

รหัสโครงการ: RSA5480026

ชื่อโครงการ: การพัฒนาระเบียบวิธีการเรียนรู้ด้วยเครื่องเพื่อใช้ในการเลือกเครื่องหมายทางพันธุกรรม
ซึ่งบ่งชี้ลักษณะจำเพาะที่สนใจจากข้อมูลสnpทั่วจีโนม

ชื่อนักวิจัย และสถาบัน: นายศิษฏ์ ทองสิมา ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ

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

ระยะเวลาโครงการ: 3 ปี

เครื่องหมายทางพันธุกรรมแบบสnp คือ ความหลากหลายทางพันธุกรรมที่เป็นตัวกำหนดความแตกต่างกันในสิ่งมีชีวิตแต่ละชนิด งานวิจัยมากมายมีการใช้สnpเป็นเครื่องมือในการตรวจสอบความสัมพันธ์ไปกับการเกิดโรคที่ซับซ้อน การใช้วิธีการทางสถิติเพื่อค้นหาสnpที่มีความสัมพันธ์กับการเกิดโรคเหล่านั้นไม่สามารถบรรลุข้อประสงค์ได้กับทุกๆกลุ่มประชากร เนื่องจากปัญหาของการมีอยู่ของโครงสร้างประชากร(population structure) ในข้อมูลที่จะเป็นตัวลดสัญญาณของนัยสำคัญทางสถิติได้

คณะผู้วิจัยได้พัฒนาอัลกอริธึมที่ชื่อ i2pPCA ที่ใช้กลไกการหยุดชื่อว่า "EigenDev" เพื่อควบคุมการทำซ้ำของอัลกอริธึม ซึ่งมีความสามารถในการตรวจพบและกำจัดโครงสร้างประชากรในข้อมูลสnpได้ เราได้ประยุกต์ใช้อัลกอริธึม i2pPCA กับข้อมูลสnpของคนไทยที่นำไปรวมกับข้อมูลสnpของประชากรอื่นจากทั่วโลกอีกจำนวน 40 ประชากร อัลกอริธึมนี้สามารถตรวจหาโครงสร้างประชากร และแยกกลุ่มตามโครงสร้าง ทำได้สามารถบอกได้ว่าคนไทยประกอบไปด้วย 4 กลุ่มย่อย ซึ่งเป็นข้อบ่งชี้ว่าในการศึกษาความสัมพันธ์ไปกับการเกิดโรคในอนาคต จำเป็นจะต้องคำนึงถึงการมีอยู่ของโครงสร้างประชากรในคนไทยด้วย นอกจากนี้คณะผู้วิจัยยังได้พัฒนาอัลกอริธึมที่ชื่อ iLOCi สำหรับใช้ค้นหาปฏิสัมพันธ์ระหว่างสnpที่สัมพันธ์กับการเกิดโรคที่ซับซ้อน ซึ่ง iLOCi สามารถค้นหาสnpเหล่านั้นเจอเมื่อประยุกต์ใช้กับข้อมูล 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 : สnp, 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 abnormalities 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