rs1553460 (Chr4:17804959)	978
	rs4471699 (Chr16:30227808)	22
HT	rs10843660 (Chr12:30259724)	999
RA	rs3785579 (Chr17:62472963)	1000
T1D	rs9273363 (Chr6:32734250)	1000
T2D	rs7077039 (Chr10:114779067)	833
	rs10787472 (Chr10:114771287)	54
	rs11196208 (Chr10:114801306)	39
	rs11196205 (Chr10:114797037)	30
	rs10885409 (Chr10:114798062)	22
	rs4074720 (Chr10:114738487)	17
Hub genes from gene-only SNP analyses		
Disease	**Hub genes**	**# Interacting genes**
BD	*CENPN*: centromere protein N	653
CAD	*CACNG1*: calcium channel, voltage-dependent, gamma subunit 1	709
CD	*ATG16L1*: ATG16 autophagy related 16-like 1 (*S. cerevisiae*)***	256
	IL23R: interleukin 23 receptor ***	20
HT	*tcag7.23*: similar to ribosomal protein L18; 60S ribosomal protein L18	170
	BCAT1: branched chain aminotransferase 1, cytosolic ***	57
	SAMD4A: sterile alpha motif domain containing 4A *	27
	GAB1: GRB2-associated binding protein 1 *	25
	RHOJ: ras homolog gene family, member J	20
	LYPD5: LY6/PLAUR domain containing 5 *	12
RA	*CACNG1*: calcium channel, voltage-dependent, gamma subunit 1	676
T1D	*HLA-DQB1*: major histocompatibility complex, class II, DQ beta 1**	686
T2D	*TCF7L2*: transcription factor 7-like 2 (T-cell specific, HMG-box)***	481

* Genes associated with disease SNPs that were previously reported in WTCCC original paper

** Genes previously reported to be disease-associated in HuGE Navigator database

*** Genes previously reported to be disease-associated in both WTCCC paper and HuGE Navigator database

Table 4 The disease association of iLOCi selected genes from gene-only SNP analyses

Disease	**# iLOCi genes in top 1000 SNP pairs**	**Reported in WTCCC single SNP analyses**		**Reported in HuGE Navigator database**	
		# Analyzed genes (# SNPs)	**# iLOCi genes**	**# Analyzed genes (# SNPs)**	**# iLOCi genes**
BD	654	42 (1757)	8	665 (16598)	52
CAD	710	29 (2097)	3	735 (11564)	37
CD	279	54 (1651)	4	531 (7181)	10
HT	595	32 (3164)	19	1240 (22004)	64
RA	677	34 (822)	4	503 (5902)	19
T1D	687	39 (1153)	5	512 (6924)	29
T2D	486	29 (1289)	5	2456 (41244)	110

The table displays the number of previously reported disease-associated genes which were found in all analyzed genes and in the set of genes involved in top 1000 interaction pairs. The reported disease genes are shown for both the genes associated with disease SNPs from WTCCC paper [3] and the ones reported in HuGE Navigator database [33].

(see Table 4, Additional File 4). On the other hand, the majority of the disease interacting genes among these pairs reported by iLOCi are novel. Moreover, most of these genes were not reported in the original WTCCC study (Table 4). To evaluate the biological significance of the novel genes among these pairs, the ToppGene candidate gene prioritization tool was employed. The full results are shown in Additional Files 3 and 4. Among the novel genes identified by iLOCi, it was observed that some well known disease pathways from KEGG [34] contain several of these genes (see Additional File 5). For instance, the 'neuroactive ligand-receptor interaction' pathway in BD contains 4 novel genes in addition to 11 previously reported genes (Figure 7). Other prominent disease pathways include 'cytokine-cytokine receptor interaction' for CAD (Figure 8) and 'type I diabetes mellitus' for T1D (Figure 9).

Discussion

In this study, we have developed a new pairwise SNP-interaction prioritization algorithm for GWAS. We hypothesized that by first accounting for pairwise marker dependencies among case and control groups, it would be possible to observe true disease interactions above the noise of dependent markers unrelated to disease, as was proposed in earlier studies of LD contrast (see Background).

In GWAS data, it is well known that LD generates strong pairwise dependency signals that are used to identify disease associated SNPs by imputation. However, this type of signal predominates pairwise markers in analysis of gene interactions. For example, in the approach used by Wan et al. [21], the majority of the interactions identified for all seven WTCCC datasets can be attributed to LD effect, i.e., the interacting

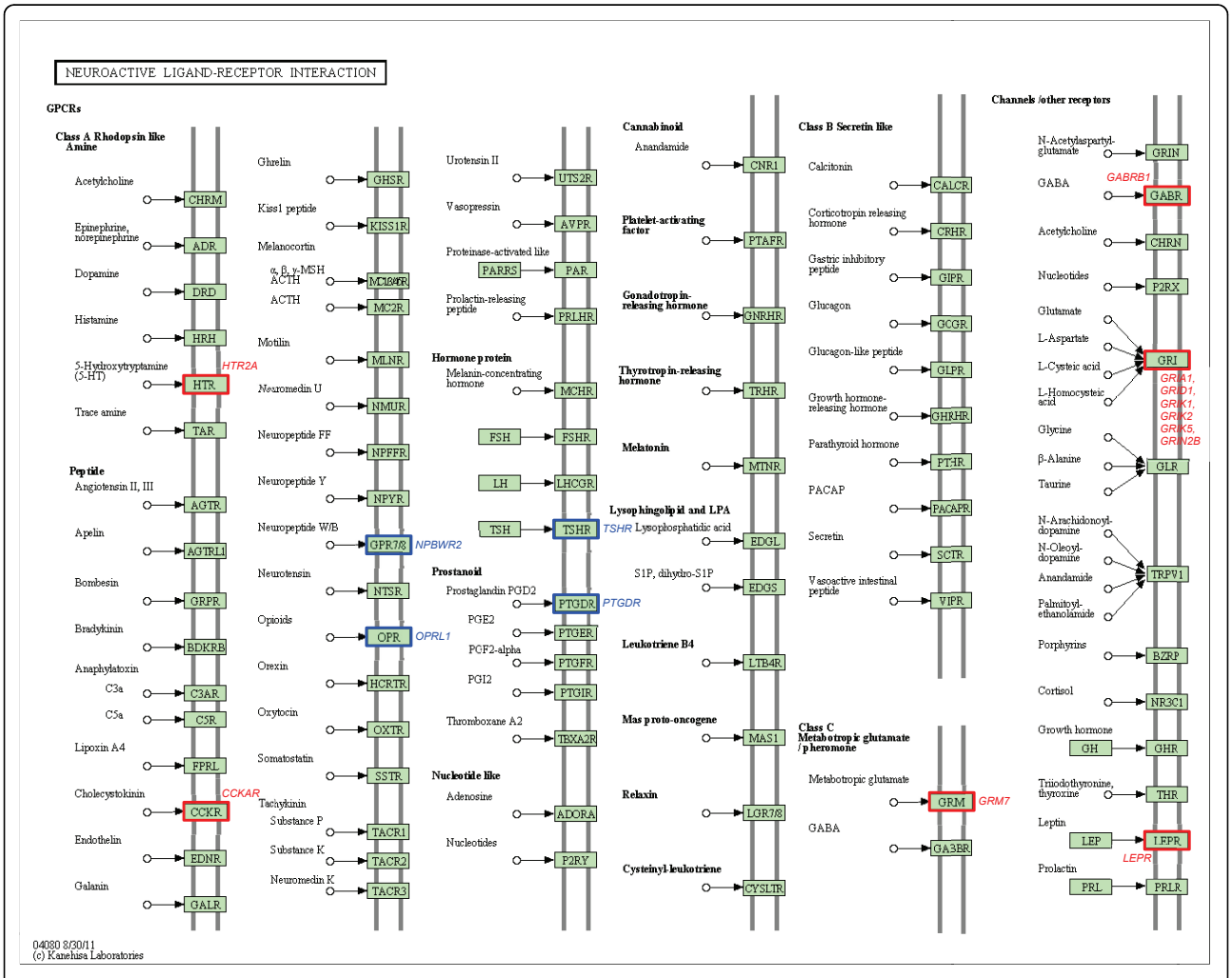


Figure 7 The iLOCi detected genes from BD and their maps in 'neuroactive ligand-receptor interaction' pathway. The KEGG pathway diagram [34] shows the mapping of BD-associated genes identified among 1000 top ranked iLOCi pairs in 'neuroactive ligand-receptor interaction' KEGG pathway. The gene families containing the genes previously reported in HuGE Navigator database and the novel disease genes are highlighted in the red boxes and the blue boxes, respectively, with their associated gene names.

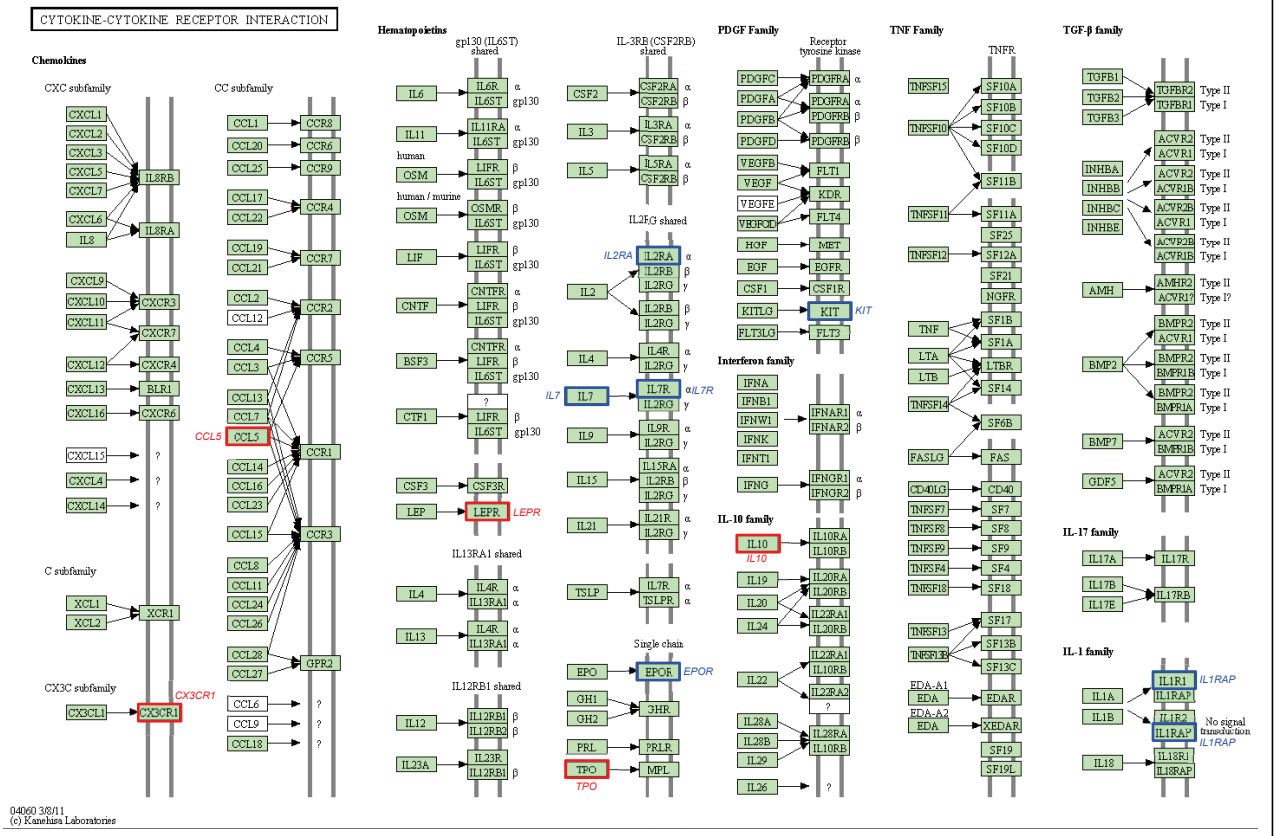


Figure 8 The iLOCi detected genes from CAD and their maps in 'cytokine-cytokine receptor interaction' pathway. The KEGG pathway diagram [34] shows the mapping of CAD-associated genes identified among 1000 top ranked iLOCi pairs in 'cytokine-cytokine receptor interaction' pathway. The gene families containing the genes previously reported in the HuGE Navigator database and novel disease genes are highlighted in the red boxes and the blue boxes, respectively, with their associated gene names.

SNPs are within 1Mb of each other in the same genomic region. To validate our approach correcting for pairwise dependencies unrelated to disease SNP interactions, extensive tests were performed on simulated data. For a simple model with only one interacting pair, the top ranked iLOCi pair is correctly identified as the disease marker pair. When testing for multiple interacting pairs, iLOCi has high accuracy under the conditions of high heritability and informativeness, i.e., low MAF. On the other hand, low heritability and/or informativeness leads to type I error as observed by ROC plot. In general, the ρ_{diff} scores reflect the degree of heritability and informativeness. Hence, it is not possible to use a single ρ_{diff} cutoff for identifying disease interactions in the real case when the heritability and informativeness are unknown.

From analyses of real GWAS data, it was found that the ρ_{diff} distributions for all seven diseases could be represented by a single kernel density function with Weibull distribution. However, the range of ρ_{diff} values varies among the diseases and follow the known heritability pattern, i.e., HT has the lowest heritability and lowest top ρ_{diff} score, while T1D has the highest heritability and highest top ρ_{diff} score (Table 2). Although it is possible to calculate *P*-values of the interacting pairs and use them as cutoffs for prioritization, we consider the use of *P*-value cutoffs inappropriate. For example, a *P*-value of 1e-5 (corresponding to ρ_{diff} values of approximately 0.2 or greater) would give approximately 16 million significant pairs for T1D and 200,000 pairs for HT. The same phenomenon of unacceptable type I error was found by others when using FastEpistasis for analysis of real datasets. It is debatable whether Bonferroni correction is valid since the tests are not independent, as shown by the heavy-tailed distributions of ρ_{diff} . Current methods for correction of type I error by false discovery rate are also likely to be impractical because of the requirement for permutation testing.

Instead of using *P*-value significance thresholds, we used the top ranked 1000 SNP pairs for prioritization, which account for a very small portion (<0.0001%) of all possible pairs. Rather than attempting to identify all gene interactions, which practically can not be found [35], we limit the prioritization to the top ranked pairs

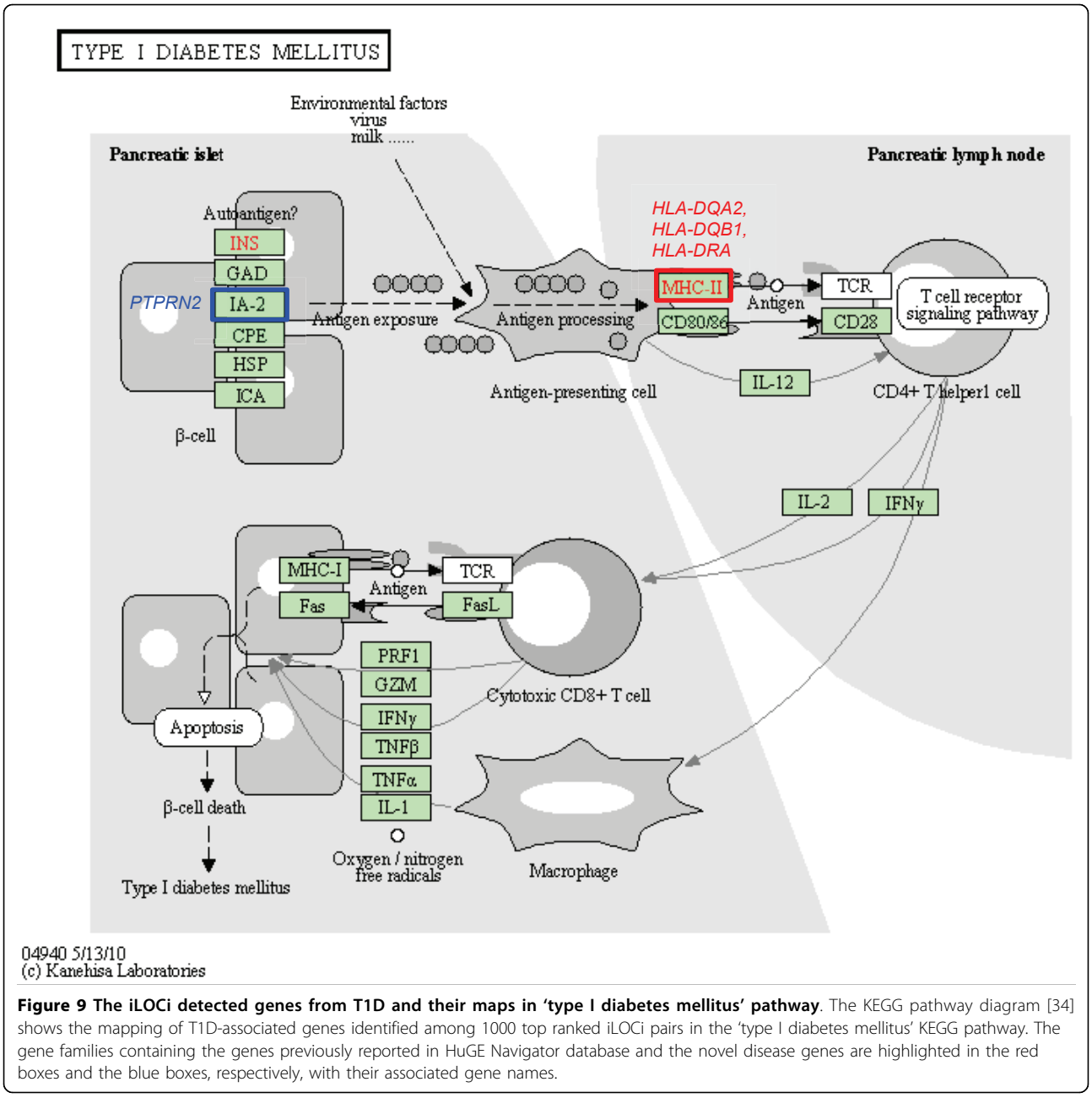


Figure 9 The iLOCi detected genes from T1D and their maps in 'type I diabetes mellitus' pathway. The KEGG pathway diagram [34] shows the mapping of T1D-associated genes identified among 1000 top ranked iLOCi pairs in the 'type I diabetes mellitus' KEGG pathway. The gene families containing the genes previously reported in HuGE Navigator database and the novel disease genes are highlighted in the red boxes and the blue boxes, respectively, with their associated gene names.

that are most likely to contain the genetic interactions which are informative of the disease etiology, i.e., disease pathways. From the full SNP set analysis, several hub SNPs were identified for each disease which interact with many other SNPs. For some diseases such as T1D, these hub SNPs map to well-known disease associated genes. However, hub SNPs for BD, HT, and CD do not map to genes. These hub SNPs may mediate interactions at an unknown gene regulatory level, e.g. as non-coding RNAs, miRNAs or cis-regulatory elements. Since our knowledge of gene regulation is far from complete [36], we repeated the iLOCi analysis on the gene-only SNPs subset. By restricting the analysis to SNP pairs in genes only, the ToppGene systems approach for gene prioritization was appropriate, as used by others for GWAS data [37-39].

Gene-based prioritization of the interacting SNP pairs revealed significant representation of previously described disease associated genes. Therefore, we are confident that the novel genes found among the prioritized SNP pairs are novel disease-associated genes. For each disease, hub genes were found which pair with many other genes. Some of these disease hub genes are known and have been replicated as disease genes by

conventional single-SNP GWAS, including the MHC gene *HLADQB1* for T1D and *TCF7L2* for T2D. However, some hub genes have not been reported previously, e.g. the *CACNG1* gene for RA. This gene's SNP shows a modest *P*-value (>1e-4) for association by single SNP analysis [3]; therefore, the disease association of this SNP is dependent on multiple interactions with other loci. For each disease, including those with low heritability such as HT, we are able to suggest novel genes and pathways for further investigation, including re-analysis of other GWAS datasets for the same diseases.

Conclusions

In this article, we introduce a novel SNP interaction prioritization method, called iLOCi. The algorithm is computationally efficient, and thus suitable for exhaustive search for interactions along markers in a typical GWAS dataset. We have shown that the approach taken by iLOCi in which marker dependencies unrelated to disease are accounted for reveal genetic interactions of biological relevance to complex disease.

Additional material

Additional file 1: The mathematical details of ρ_{diff} value and its relation with LD (iLOCi_details.pdf). This file includes the mathematical details of iLOCi formula and its relationship with the allele-based LD calculation.

Additional file 2: Penetrance tables for dataset simulation (Penetrance_tables.pdf). This file includes the penetrance models used for dataset simulation of two-locus and high-order ineractions.

Additional file 3: Top 1000 SNP pairs from analyses of complete SNP set of WTCCC (TopPairs_Complete.xls). This file includes the list of top 1000 SNP pairs with their associated genes obtained from the iLOCi analyses of all SNPs passing the quality control step. The evidences for disease association of each identified gene as reported in WTCCC original paper and HuGE Navigator database are also shown. The genes identified as candidate disease genes from ToppGene prioritization are indicated with their rank numbers and *P*-values.

Additional file 4: Top 1000 SNP pairs from analyses of gene-only SNP set of WTCCC (TopPairs_GeneOnly.xls). This file includes the list of top 1000 SNP pairs with their associated genes obtained from the iLOCi analyses of gene-only SNPs. The evidences for disease association of each identified gene as reported in WTCCC original paper and HuGE Navigator database are also shown. The genes identified as candidate disease genes from ToppGene prioritization are indicated with their rank numbers and *P*-values.

Additional file 5: Pathway enrichment analysis of WTCCC datasets (Pathway_analysis.xls). This file includes the list of enriched biological pathways obtained from ToppGene program using the training sets of HuGE Navigator disease-associated genes. The pathway *P*-value is reported along with the list of iLOCi identified genes associated with such pathway. For each pathway, the number of genes previously reported in HuGE Navigator database, reported in WTCCC paper, and the novel disease genes, is shown.

Acknowledgements

JP is supported by the new researcher grant from the Thailand Research Fund and National Center for Genetic Engineering and Biotechnology (grant number TRG5580011). ST would like to acknowledge the TRF grant number RSA5480026 and the Research Chair Grant 2011 from the National Science and Technology Development Agency (NSTDA), Thailand that partially support this work. ST was supported in part by the office of the higher education commission and Mahidol University under the national research university initiative. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

This article has been published as part of *BMC Genomics* Volume 13 Supplement 7, 2012: Eleventh International Conference on Bioinformatics (InCoB2012): Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S7.

Author details

[1]National Center for Genetic Engineering and Biotechnology, Pathumthani, 12120, Thailand. [2]National Electronics and Computer Technology Center, Pathumthani, 12120, Thailand.

Authors' contributions

JP designed the algorithm and the experiments, generated simulated data, analyzed test results, and wrote the manuscript. CN performed most experiments on simulated and real datasets. AI designed the algorithm and performed the mathematical proof of formula. SK implemented iLOCi program. AA designed the algorithm. CB performed the functional analysis of real dataset. PJS wrote the manuscript and discussed the test results. ST designed the algorithm, discussed the test results, and wrote the manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 13 December 2012

References

1. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, *et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci.** *Nature* 2007, **447(7148)**:1087-1093.
2. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, *et al*: **Genomewide association analysis of coronary artery disease.** *N Engl J Med* 2007, **357(5)**:443-453.
3. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447(7145)**:661-678.
4. Manolio TA, Brooks LD, Collins FS: **A HapMap harvest of insights into the genetics of common disease.** *J Clin Invest* 2008, **118(5)**:1590-1605.
5. Moore JH, Asselbergs FW, Williams SM: **Bioinformatics challenges for genome-wide association studies.** *Bioinformatics* 2010, **26(4)**:445-455.
6. Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB: **Detection of gene × gene interactions in genome-wide association studies of human population data.** *Hum Hered* 2007, **63(2)**:67-84.
7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81(3)**:559-575.
8. Zhao J, Jin L, Xiong M: **Test for interaction between two unlinked loci.** *Am J Hum Genet* 2006, **79(5)**:831-845.
9. Yang Y, Houle AM, Letendre J, Richter A: **RET Gly691Ser mutation is associated with primary vesicoureteral reflux in the French-Canadian population from Quebec.** *Hum Mutat* 2008, **29(5)**:695-702.
10. Millstein J, Conti DV, Gilliland FD, Gauderman WJ: **A testing framework for identifying susceptibility genes in the presence of epistasis.** *Am J Hum Genet* 2006, **78(1)**:15-27.
11. Zhang Y, Liu JS: **Bayesian inference of epistatic interactions in case-control studies.** *Nat Genet* 2007, **39(9)**:1167-1173.
12. Ueki M, Cordell HJ: **Improved statistics for genome-wide interaction analysis.** *PLoS Genet* 2012, **8(4)**:e1002625.

13. Wu X, Dong H, Luo L, Zhu Y, Peng G, Reveille JD, Xiong M: **A novel statistic for genome-wide interaction analysis.** *PLoS Genet* 2010, **6(9)**: e1001131.
14. Hunter DJ, Kraft P: **Drinking from the fire hose–statistical issues in genomewide association studies.** *N Engl J Med* 2007, **357(5)**:436-439.
15. Cordell HJ: **Detecting gene-gene interactions that underlie human diseases.** *Nat Rev Genet* 2009, **10(6)**:392-404.
16. McKinney BA, Reif DM, Ritchie MD, Moore JH: **Machine learning for detecting gene-gene interactions: a review.** *Appl Bioinformatics* 2006, **5(2)**:77-88.
17. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.** *Am J Hum Genet* 2001, **69(1)**:138-147.
18. Yoshida M, Koike A: **SNPInterForest: a new method for detecting epistatic interactions.** *BMC Bioinformatics* 2011, **12**:469.
19. Yang C, He Z, Wan X, Yang Q, Xue H, Yu W: **SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies.** *Bioinformatics* 2009, **25(4)**:504-511.
20. Wan X, Yang C, Yang Q, Xue H, Tang NL, Yu W: **Predictive rule inference for epistatic interaction detection in genome-wide association studies.** *Bioinformatics* 2010, **26(1)**:30-37.
21. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W: **BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies.** *Am J Hum Genet* 2010, **87(3)**:325-340.
22. Ueki M, Tamiya G: **Ultrahigh-dimensional variable selection method for whole-genome gene-gene interaction analysis.** *BMC Bioinformatics* 2012, **13(1)**:72.
23. Hedrick PW: **Genetics of populations.** Sudbury, Boston, Toronto, London, Singapore: Jones and Bartlett Publishers;, 3 2005.
24. Wang T, Zhu X, Elston RC: **Improving power in contrasting linkage-disequilibrium patterns between cases and controls.** *Am J Hum Genet* 2007, **80(5)**:911-920.
25. Zaykin DV, Meng Z, Ehm MG: **Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method.** *Am J Hum Genet* 2006, **78(5)**:737-746.
26. Embrechts P, Klüppelberg C, Mikosch T (eds.): **Modelling Extremal Events for Insurance and Finance.** Berlin: Springer Verlag;, 1 1997.
27. Schupbach T, Xenarios I, Bergmann S, Kapur K: **FastEpistasis: a high performance computing solution for quantitative trait epistasis.** *Bioinformatics* 2010, **26(11)**:1468-1469.
28. Dudek SM, Motsinger AA, Velez DR, Williams SM, Ritchie MD: **Data simulation software for whole-genome association and other studies in human genetics.** *Pac Symp Biocomput* 2006, 499-510.
29. Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH: **A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction.** *Genet Epidemiol* 2007, **31(4)**:306-315.
30. Moore J, Hahn L, Ritchie M, Thornton T, White B: **Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics.** *Proceedings of the Genetic and Evolutionary Computation Conference: July 9-13, 2002 2002; New York, USA* Morgan Kaufman; 2002, 1150-1155.
31. Neuman RJ, Rice JP: **Two-locus models of disease.** *Genet Epidemiol* 1992, **9**:347-365.
32. Chen J, Bardes EE, Aronow BJ, Jegga AG: **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.** *Nucleic Acids Res* 2009, , **37 Web Server:** W305-311.
33. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: **A navigator for human genome epidemiology.** *Nat Genet* 2008, **40(2)**:124-125.
34. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Res* 2012, , **40 Database:** D109-114.
35. Zuk O, Hechter E, Sunyaev SR, Lander ES: **The mystery of missing heritability: Genetic interactions create phantom heritability.** *Proc Natl Acad Sci USA* 2012, **109(4)**:1193-1198.
36. Esteller M: **Non-coding RNAs in human disease.** *Nat Rev Genet* 2011, **12(12)**:861-874.
37. Dick DM, Aliev F, Krueger RF, Edwards A, Agrawal A, Lynskey M, Lin P, Schuckit M, Hesselbrock V, Nurnberger J Jr, *et al*: **Genome-wide association study of conduct disorder symptomatology.** *Mol Psychiatry* 2010, **16(8)**:800-808.
38. Edwards AC, Aliev F, Bierut LJ, Bucholz KK, Edenberg H, Hesselbrock V, Kramer J, Kuperman S, Nurnberger JI Jr, Schuckit MA, *et al*: **Genome-wide association study of comorbid depressive syndrome and alcohol dependence.** *Psychiatr Genet* 2012, **22(1)**:31-41.
39. Lascorz J, Forsti A, Chen B, Buch S, Steinke V, Rahner N, Holinski-Feder E, Morak M, Schackert HK, Gorgens H, *et al*: **Genome-wide association study for colorectal cancer identifies risk polymorphisms in German familial cases and implicates MAPK signalling pathways in disease susceptibility.** *Carcinogenesis* 2010, **31(9)**:1612-1619.

doi:10.1186/1471-2164-13-S7-S2
Cite this article as: Piriyapongsa *et al.*: iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies. *BMC Genomics* 2012 **13**(Suppl 7):S2.

OPEN ACCESS Freely available online

PLOS ONE

Insight into the Peopling of Mainland Southeast Asia from Thai Population Genetic Structure

Pongsakorn Wangkumhang[1,2☯], Philip James Shaw[1☯], Kridsadakorn Chaichoompu[1], Chumpol Ngamphiw[1,2], Anunchai Assawamakin[3], Manit Nuinoon[4], Orapan Sripichai[5], Saovaros Svasti[5], Suthat Fucharoen[5], Verayuth Praphanphoj[6], Sissades Tongsima[1*]

1 National Center for Genetic Engineering and Biotechnology (BioTeC), Khlong Luang, Pathum Thani, Thailand, **2** Inter-Department Program of Biomedical Sciences, Chulalongkorn University, Pathumwan, Bangkok, Thailand, **3** Faculty of Pharmacy, Mahidol University, Rajathevi, Bangkok, Thailand, **4** School of Allied Health Sciences and Public Health, Walailak University, Thai Buri, Nakhon Sri Thammarat, Thailand, **5** Thalassemia Research Center, Mahidol University, Salaya, Nakhon Pathom, Thailand, **6** Center for Medical Genetics Research, Rajanukul Institute, Dindaeng, Bangkok, Thailand

Abstract

There is considerable ethno-linguistic and genetic variation among human populations in Asia, although tracing the origins of this diversity is complicated by migration events. Thailand is at the center of Mainland Southeast Asia (MSEA), a region within Asia that has not been extensively studied. Genetic substructure may exist in the Thai population, since waves of migration from southern China throughout its recent history may have contributed to substantial gene flow. Autosomal SNP data were collated for 438,503 markers from 992 Thai individuals. Using the available self-reported regional origin, four Thai subpopulations genetically distinct from each other and from other Asian populations were resolved by Neighbor-Joining analysis using a 41,569 marker subset. Using an independent Principal Components-based unsupervised clustering approach, four major MSEA subpopulations were resolved in which regional bias was apparent. A major ancestry component was common to these MSEA subpopulations and distinguishes them from other Asian subpopulations. On the other hand, these MSEA subpopulations were admixed with other ancestries, in particular one shared with Chinese. Subpopulation clustering using only Thai individuals and the complete marker set resolved four subpopulations, which are distributed differently across Thailand. A Sino-Thai subpopulation was concentrated in the Central region of Thailand, although this constituted a minority in an otherwise diverse region. Among the most highly differentiated markers which distinguish the Thai subpopulations, several map to regions known to affect phenotypic traits such as skin pigmentation and susceptibility to common diseases. The subpopulation patterns elucidated have important implications for evolutionary and medical genetics. The subpopulation structure within Thailand may reflect the contributions of different migrants throughout the history of MSEA. The information will also be important for genetic association studies to account for population-structure confounding effects.

Citation: Wangkumhang P, Shaw PJ, Chaichoompu K, Ngamphiw C, Assawamakin A, et al. (2013) Insight into the Peopling of Mainland Southeast Asia from Thai Population Genetic Structure. PLoS ONE 8(11): e79522. doi:10.1371/journal.pone.0079522

Editor: David Caramelli, University of Florence, Italy

Received June 20, 2013; **Accepted** September 23, 2013; **Published** November 4, 2013

Funding: ST is funded by the Thailand Research Fund (grant no. RSA5480026) and the Research Chair Grant National Science and Technology Development Agency. VP was funded by the department of mental health, Ministry of Public Health, Thailand. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: sissades@biotec.or.th

☯ These authors contributed equally to this work.

Introduction

The human population genetic history of Asia is complex, which is highlighted by the controversy surrounding the earliest migrations through Asia. One school of thought is that Asians are descended from two major ancestral groups, the earliest who migrated via a southern coastal route and a later group who spread across northern and eastern Asia [1]. An alternative hypothesis from genome-wide surveying of genetic variation across 73 Asian populations is that there was only one major migration pattern, in which East Asian peoples are descended from southern migrants who migrated north [2]. The controversy has been reignited following analysis of ancient human genomes from Central Asia [3] and Australia [4] which tend to support the two-wave hypothesis. The great diversity across Asia shaped by multiple migrations and population expansions throughout history will only be realized by more in-depth population genetic studies [5]. This gap in knowledge

has begun to be addressed by large-scale studies of Asian populations sampling thousands of individuals, which have revealed stratification (distinct subpopulations) among the populations of India [6], Japan [7], and China [8,9]. The degree of genetic stratification in these populations largely reflects known ethno/cultural/linguistic divisions and patterns of assumed ancestry.

Thailand lies at the heart of mainland Southeast Asia (MSEA), the region in which peoples speaking Tai-Kadai, Austroasiatic (Mon-Khmer), Sino-Tibetan, Hmong-Mien and Austronesian languages are present. The contemporary populations of this region are dominated by Tai language speakers (Thai and Laotian) and Austroasiatic speakers (Cambodian and Vietnamese). Most importantly, Thailand is located at the crossroads of ancient human migration paths between North and East Asia and Island Southeast Asia. Therefore, the genetic footprints of ancestral migrants may be present among people in this region. The earliest archaeological evidence of humans in MSEA was obtained in southern Thailand, dating to approximately 25,000 Years Before Present (YBP) [10], which is among the oldest remains documented in Southeast Asia [11]. mtDNA analysis of this specimen showed close relationship with the present-day Semang population in Peninsula Malaysia [12]. The Semang are an aboriginal “Negrito” people (distinguished by their darker skin pigmentation, different hair morphology, and short average stature), who may have been living continuously in Southeast Asia since the earliest Asian migration to Australia 60-75,000 YBP [13]; other Negrito populations elsewhere in Southeast Asia have a similarly ancient origin [14,15]. The southern part of Thailand was thus first populated by “Australo-Melanesian” [13] ancestral people. On the other hand, it is not clear how extensively populated MSEA was at this time, since archaeological evidence for communities and settlement prior to the Bronze Age (approximately 4500 YBP) in MSEA is sparse [16]. Bellwood (1993) argued that the earliest humans in MSEA would have been restricted to the coastal regions and not penetrated inland as the environment was not suitable for a foraging lifestyle [17]. Therefore, it is likely that the earliest populations of significance in MSEA were established by Austric agriculturalist people, the ancestors of Austroasiatic and Austronesians, who may have originated in Southern China. These migrants spread along river basins in MSEA reaching the Malaysian Peninsula in the Neolithic period [16]. Mitochondrial DNA study of Bronze and Iron age human remains from central Thailand was concordant with the presence of autochthonous Austric people in central Thailand [18]. Tai people migrated from southern China into northern Thailand more recently, establishing settlements in Thailand alongside the autochthonous Austrics. Eventually, the Tai became dominant, establishing control over northern Thailand from the 8th Century AD [19]. Later Tai domination covering much of present-day Thailand was evidenced by the Sukhothai dynasty (established 13th Century AD) and the Ayutthaya dynasty (established 15th Century AD), although the southern region of Thailand was essentially autonomous and ruled by Malay vassals until the 19th Century AD. During this most recent phase of Thai history, a large influx of migrants from southern China occurred [20]. Within the same period, other MSEA populations also experienced similar patterns of immigration and assimilation of southern Chinese, with Chinese influence greatest in Vietnam [21].

Despite the strategic location of Thailand in MSEA, there has been no large-scale study of its population’s genetic variation. Previous studies of human genetic diversity in Thailand were done with limited marker sets [22,23], and/or limited sampling (restricted to ethnic minorities); [2,22,24-28]. To better our understanding of mainland Southeast Asian and Thai population genetics, we undertook a study of Thai population genetic structure. The Thai population dataset comprises 992 individuals genotyped for 552,386 autosomal SNP markers. We found that the Thai population is genetically distinct from other Asian populations, but there is evidence of shared ancestry supporting the known origins and historical migration patterns across MSEA. Four Thai subpopulations were resolved which are distributed differently across Thailand. Interestingly, the most highly differentiated markers which can distinguish the four Thai subpopulations include several within genes which are known to affect traits such as skin pigmentation and susceptibility to common diseases.

Methods

Ethical statement

The recruitment of human subjects was approved by the ethical review committee for research in human subjects (Mental Health and Psychiatry): Ministry of Public Health, Thailand (CCA No. Si 32/2009).

Three SNP genotyping datasets were analyzed in this study. The first dataset is from a worldwide population study of 850 individuals from 40 populations published in [29]. The genotypic data from this dataset were obtained using the Affymetrix Human SNP Array 6.0 comprising 246,554 SNPs that passed quality control (after removal of markers that deviate from Hardy-Weinberg Equilibrium (HWE) ($P< 5.5\times10^{-8}$) and missing data >10%). The second dataset is a case-control association study to identify genetic factors of major depressive disorder. Human subjects for genotyping were recruited according to the ethical statement mentioned above. The dataset comprises 374 individuals (186 cases and 188 controls) collected from North, Northeastern, Central and Southern regions of Thailand. The DNA samples were genotyped using the Illumina Human 610-Quad BeadChips Array at RIKEN, Japan. The total number of genotyped SNPs is 593,542. SNPs were filtered to remove markers in high LD (linkage disequilibrium $r^2 > 0.5$), high deviation from HWE ($P<10^{-3}$) and missing data >5% using the PLINK tool. After filtering, 438,503 SNP markers remained for further analyses. Disease association test was performed using the PLINK tool. No marker passed the threshold for Bonferroni-corrected significance ($P<10^{-7}$). The top 50 ranked markers are shown in Table S1. The third dataset is a case-control study to identify modifying genetic factors that cause patients with β^0-thalassemia/hemoglobin E with different spectrums of disease severities. The study collected 383 severe patients and 235 mild patients and performed case-control association. The data

and association study were previously published in [30]. Genotyping was done using the same platform as with the second dataset, i.e. 610-Quad BeadChips Array for a total number of 593,542 SNPs. Note that both datasets 2 and 3 were from two *independent* case-control association studies of Thais where individuals' samples were collected from different regions in Thailand by different Principal Investigators. For datasets 2 and 3, individuals were asked to assign a geographical label for themselves (North, South, Northeast or Central) based on their place of birth, or their parent's place of birth. We tested for systematic differences of allele frequency caused by sampling bias between datasets 2 and 3 for 438,503 SNPs. A Bonferroni corrected P-value of 10-7 was used as the significance threshold. In accordance with PLoS policy on data availability, requests to access datasets 2 and 3 should be sent to Dr. Verayuth Prapanpoj and Prof. Suthat Fucharoen, respectively.

Population analyses

The analyses were done in two stages. First we observed the relationship between Thais and other related populations. The common polymorphic SNPs from all three datasets (41,569 SNPs) were used for population structure analysis. This marker set includes only SNPs that have the same reference SNP identification code (rs-id) between the Affymetrix and Illumina SNP array platforms. For some of these SNPs in common, the SNP calling on one platform is the complement of the other platform, i.e., A/G versus T/C. In these cases, the Affymetrix SNP calls were complemented to be the same as Illumina's. Common SNPs in which the base identity of the variant SNP was ambiguous on either platform were excluded. Finally to ensure that no hidden technical bias may exist between the two platforms for the common marker set, minor allele frequencies (MAF) for each SNP were calculated from a control population with 136 samples from Affymetrix [29] and 1,182 samples from Illumina [31] platforms, respectively. The scatter plot and the calculated correlation coefficient of MAFs do not show any evidence of biased MAFs (Figure S1).

Population structure was analyzed first by bootstrapping neighbor-joining (NJ) tree of the three combined datasets (1,842 individuals genotyped for 41,569 markers common among the two genotyping platforms) using the *seqboot*, *gendist*, *consense* and *neighbor* programs within the PHYLIP program suite (with default parameters) [32]. Allele frequencies of each population were calculated using *seqboot* (individuals with the same label were assumed to belong to the same population). The dissimilarity matrix was calculated from the matrix of allele frequencies using the *gendist* program. The *neighbor* module was used to construct NJ-trees from these matrices. Finally, *consense* was used to generate the consensus tree with bootstrapping values using the Pygmy population as an out-group. The unrooted phylogram was plotted using Dendroscope [33].

The ipPCA program [34,35] was used with stopping criterion EigenDev=0.21 [35] to assign 1,842 individuals genotyped for 41,569 markers into subpopulations in an unsupervised manner disregarding the population labels for each individual. The data matrices were generated with each row representing a SNP profile for an individual and each column representing a SNP genotype (0: homozygous wild type, 1: heterozygous and 2: homozygous variant). The ADMIXTURE [36] program was used to estimate individual ancestries of each individual from the same SNP genotypic data from K=2 to K=10 ancestors. ADMIXTURE uses the same maximum likelihood principle of STRUCTURE [37] to infer the ratio of assumed ancestors for each individual. The admixture ratios of individuals were plotted using the 'bar' function in MATLAB version 2009b on Linux operating system.

High-resolution study of population substructure within the Thai population was performed on the combined datasets 2 and 3 (992 individuals genotyped for 438,503 SNPs). Subpopulations were assigned using ipPCA with stopping criterion EigenDev=0.21. ADMIXTURE was used to estimate individual ancestries from K=2 to K=4 ancestors. Genome-wide Fst values [37] were calculated among all pair-wise combination of ipPCA assigned subpopulations using the Arlequin software with default settings [38], and the significance tested by permutation testing option for 1023 permutations. Fst values for each of the 438,503 SNPs among all pair-wise combination of ipPCA assigned subpopulations were calculated using the Arlequin software. The SNPs were then ranked according to Fst values in all pairwise subpopulation comparisons.

Results

In order to frame the Thai population in a worldwide context, the Thai genetic data were combined with the worldwide population data published in [29]. The combined dataset of 1,842 individuals was analyzed using the 41,569 SNP markers common to the two different microarray platforms (File S1). Consensus neighbor-joining (NJ) unrooted tree of populations assigned using the ethno-geographical information (Figure 1) reveals that the Southeast and East Asian populations are distinct from the rest of the world. Moreover, all the Southeast and East Asian populations occupy *distinctive* positions (clades with 100% bootstrap support) from other populations except for Thai Moken and Cambodian people who occupy positions in the tree with weaker bootstrap support. It is striking that the Thai subpopulations (according to the regional geographic origins) are also distinct.

Next, subpopulation genetic structure was analyzed using the ipPCA algorithm [34,35]. Subpopulation assignment of individuals by this algorithm is performed using an unsupervised clustering approach that does not use the individuals' ethno-geographical information. The subpopulations resolved by this algorithm are genetically homogeneous with no significant variation from that expected for a random collection of unrelated individuals. The resulting 24 subpopulations assigned by ipPCA generally reflected the individual ethno-geographical labels in agreement with the pattern from the consensus NJ tree (Figure 2), but with some interesting discrepancies. Mainland Thais were assigned to four subpopulations (SP19-22) together with some of the Thai Moken individuals from Xing's dataset. However, Thai Mokens were assigned exclusively to SP23. Interestingly, all

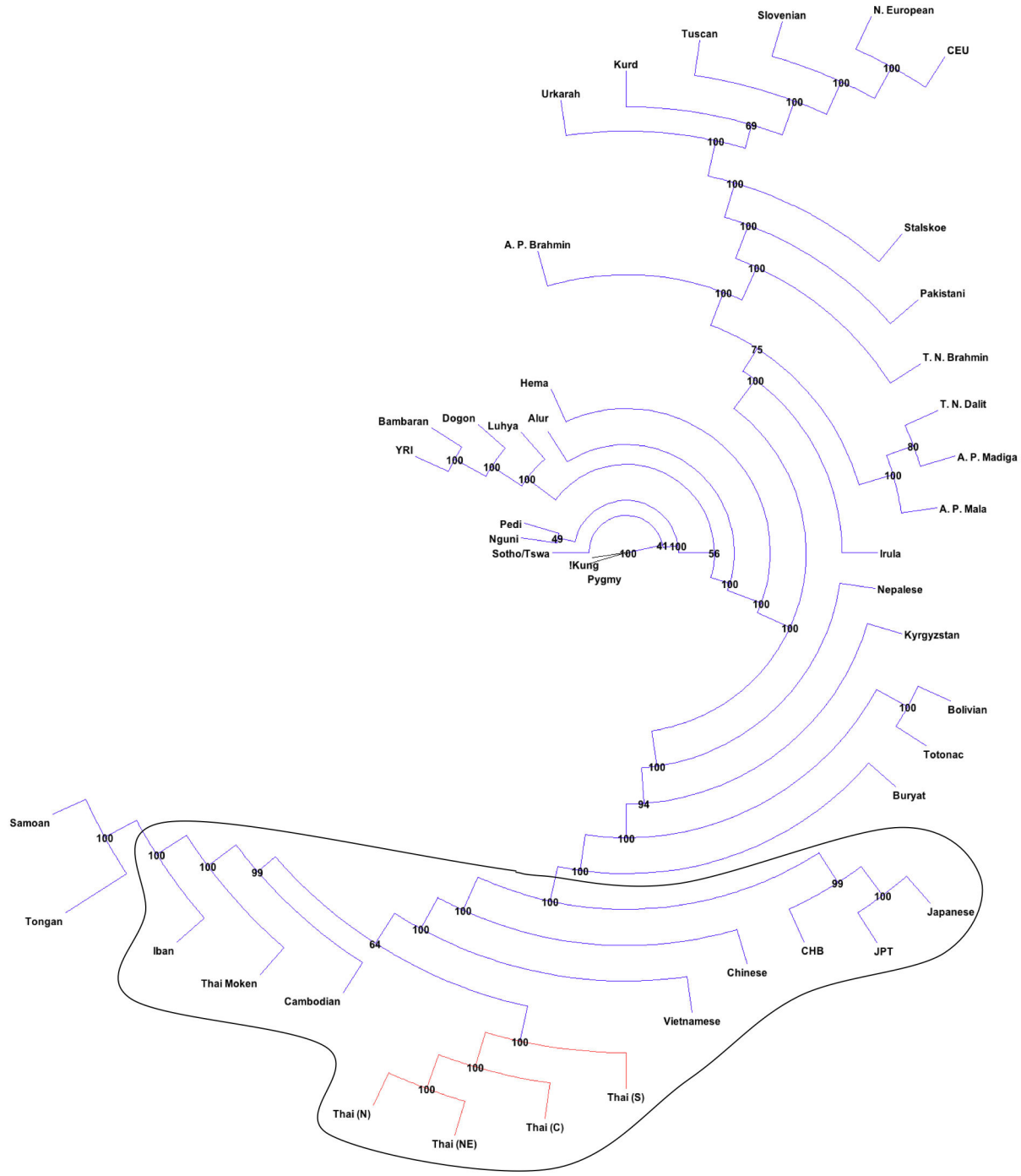


Figure 1. Consensus population Neighbor-Joining unrooted Tree. An amalgamated worldwide dataset of 1842 individuals genotyped for 41,569 SNPs was analyzed by PHYLIP. The minor allele frequencies for each population were calculated and used as input to produce the dissimilarity matrix using Nei's approach for unrooted NJ tree. The data were comprised of 850 individuals from 40 populations (dataset no.1; [29]), 618 Thai individuals (dataset no. 2; [30]) and 374 Thai individuals (dataset no. 3; this study). The Thai individuals from datasets no. 2 and 3 were assumed to belong to the same population and then separated into regional subpopulations based on self-reported origins: Thai (C), Thai (NE), Thai (N) and Thai (S). The other population labels are the same as those reported previously in [29], except "Thai" which has been re-labeled as "Thai-Moken". The consensus tree from 100 bootstrap replicates is shown, and the bootstrap values are indicated on each node of the tree. Southeast and East Asian populations are ringed and the clades separating Thai subpopulations are in red.

doi: 10.1371/journal.pone.0079522.g001

Vietnamese individuals were assigned with Thais in SP21 and SP22 and all Cambodians were assigned with Thais in SP19, 20 and 22. Some Chinese individuals were also assigned to SP22 with Thais, Vietnamese and Cambodians. Another important observation is that among the predominantly Thai subpopulations SP19-22, there appears to be regional bias. For instance, SP19 contained the majority of Southern Thais, while SP20 contained the majority of Northeastern Thais and SP21 the majority of Northern Thais. SP22 is dominated by Central Thais, although this subpopulation constitutes only a minority of the total of Central Thais. 20 Thai individuals appeared as genetically distinct "outliers" that could not be assigned to a specific subpopulation and were separated by ipPCA at different iterations of the algorithm (see Figure S2).

Next, admixture ratios of inferred ancestry (K=2 to 10) for each individual (ipPCA outliers excluded) were determined using the ADMIXTURE program [36]. When individuals are grouped according to their subpopulation assignments made by ipPCA, subpopulation-distinctive admixture patterns were observed at K=7 (Figure 3). Analysis with higher K ancestral clusters was not much more informative, since no new major ancestral components of any subpopulation were apparent. SP19-22 containing mostly Thai individuals were assigned with one major ancestral component (blue) and two minor components (pink and yellow) at K=7. The major blue component is also a major component of SP24 (Iban individuals) and to a lesser extent SP18 (mostly Chinese individuals).

Next, having shown substructure among the mainland Thai population with relatively few markers, a higher resolution analysis of 992 Thai individuals was performed using 438,503 SNP markers. Subpopulation assignment by ipPCA revealed four subpopulations labeled SPA, B, C and D (Figure 4). 20 outlier individuals could not be assigned to these four subpopulations (Figure S3), and were excluded from further analysis. The assignment of individuals to the four subpopulations SPA, B, C and D was correspondent with SP19, 20, 21, and 22, respectively from low-resolution ipPCA (Figure 2), with minor discrepancies (Table S2). Regional bias in subpopulation assignment was apparent, with predominance of South individuals in SPA, Northeast individuals in SP-B, and North individuals in SPC. SPD contains predominantly Central individuals, although this subpopulation does not constitute the majority of Central individuals. The level of variance in allele frequency among subpopulations SPA, B, C and D was determined by Fst analysis, and all pairwise comparisons were significant as shown by permutation testing (Table 1). Therefore, the population substructure found by ipPCA was cross-validated by Fst analysis. An alternative explanation for the substructure among the Thai samples is that the patterns reflect the individual's disease status or an artifact of the sample collection rather than general population structure. To test this hypothesis, deviation of minor allele frequency of the Thalassemia dataset was compared with the Major depressive disorder dataset from the expected ratio for all markers (438,503) by chi-squared analysis. No markers showed significant deviation (Table S3), indicating that the amalgamation of two datasets carried no bias for population structure analysis. Admixture analysis of these individuals with 438,503 SNP markers shows that each subpopulation has distinct patterns of admixture ratios at K=3; the fourth ancestral component is not informative as it carries only a tiny proportion of the ancestry in almost all individuals (Figure 5).

Having demonstrated substructure among the Thai population, an investigation of the genomic regions most diverged among the subpopulations was performed. The markers were ranked according to their Fst values in pairwise subpopulation comparisons (Table S4). Among the top-ranked markers with highest Fst between subpopulations, several were present in genes, and a few have been reported previously to affect phenotypic traits such as skin pigmentation and susceptibility to disease in other populations (Table 2). SPA is distinguished by high frequencies of SNPs in the OCA2 and SLC24A5 genes, and these markers are strongly associated across different populations with skin pigmentation [38]. The same markers are present at lowest frequency in SPC compared with SPA, B and D. SPB is distinguished by high frequency of the rs987870 SNP, which present in the HLA-DPB1 gene and is associated with pediatric asthma in different Asian populations [39]. SPD is distinguished by high frequency of several SNPs previously reported to be associated with disease in East Asian populations, including SNPs in the ADH4, ALDH2, BRAP and PANK4 genes which are associated with upper aerodigestive tract cancer, metabolic effect of alcohol, metabolic syndrome and type 2 diabetes, respectively [40-43]. Although some of the markers that distinguish the Thai subpopulations have phenotypic associations in other populations, phenotypic associations for the majority of distinguishing markers have not been reported.

Discussion

In this study, we have attempted to fill an important gap in the knowledge about human population genetics in MSEA. Consensus NJ tree (Figure 1) and ipPCA subpopulation assignment using a limited marker set (Figure 2) showed that genetically distinct groups exist among Eurasian peoples that are broadly aligned with ethno-linguistic labels. Among these populations though, there were some unexpected patterns. Five subpopulations of Thais were clearly distinct by NJ tree and ipPCA assignment, including a subpopulation of Thai individuals from the Xing dataset (SP23, Figure 2). The Thai individuals in SP23 were sampled from the Moken minority ethnic group, who are distinct from majority Thais in that they have lived continuously in coastal areas of Southern Thailand for several generations and speak their own Austronesian language [29]. The distinct ethnic identity of the Moken may thus have acted as a barrier to gene flow and led to genetic divergence from the majority of Thai people. The existence of the other four Thai subpopulations was unexpected as there are no ethno/linguistic distinguishing labels among these individuals. Geographical origin could partly explain the divergence of these subpopulations, with South, North and Northeastern Thais predominating SP19, 20 and 21 respectively. Central individuals comprised the majority of SP22, but this subpopulation was only a minority of the total of

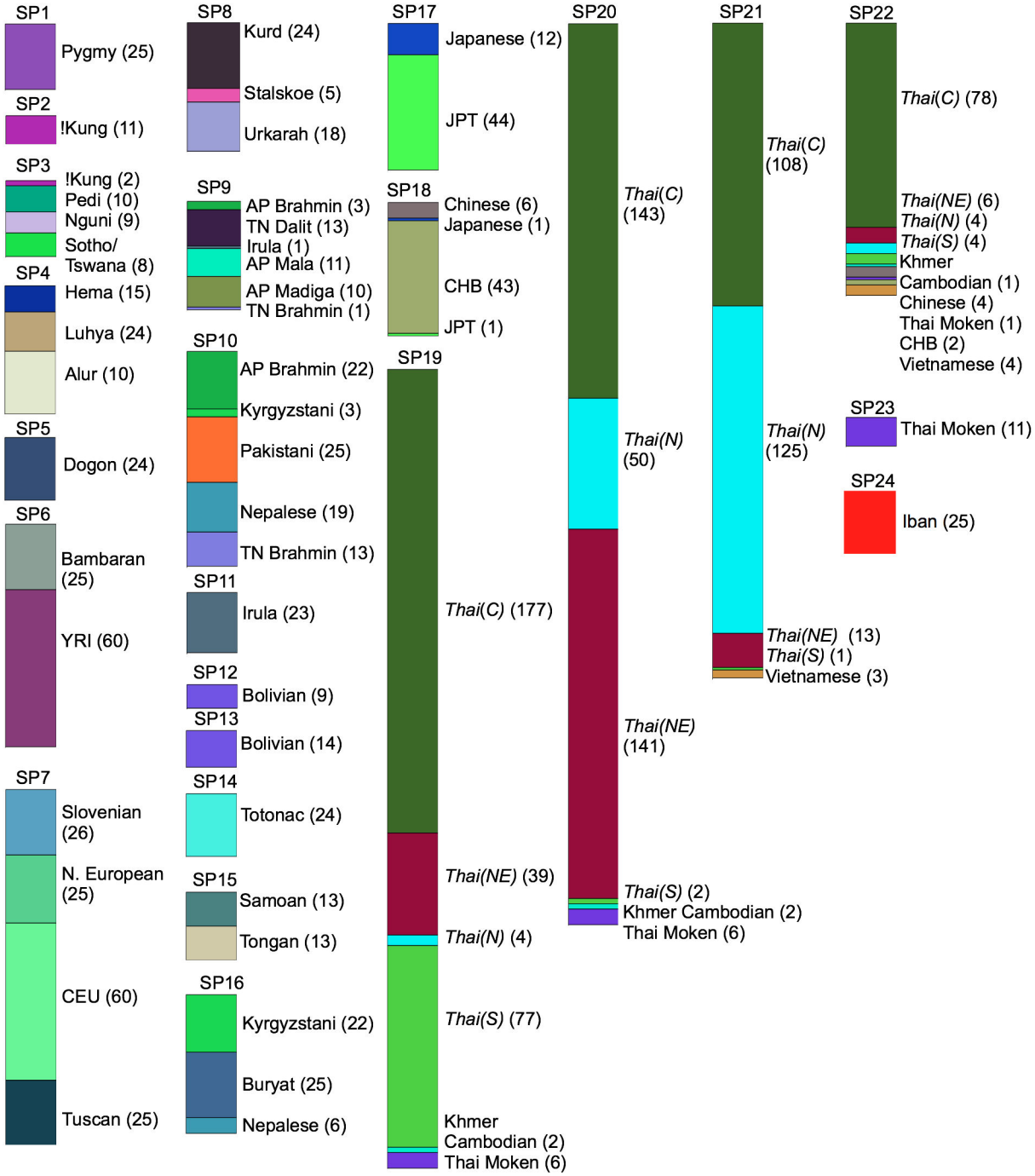


Figure 2. ipPCA subpopulation assignment. The amalgamated worldwide dataset of 1842 individuals was analyzed by ipPCA. The Thai ethno/geographical labels pertaining to datasets 2 and 3 are italicized; all other labels are the same as those shown in Figure 1. Individuals were assigned into 24 genetically distinct subpopulations (SP1 to 24) by ipPCA. 20 Thai individuals that could not be assigned to subpopulations are not shown. The height of each subpopulation bar is proportional to the number of assigned individuals.

doi: 10.1371/journal.pone.0079522.g002

Central individuals. Also surprising was the genetic similarity of other MSEA peoples with Thais, i.e., Cambodians were assigned with Thais in SP19, 20 and 22, while Vietnamese were assigned with Thais in SP21 and SP22 (with some

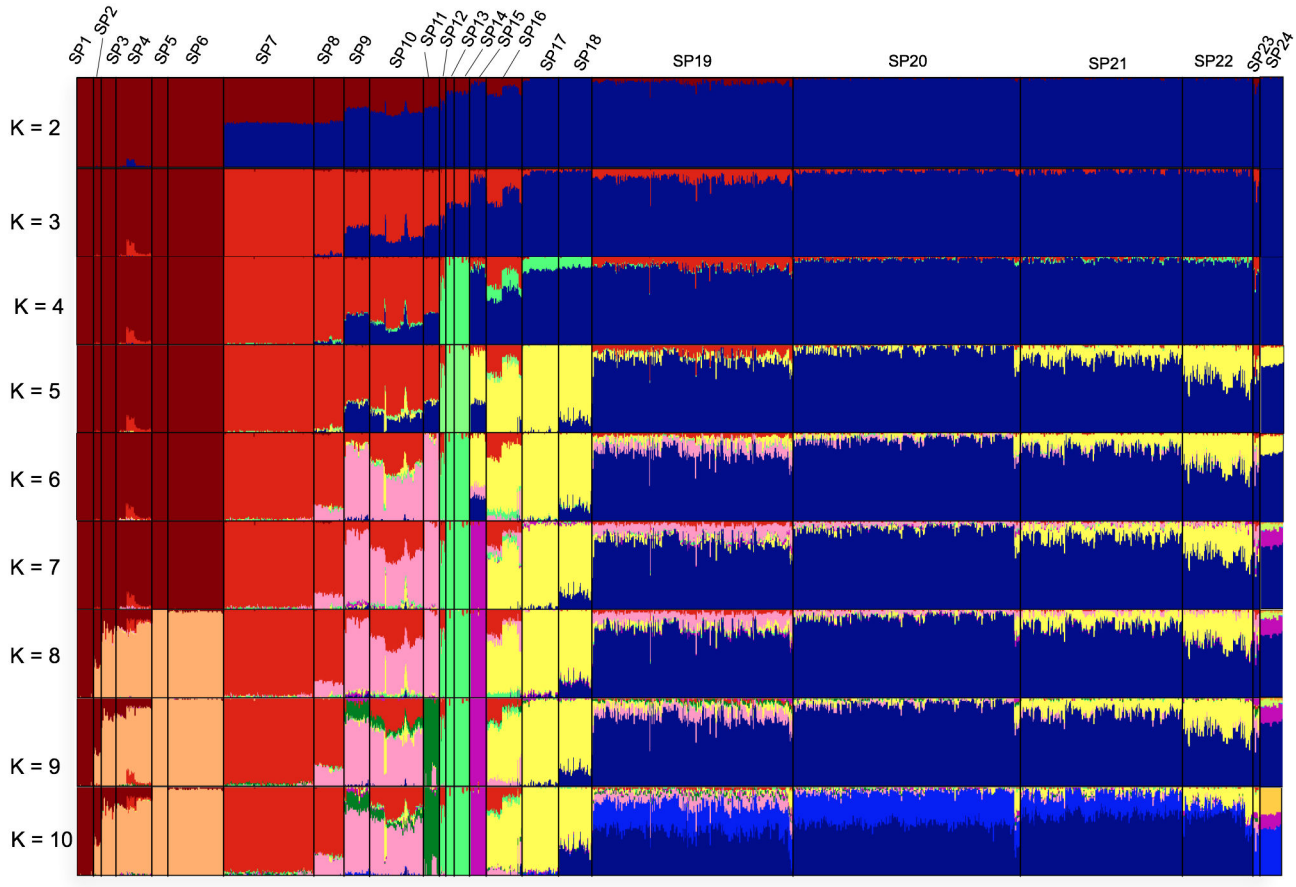


Figure 3. Ancestry analysis by ADMIXTURE. The amalgamated worldwide dataset of 1842 individuals was analyzed by the ADMIXTURE program. The number of K ancestral clusters was varied from 2 to 10. Individuals were grouped according to the subpopulation assignments made by ipPCA (Figure 2). The ordering of individuals within each subpopulation group is arbitrary.
doi: 10.1371/journal.pone.0079522.g003

Chinese also). Although the sampling of Cambodians and Vietnamese was much lower than Thais, the patterns suggest that the subpopulation structure within Thailand is representative of MSEA.

From the Admixture analysis at K=7, MSEA people in SP19-22 were shown to be represented by one major ancestral component (Figure 3). This component could represent the ancestry of autochthonous Austroasiatic people present in MSEA before the Tai expansion (see Introduction). This ancestry is also a major component of SP24 which is comprised of Austronesian-speaking Iban from the Peninsula Malaysia. Previous genetic analysis of Iban showed close association with MSEA people, suggesting that the ancestors of Iban were from MSEA [44]. The MSEA ancestors of the Iban and other Austronesians in MSEA were probably Austric-speaking migrants who migrated from central Thailand to the Malaysian Peninsula [45]. The most common mtDNA haplotypes in the Austronesian-speaking Thai Moken are also found in aboriginal peoples of the Malaysian Peninsula [46], and these Malay aborigines speak Austronesian and Austroasiatic languages. Among other Austronesian-speaking minorities in MSEA, the Cham group in Vietnam also has a closer genetic affiliation with Austroasiatic populations in MSEA than with Austronesian populations from Island Southeast Asia [47].

Four genetically distinct Thai subpopulations were assigned using 438,503 SNPs with essentially the same assignment as with the smaller marker set. The minor discrepancy between the two ipPCA analyses performed with different numbers of markers is clustering error since the ability to resolve population structure is dependent on the number of markers available [48]. Even with a larger marker set, a small number of Thai individuals could not be assigned to subpopulations by ipPCA and instead separated as outliers at various clustering steps of ipPCA (Figures S2 and S3). These outlier individuals may constitute individuals with recent non-SE Asian ancestry, or unaccounted for familial relationship. Such outlier individuals are likely to be present in any large population study and are typically excluded [49,50]. Among the four geographical regions of Thailand, the Central region is the most diverse in that no one subpopulation is dominant. In contrast, the other regions are more genetically homogeneous. The high diversity

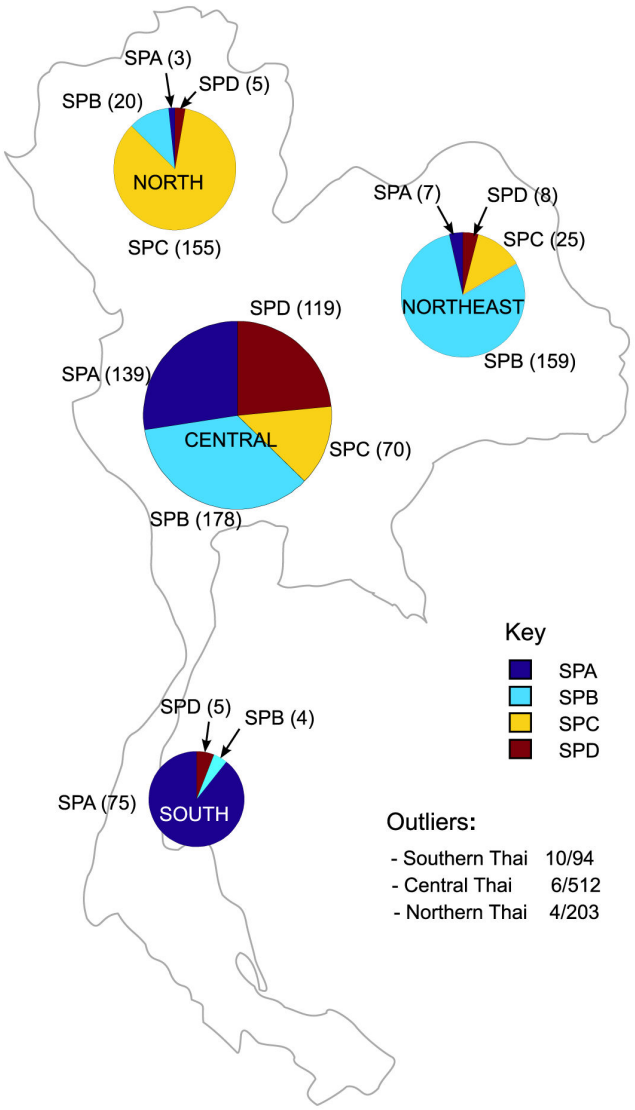


Figure 4. High-resolution ipPCA assignment of 992 Thai individuals. 992 Thai individuals from datasets no. 2 and 3 were combined and analyzed by ipPCA utilizing 438,503 SNP markers. Four subpopulations (SPA, B, C and D) were resolved by ipPCA, whereas 20 individuals could not be assigned to a subpopulation and are separated as "Outliers". The proportions of individuals assigned to each subpopulation are shown for each geographical region based on the available information of self-reported origin (North, Northeast, Central, and South).

doi: 10.1371/journal.pone.0079522.g004

of the Central region is likely because of recent migration, as this region has been the economic center of the country since the 15th Century AD Ayutthaya period. Although SP22/SPD constitutes a minority of Central Thais, SP22/SPD individuals are concentrated in this region. Several Chinese, Vietnamese and a Cambodian individual were assigned by ipPCA with Thais in SP22. One explanation for this pattern, given the

Table 1. Pairwise Fst analysis of Thai subpopulations.

	SP-A	SP-B	SP-C	SP-D
SP-A	0	0.0020*	0.0032*	0.0034*
SP-B		0	0.0015*	0.0025*
SP-C			0	0.0023*
SP-D				0

* Significance tests were performed with 1023 permutations and their resulting P-value < 0.01

doi: 10.1371/journal.pone.0079522.t001

modern history of MSEA is that the Thais, Vietnamese and Cambodian in SP-22 may be descendants of recent Chinese migrants. In support of this conjecture, Admixture analysis showed that these individuals share a prominent ancestry with predominantly Chinese SP18 individuals (yellow component in Figure 3). Moreover, among the top-ranked SNP markers which are present at high frequency in SP-D and distinguish it from the other four Thai subpopulations, three (rs671, rs3782886 and rs7535528) have previously been reported to be associated with disease in the Chinese [41-43]. The documented rapid expansion and assimilation of very recent (within 200 years) Chinese immigrants into Thailand (see Introduction) has thus created a sizeable genetically distinct Sino-Thai subpopulation. Other evidence to support a subpopulation of Sino-Thai includes the presence of an "EAsian" *Helicobacter pylori* haplotype among Thais, which is also found in Malays of recent Chinese descent [51].

The predominantly southern Thai subpopulation SP19/SPA is distinguishable from the other Thai subpopulations by the presence of minor ADMIXTURE-inferred ancestry at K=7 (pink component, Figure 3). This ancestry is a major component of subpopulations SP8-11 comprised of predominantly South and Central Asians. This ancestry in the SP19/SPA Thais may be the signal of earliest Australo-Melanesian ancestors who came from South and Central Asia and migrated via Southeast Asia to Australia. Other genetic evidence of these very early ancestors was reported in [28], who found that the Sakai from southern Thailand were the most diverged ethnic group from other Thais. The Sakai are a very small ethnic group living near the Malaysian border and have a Negrito appearance and speak their own Austroasiatic language similar to Semang Negritos in Malaysia [52]. Among the top-ranked SNP markers which are present at higher frequency in SP-A and distinguish it from the other four Thai subpopulations, two are in genes, namely SLC24A5 and OCA2, known to be associated with skin pigmentation in different populations. However, the association of skin pigmentation with these marker among Asian populations is weak, e.g., as shown among different aboriginal populations of Peninsula Malaysia [53]. The differences in allele frequencies for these markers, and others (Table 2), are thus not likely to reflect signals of selection among Thai subpopulations.

The other Thai subpopulations SP20/SPB and SP21/SPC are the two largest. Among the three SNPs which distinguish SPB from the other Thai subpopulations, one at higher frequency in the HLA-DPB1 gene has been reported to confer

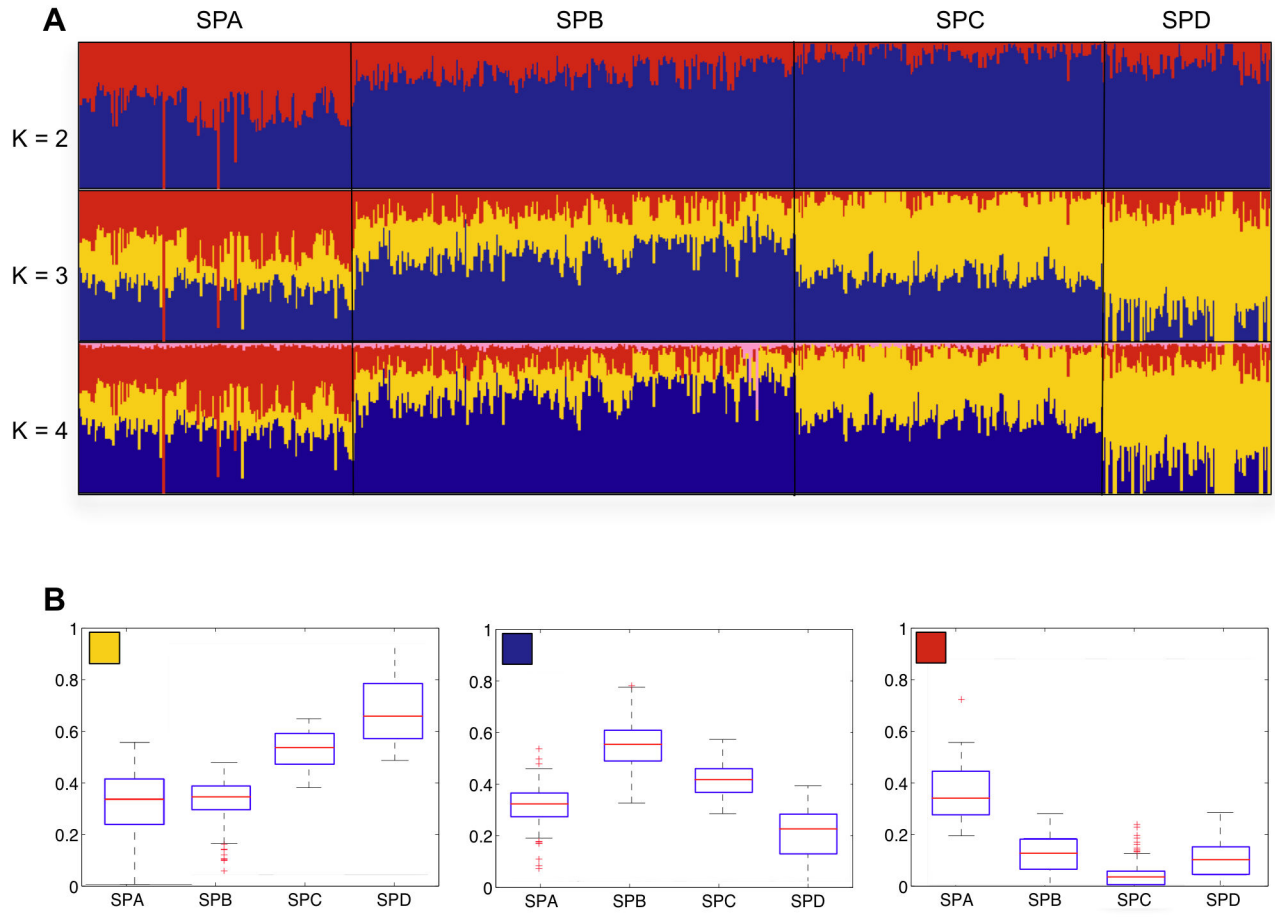


Figure 5. High-resolution ancestry analysis of 972 Thais. A) 972 individuals from datasets no. 2 and 3 (ipPCA Outliers removed) were combined and analyzed by ADMIXTURE utilizing 438,503 SNP markers. The individuals were grouped according to the subpopulation assignments made by ipPCA shown in Figure 4.
B) Box and whiskers plots for K=3 ADMIXTURE-inferred ancestral components (blue, yellow and red) of ipPCA-assigned subpopulations SPA, B, C and D.
doi: 10.1371/journal.pone.0079522.g005

a pediatric asthma risk (Table 2). Although the MAF differences among disease associated SNPs appear small among Thai subpopulations, they collectively may nonetheless have important consequences for GWAS. It is well-known that cases and controls must be drawn from a similar genetic background for GWAS, otherwise spurious associations will result [54]. We propose that future GWAS for the Thai population must take into account of the subpopulation background to avoid population structure confounding effects such as spurious associations and loss of power to detect subpopulation-specific disease associations. Regional grouping of samples may not be effective, particularly for the Central region where no one subpopulation is in the majority.

Conclusions

This study has elucidated the Thai population structure, revealing four major subpopulations. A major ancestry is common across these subpopulations, which is probably the signal of Austric ancestors who originally settled across most of MSEA. The more recent expansion of Tai-Kadai language throughout MSEA was thus accompanied by assimilation, rather than displacement of the indigenous people. On the other hand, the most recent assimilation of southern Chinese migrants has created shifts in population structure, with one example being the presence of a distinctive Sino-Thai subpopulation that is concentrated in the Central region of Thailand (but which is not in the majority).

Further sampling of genetic variation in other MSEA populations, particularly Vietnamese and Cambodians may shed further light on this pattern.

Table 2. Top-ranked SNPs with highest Fst between subpopulations with known phenotypic association.

									Subpopulation minor allele frequencies			
Fst[a] value (SPx-SPy)	Rank[b]	rsID	Chr	Position	Allele	Region	Gene	Reported Phenotypic association	SPA	SPB	SPC	SPD
0.023 (SPA-SPB)	109	rs4778220	15	25872900	T/G	intron	OCA2	hair color and skin pigmentation	0.1	0.03	0.02	0.02
0.046 (SPA-SPC)	12	rs1426654	15	46213776	AG	coding	SLC24A5	skin pigmentation	0.14	0.04	0.02	0.04
0.021 (SPB-SPC)	29	rs987870	6	33150858	T/C	Flanking 5'UTR	HLA-DPB1	pediatric asthma	0.16	0.24	0.13	0.12
0.037 (SPA-SPD)	65	rs3805322	4	1E+08	A/G	intron	ADH4	upper aerodigestive tract cancer	0.17	0.19	0.21	0.35
0.042 (SPB-SPD)	6	rs671	12	1.11E+08	T/C	coding	ALDH2	metabolic effect of alcohol	0.1	0.06	0.06	0.19
0.041 (SPB-SPD)	16	rs3782886	12	1.11E+08	A/G	coding	BRAP	metabolic syndrome	0.1	0.06	0.06	0.18
0.038 (SPB-SPD)	22	rs7535528	1	2434274	T/C	coding	PANK4	type II diabetes	0.22	0.18	0.21	0.36
0.046 (SPA-SPC)	13	rs2517646	6	30230554	T/C	intron	TRIM10	highly differentiated SNP between Chinese subpopulations	0.23	0.12	0.08	0.1
0.045 (SPA-SPC)	17	rs11130248	3	50327204	A/G	Flanking 5'UTR	COL4A1	susceptibility loci for keloid in the Japanese population	0.21	0.12	0.07	0.12
0.044 (SPA-SPC)	20	rs2291652	3	1.97E+08	T/C	coding	MUC3	endometriosis-related infertility	0.27	0.16	0.11	0.19
0.048 (SPA-SPD)	20	rs1165153	6	2592768	T/C	intron	SLC17A1	development of gout	0.38	0.36	0.27	0.18
0.035 (SPB-SPD)	33	rs103294	19	59489660	T/C	Flanking 3'UTR	LILRA3	prostate cancer	0.27	0.15	0.17	0.31

[a] Fst is the value between the specified pair-wise subpopulation comparison shown in parenthesis.

[b] Rank value refers to the rank of Fst value for the same pair-wise subpopulation comparison (see Table S4 for complete ranked list)

doi: 10.1371/journal.pone.0079522.t002

Supporting Information

File S1. List of SNP-ids for the 41,569 SNP markers common to the Illumina Human 610-Quad BeadChips Array and the Affymetrix Human SNP Array 6.0 platforms.
(ZIP)

Figure S1. MAF correlation of 41,569 SNPs between Illumina and Affymetrix platforms. MAFs for each SNP were calculated from a control population of European ancestry with 136 samples from Affymetrix [29] and 1,182 samples from Illumina [31] platforms, respectively. The calculated correlation coefficient is indicated by the red line.
(TIFF)

Figure S2. ipPCA clustering decision tree for analysis of combined datasets 1, 2 and 3 (worldwide datasets). The terminal nodes boxed in red represent ipPCA resolved subpopulations labeled SP1-24. The internal nodes represent groups of individuals with unresolved population structure. Terminal nodes marked with asterisks represent outlier individuals. The EigenDev value for each iteration of ipPCA is shown in each node; values >0.21 indicate the present of substructure.
(PDF)

Figure S3. ipPCA clustering decision tree for analysis of combined datasets 2 and 3 (Thai individuals). The terminal nodes boxed in red and labeled as SPA, SPB, SPC, and SPD represent ipPCA resolved subpopulations. Terminal nodes marked with asterisks represent outlier individuals. The numbers of individuals for each regional origin label (Thai C, S, NE and N) are indicated in each node. The intermediate nodes represent groups of individuals with unresolved population structure. The EigenDev value for each iteration of ipPCA is shown in each node; values >0.21 indicate the present of substructure.
(TIFF)

Table S1. Major depressive disorder GWAS top 50 associated SNP data.
(XLSX)

Table S2. Correspondence of individual ipPCA-assignments of SP19-22 with SPA-D.
(XLSX)

Table S3. Top 50 rank SNP from Chi-squared analysis between Thalassemia dataset and the Major depressive disorder dataset from the expected ratio for all markers.
(XLSX)

Table S4. Top 200 ranked SNPs based on Fst values for all pair-wise comparisons between SPA, SPB, SPC and SPD.
(XLSX)

Acknowledgements

We acknowledge personnel from departments of mental health, medical schools at Khonkhaen University, Chiangmai University, Prince of Songkhla Nakarin University, Ramatibodi Hospital and Siriraj Hospital, Thailand, for recruiting participants in the depression study. We thank Dr. Surakameth Mahasirimongkol from the Department of Medical Science, Ministry of Health, Thailand, for coordination of sample genotyping at the center for Genomic Medicine RIKEN, Yokohama, Japan. We also thank Dr. Jonathan Chan and Ms. Sattara Hattirat for their discussions and preliminary data analyses. Furthermore, we acknowledge the support from the Center for Excellence in Molecular Genetics of Cancer and Human Diseases, Department of Anatomy, Faculty of Medicine, Chulalongkorn University. Finally, we acknowledge the NIH GWAS Data Repository and the Joint Addiction, Aging, and Mental Health DAC (JAAMH) for providing data from the dbGaP accession number phs000168.v1.p1.

Author Contributions

Conceived and designed the experiments: ST PJS AA. Performed the experiments: MN SS OS SF VP. Analyzed the data: PW CN KC PJS ST. Wrote the manuscript: PW PJS ST.

References

1. Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. xi. Princeton, N.J.: Princeton University Press. p. 518, p.A paragraph return was deleted
2. Consortium HP-AS, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, et al. (2009) Mapping human genetic diversity in Asia. Science 326: 1541-1545.
3. Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR et al. (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet 89: 516-528. doi: 10.1016/j.ajhg.2011.09.005. PubMed: 21944045.
4. Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S et al. (2011) An Aboriginal Australian genome reveals separate human dispersals into Asia. Science 334: 94-98. doi:10.1126/science.1211177. PubMed: 21940856.
5. Stoneking M, Delfin F (2010) The human genetic history of East Asia: weaving a complex tapestry. Curr Biol 20: R188-R193. doi:10.1016/j.cub.2009.11.052. PubMed: 20178766.
6. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. Nature 461: 489-494. doi: 10.1038/nature08365. PubMed: 19779445.
7. Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N et al. (2008) Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. Am J Hum Genet 83: 445-456. doi:10.1016/j.ajhg.2008.08.019. PubMed: 18817904.
8. Xu S, Yin X, Li S, Jin W, Lou H et al. (2009) Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am J Hum Genet 85: 762-774. doi:10.1016/j.ajhg.2009.10.015. PubMed: 19944404.
9. Chen J, Zheng H, Bei JX, Sun L, Jia WH et al. (2009) Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet 85: 775-785. doi:10.1016/j.ajhg.2009.10.016. PubMed: 19944401.
10. Matsumura H, Pookajorn S (2005) A morphometric analysis of the Late Pleistocene Human Skeleton from the Moh Khiew Cave in Thailand. Homo 56: 93-118. doi:10.1016/j.jchb.2005.05.004. PubMed: 16130834.
11. Matsumura H, Hudson MJ (2005) Dental perspectives on the population history of Southeast Asia. Am J Phys Anthropol 127: 182-209. doi:10.1002/ajpa.20067. PubMed: 15558609.
12. Oota H, Kurosaki K, Pookajorn S, Ishida T, Ueda S (2001) Genetic study of the Paleolithic and Neolithic Southeast Asians. Hum Biol 73: 225-231. doi:10.1353/hub.2001.0023. PubMed: 11446426.
13. Hill C, Soares P, Mormina M, Macaulay V, Meehan W et al. (2006) Phylogeography and ethnogenesis of aboriginal Southeast Asians. Mol Biol Evol 23: 2480-2491. doi:10.1093/molbev/msl124. PubMed: 16982817.
14. Thangaraj K, Chaubey G, Reddy AG, Singh VK, Singh L (2006) Unique origin of Andaman Islanders: insight from autosomal loci. J Hum Genet 51: 800-804. doi:10.1007/s10038-006-0026-0. PubMed: 16924390.
15. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D et al. (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308: 1034-1036. doi:10.1126/science.1109792. PubMed: 15890885.
16. Higham C (1996) The Bronze Age of Southeast Asia. xvi. Cambridge, England ; New York: Cambridge University Press. 381 pp.
17. Bellwood PS, Fox JJ, Tryon DT (1996) The Austronesians: historical and comparative perspectives. Canberra: Department of Anthropology. : Research School of Pacific and Asian Studies. 359 p
18. Lertrit P, Poolsuwan S, Thosarat R, Sanpachudayan T, Boonyarit H et al. (2008) Genetic history of Southeast Asian populations as revealed by ancient and modern human mitochondrial DNA analysis. Am J Phys Anthropol 137: 425-440. doi:10.1002/ajpa.20884. PubMed: 18615504.
19. Schliesinger J (2001) Tai groups of Thailand. Bangkok, Thailand: White Lotus Press.
20. Baker CJ, Pasuk P (2009) A history of Thailand. Cambridge ; New York: Cambridge University Press. 315,
21. Ooi KG (2004) Southeast Asia : a historical encyclopedia, from Angkor Wat to East Timor. Santa Barbara, CA: ABC-CLIO.
22. Kutanan W, Kampuansai J, Colonna V, Nakbunlung S, Lertvicha P et al. (2011) Genetic affinity and admixture of northern Thai people along their migration route in northern Thailand: evidence from autosomal STR loci. J Hum Genet 56: 130-137. doi:10.1038/jhg.2010.135. PubMed: 21107341.
23. Mahasirimongkol S, Chantratita W, Promso S, Pasomsab E, Jinawath N et al. (2006) Similarity of the allele frequency and linkage disequilibrium pattern of single nucleotide polymorphisms in drug-related gene loci between Thai and northern East Asian populations: implications for tagging SNP selection in Thais. J Hum Genet 51: 896-904. doi:10.1007/s10038-006-0041-1. PubMed: 16957813.
24. Listman JB, Malison RT, Sughondhabirom A, Yang BZ, Raaum RL et al. (2007) Demographic changes and marker properties affect detection of human population differentiation. BMC Genet 8: 21. doi: 10.1186/1471-2156-8-21. PubMed: 17498298.
25. Xu S, Kangwanpong D, Seielstad M, Srikummool M, Kampuansai J et al. (2010) Genetic evidence supports linguistic affinity of Mlabri--a hunter-gatherer group in Thailand. BMC Genet 11: 18. doi: 10.1186/1471-2156-11-18. PubMed: 20302622.
26. Zimmermann B, Bodner M, Amory S, Fendt L, Rock A et al. (2009) Forensic and phylogeographic characterization of mtDNA lineages from northern Thailand (Chiang Mai). Int J Leg Med 123: 495-501. doi: 10.1007/s00414-009-0373-4.
27. Besaggio D, Fuselli S, Srikummool M, Kampuansai J, Castrì L et al. (2007) Genetic variation in Northern Thailand Hill Tribes: origins and relationships with social structure and linguistic differences. BMC Evol Biol 7 Suppl 2: S12. doi:10.1186/1471-2148-7-12. PubMed: 17767728.
28. Fucharoen G, Fucharoen S, Horai S (2001) Mitochondrial DNA polymorphisms in Thailand. J Hum Genet 46: 115-125. doi:10.1007/s100380170098. PubMed: 11310578.
29. Xing J, Watkins WS, Shlien A, Walker E, Huff CD et al. (2010) Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. Genomics 96: 199-210. doi:10.1016/j.ygeno.2010.07.004. PubMed: 20643205.
30. Nuinoon M, Makarasara W, Mushiroda T, Setianingsih I, Wahidiyat PA et al. (2010) A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E. Hum Genet 127: 303-314. doi:10.1007/s00439-009-0770-2. PubMed: 20183929.
31. Lee JH, Cheng R, Graff-Radford N, Foroud T, Mayeux R et al. (2008) Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: implication of additional loci. Arch Neurol 65: 1518-1526. doi:10.1001/archneur.65.11.1518. PubMed: 19001172.

32. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17: 368-376. doi:10.1007/BF01734359. PubMed: 7288891.
33. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460. doi:10.1186/1471-2105-8-460. PubMed: 18034891.
34. Intarapanich A, Shaw PJ, Assawamakin A, Wangkumhang P, Ngamphiw C et al. (2009) Iterative pruning PCA improves resolution of highly structured populations. BMC Bioinformatics 10: 382. doi:10.1186/1471-2105-10-382. PubMed: 19930644.
35. Limpiti T, Intarapanich A, Assawamakin A, Shaw PJ, Wangkumhang P et al. (2011) Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and structure. BMC Bioinformatics 12: 255. doi:10.1186/1471-2105-12-255. PubMed: 21699684.
36. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655-1664. doi:10.1101/gr.094052.109. PubMed: 19648217.
37. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959. PubMed: 10835412.
38. Giardina E, Pietrangeli I, Martínez-Labarga C, Martone C, de Angelis F et al. (2008) Haplotypes in SLC24A5 Gene as Ancestry Informative Markers in Different Populations. Curr Genomics 9: 110-114. doi:10.2174/138920208784139528. PubMed: 19440451.
39. Noguchi E, Sakamoto H, Hirota T, Ochiai K, Imoto Y et al. (2011) Genome-wide association study identifies HLA-DP as a susceptibility gene for pediatric asthma in Asian populations. PLOS Genet 7: e1002170. PubMed: 21814517.
40. Oze I, Matsuo K, Suzuki T, Kawase T, Watanabe M et al. (2009) Impact of multiple alcohol dehydrogenase gene polymorphisms on risk of upper aerodigestive tract cancers in a Japanese population. Cancer Epidemiol Biomarkers Prev 18: 3097-3102. doi:10.1158/1055-9965.EPI-09-0499. PubMed: 19861527.
41. Tan A, Sun J, Xia N, Qin X, Hu Y et al. (2012) A genome-wide association and gene-environment interaction study for serum triglycerides levels in a healthy Chinese male population. Hum Mol Genet 21: 1658-1664. doi:10.1093/hmg/ddr587. PubMed: 22171074.
42. Wu L, Xi B, Hou D, Zhao X, Liu J et al. (2013) The single nucleotide polymorphisms in BRAP decrease the risk of metabolic syndrome in a Chinese young adult population. Diabetes Vasc Dis Res 10: 202-207. doi:10.1177/1479164112455535.
43. Li Y, Wu GD, Zuo J, Meng Y, Fang FD (2005) [Screening susceptibility genes of type 2 diabetes in Chinese population by single nucleotide polymorphism analysis]. Zhongguo Yi Xue Ke Xue Yuan Xue Bao 27: 274-279. PubMed: 16038259.
44. Simonson TS, Xing J, Barrett R, Jerah E, Loa P et al. (2011) Ancestry of the Iban is predominantly Southeast Asian: genetic evidence from autosomal, mitochondrial, and Y chromosomes. PLOS ONE 6: e16338. doi:10.1371/journal.pone.0016338. PubMed: 21305013.
45. Jinam TA, Hong LC, Phipps ME, Stoneking M, Ameen M et al. (2012) Evolutionary history of continental southeast Asians: "early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. Mol Biol Evol 29: 3513-3527. doi:10.1093/molbev/mss169. PubMed: 22729749.
46. Dancause KN, Chan CW, Arunotai NH, Lum JK (2009) Origins of the Moken Sea Gypsies inferred from mitochondrial hypervariable region and whole genome sequences. J Hum Genet 54: 86-93. doi:10.1038/jhg.2008.12. PubMed: 19158811.
47. Peng MS, Quang HH, Dang KP, Trieu AV, Wang HW et al. (2010) Tracing the Austronesian footprint in Mainland Southeast Asia: a perspective from mitochondrial DNA. Mol Biol Evol 27: 2417-2430. doi:10.1093/molbev/msq131. PubMed: 20513740.
48. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLOS Genet 2: e190. doi:10.1371/journal.pgen.0020190. PubMed: 17194218.
49. Trust Wellcome. Case Control C (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678
50. Luca D, Ringquist S, Klei L, Lee AB, Gieger C et al. (2008) On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82: 453-463. doi:10.1016/j.ajhg.2007.11.003. PubMed: 18252225.
51. Breurec S, Guillard B, Hem S, Brisse S, Dieye FB et al. (2011) Evolutionary history of Helicobacter pylori sequences reflect past human migrations in Southeast Asia. PLOS ONE 6: e22058. doi:10.1371/journal.pone.0022058. PubMed: 21818291.
52. Benjamin G, Chou C (2002) Tribal communities in the Malay world : historical, cultural, and social perspectives. Leiden, the Netherlands. Singapore: International Institute for Asian Studies; Institute of Southeast Asian Studies. 489,
53. Ang KC, Ngu MS, Reid KP, Teh MS, Aida ZS et al. (2012) Skin color variation in Orang Asli tribes of Peninsular Malaysia. PLOS ONE 7: e42752. doi:10.1371/journal.pone.0042752. PubMed: 22912732.
54. Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73: 1162-1169. doi:10.1086/379378. PubMed: 14574645.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 148014, 9 pages
http://dx.doi.org/10.1155/2013/148014

Research Article

Biomarker Selection and Classification of "-*Omics*" Data Using a Two-Step Bayes Classification Framework

Anunchai Assawamakin,[1] Supakit Prueksaaroon,[2] Supasak Kulawonganunchai,[3] Philip James Shaw,[3] Vara Varavithya,[4] Taneth Ruangrajitpakorn,[5] and Sissades Tongsima[3]

[1] *Department of Pharmacology, Faculty of Pharmacy, Mahidol University, 447 Sri-Ayuthaya Road, Rajathevi, Bangkok 10400, Thailand*
[2] *Department of Electrical and Computer Engineering, Faculty of Engineering, Thammasat University, 99 Phahonyothin Road, Khlong Nueng, Khlong Luang, Pathum Thani 12120, Thailand*
[3] *National Center for Genetic Engineering and Biotechnology, 113 Thailand Science Park, Phahonyothin Road, Khlong Nueng, Khlong Luang, Pathum Thani 12120, Thailand*
[4] *Department of Electrical and Computer Engineering, King Mongkut University of Technology North Bangkok, 1518 Piboonsongkarm Road, Bangkok 10800, Thailand*
[5] *Language and Semantic Technology Laboratory, National Electronic and Computer Technology Center, 112 Thailand Science Park, Phahonyothin Road, Khlong Nueng, Khlong Luang, Pathum Thani 12120, Thailand*

Correspondence should be addressed to Sissades Tongsima; sissades@biotec.or.th

Received 22 April 2013; Revised 4 July 2013; Accepted 6 August 2013

Academic Editor: Florencio Pazos

Identification of suitable biomarkers for accurate prediction of phenotypic outcomes is a goal for personalized medicine. However, current machine learning approaches are either too complex or perform poorly. Here, a novel two-step machine-learning framework is presented to address this need. First, a Naïve Bayes estimator is used to rank features from which the top-ranked will most likely contain the most informative features for prediction of the underlying biological classes. The top-ranked features are then used in a Hidden Naïve Bayes classifier to construct a classification prediction model from these filtered attributes. In order to obtain the minimum set of the most informative biomarkers, the bottom-ranked features are successively removed from the Naïve Bayes-filtered feature list one at a time, and the classification accuracy of the Hidden Naïve Bayes classifier is checked for each pruned feature set. The performance of the proposed two-step Bayes classification framework was tested on different types of *-omics* datasets including gene expression microarray, single nucleotide polymorphism microarray (SNParray), and surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) proteomic data. The proposed two-step Bayes classification framework was equal to and, in some cases, outperformed other classification methods in terms of prediction accuracy, minimum number of classification markers, and computational time.

1. Introduction

In recent years, the advent of technologies such as microarrays, proteomics, and next-generation sequencing has transformed life science. The data from these experimental approaches provide a comprehensive picture of the complexity of biological systems at different levels. Within each of these "*-omics*" data strata, there exists a small amount of information relevant to particular biological questions, for example, indicative markers or biomarkers (for short) that can accurately predict (classify) phenotypic outcomes. Various machine learning techniques have been proposed to identify biomarkers that can accurately predict phenotypic classes by learning the cryptic pattern from *-omics* data [1]. There are three main categories of machine learning methods for biomarker selection and phenotypic classification, namely, *filter*, *wrapper*, and *embedded* [2]. These methods differ in the degree of computational complexity and prediction accuracy outcomes.

Filtering methods are the least computationally complex and are used to identify a subset of the most informative features from *-omics* data to assist the following classification

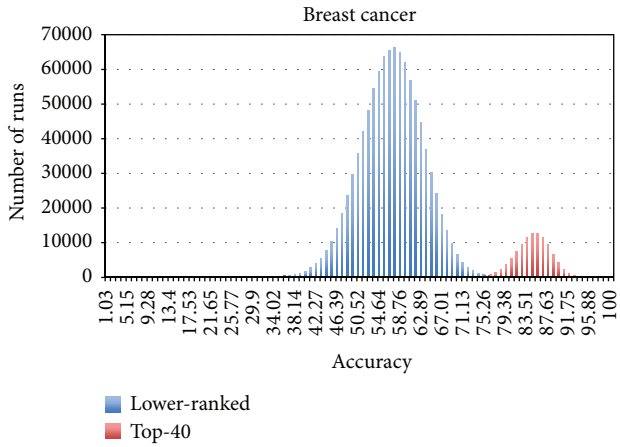


Figure 1: Empirical testing of NB selection using breast cancer dataset. Training breast cancer dataset was sampled 1 million times for lower-ranked marker set and 100,000 times for the top 40-ranked marker set.

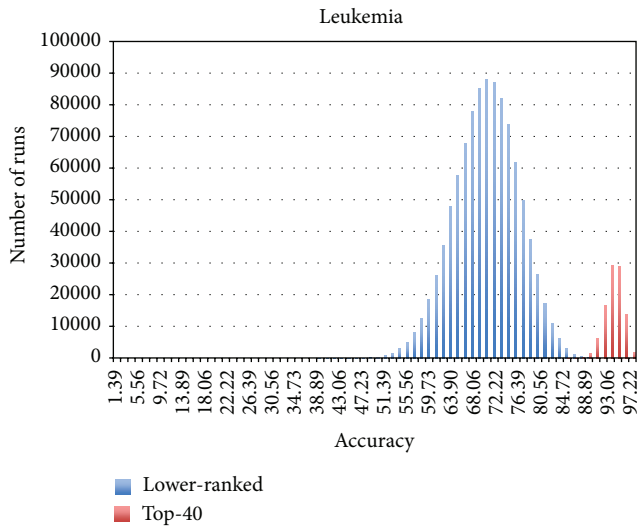


Figure 2: Empirical testing of NB selection using leukemia dataset. Training leukemia dataset was sampled 1 million times for lower-ranked marker set and 100,000 times for the top 40-ranked marker set.

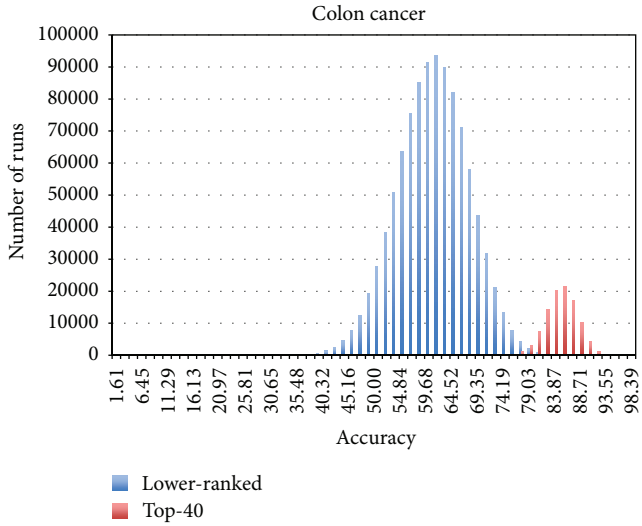


Figure 3: Empirical testing of NB selection using colon cancer dataset. Training colon cancer dataset was sampled 1 million times for lower-ranked marker set and 100,000 times for the top 40-ranked marker set.

process. These approaches operate by generating a value for each marker according to their degree of correlation with a given phenotype (class label), and then markers are ranked. However, filter methods are subject to selection of redundant biomarkers; furthermore, these methods cannot explore solutions that require more than one marker to predict the underlying classes. A common filter method is the well-known Student's t-test, which is popular because of its simplicity [7].

Wrapper methods iteratively perform combinatorial biomarker search aiming to optimize the predictive power of a classification model. Since this combinatorial optimization process is computationally complex, NP-hard problem, many heuristic have been proposed, for example, [8], to reduce the search space and thus reduce the computational burden of the biomarker selection.

Similar to wrapper methods, embedded methods attempt to perform feature selection and classification simultaneously. Embedded methods, however, integrate feature selection into the construction of classification models. Recursive feature elimination support vector machine (SVM-RFE) is a widely used technique for analysis of microarray data [9, 10]. The SVM-RFE procedure constructs a classification model using all available features, and the least informative features for that particular model are eliminated. The process of classification model building and feature elimination is repeated until a model using the predetermined minimum number of features is obtained. This approach is thus computationally impractical when a large number of features are considered, since many iterations of the algorithm are required.

Another approach for performing class prediction is Naïve Bayes (NB). The NB learning model relies on Bayes probability theory, in which attributes are used to build a statistical estimator for predicting classes. NB is the simplest form of the general Bayesian network in which all attributes are assumed to be independent. This assumption is not valid for biological systems, in which complex networks of interactions exist, that is, gene regulation; hence, NB has not received much attention for predicting biological classes. Nevertheless, modified Bayesian classification approaches which account for dependencies among features can accurately predict biological classes. Notable examples include selective Bayesian classifiers (SCB) [11], tree-augmented Naïve Bayes (TAN), and averaged one-dependence estimators (AODE) [12]. The Hidden Naïve Bayes (HNB) classifier approach has recently been claimed to show significant improvement over other NB techniques [13]. HNB uses a discrete structural model and hence requires the discretization for preprocessing with continuous signal attributes, for example, expression microarray data.

Table 1: Actual performance results on breast cancer (KRBDSR).

Criterion	Filter	Wrapper methods				Hybrid methods				
	Fisher's ratio	RFE-LNW-GD	RFE-SVM	RFE-LSSVM	RFE-RR	RFE-FLDA	RFE-LNW1	RFE-LNW2	RFE-FSVs-7DK	NB-HNB
Accuracy	0.88	0.78	0.76	0.75	0.74	0.75	0.82	0.88	0.85	0.91
Sensitivity, specificity	0.83, 0.90	0.77, 0.81	0.68, 0.80	0.68, 0.80	0.68, 0.77	0.69, 0.80	0.74, 0.88	0.82, 0.90	0.84, 0.86	0.91, 0.91
Number of genes selected	35	26	33	36	39	28	35	33	21	25

Table 2: Actual performance results on leukemia (KRBDSR).

Criterion	Filter	Wrapper methods				Hybrid methods				
	Fisher's ratio	RFE-LNW-GD	RFE-SVM	RFE-LSSVM	RFE-RR	RFE-FLDA	RFE-LNW1	RFE-LNW2	RFE-FSVs-7DK	NB-HNB
Accuracy	0.99	0.99	0.99	0.99	0.48	0.997	0.96	0.99	0.98	1.00
Sensitivity, specificity	0.95, 1.00	1.00, 0.99	0.95, 1.00	0.98, 0.99	1.00, 0.31	0.99, 1.00	0.90, 0.98	0.95, 1.00	0.91, 1.00	1.00, 1.00
Number of genes selected	4	5	4	30	6	5	4	4	3	14

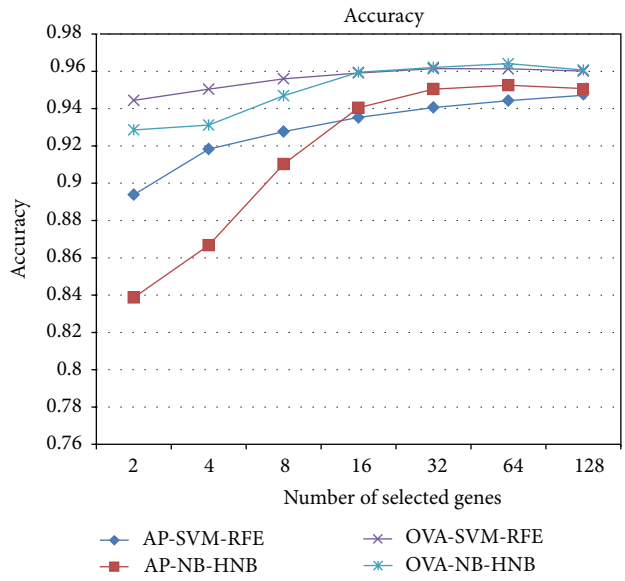


Figure 4: Comparison of average accuracy results over all datasets (Avg), 35 All-Paired datasets (AP) and 9 One-Versus-All (OVA) datasets.

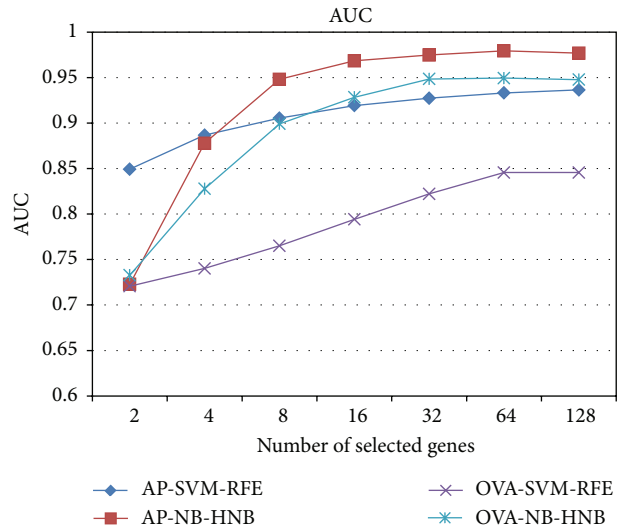


Figure 5: AUC metrics comparing different approaches.

In this paper, a hybrid statistic-based machine learning approach is suggested that utilizes a two-step heuristic to dramatically reduce the computational time required by HNB, while maintaining high-prediction accuracy when comparing with the other state-of-the-art machine learning techniques. Our proposed two-step framework includes (1) attribute filtering using Naïve Bayes (NB) to extract the most informative features and thus greatly reduce the number of data dimensions and (2) the subsequent higher order classification using Hidden Naïve Bayes (HNB). HNB can be used to construct a high-dimensional classification model that takes into account dependencies among the attributes for analysis of complex biological -omics datasets containing dependencies of features. The performance of the proposed two-step Bayes classification framework was evaluated using datasets from SNParray, cDNA expression microarray, and SELDI-TOF proteomics. The proposed framework was compared with SVM-RFE in terms of classification accuracy, area under the ROC curve (AUC), sensitivity, specificity, and the number of informative biomarkers used for classification.

2. Results and Discussion

In order to understand how a two-step Bayes classification framework can be used to analyze -omics data, the experiments in this section were performed in three different

Table 3: Actual performance results on colon cancer (KRBDSR).

Criterion	Filter	Wrapper methods				Hybrid methods				
	Fisher's ratio	RFE-LNW-GD	RFE-SVM	RFE-LSSVM	RFE-RR	RFE-FLDA	RFE-LNW1	RFE-LNW2	RFE-FSVs-7DK	NB-HNB
Accuracy	0.90	0.87	0.87	0.91	0.83	0.89	0.91	0.89	0.91	0.93
Sensitivity, specificity	0.92, 0.88	0.89, 0.85	0.92, 0.79	0.97, 0.81	0.77, 0.91	0.93, 0.84	0.93, 0.88	0.93, 0.84	0.93, 0.89	0.93, 0.90
Number of genes selected	16	17	16	22	19	14	10	15	12	23

scenarios. First, we need to know if Naïve Bayes (NB) filtering can select good (highly informative) candidate biomarkers, for example, SNPs, genes, or proteins for construction of an accurate classification model. Secondly, we need to demonstrate that the two-step Bayes classification framework is at least as good as a state-of-the-art method such as SVM-RFE. Standard performance metrics were used to carry out the head-to-head comparison. Finally, we show how the two-step Bayes classification framework can also be applied to other kinds of -omics datasets, in which SNP genotyping dataset and proteomic profiles from SELDI-TOF were analyzed.

2.1. Evaluation of Naïve Bayes Filtering. First, we hypothesized that the Naïve Bayes (NB) ranking module can precisely extract the most informative biomarkers to maximize the accuracy of the corresponding classification model. To our knowledge, the use of NB as a filter method for identifying highly informative markers is novel. NB allows us to interrogate each marker separately if it can predict the class outcomes with high confidence. The marker can be combined with other informative markers and collectively improve the prediction accuracy in successive multifeature classification HNB step. The experiments were performed using three microarray datasets, namely, breast cancer (24481 genes), leukemia (7129 genes), and colon cancer (2000 genes), from the Kent Ridge Biomedical Data Set Repository (KRBDSR) [3]. The NB and HNB modules from the popular open source machine learning software, Waikato Environment for Knowledge Analysis (Weka) [14], were employed for the two-step Bayes classification framework. The NB module was used to select the top features (genes), whose prediction accuracies are greater than or equal to 75%. Using this criterion, approximately 40 genes were selected by the NB filtering module as the top-ranked informative markers. From empirical testing of several datasets, we have found that this filtering criterion is broadly applicable for reducing the number of markers to a level practical for the subsequent HNB module, without reducing the accuracy of the final HNB classification. The *sampling-with-replacement* of 20 markers was done from both the top 40 group as well as the remaining unselected markers in the three datasets. The classification accuracy of each sampling was tested using the Hidden Naïve Bayes (HNB) module with 10-fold cross-validation classification available in Weka. Twenty genes were sampled from the selected top 40 and the unselected lower-ranked genes for 100,000 and 1 million times, respectively. The frequencies for each classification accuracy event were recorded. The results for the breast cancer, leukemia, and colon cancer data are shown in Figures 1, 2, and 3, respectively. Most importantly, sampling from the top 40 NB-selected genes gives the highest prediction accuracy, and the density distribution plots from the selected top 40 and unselected lower-ranked genes give minimal or no overlap. These results suggest that the NB filtering module is effective for selection of the most informative markers to be used in the following classification model construction by HNB. The threshold of top-ranked m-genes could be optimized for each type of dataset; that is, more or fewer than 40 markers may give slightly better prediction accuracy in the final HNB constructed model. However, in this paper, we did not exhaustively test different m-thresholds, as our focus is more to demonstrate the NB-HNB combination approach.

When the top NB selected genes were used for classification by HNB, the prediction accuracy was excellent for the leukemia dataset (average accuracy 92.90%; range 100% to 87.5%) and good for the breast (average 84.67%; range 96.90–70.10%) and colon cancer datasets (average 86.53%; range 96.77–70.97%). In contrast, the HNB prediction accuracy using markers from the lower-ranked unselected genes was markedly poor: breast cancer average prediction accuracy 57.16% (range 84.54–27.84%), leukemia average accuracy 72.14% (range 97.22–40.28%), and colon cancer average accuracy 50.16% (range 53.16–30.65%).

It should be noted that NB filtering is not a good realistic statistical model because of the underlying independency assumption among the features (see Section 4.2). In other words, the top NB selected attributes may not always contain the optimal set of features for classification. Nonetheless, when feeding the NB top-ranked attributes to the successive HNB step, HNB was able to better construct a higher order interaction prediction model from these features without exhaustively searching for all different combinations.

2.2. Head-to-Head Comparison with SVM-RFE. In order to clearly demonstrate the performance of the two-step Bayes classification framework, a head-to-head performance evaluation between the state-of-the-art machine learning technique, recursive feature elimination support vector machine (SVM-RFE), and our proposed framework was performed. There are 42 previously published SVM-RFE analyses for comparison (see full listing in Section 4). The performance of the two-step Bayes classification framework was compared with the results published in [15]. Nine different machine learning techniques, grouped as filtering, wrapper, and hybrid methods, were compared using breast cancer,

Table 4: Performance comparison between NB-HNB and SVM-RFE on GEMLeR datasets.

Data	NB-HNB		SVM-RFE	
	Accuracy	Number of genes selected	Accuracy	Number of genes selected
AP_Breast_Colon	0.96	22	0.96	8
AP_Breast_Kidney	0.96	17	0.96	8
AP_Breast_Lung	0.94	27	0.94	16
AP_Breast_Omentum	0.95	25	0.96	32
AP_Breast_Ovary	0.96	17	0.96	16
AP_Breast_Prostate	0.99	28	0.99	8
AP_Breast_Uterus	0.96	27	0.95	8
AP_Colon_Kidney	0.97	10	0.98	32
AP_Colon_Lung	0.95	17	0.94	32
AP_Colon_Omentum	0.95	18	0.94	32
AP_Colon_Ovary	0.95	11	0.94	16
AP_Colon_Prostate	0.98	20	0.98	8
AP_Colon_Uterus	0.96	10	0.95	16
AP_Endometrium_Breast	0.97	20	0.97	32
AP_Endometrium_Colon	0.95	21	0.97	32
AP_Endometrium_Kidney	0.98	17	0.98	32
AP_Endometrium_Lung	0.94	27	0.95	32
AP_Endometrium_Omentum	0.92	14	0.9	32
AP_Endometrium_Ovary	0.91	12	0.92	32
AP_Endometrium_Prostate	0.98	20	0.99	4
AP_Endometrium_Uterus	0.9	14	0.76	256
AP_Lung_Kidney	0.96	7	0.96	32
AP_Lung_Uterus	0.93	22	0.93	32
AP_Omentum_Kidney	0.97	18	0.98	16
AP_Omentum_Lung	0.94	24	0.9	128
AP_Omentum_Ovary	0.98	27	0.76	4
AP_Omentum_Prostate	0.98	30	0.98	16
AP_Omentum_Uterus	0.91	15	0.88	16
AP_Ovary_Kidney	0.97	14	0.97	32
AP_Ovary_Lung	0.94	15	0.93	32
AP_Ovary_Uterus	0.88	21	0.89	64
AP_Prostate_Kidney	0.98	20	0.98	2
AP_Prostate_Lung	0.98	14	0.98	4
AP_Prostate_Ovary	0.98	19	0.98	2
AP_Prostate_Uterus	0.97	28	0.99	2
AP_Uterus_Kidney	0.96	12	0.97	32
Average	0.954	18.89	0.94	30.5
Standard deviation	0.02568		0.05357	
OVA_Breast	0.94	15	0.96	32
OVA_Colon	0.96	19	0.97	16
OVA_Endometrium	0.97	6	0.96	2
OVA_Kidney	0.98	20	0.98	8
OVA_Lung	0.97	24	0.97	4
OVA_Omentum	0.95	3	0.95	2
OVA_Ovary	0.92	10	0.93	32
OVA_Prostate	0.99	13	0.997	2
OVA_Uterus	0.97	21	0.93	32
Average	0.96	14.55	0.96	14.44
Standard deviation	0.02147		0.02198	

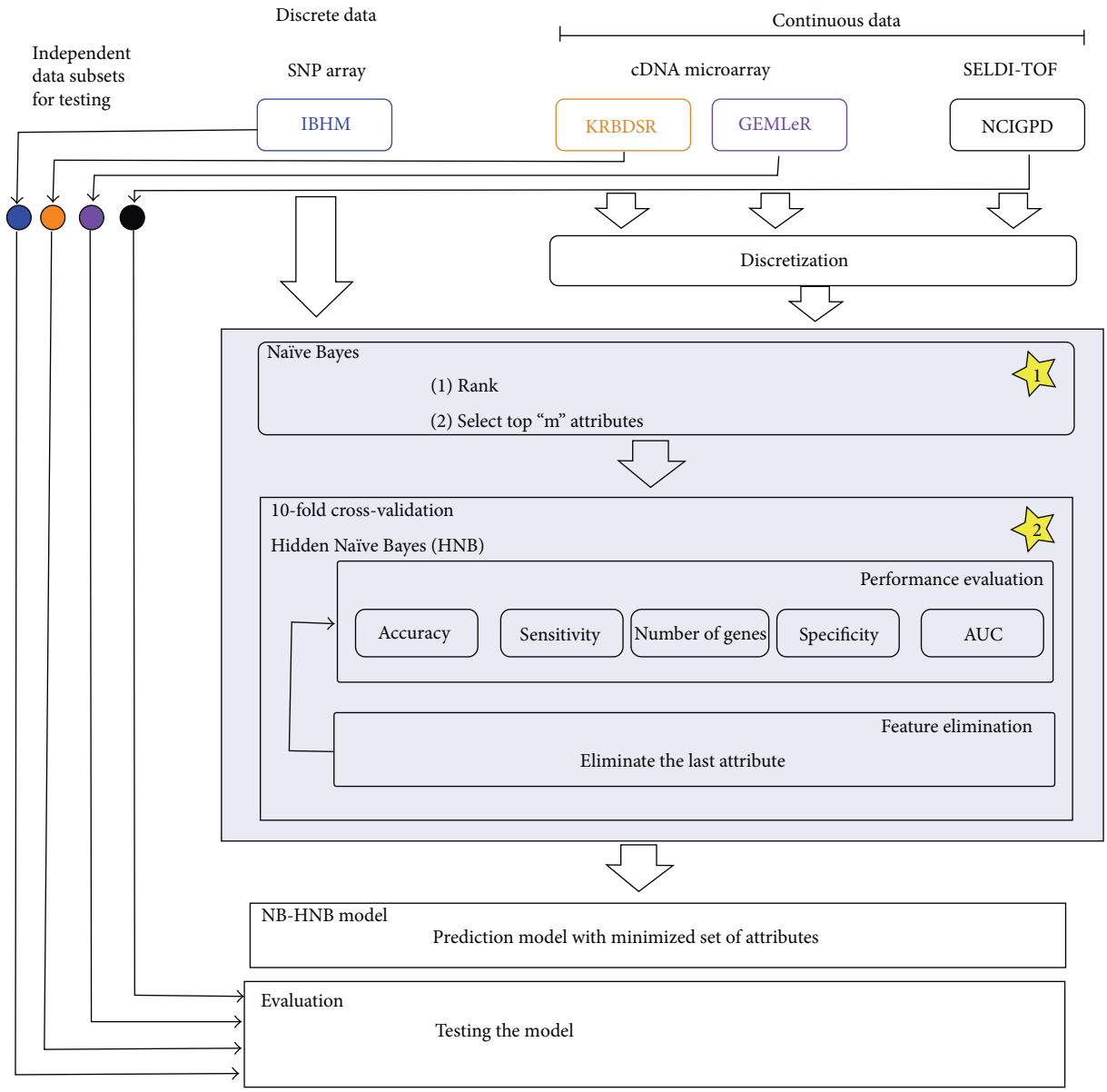


Figure 6: The overall two-step Bayes classification framework.

Table 5: Actual performance result on SNPs data (Bovine) from IBHM.

	Accuracy	Sensitivity	Specificity	Number of selected SNP
NB-HNB	0.92	0.92	0.99	33

Table 6: Actual performance result of NB-HNB from SELDI-TOF.

	Accuracy	Sensitivity	Specificity	Number of selected genes
Prostate	0.86	0.86	0.89	8
Ovarian	0.98	0.98	0.97	8

leukemia, and colon cancer datasets from KRBDSR. The criteria used to measure the performance of different methods include prediction accuracy, sensitivity, specificity, and the number of selected genes. We tested the proposed two-step Bayes classification framework against these datasets and augmented our performance in conjunction with the tables published in [15]. Tables 1, 2, and 3 show the results from our proposed framework (NB-HNB) in comparison with other methods. NB-HNB outperformed other machine learning methods in terms of prediction accuracy, sensitivity, and specificity. The greater marker requirement of NB-HNB indicates that the Naïve Bayes filtering probably did not rank the top dependent features that can optimally construct an accurate classification model in the correct order. Hence to

Table 7: Summary of the information about each dataset, for example, sample sizes, number of attributes.

SNP array		cDNA microarray				SELDI-TOF	
IBHM [3]		KRBDSR [4]		GEMLeR [5]		NCICPD [6]	
Data	Number of SNP (Number of samples)	Data	Number of genes (Number of samples)	Data	Number of genes (Number of samples)	Data	Number of genes (Number of samples)
Bovine	9239 (497)	Leukemia	7129 (72)	Colon	10935 (286)	Ovarian	15154 (253)
		Colon cancer	2000 (62)	Breast	10935 (344)	Prostate	15154 (266)
		Breast cancer	24481 (78)	Endometrium	10935 (61)		
		Lymphoma	4026 (47)	Kidney	10935 (260)		
		Prostate	12600 (102)	Lung	10935 (126)		
		Lung cancer	7129 (96)	Omentum	10935 (77)		
		Nervous	7129 (60)	Ovary	10935 (198)		
				Prostate	10935 (69)		
				Uterus	10935 (124)		

achieve 100% accuracy from the training set, HNB required more genes to classify.

Since the three datasets from KRBDSR are insufficient to demonstrate the performance of our two-step Bayes classification framework, we compared the NB-HNB framework against SVM-RFE using 45 microarray datasets from GEMLeR. The performance results were recorded in terms of (1) classification accuracy, (2) area under the ROC curve (AUC), (3) sensitivity, (4) specificity, and (5) the number of informative biomarkers used for classification. The comparison results of all experiments, including 36 all-possible pairs (AP) datasets and 9 one-tissue-type versus all-other-types (OVA) datasets, are shown in Table 4. In summary, NB-HNB outperformed SVM-RFE on most performance metrics. Figure 4 presents the average classification accuracy versus the number of selected genes. For all datasets, the accuracy of NB-HNB is better when the number of selected genes is larger than 16. A similar pattern is also observed when comparing AUC between the two approaches (Figure 5). Moreover, the accuracy and AUC do not vary much across different datasets since the standard deviations (Table 4) between NB-HNB and SVM-RFE are similar.

2.3. Experiments on Other Types of -Omics Datasets. We tested whether HNB could also be applied for class prediction from SNP genotyping and SELDI-TOF proteomics datasets. For the bovine dataset, NB-HNB was able to achieve 92% accuracy with 92% sensitivity and as high as 99% specificity using only 33 SNPs, as shown in Table 5. NB-HNB can also be applied to classify cancer proteomics data obtained from SELDI-TOF experiments. For prostate cancer, NB-HNB was able to reach 86% accuracy with 86% sensitivity and 89% specificity using only 8 protein markers. The performance is even better with ovarian cancer, in which NB-HNB demonstrated 98% accuracy at 98% sensitivity and 97% specificity using only 8 protein markers, as shown in Table 6.

3. Conclusions

The proposed two-step Bayes classification framework outperformed SVM-RFE in all previously reported experiments. Furthermore, we demonstrated that this two-step Bayes classification framework could address the biomarker selection and classification problem beyond the analysis of expression microarray data. Since the two-step Bayes classification framework utilizes Naïve Bayes filtering prior to HNB classification, the complexity of this classification framework is very low permitting analysis of data with many features.

4. Material and Methods

4.1. Datasets. The datasets used in the experiments comprise three groups: (1) genomic (2) transcriptomic, and (3) proteomic categories. The first category is SNP genotyping data obtained from the International Bovine HapMap (IBHM) [3] consortium containing 230 individual samples from 19 cattle breeds, each of which has 9,239 SNPs. For the transcriptomic datasets, microarray gene expression data were downloaded from two main repositories: the Gene Expression Machine Learning Repository (GEMLeR) [5] and the Kent Ridge Biomedical Data Set Repository (KRBDSR) [4]. GEMLeR contains microarray data from 9 different tissue types including colon, breast, endometrium, kidney, lung, omentum, ovary, prostate, and uterus. Each microarray sample is classified as tumor or normal. The data from this repository were collated into 36 possible pairings of

two tissue types, termed all-possible pairs (AP) datasets and 9 one-tissue-type versus all-other-types (OVA) datasets where the second class is labeled as "other." All GEMLeR microarray datasets have been analyzed by SVM-RFE, the results of which are available from the same resource. The datasets from KRBDSR contain 7 case-control microarray experiments (tumor versus normal). However, the SVM-RFE results are available only for five datasets from [8, 15, 17], namely, leukemia, colon cancer, breast cancer, lymphoma, and prostate cancer. Ovarian and prostate cancer SELDI-TOF proteomic datasets were obtained from the National Cancer Institute Clinical Proteomics Database (NCICPD) [6]. The information about each dataset, that is, sample size and number of features, is summarized in Table 7.

4.2. Methods. The two-step Bayes classification framework is composed of two modules: Naïve Bayes (NB) filtering and Hidden Naïve Bayes (HNB) classification. Figure 6 shows the overall two-step Bayes classification framework. For continuous signal data (e.g., cDNA expression microarray), the data must first be preprocessed by (feature) discretization [20]. The process simply involves processing the data into a series of bins according to the range of values in the dataset. Ten bins were used to group the continuous microarray data by loosely setting each interval (bin) to have the same range. This was done using the Weka discretize module with the following settings: $-B = 10$ and $-M = -1.0$ where $-B$ specifies the number of bins and $-M$ indicates the weight of instances per interval to create bins of equal interval size. For evaluation of the performance of the NB-HNB model, an independent test dataset is required. We obtained test dataset by randomly selecting 10% of the data from the original dataset that is reserved as a blind dataset (i.e., the data that are analyzed only once using the final classification model) while the rest are used as training data for feature selection and the model classification.

The number of m top-ranked features for NB filtering is selected by the user, who inputs the cutoff for individual marker prediction accuracy. From our empirical studies, the top-ranked 40 features provide 75% or greater prediction accuracy. Therefore, we chose this cutoff as the number of markers which can be practically used for HNB processing on a typical desktop computer containing 4 GB RAM with multicore architecture. Obviously with greater computing power, more features could be chosen for higher accuracy. From the NB filtered list of features, an HNB classification model is constructed. The lowest-ranked feature is then removed and another HNB classifier model constructed, which is compared with the previous model for classification accuracy. The process of model building and feature elimination is repeated until the minimum feature subset is obtained which gives a classifier model with the maximum prediction accuracy.

Intuitively, NB filtering operates by constructing a density estimator using standard Naïve Bayes. The class c of sample E with attributes can be classified by

$$c(E) = \arg\max_{c \in C} P(c)\, P(a_1, a_2, \ldots, a_n \mid c). \quad (1)$$

Naïve Bayes assumes that all attributes are independent for a given class. We can then simply represent the above equation by

$$c(E) = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(a_i \mid c). \quad (2)$$

The filtering step is performed to quickly extract all the informative features. The ranking is done by sorting the value $P(c)P(a_i \mid c)$, which can be run very quickly by simply counting the number of feature occurrences in each of the corresponding classes; the time complexity is thus $O(n)$. This, however, does not guarantee that the top-ranked features will contain the optimal set of features that will give the most accurate classification model. The more realistic approach would be to consider all possible dependencies amongst features. However, it has been known that building an optimal Bayesian network classifier is NP hard. To overcome this limitation, we proposed that Hidden Naïve Bayes (HNB) should be used to construct the more realistic classification model from the set of NB filtered attributes.

Instead of building a complete Bayesian graph, which is intractable, HNB is used to construct the dependencies between attributes A_i with a hidden parent A_{hp_i}. The modification with the dependency from the hidden parent makes HNB become more realistic by adjusting the weight influenced by all other attributes. A classifier of a sample E with attributes $[a_1, a_2, \ldots, a_n]$ can be represented by

$$c(E) = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P\left(a_i \mid a_{hp_i}, c\right), \quad (3)$$

where

$$P\left(a_i \mid a_{hp_i}, c\right) = \sum_{j=1, j \neq i}^{n} W_{ij} \times P\left(a_i \mid a_j, c\right),$$
$$W_{ij} = \frac{I_P\left(A_i; A_j \mid C\right)}{\sum_{j=1, j \neq i}^{n} I_P\left(A_i; A_j \mid C\right)}, \quad (4)$$

where the conditional mutual information $I_P(A_i; A_j \mid C)$ can be computed as

$$I_P\left(A_i; A_j \mid C\right) = \sum_{a_i, a_j, c} P\left(a_i, a_j, c\right) \log\left(\frac{P\left(a_i, a_j \mid c\right)}{P\left(a_i \mid c\right) P\left(a_j \mid c\right)}\right). \quad (5)$$

Authors' Contribution

Supakit Prueksaaroon, Philip James Shaw, Taneth Ruangrajitpakorn, and Sissades Tongsima wrote the paper. Anunchai Assawamakin, Supasak Kulawonganunchai, and Supakit Prueksaaroon designed and conducted the experiments presented in this work. Anunchai Assawamakin, Supasak Kulawonganunchai, Vara Varavithya, and Sissades Tongsima conceived the initial idea. Supakit Prueksaaroon, Anunchai Assawamakin, and Sissades Tongsima analyzed the results.

Conflicts of Interests

The authors declare no conflict of interests.

Acknowledgments

Supakit Prueksaaroon would like to acknowledge the National Electronics and Computer Technology Center (NECTEC) and Faculty of Engineering, Thammasat University for partially supporting this work during his tenure at NECTEC, while the completion of this work was done at Thammasat University. Sissades Tongsima received the support from the National Center for Genetic Engineering and Biotechnology (BIOTEC) platform technology, The Research Chair Grant 2011 from the National Science and Technology Development Agency (NSTDA), and Thailand and TRF Career Development Grant RSA5480026. Furthermore, Sissades Tongsima acknowledges the computing infrastructure supported by the National Infrastructure program (C2-14) under National Science and Technology Development Agency (NSTDA). Finally, the authors would like to thank donors who participated in this study. Anunchai Assawamakin and Supakit Prueksaaroon are co-first authors.

References

[1] Z.-Z. Hu, H. Huang, C. H. Wu et al., "Omics-based molecular target and biomarker identification," *Methods in Molecular Biology*, vol. 719, pp. 547–571, 2011.

[2] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2009.

[3] The Bovine HapMap Consortium, "Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds," *Science*, vol. 324, no. 5926, pp. 528–532, 2009.

[4] Kent Ridge Biomedical Data Set Repository, http://datam.i2r.a-star.edu.sg/datasets/krbd/.

[5] Gene Expression Machine Learning Repository (GEMLeR), http://gemler.fzv.uni-mb.si/download.php.

[6] National Cancer Institute Clinical Proteomics Database (NCI-CPD), http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp.

[7] Y. Leung and Y. Hung, "A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 1, pp. 108–117, 2010.

[8] A. Mohammadi, M. H. Saraee, and M. Salehi, "Identification of disease-causing genes using microarray data mining and gene ontology," *BMC Medical Genomics*, vol. 4, article 12, supplement 2, 2011.

[9] Y. Ding and D. Wilkins, "Improving the performance of SVM-RFE to select genes in microarray data," *BMC Bioinformatics*, vol. 7, no. 2, article S12, 2006.

[10] S. Balakrishnan, R. Narayanaswamy, N. Savarimuthu, and R. Samikannu, "Svm ranking with backward search for feature selection in type II diabetes databases," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '08)*, pp. 2628–2633, October 2008.

[11] R. Blanco, I. Inza, M. Merino, J. Quiroga, and P. Larrañaga, "Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS," *Journal of Biomedical Informatics*, vol. 38, no. 5, pp. 376–388, 2005.

[12] G. I. Webb, J. R. Boughton, and Z. Wang, "Not so naive Bayes: aggregating one-dependence estimators," *Machine Learning*, vol. 58, no. 1, pp. 5–24, 2005.

[13] L. Jiang, H. Zhang, and Z. Cai, "A novel bayes model: hidden naive bayes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 10, pp. 1361–1371, 2009.

[14] http://www.cs.waikato.ac.nz/ml/weka/.

[15] M. Zervakis, M. E. Blazadonakis, G. Tsiliki, V. Danilatou, M. Tsiknakis, and D. Kafetzopoulos, "Outcome prediction based on microarray analysis: a critical perspective on methods," *BMC Bioinformatics*, vol. 10, article 53, 2009.

[16] G. Stiglic, J. J. Rodriguez, and P. Kokol, "Finding optimal classifiers for small feature sets in genomics and proteomics," *Neurocomputing*, vol. 73, no. 13–15, pp. 2346–2352, 2010.

[17] K. Duan and J. C. Rajapakse, "SVM-RFE peak selection for cancer classification with mass spectrometry data," in *Advances in Bioinformatics and Computational Biology*, P. Chen and L. Wong, Eds., pp. 191–200, Imperial College Press, London, UK, 2005.