

การแทนยีนโดยสลิปตัวแทนไม่มีผลต่อการวิเคราะห์การได้มากขึ้นจากเซตของยีนในบาทวิถี

การให้สัญญาณจากฐานข้อมูล KEGG ในการศึกษาความสัมพันธ์ทั้งจีโนม

Representing Genes by Tag SNPs Has No Effects on Gene Set Enrichment Analysis of

KEGG Signalling Pathways in Genome-Wide Association Studies

เจษฎา วีระเดชกำพล (Jessada Weeradetkumpon)* ดร.ประภาส จงสถิตย์วัฒนา (Dr.Prabhas Chongstitvatana)**

ดร.ณชล ไชยรัตน์ (Dr.Nachol Chaiyaratana)***

บทคัดย่อ

บทความนี้นำเสนอการเปรียบเทียบระหว่างการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสลิปทั้งหมดและข้อมูลสลิปตัวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนม ชุดการวัดเปรียบเทียบสมรรถนะได้สร้างจากเจ็ดเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมจากการศึกษาความสัมพันธ์ทั้งจีโนมของเจ็ดโรคซับซ้อนโดย Wellcome Trust Case Control Consortium หนึ่งในสลิปได้รับการคัดเลือกสำหรับใช้เป็นตัวแทนยีนโดยการหาค่าสูงสุดของค่าสถิติทดสอบแนวนอนเมเยียงคอราน-อาร์มิตาจเป็นเงื่อนไขการคัดเลือก ถึงแม้ว่ามีการคำนวณค่าสถิติทดสอบสำหรับแต่ละสลิป ค่าสถิติทดสอบสำหรับสลิปตัวแทนซึ่งได้รับการคัดเลือกโดยใช้ Tagger จะใช้เป็นค่าสถิติทดสอบสำหรับสลิปที่มีตัวแทนด้วย การวิเคราะห์บาทวิถีกระทำโดยใช้ GSEA-SNP ซึ่งเป็นเทคนิคที่ได้รับการพัฒนาต่อจากเทคนิคการวิเคราะห์การได้มากขึ้นจากเซตของยีนหรือ GSEA บาทวิถีการให้สัญญาณจาก Kyoto Encyclopedia of Genes and Genomes (KEGG) เป็นเป้าหมายสำหรับการตรวจจับความสัมพันธ์ โดยรวมการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสลิปทั้งหมดให้ผลการวิเคราะห์ไม่แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสลิปตัวแทน

ABSTRACT

This article presents a comparison between pathway analysis of all single nucleotide polymorphisms (SNPs) and tag SNPs from genome-wide association studies. Seven case-control datasets from genome-wide association studies of seven complex diseases investigated by the Wellcome Trust Case Control Consortium were used to form benchmark suites. A SNP was chosen to represent each gene where the chosen criterion was based on the maximisation of Cochran-Armitage trend test statistics. Although Cochran-Armitage trend tests were performed on all SNPs, the test statistics of tag SNPs selected by Tagger were also assigned to their tagged SNPs. GSEA-SNP, which is an extension of gene set enrichment analysis (GSEA), was the chosen pathway analysis technique. Signalling pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) were the targets for association detection. Overall, the pathway analyses of all SNPs were similar to those of tag SNPs.

คำสำคัญ: การวิเคราะห์บาทวิถี การศึกษาความสัมพันธ์ทั้งจีโนม สลิปตัวแทน

Keywords: Pathway analysis, Genome-wide association study, Tag SNP

*นิสิต หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

**ศาสตราจารย์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

***ศาสตราจารย์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

บทนำ

การศึกษาความสัมพันธ์ทั้งจีโนม (Genome-Wide Association Study หรือ GWAS) เป็นหนึ่งในตัวเลือกสำหรับการศึกษาความสัมพันธ์ทางพันธุกรรม (Genetic Association Study) ของโรคซับซ้อน (Complex Disease) ในช่วงทศวรรษที่ผ่านมา (Tam et al., 2019) เนื่องจากค่าใช้จ่ายการเก็บข้อมูลจีโนไทป์ (Genotype) ลดลงและความสามารถในการคำนวณโดยใช้คอมพิวเตอร์เพิ่มขึ้น ด้วยเหตุนี้จึงมีความเป็นไปได้ที่จะค้นพบยีน (Gene) ที่สัมพันธ์กับโรคซับซ้อนมากขึ้น อย่างไรก็ตาม การศึกษาความสัมพันธ์ทั้งจีโนมต้องพิจารณาข้อมูลสลับ (Single Nucleotide Polymorphism หรือ SNP) จำนวนมากและใช้การวิเคราะห์ที่แม่นยำ

ข้อมูลสลับจาก International HapMap Project (The International HapMap Consortium, 2005) ทำให้การออกแบบสลับชิป (SNP Chip) สำหรับการศึกษาความสัมพันธ์ทั้งจีโนมเป็นไปได้ การออกแบบสลับชิปสามารถแบ่งเป็นสองวิธี ได้แก่ การออกแบบสลับชิปโดยอาศัยสลับทั้งหมดและการออกแบบสลับชิปโดยอาศัยสลับตัวแทน (Tag SNP) การออกแบบสลับชิปโดยอาศัยสลับทั้งหมดใช้คุณภาพการเก็บข้อมูลจีโนไทป์ในการคัดเลือกสลับชิป ส่งผลให้ข้อมูลสลับชิปที่ได้จากสลับชิปมีลักษณะกระจายในจีโนม (Genome) อย่างสุ่ม ตัวอย่างของสลับชิปที่ได้รับการออกแบบด้วยวิธีนี้คือสลับชิปขนาด 111,000 และ 500,000 สลับของ Affymetrix ในทางตรงกันข้าม การออกแบบสลับชิปโดยอาศัยสลับตัวแทนสนใจเฉพาะสลับชิปตัวแทนซึ่งเป็นสลับชิปที่มีสหสัมพันธ์ (Correlation) หรือความไม่สมดุลการเชื่อมโยง (Linkage Disequilibrium) กับสลับชิปที่มีตัวแทน (Tagged SNP) ส่งผลให้ข้อมูลสลับชิปที่ได้จากสลับชิปที่มีสหสัมพันธ์กับข้อมูลสลับชิปที่ไม่ได้จากสลับชิป ตัวอย่างของสลับชิปที่ได้รับการออกแบบด้วยวิธีนี้คือสลับชิปขนาด 317,000 และ 555,000 สลับของ Illumina (Wallace et al., 2007)

การวิเคราะห์ข้อมูลสลับชิปจากการศึกษาความสัมพันธ์ทั้งจีโนมสามารถกระทำโดยการวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้ง (Single-Locus Analysis) และการวิเคราะห์ครั้งละหลายตำแหน่งที่ตั้ง (Multi-locus Analysis) (Heidema et al., 2006; Lewis, 2002; Montana, 2006) การวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้งเป็นการวิเคราะห์ที่ไม่ซับซ้อนและผลการวิเคราะห์ที่ได้ตีความง่าย อย่างไรก็ตาม การวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้งเหมาะสมสำหรับกรณีซึ่งสลับชิปที่สัมพันธ์กับโรคซับซ้อนมีผลหลัก (Main Effect) หรือผลหนึ่งตำแหน่งที่ตั้งแบบขอบ (Marginal Single-Locus Effect) เท่านั้น ข้อจำกัดดังกล่าวทำให้มีโอกาสการไม่ตรวจจับบางสลับชิปที่สัมพันธ์กับโรคซับซ้อน ในทางตรงกันข้าม การวิเคราะห์ครั้งละหลายตำแหน่งที่ตั้งไม่มีข้อจำกัดดังกล่าวซึ่งส่งผลให้การวิเคราะห์ครั้งละหลายตำแหน่งที่ตั้งสามารถตรวจจับอิพิสเตซิส (Epistasis) (Van Steen, 2012; Wongseree et al., 2009) ความต่างแบบกันทางพันธุกรรม (Genetic Heterogeneity) (Setsirichok et al., 2013) และผลแฮปโลไทป์ (Haplotype Effect) (Epstein, Satten, 2003) อย่างไรก็ตาม การวิเคราะห์ครั้งละหลายตำแหน่งที่ตั้งต้องใช้ทรัพยากรในการคำนวณมากกว่าการวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้ง และผลการวิเคราะห์ที่ได้ตีความยากกว่าผลการวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้ง

การคัดเลือกสลับชิปที่สัมพันธ์กับโรคซับซ้อนโดยการวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้งและหลายตำแหน่งที่ตั้งสามารถพิจารณาเป็นการคัดเลือกลักษณะประจำ (Attribute Selection) หรือการคัดเลือกตัวแปร (Variable Selection) จากมุมมองการรู้จำแบบ (Pattern Recognition) (Saeys et al., 2007) นอกจากการคัดเลือกสลับชิปที่สัมพันธ์กับโรคซับซ้อนโดยตรงแล้ว การวิเคราะห์บาทวิถี (Pathway Analysis) (Wang et al., 2010) เป็นอีกการวิเคราะห์ซึ่งได้รับความสนใจในการศึกษาความสัมพันธ์ทั้งจีโนม การวิเคราะห์บาทวิถีใช้การจัดกลุ่มสลับชิปสำหรับใช้เป็นตัวแทนยีนตามบาทวิถีชีวภาพ (Biological Pathway) และมีเป้าหมายคือการตรวจจับบาทวิถีชีวภาพที่สัมพันธ์กับโรคซับซ้อน ดังนั้นการวิเคราะห์บาทวิถีจึงสามารถพิจารณาเป็นการวิเคราะห์ครั้งละหลายตำแหน่งที่ตั้งซึ่งสนใจกลุ่มเฉพาะของสลับชิปสำหรับใช้เป็นตัวแทนยีนเท่านั้นและเป็นการคัดเลือกลักษณะประจำเช่นกัน

หลายเทคนิคการวิเคราะห์บาทวิถีสำหรับการศึกษาความสัมพันธ์ทั้งจีโนมได้รับการพัฒนาจากเทคนิคการวิเคราะห์บาทวิถีสำหรับการวิเคราะห์การแสดงออกของยีน (Gene Expression Analysis) (Wang et al., 2010) GSEA-SNP เป็นหนึ่งในเทคนิคดังกล่าว (Holden et al., 2008) โดย GSEA-SNP ได้รับการพัฒนาจากเทคนิคการวิเคราะห์การได้มากขึ้นจากเซตของยีน (Gene Set Enrichment Analysis หรือ GSEA) (Subramanian et al., 2005) ตามปกติแล้ว ข้อมูลการแสดงออกของยีนที่ได้จากหนึ่งเซตของโพรบ (Probeset) พอเพียงสำหรับการใช้เป็นตัวแทนหนึ่งยีนในการวิเคราะห์โดยใช้ GSEA ดังนั้นข้อมูลสนิปที่ได้จากหนึ่งสนิปจึงเพียงพอสำหรับการใช้เป็นตัวแทนหนึ่งยีนในการวิเคราะห์โดยใช้ GSEA-SNP

ถึงแม้ว่าการศึกษาความสัมพันธ์ทั้งจีโนมต้องพิจารณาข้อมูลสนิปจำนวนมาก แต่การวิเคราะห์บาทวิถีจำเป็นต้องใช้ข้อมูลหนึ่งสนิปที่อยู่เหนือหรือใกล้ยีนเพื่อใช้เป็นตัวแทนแต่ละยีนเท่านั้น ส่งผลให้มีข้อมูลสนิปจำนวนมากที่ไม่ได้ใช้ในการวิเคราะห์บาทวิถี ดังนั้นจึงมีความเป็นไปได้ว่า การใช้ข้อมูลสนิปตัวแทนซึ่งได้รับการคัดเลือกจากสนิปทั้งหมดในการศึกษาความสัมพันธ์ทั้งจีโนมพอเพียงสำหรับการวิเคราะห์บาทวิถี นั่นคือการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนมให้ผลการวิเคราะห์ที่ไม่แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปตัวแทนเท่านั้น ภายใต้เงื่อนไขการมีอยู่ของข้อมูลความไม่สมดุลการเชื่อมโยงระหว่างสนิปตัวแทนและสนิปที่มีตัวแทน การทดสอบแนวคิดนี้เป็นประโยชน์ต่อการวิเคราะห์ข้อมูลสนิปจากการศึกษาความสัมพันธ์ทั้งจีโนมในฐานะข้อมูลสาธารณะ เช่น Database of Genotypes and Phenotypes (dbGaP) (National Center for Biotechnology Information, 2021) โดยเฉพาะเมื่อใช้สนิปชิปขนาด 317,000 และ 555,000 สนิปของ Illumina ในการเก็บข้อมูลจีโนมไทป์ เนื่องจากข้อมูลสนิปที่ได้จากสนิปชิปของ Illumina มีสหสัมพันธ์กับข้อมูลสนิปที่ไม่ได้จากสนิปชิปดังกล่าวข้างต้น

บทความนี้สนใจการเปรียบเทียบระหว่างการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดและข้อมูลสนิปตัวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนม ข้อมูลที่ใช้ในการเปรียบเทียบคือข้อมูลจากการศึกษาความสัมพันธ์ทั้งจีโนมโดย Wellcome Trust Case Control Consortium (WTCCC) ซึ่งการเก็บข้อมูลจีโนมไทป์ใช้สนิปชิปขนาด 500,000 สนิปของ Affymetrix (The Wellcome Trust Case Control Consortium, 2007) ส่งผลให้ได้ข้อมูลสนิปซึ่งมีลักษณะกระจายในจีโนมอย่างสุ่ม ดังนั้นการคัดเลือกสนิปตัวแทนจากสนิปทั้งหมดซึ่งกระทำโดยใช้ Tagger (de Bakker et al., 2005) จึงมีลักษณะไม่แตกต่างจากการคัดเลือกสนิปสำหรับการออกแบบสนิปชิปของ Illumina การวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดและข้อมูลสนิปตัวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนมกระทำโดยใช้ GSEA-SNP บาทวิถี (Pathway) ที่สนใจคือบาทวิถีการให้สัญญาณ (Signalling Pathway) จาก Kyoto Encyclopedia of Genes and Genomes (KEGG) (KEGG: Kyoto Encyclopedia of Genes and Genomes [KEGG], 2021)

วัตถุประสงค์การวิจัย

บทความนี้มีวัตถุประสงค์เพื่อทดสอบว่า การวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนมให้ผลการวิเคราะห์แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปตัวแทนเท่านั้นหรือไม่

วิธีการวิจัย

เซตข้อมูลและการจัดการข้อมูล

เซตข้อมูลที่ใช้คือเจ็ดเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุม (Case-Control Dataset) จากการศึกษารวมสัมพันธ์ทั้งจีโนมของเจ็ดโรคซับซ้อนโดย WTCCC แต่ละเซตข้อมูลประกอบด้วยตัวอย่างกลุ่มกรณี (Case Sample) จากบุคคลเป็นโรค (Affected Individual) ในสหราชอาณาจักรซึ่งเป็นหนึ่งในเจ็ดโรคซับซ้อน ได้แก่ โรคอารมณ์สองขั้ว (Bipolar Disorder หรือ BD) โรคหลอดเลือดแดงโคโรนารี (Coronary Artery Disease หรือ CAD) โรคโครห์น (Crohn's Disease หรือ CD) ความดัน

ตารางที่ 1 จำนวนตัวอย่างของกลุ่มกรณีและกลุ่มควบคุม

ชื่อข้อมูล	NBS	58C	BD	CAD	CD	HT	RA	T1D	T2D
จำนวนตัวอย่าง	1,458	1,480	1,868	1,962	1,748	1,952	1,860	1,963	1,924

เลือดสูง (Hypertension หรือ HT) โรคข้ออักเสบรูมาตอยด์ (Rheumatoid Arthritis หรือ RA) เบาหวานชนิดที่ 1 (Type 1 Diabetes หรือ T1D) และเบาหวานชนิดที่ 2 (Type 2 Diabetes หรือ T2D) นอกจากนี้ แต่ละเซตข้อมูลประกอบด้วยตัวอย่างกลุ่มควบคุม (Control Sample) จากบุคคลไม่เป็นโรค (Unaffected Individual) ตัวอย่างกลุ่มควบคุมประกอบด้วยตัวอย่างจากหน่วยบริการเลือดสหราชอาณาจักร (UK Blood Services หรือ NBS) และตัวอย่างจากบุคคลที่เกิดในสหราชอาณาจักรในปี ค.ศ. 1958 (1958 British Birth Cohort หรือ 58C) จำนวนตัวอย่างของทั้งสองกลุ่มได้สรุปในตารางที่ 1

ทุกข้อมูลมี 469,612 สนิป การเก็บข้อมูลจีโนมโทป์ใช้สนิปชิป Affymetrix GeneChip Human Mapping 500K Array Set ข้อมูลจีโนมโทป์ผ่านการควบคุมคุณภาพโดย WTCCC (The Wellcome Trust Case Control Consortium, 2007) บทความนี้สนใจเฉพาะสนิปซึ่งมีค่าความถี่ส่วนน้อย (Minor Allele Frequency หรือ MAF) ในตัวอย่างกลุ่มควบคุมมากกว่าหรือเท่ากับ 0.05 และสามารถระบุตำแหน่งในจีโนมได้เท่านั้น ซึ่งส่งผลให้สามารถคำนวณค่า r^2 (Hill, Robertson, 1968) สำหรับการอธิบายความไม่สมดุลการเชื่อมโยงระหว่างคู่สนิปอย่างมีความเชื่อถือได้ หลังจากกำจัดสนิปในข้อมูลซึ่งไม่สอดคล้องกับเงื่อนไขแล้วเหลือสนิปสำหรับการทดลองทั้งหมด 367,623 สนิป

การคัดเลือกสนิปตัวแทนโดยใช้ Tagger

Tagger เป็นโปรแกรมสำหรับการคัดเลือกสนิปตัวแทนโดยไม่ใช้บล็อกของแฮปโลไทป์ (Haplotype Block-Free Approach) (de Bakker et al., 2005) Tagger สามารถคัดเลือกสนิปตัวแทนโดยการพิจารณาความสัมพันธ์ระหว่างอัลลีล (Allele) ของคู่สนิป การคัดเลือกสนิปตัวแทนใช้ขั้นตอนวิธีละโมบ (Greedy Algorithm) ซึ่งอาศัยค่า r^2 สำหรับการอธิบายความไม่สมดุลการเชื่อมโยงระหว่างคู่สนิป ขั้นตอนวิธีละโมบเริ่มต้นด้วยการระบุสนิปตัวแทนซึ่งมีจำนวนสนิปเชื่อมโยง (Linked SNP) กับสนิปดังกล่าวสูงสุดโดยอิงจากขีดเริ่มเปลี่ยน r^2 (r^2 Threshold) สนิปตัวแทนนี้และสนิปเชื่อมโยงของสนิปตัวแทนนี้จะได้รับการรวมไว้ในหนึ่งผลแบ่งกัน (Partition) ถ้ามีสนิปอื่นในผลแบ่งกันซึ่งเชื่อมโยงกับสนิปที่เหลือในผลแบ่งกัน แล้วสนิปนี้จะเป็นสนิปตัวแทนเช่นกัน อย่างไรก็ตาม หนึ่งสนิปตัวแทนเพียงพอสำหรับหนึ่งผลแบ่งกัน จากนั้นขั้นตอนวิธีละโมบจะระบุสนิปตัวแทนจากสนิปที่เหลือในลักษณะเดียวกัน ถ้ามีสนิปซึ่งไม่เชื่อมโยงกับสนิปอื่น แล้วสนิปนี้จะเป็นสนิปตัวแทนซึ่งอยู่ในผลแบ่งกันของตัวเอง (Carlson et al., 2004) ในบทความนี้ ระยะทางสูงสุดระหว่างสนิปสำหรับการคำนวณค่า r^2 คือ 500 กิโลเบส (Kilobase)

การทดสอบแนวโน้มเอียงคอคราน-อาร์มีเทจ

การทดสอบแนวโน้มเอียงคอคราน-อาร์มีเทจ (Cochran-Armitage Trend Test หรือ CA Trend Test) เป็นหนึ่งในการทดสอบเชิงสถิติซึ่งได้รับความนิยมมากที่สุดในการศึกษาความสัมพันธ์ทางพันธุกรรม (Sasieni, 1997) พิจารณาเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมซึ่งมีจำนวนตัวอย่างตามจีโนมโทป์ที่สนิปดังแสดงในตารางที่ 2 ค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มีเทจ (Cochran-Armitage Trend Test Statistic) หรือ T_{CA} สามารถนิยามโดย

$$T_{CA} = \frac{N}{R(N-R)} \frac{(N \sum_{i=0}^2 r_i x_i - R \sum_{i=0}^2 n_i x_i)^2}{N \sum_{i=0}^2 n_i x_i^2 - (\sum_{i=0}^2 n_i x_i)^2} \quad (1)$$

โดยที่ x_i เป็นตัวถ่วงน้ำหนักสำหรับจีโนมโทป์ i สามแบบจำลองทางพันธุกรรม (Genetic Model) ที่สนใจในบทความนี้ได้แก่แบบจำลองลักษณะบวก (Additive Model) แบบจำลองลักษณะเด่น (Dominant Model) และแบบจำลองลักษณะด้อย (Recessive Model) การทดสอบแนวโน้มเอียงคอคราน-อาร์มีเทจใช้ตัวถ่วงน้ำหนัก $x_0 = 0$, $x_1 = 1$ และ $x_2 = 2$ สำหรับ

ตารางที่ 2 การแจกแจงตัวอย่างกลุ่มกรณีและตัวอย่างกลุ่มควบคุมในเซตข้อมูลตามจีโนไทป์ที่สนิป

สถานะ	จีโนไทป์ที่สนิป			จำนวนตัวอย่าง
	จีโนไทป์ของ	จีโนไทป์ของ	จีโนไทป์ของ	
	พันธุ์ป่าโฮโมไซโกต (Homozygous Wild-Type Genotype)	เฮเทอโรไซโกต (Heterozygous Genotype)	พันธุ์กลายโฮโมไซโกต (Homozygous Variant Genotype)	
กลุ่มกรณี	r_0	r_1	r_2	R
กลุ่มควบคุม	s_0	s_1	s_2	S
ทั้งหมด	n_0	n_1	n_2	N

การทดสอบผลลักษณะบวก (Additive Effect) ตัวถ่วงน้ำหนัก $x_0 = 0$, $x_1 = 1$ และ $x_2 = 1$ สำหรับการทดสอบผลลักษณะเด่น (Dominant Effect) และตัวถ่วงน้ำหนัก $x_0 = 0$, $x_1 = 0$ และ $x_2 = 1$ สำหรับการทดสอบผลลักษณะด้อย (Recessive Effect) ค่าสถิติทดสอบแนวมัยเอียงคอคราน-อาร์มีเทจเป็นไปตามการแจกแจงไคกำลังสอง (χ^2 Distribution) ซึ่งมีหนึ่งระดับขั้นความเสรี (Degree of Freedom)

การคัดเลือกสนิปสำหรับใช้เป็นตัวแทนยีน

ตามปกติแล้ว มีหลายสนิปที่อยู่ในหรือใกล้ยีน Wang et al. (2007) แนะนำว่า สนิปซึ่งมีค่าสถิติทดสอบสุดขีด (Extreme) ที่สุดเมื่อเทียบกับสนิปที่อยู่ในหรือใกล้ยีนเดียวกันสามารถใช้เป็นตัวแทนยีนในการศึกษากรณีสวน-กรณีสวน (Case-Control Study) ค่าสถิติทดสอบแนวมัยเอียงคอคราน-อาร์มีเทจคือค่าสถิติทดสอบที่ใช้ในบทความนี้ สามค่าสถิติทดสอบแนวมัยเอียงคอคราน-อาร์มีเทจสำหรับการทดสอบผลลักษณะบวก ผลลักษณะเด่น และผลลักษณะด้อยจะได้รับการคำนวณสำหรับแต่ละสนิปในเซตข้อมูลกรณีสวน-กรณีสวน แบบจำลองทางพันธุกรรมที่ได้รับการเลือกสำหรับแต่ละสนิปคือแบบจำลองทางพันธุกรรมที่ให้ค่าสถิติทดสอบแนวมัยเอียงคอคราน-อาร์มีเทจสูงสุด กรณีที่สนใจข้อมูลสนิปทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนม ค่าสถิติทดสอบแนวมัยเอียงคอคราน-อาร์มีเทจสำหรับแต่ละสนิปต้องได้รับการคำนวณในทางตรงกันข้าม กรณีที่สนใจข้อมูลสนิปตัวแทนเท่านั้น ค่าสถิติทดสอบแนวมัยเอียงคอคราน-อาร์มีเทจสำหรับสนิปที่มีตัวแทนซึ่งเชื่อมโยงกับสนิปตัวแทนและเปรียบเสมือนสนิปที่ไม่ได้จากสนิปซิปจะเท่ากับค่าสถิติทดสอบแนวมัยเอียงคอคราน-อาร์มีเทจสำหรับสนิปตัวแทน สนิปสำหรับใช้เป็นตัวแทนยีนคือสนิปซึ่งมีค่าสถิติทดสอบแนวมัยเอียงคอคราน-อาร์มีเทจที่ได้รับการเลือกสูงสุดเมื่อเทียบกับสนิปที่อยู่ในหรือใกล้ยีนเดียวกัน สนิปที่อยู่ใกล้ยีนคือสนิปที่มีตำแหน่งไม่เกิน 500 กิโลเบสเมื่อนับย้อนหลังจากตำแหน่งเริ่มการถอดรหัส (Transcription Start Site) หรือเมื่อนับไปข้างหน้าจากตำแหน่งเลิกการถอดรหัส (Transcription Termination Site) (Freytag et al., 2013; Wang et al., 2007) การกำหนดขีดเริ่มเปลี่ยนระยะทางระหว่างตำแหน่งในจีโนมข้างต้นสอดคล้องกับข้อเสนอแนะการระบุยีนให้กับสนิปในการวิเคราะห์บาทวิถีสำหรับการศึกษาความสัมพันธ์ทั้งจีโนม (Brodie et al., 2016) เนื่องจากการเก็บข้อมูลจีโนไทป์ใช้สนิปชิป Affymetrix GeneChip Human Mapping 500K Array Set ซึ่งประกอบด้วยสนิปชิป Affymetrix GeneChip Human Mapping 250K Nsp Array และสนิปชิป Affymetrix GeneChip Human Mapping 250K Sty Array การระบุตำแหน่งสนิปและยีนในจีโนมจึงกระทำโดยใช้สองไฟล์บรรณนิทัศน์ของ NetAffx (NetAffx Annotation File) ล่าสุดสำหรับสองสนิปชิปนี้ (Affymetrix, 2017)

GSEA-SNP

GSEA-SNP เป็นเทคนิคที่ได้รับการพัฒนาต่อจากเทคนิคการวิเคราะห์การได้มากขึ้นจากเซตของยีนหรือ GSEA สำหรับการวิเคราะห์การแสดงออกของยีน (Subramanian et al., 2005) การพัฒนาดังกล่าวส่งผลให้ GSEA-SNP เหมาะสมสำหรับการศึกษาความสัมพันธ์ทั้งจีโนม (Holden et al., 2008) GSEA-SNP สามารถระบุเซตของยีน (Gene Set) ใน

บาทวิถีสัมพันธ์กับโรคซับซ้อนที่สนใจอย่างมีนัยสำคัญทางสถิติหรือไม่โดยใช้การคำนวณคะแนนการได้มากขึ้น (Enrichment Score) และการทดสอบการเรียงสับเปลี่ยน (Permutation Test) การทำงานของ GSEA-SNP สามารถอธิบายได้ดังนี้

พิจารณาเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมซึ่งประกอบด้วยหลายสนิปจาก N_G ยีน ถึงแม้ว่าการวิเคราะห์เซตข้อมูลซึ่งแต่ละยีนมีหลายสนิปสามารถกระทำได้โดยใช้ GSEA-SNP ในบทความนี้ แต่ละยีนมีหนึ่งสนิปซึ่งได้รับการคัดเลือกสำหรับใช้เป็นตัวแทนยีนตั้งที่กล่าวข้างต้น (Wang et al., 2007) ยีนทั้งหมด N_G ยีนจะได้รับการเรียงลำดับตามค่าสถิติทดสอบแนวโน้มเอียงคอคอราน-อาร์มิเทจของสนิปสำหรับใช้เป็นตัวแทนยีนจากค่าสูงสุดไปค่าต่ำสุด สำหรับเซตของยีน L ซึ่งประกอบด้วย N_L ยีน คะแนนการได้มากขึ้นสำหรับเซตของยีนนี้หรือ $ES(L)$ สามารถนิยามโดย

$$ES(L) = \sum_{j \in L} \frac{c_j^\alpha}{N_C} - \sum_{j \notin L} \frac{1}{N_C - N_L} \quad (2)$$

โดยที่

$$i^* = \operatorname{argmax}_{1 \leq i \leq N_G} \left| \sum_{j \leq i} \frac{c_j^\alpha}{N_C} - \sum_{j \leq i} \frac{1}{N_C - N_L} \right| \quad (3)$$

c_j คือค่าสถิติทดสอบแนวโน้มเอียงคอคอราน-อาร์มิเทจของสนิปสำหรับใช้เป็นตัวแทนยีน g_j , α คือพารามิเตอร์ถ่วงน้ำหนักซึ่งได้รับการกำหนดให้เท่ากับ 1 (Subramanian et al., 2005; Wang et al., 2007) และ $N_C = \sum_{j \in L} c_j^\alpha$ คะแนนการได้มากขึ้นสะท้อนถึงค่าเบี่ยงเบนสูงสุดจากศูนย์ของผลรวมค่าสถิติทดสอบแนวโน้มเอียงคอคอราน-อาร์มิเทจสำหรับยีนในเซตของยีนที่ไต่ระหว่างการแหว่งผ่าน (Traversal) ตามรายการยีนซึ่งได้รับการเรียงลำดับตามค่าสถิติทดสอบแนวโน้มเอียงคอคอราน-อาร์มิเทจ

หลังจากการคำนวณคะแนนการได้มากขึ้น การทดสอบว่า เซตของยีนสัมพันธ์กับโรคซับซ้อนที่สนใจอย่างมีนัยสำคัญทางสถิติหรือไม่สามารถกระทำได้โดยใช้การทดสอบการเรียงสับเปลี่ยน การทดสอบการเรียงสับเปลี่ยนในบทความนี้ใช้ 1,000 เซตข้อมูลเรียงสับเปลี่ยน (Permutation Replicate) ซึ่งแต่ละเซตข้อมูลเรียงสับเปลี่ยนสร้างจากเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมที่ได้รับการเรียงสับเปลี่ยนเชิงสุ่มสถานะกรณีและสถานะควบคุมของตัวอย่างในเซตข้อมูล ในขณะที่จำนวนตัวอย่างกลุ่มกรณีและจำนวนตัวอย่างกลุ่มควบคุมเป็นจำนวนเดิม จากนั้นคะแนนการได้มากขึ้นจะได้รับการคำนวณโดยใช้แต่ละเซตข้อมูลเรียงสับเปลี่ยน ค่าความน่าจะเป็นหรือค่าพี (p-value) ที่ได้จาก GSEA-SNP คือผลหารระหว่างจำนวนเซตข้อมูลเรียงสับเปลี่ยนซึ่งคะแนนการได้มากขึ้นสุดขีดกว่าหรือเท่ากับคะแนนการได้มากขึ้นซึ่งคำนวณโดยใช้เซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมและจำนวนเซตข้อมูลเรียงสับเปลี่ยนทั้งหมด

ตามปกติแล้ว หลายเซตของยีนจะได้รับการพิจารณาว่า แต่ละเซตของยีนสัมพันธ์กับโรคซับซ้อนที่สนใจอย่างมีนัยสำคัญทางสถิติหรือไม่ ดังนั้นการแก้สำหรับการทดสอบหลายสมมติฐาน (Correction for Multiple Hypothesis Testing) จึงจำเป็นสำหรับ GSEA-SNP ในบทความนี้ อัตราการค้นพบเท็จ (False Discovery Rate หรือ FDR) เป็นค่าที่สนใจหลังการแก้สำหรับการทดสอบหลายสมมติฐาน อัตราการค้นพบเท็จสามารถคำนวณโดยใช้เซตข้อมูลเรียงสับเปลี่ยนดังนี้ ค่าเฉลี่ยของคะแนนการได้มากขึ้นซึ่งคำนวณโดยใช้ทุกเซตข้อมูลเรียงสับเปลี่ยนจะได้รับการคำนวณสำหรับแต่ละเซตของยีน จากนั้นคะแนนการได้มากขึ้นซึ่งคำนวณโดยใช้เซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมและเซตข้อมูลเรียงสับเปลี่ยนสำหรับแต่ละเซตของยีนจะได้รับการทำให้เป็นบรรทัดฐาน (Normalisation) โดยการหารด้วยค่าเฉลี่ย อัตราการค้นพบเท็จสำหรับเซตของยีน L^* ที่สนใจหรือ $FDR(L^*)$ สามารถคำนวณได้จาก

$$FDR(L^*) = \frac{\text{percentage of all pairs } (L, \pi) \text{ with } NES(L, \pi) \text{ more extreme than or equal to } NES^*}{\text{percentage of gene set } L \text{ with } NES(L) \text{ more extreme than or equal to } NES^*} \quad (4)$$

โดยที่ π คือตัวแปรที่ใช้ระบุเซตข้อมูลเรียงสับเปลี่ยน NES คือคะแนนการได้มากขึ้นที่ได้รับการทำให้เป็นบรรทัดฐาน (Normalised Enrichment Score) และ NES^* คือคะแนนการได้มากขึ้นที่ได้รับการทำให้เป็นบรรทัดฐานสำหรับเซตของยีน L^* (Wang et al., 2007)

บาทวิถีการให้สัญญาณและบาทวิถีเป้าหมาย

เซตของยีนในบาทวิถีที่สนใจในบทความนี้คือเซตของยีนในบาทวิถีการให้สัญญาณจาก KEGG มีบาทวิถีการให้สัญญาณทั้งหมด 223 บาทวิถี นอกจากนี้ หลักฐานการศึกษาความสัมพันธ์ทางพันธุกรรมแสดงให้เห็นว่า บางบาทวิถีการให้สัญญาณสัมพันธ์กับแต่ละโรคซับซ้อนที่สนใจ (KEGG, 2021; O' Dushlaine et al., 2011) บาทวิถีการให้สัญญาณเหล่านี้คือบาทวิถีเป้าหมาย (Target Pathway) สำหรับการทดสอบสมรรถนะของ GSEA-SNP ในการระบุว่า เซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับแต่ละโรคซับซ้อนอย่างมีนัยสำคัญทางสถิติ บาทวิถีเป้าหมายสำหรับแต่ละโรคซับซ้อนได้แสดงในตารางที่ 3

ตารางที่ 3 บาทวิถีการให้สัญญาณจาก KEGG ซึ่งเป็นบาทวิถีเป้าหมายสำหรับแต่ละโรคซับซ้อน

โรคซับซ้อน	บาทวิถีเป้าหมาย		
	KEGG ID	บาทวิถีการให้สัญญาณ	เอกสารอ้างอิง
BD	hsa04514	Cell adhesion molecules (CAMs)	O' Dushlaine et al. (2011)
	hsa04530	Tight junction	O' Dushlaine et al. (2011)
CAD	hsa04022	cGMP-PKG signalling pathway	KEGG (2021)
	hsa04310	Wnt signalling pathway	KEGG (2021)
	hsa04928	Parathyroid hormone synthesis, secretion and action	KEGG (2021)
CD	hsa04060	Cytokine-cytokine receptor interaction	KEGG (2021)
	hsa04140	Regulation of autophagy	KEGG (2021)
	hsa04621	NOD-like receptor signalling pathway	KEGG (2021)
	hsa04630	Jak-STAT signalling pathway	KEGG (2021)
	hsa05321	Inflammatory bowel disease	KEGG (2021)
HT	hsa04925	Aldosterone synthesis and secretion	KEGG (2021)
	hsa04960	Aldosterone-regulated sodium reabsorption	KEGG (2021)
RA	hsa05323	Rheumatoid arthritis	KEGG (2021)
T1D	hsa04060	Cytokine-cytokine receptor interaction	KEGG (2021)
	hsa04151	PI3K-Akt signalling pathway	KEGG (2021)
	hsa04612	Antigen processing and presentation	KEGG (2021)
	hsa04630	Jak-STAT signalling pathway	KEGG (2021)
	hsa04940	Type I diabetes mellitus	KEGG (2021)
T2D	hsa03320	PPAR signalling pathway	KEGG (2021)
	hsa04110	Cell cycle	KEGG (2021)
	hsa04115	p53 signalling pathway	KEGG (2021)
	hsa04141	Protein processing in endoplasmic reticulum	KEGG (2021)
	hsa04310	Wnt signalling pathway	KEGG (2021)
	hsa04330	Notch signalling pathway	KEGG (2021)
	hsa04350	TGF-beta signalling pathway	KEGG (2021)
	hsa04911	Insulin secretion	KEGG (2021)
	hsa04930	Type II diabetes mellitus	KEGG (2021)
hsa04972	Pancreatic secretion	KEGG (2021)	

สังเกตว่า ไม่มีบาทวิถีการให้สัญญาณที่สัมพันธ์กับโรคอารมณ์สองขั้วใน KEGG (2021) นอกจากนี้ ไม่มียีนที่สัมพันธ์กับโรคอารมณ์สองขั้วใน KEGG (2021) เช่นกัน อย่างไรก็ตาม การวิเคราะห์บาทวิถีโดยใช้เซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมจากการศึกษาความสัมพันธ์ทั้งจีโนมของโรคอารมณ์สองขั้วโดย WTCCC แสดงให้เห็นว่า มีสองบาทวิถีการให้สัญญาณ ได้แก่ บาทวิถี Cell Adhesion Molecules (CAMs) (hsa04514) และบาทวิถี Tight Junction (hsa04530) ที่สัมพันธ์โรคอารมณ์สองขั้ว (O' Dushlaine et al., 2011) ดังนั้นสองบาทวิถีนี้จึงเป็นบาทวิถีเป้าหมายสำหรับโรคอารมณ์สองขั้ว

ผลการวิจัย

ในบทความนี้ ชุดการวัดเปรียบเทียบสมรรถนะ (Benchmark Suite) ได้สร้างจากเจ็ดเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมจากการศึกษาความสัมพันธ์ทั้งจีโนมของโรคซึมเศร้าโดย WTCCC เนื่องจากชิป Affymetrix GeneChip Human Mapping 500K Array Set ประกอบด้วยชิป Affymetrix GeneChip Human Mapping 250K Nsp Array และชิป Affymetrix GeneChip Human Mapping 250K Sty Array ชิปในแต่ละเซตข้อมูลจึงสามารถแบ่งเป็นสองส่วนไม่ซ้อนเหลื่อม ส่งผลให้มีสามชุดการวัดเปรียบเทียบสมรรถนะ ได้แก่ ชุดการวัดเปรียบเทียบสมรรถนะ 250K Nsp Array ชุดการวัดเปรียบเทียบสมรรถนะ 250K Sty Array และชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set ชีตเริ่มเปลี่ยน r^2 สำหรับการคัดเลือกชิปตัวแทนจากชิปในตัวอย่างกลุ่มควบคุมโดยใช้ Tagger ที่สนใจคือ 0.8 และ 0.9 จำนวนชิปทั้งหมดและจำนวนชิปตัวแทนในแต่ละชุดการวัดเปรียบเทียบสมรรถนะได้แสดงในตารางที่ 4

ดังที่กล่าวข้างต้น ชิปซึ่งมีค่าสถิติทดสอบแนวโน้มเอียงคออราเนอ-อาร์มีเทจสูงสุดเมื่อเทียบกับชิปที่อยู่ในหรือใกล้เคียงกันจะได้รับการคัดเลือกสำหรับใช้เป็นตัวแทนยีน เงื่อนไขดังกล่าวส่งผลให้ไม่สามารถระบุชิปสำหรับใช้เป็นตัวแทนยีนได้ครบทุกยีน ดังนั้นจำนวนชิปสำหรับใช้เป็นตัวแทนยีนรวมทั้งจำนวนยีนที่มีชิปสำหรับใช้เป็นตัวแทนยีนจึงน้อยกว่าจำนวนยีนที่ระบุในไฟล์บรรณนิทัศน์ของ NetAffx จำนวนชิปสำหรับใช้เป็นตัวแทนยีน จำนวนยีนที่มีชิปสำหรับใช้เป็นตัวแทนยีน และจำนวนยีนที่ระบุในไฟล์บรรณนิทัศน์ของ NetAffx ในแต่ละชุดการวัดเปรียบเทียบสมรรถนะได้แสดงในตารางที่ 5 สังเกตว่ามากกว่าหนึ่งยีนมีชิปสำหรับใช้เป็นตัวแทนยีนเป็นชิปเดียวกัน ส่งผลให้จำนวนชิปสำหรับใช้เป็นตัวแทนยีนน้อยกว่าจำนวนยีนที่มีชิปสำหรับใช้เป็นตัวแทนยีน นอกจากนี้ เนื่องจากการคัดเลือกชิปตัวแทนไม่มีผลต่อจำนวนชิปสำหรับใช้เป็นตัวแทน

ตารางที่ 4 จำนวนชิปทั้งหมดและจำนวนชิปตัวแทนในแต่ละชุดการวัดเปรียบเทียบสมรรถนะ

ชุดการวัดเปรียบเทียบสมรรถนะ	จำนวนชิปทั้งหมด	จำนวนชิปตัวแทนเมื่อชีตเริ่มเปลี่ยน $r^2 = 0.9$	จำนวนชิปตัวแทนเมื่อชีตเริ่มเปลี่ยน $r^2 = 0.8$
250K Nsp Array	197,764	135,783	122,810
250K Sty Array	169,859	125,497	114,913
500K Array Set	367,623	224,324	195,847

ตารางที่ 5 จำนวนชิปสำหรับใช้เป็นตัวแทนยีน จำนวนยีนที่มีชิปสำหรับใช้เป็นตัวแทนยีน และจำนวนยีนที่ระบุในไฟล์บรรณนิทัศน์ของ NetAffx ในแต่ละชุดการวัดเปรียบเทียบสมรรถนะ

ชุดการวัดเปรียบเทียบสมรรถนะ	จำนวนชิปสำหรับใช้เป็นตัวแทนยีน	จำนวนยีนที่มีชิปสำหรับใช้เป็นตัวแทนยีน	จำนวนยีนที่ระบุในไฟล์บรรณนิทัศน์ของ NetAffx
250K Nsp Array	12,640	15,854	15,860
250K Sty Array	13,477	16,852	16,856
500K Array Set	14,814	18,239	18,245

ตารางที่ 6 บาทวิถีเป้าหมายซึ่งผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 250K Nsp Array โดยใช้ GSEA-SNP ระบุว่า เซตของยีนในบาทวิถีสัมพันธ์กับแต่ละโรคซับซ้อน

โรคซับซ้อน	บาทวิถีเป้าหมายที่ระบุจาก การใช้สลิปทั้งหมด	บาทวิถีเป้าหมายที่ระบุจาก การใช้สลิปตัวแทนเมื่อ ขีดเริ่มเปลี่ยน $r^2 = 0.9$	บาทวิถีเป้าหมายที่ระบุจาก การใช้สลิปตัวแทนเมื่อ ขีดเริ่มเปลี่ยน $r^2 = 0.8$
BD	-	-	-
CAD	-	-	-
CD	-	-	-
HT	-	-	-
RA	hsa05323	hsa05323	hsa05323
T1D	hsa04612, hsa04940	hsa04612, hsa04940	hsa04612, hsa04940
T2D	-	-	-

ตารางที่ 7 บาทวิถีเป้าหมายซึ่งผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 250K Sty Array โดยใช้ GSEA-SNP ระบุว่า เซตของยีนในบาทวิถีสัมพันธ์กับแต่ละโรคซับซ้อน

โรคซับซ้อน	บาทวิถีเป้าหมายที่ระบุจาก การใช้สลิปทั้งหมด	บาทวิถีเป้าหมายที่ระบุจาก การใช้สลิปตัวแทนเมื่อ ขีดเริ่มเปลี่ยน $r^2 = 0.9$	บาทวิถีเป้าหมายที่ระบุจาก การใช้สลิปตัวแทนเมื่อ ขีดเริ่มเปลี่ยน $r^2 = 0.8$
BD	-	-	-
CAD	-	-	-
CD	hsa05321	hsa05321	hsa05321
HT	-	-	-
RA	hsa05323	hsa05323	hsa05323
T1D	hsa04612, hsa04940	hsa04612, hsa04940	hsa04612, hsa04940
T2D	-	-	-

ตารางที่ 8 บาทวิถีเป้าหมายซึ่งผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set โดยใช้ GSEA-SNP ระบุว่า เซตของยีนในบาทวิถีสัมพันธ์กับแต่ละโรคซับซ้อน

โรคซับซ้อน	บาทวิถีเป้าหมายที่ระบุจาก การใช้สลิปทั้งหมด	บาทวิถีเป้าหมายที่ระบุจาก การใช้สลิปตัวแทนเมื่อ ขีดเริ่มเปลี่ยน $r^2 = 0.9$	บาทวิถีเป้าหมายที่ระบุจาก การใช้สลิปตัวแทนเมื่อ ขีดเริ่มเปลี่ยน $r^2 = 0.8$
BD	-	-	-
CAD	-	-	-
CD	hsa05321	-	-
HT	-	-	-
RA	hsa05323	hsa05323	hsa05323
T1D	hsa04612, hsa04940	hsa04612, hsa04940	hsa04612, hsa04940
T2D	-	-	-

ยีน จำนวนสลับสำหรับใช้เป็นตัวแทนยีนจึงมีจำนวนเท่ากันไม่ว่าจะสนใจสลับทั้งหมดหรือสลับตัวแทนในแต่ละชุดการวัดเปรียบเทียบสมรรถนะ

กำหนดให้เซตของยีนในบาทวิถีสัมพันธ์กับโรคซับซ้อนที่สนใจอย่างมีนัยสำคัญทางสถิติคือเซตของยีนซึ่งผลการวิเคราะห์โดยใช้ GSEA-SNP มีอัตราการค้นพบที่น้อยกว่าหรือเท่ากับ 0.05 บาทวิถีเป้าหมายซึ่งผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 250K Nsp Array ชุดการวัดเปรียบเทียบสมรรถนะ 250K Sty Array และชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set โดยใช้ GSEA-SNP ระบุว่า เซตของยีนในบาทวิถีสัมพันธ์กับแต่ละโรคซับซ้อนได้แสดงในตารางที่ 6, 7 และ 8 ตามลำดับ

เนื่องจาก GSEA-SNP ใช้การทดสอบการเรียงสับเปลี่ยนในการประเมินค่าพีและอัตราการค้นพบเท็จ ดังนั้นการวิเคราะห์โดยใช้ GSEA-SNP จึงได้รับการทดลองซ้ำเพื่อทดสอบว่า การสุ่มในการทดสอบการเรียงสับเปลี่ยนไม่มีผลต่อการระบุเซตของยีนในบาทวิถีซึ่งสัมพันธ์กับโรคซับซ้อนที่สนใจอย่างมีนัยสำคัญทางสถิติ การทดลองซ้ำใช้สลับตัวแทนเมื่อซีตเริ่มเปลี่ยน $r^2 = 0.8$ จากเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมซึ่งกลุ่มกรณีคือกลุ่มบุคคลเป็นโรคข้ออักเสบรูมาตอยด์และเบาหวานชนิดที่ 2 ในชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set การทดลองซ้ำสนใจสองเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมดังกล่าวเนื่องจาก GSEA-SNP สามารถระบุว่า เซตของยีนในบาทวิถีเป้าหมายซึ่งมีหนึ่งบาทวิถีสัมพันธ์โรคข้ออักเสบรูมาตอยด์ และ GSEA-SNP ไม่สามารถระบุว่า เซตของยีนในบาทวิถีเป้าหมายซึ่งมีสิบบาทวิถีสัมพันธ์กับเบาหวานชนิดที่ 2 การทดลองซ้ำ 100 ครั้งยืนยันว่า การสุ่มในการทดสอบการเรียงสับเปลี่ยนไม่มีผลต่อผลการวิเคราะห์โดยใช้ GSEA-SNP

อภิปรายและสรุปผลการวิจัย

การวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะโดยใช้ GSEA-SNP แสดงให้เห็นว่า ภายใต้เงื่อนไขการมีอยู่ของข้อมูลความไม่สมดุลการเชื่อมโยงระหว่างสลับตัวแทนและสลับที่มีตัวแทนอย่างสมบูรณ์ โดยรวมการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสลับทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนมให้ผลการวิเคราะห์ไม่แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสลับตัวแทนเท่านั้น อย่างไรก็ตาม การวิเคราะห์เซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมซึ่งกลุ่มกรณีคือกลุ่มบุคคลเป็นโรคโครห์นในชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set แสดงให้เห็นว่า กรณีที่สนใจข้อมูลสลับตัวแทนเท่านั้นสมรรถนะในการระบุเซตของยีนในบาทวิถีเป้าหมายที่สัมพันธ์กับโรคซับซ้อนของ GSEA-SNP ลดลง ข้อสังเกตดังกล่าวสามารถอธิบายได้โดยพิจารณาสลับชิป Affymetrix GeneChip Human Mapping 500K Array Set ซึ่งประกอบด้วยสลับชิป Affymetrix GeneChip Human Mapping 250K Nsp Array และสลับชิป Affymetrix GeneChip Human Mapping 250K Sty Array ดังนั้น GSEA-SNP สามารถระบุว่า เซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับโรคโครห์นจากผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 250K Sty Array ในทางตรงกันข้าม GSEA-SNP ไม่สามารถระบุว่า เซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับโรคโครห์นจากผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 250K Nsp Array นั่นคือสลับที่ได้จากสลับชิป Affymetrix GeneChip Human Mapping 250K Sty Array จำเป็นสำหรับการใช้เป็นตัวแทนยีนในการระบุว่า เซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับโรคโครห์น ดังนั้นกรณีที่สนใจข้อมูลสลับทั้งหมดในชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set, GSEA-SNP สามารถระบุว่า เซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับโรคโครห์นเพราะมีสลับสำหรับใช้เป็นตัวแทนยีนเป็นสลับที่ได้จากสลับชิป Affymetrix GeneChip Human Mapping 250K Sty Array เพียงพอ ในทางตรงกันข้าม กรณีที่สนใจข้อมูลสลับตัวแทนเท่านั้นในชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set, GSEA-SNP ไม่สามารถระบุว่า เซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับโรคโครห์นเพราะขาดสลับสำหรับใช้เป็นตัวแทนยีนเป็นสลับที่ได้จากสลับชิป Affymetrix GeneChip Human Mapping 250K Sty Array ที่จำเป็น ดังนั้นสลับชิปที่ใช้ในการเก็บข้อมูลจีโนมโทปมีผลต่อข้อสรุปข้างต้น

ในบทความนี้ การคัดเลือก SNP ตัวแทนใช้ SNP ในตัวอย่างกลุ่มควบคุม ส่งผลให้มีข้อมูลความไม่สมดุลการเชื่อมโยงระหว่าง SNP ตัวแทนและ SNP ที่มีตัวแทนอย่างสมบูรณ์ ตามปกติแล้ว SNP ตัวแทนที่ใช้ในการออกแบบ SNP chip โดยอาศัย SNP ตัวแทนจะได้รับการคัดเลือก SNP ในแผง SNP อ้างอิง (Reference SNP Panel) ของประชากรเดียวกับหรือใกล้เคียงกับประชากรที่สนใจในการศึกษาความสัมพันธ์ทั้งจีโนม ข้อแตกต่างระหว่างการแจกแจงความถี่อัลลีล (Allele Frequency Distribution) ของ SNP ในแผง SNP อ้างอิงและ SNP ในตัวอย่างกลุ่มควบคุมส่งผลต่อข้อมูลความไม่สมดุลการเชื่อมโยงระหว่าง SNP ตัวแทนและ SNP ที่มีตัวแทน ดังนั้นการศึกษาลงของปัจจัยดังกล่าวต่อการทดสอบแนวคิดที่ว่า การวิเคราะห์หาวิถีโดยใช้ข้อมูล SNP ทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนมให้ผลการวิเคราะห์ที่ไม่แตกต่างจากการวิเคราะห์หาวิถีโดยใช้ข้อมูล SNP ตัวแทนเท่านั้นจึงจำเป็นต้องได้รับการศึกษา

กิตติกรรมประกาศ

บทความนี้ใช้ข้อมูลจาก Wellcome Trust Case Control Consortium รายชื่อนักวิจัยทั้งหมดที่มีส่วนร่วมในการสร้างข้อมูลอยู่ที่ www.wtccc.org.uk ทุนวิจัยสำหรับโครงการได้รับการสนับสนุนจาก Wellcome Trust ภายใต้รหัสโครงการ 076113, 085475 และ 090355

เอกสารอ้างอิง

- Affymetrix. Human Mapping 500K Array Set - Support Materials [online] 2017 [cited 2021 Jun 14]. Available from: <http://www.affymetrix.com/support/technical/byproduct.affx?product=500k>
- Brodie A, Azaria JR, Ofra Y. How far from the SNP may the causative genes be? *Nucleic Acids Res* 2016; 44(13): 6046-54.
- Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; 74(1): 106-20.
- de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nature Genet* 2005; 37(11): 1217-23.
- Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003; 73(6): 1316-29.
- Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, et al. A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. *Hum Hered* 2013; 76(2): 64-75.
- Heidema AG, Boer JMA, Nagelkerke N, Mariman ECM, van der A DL, Feskens EJM. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC Genet* 2006; 7: 23.
- Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet* 1968; 38(6): 226-31.
- Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 2008; 24(23): 2784-5.

- KEGG: Kyoto Encyclopedia of Genes and Genomes. KEGG Disease Database [online] 2021 [cited 2021 Jun 14]. Available from: <https://www.genome.jp/kegg/disease/>
- Lewis CM. Genetic association studies: design, analysis and interpretation. *Brief Bioinform* 2002; 3(2): 146-53.
- Montana G. Statistical methods in genetics. *Brief Bioinform* 2006; 7(3): 297-308.
- National Center for Biotechnology Information. dbGaP: Database of Genotypes and Phenotypes [online] 2021 [cited 2021 Jun 14]. Available from: <https://www.ncbi.nlm.nih.gov/gap/>
- O'Dushlaine C, Kenny E, Heron E, Donohoe G, Gill M, Morris D, et al. Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol Psychiatr* 2011; 16(3): 286-92.
- Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007; 23(19): 2507-17.
- Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 1997; 53(4): 1253-61.
- Setsirichok D, Tienboon P, Jaronruang N, Kittichaijaroen S, Wongseree W, Piroonratana T, et al. An omnibus permutation test on ensembles of two-locus analyses can detect pure epistasis and genetic heterogeneity in genome-wide association studies. *SpringerPlus* 2013; 2: 230.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; 102(43): 15545-50.
- Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019; 20(8): 467-84.
- The International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005; 437(7063): 1299-320.
- 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): 661-78.
- Van Steen K. Travelling the world of gene-gene interactions. *Brief Bioinform* 2012; 13(1): 1-19.
- Wallace C, Dobson RJ, Munroe PB, Caulfield MJ. Information capture using SNPs from HapMap and whole-genome chips differs in a sample of inflammatory and cardiovascular gene-centric regions from genome-wide estimates. *Genome Res* 2007; 17(11): 1596-602.
- Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 2007; 81(6): 1278-83.
- Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* 2010; 11(12): 843-54.
- Wongseree W, Assawamakin A, Piroonratana T, Sinsomros S, Limwongse C, Chaiyaratana, N. Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses. *BMC Bioinformatics* 2009; 10: 294.