ผลของการคัดเลือกสนิปตัวแทนต่อการวิเคราะห์การได้มากขึ้นจากเซตของยีนในบาทวิถีการให้สัญญาณจากฐานข้อมูล KEGG ในการศึกษาความสัมพันธ์ทั้งจีโนม

นายเจษฎา วีรเดชกำพล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

Effects of tag SNP selection on gene set enrichment analysis of KEGG signalling pathways in genome-wide association studies

Mr. Jessada Weeradetkumpon

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2021

หัวข้อวิทยานิพนธ์	ผลของการคัดเลือกสนิปตัวแทนต่อการวิเคราะห์การได้มากขึ้นจากเซตของยีนในบาทวิถีการให้สัญญาณจากฐานข้อมูล KEGG ในการศึกษาความสัมพันธ์ทั้งจีโนม
โดย	นายเจษฎา วีรเดชกำพล
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ศาสตราจารย์ ดร.ณชล ไชยรัตนะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.เศรษฐา ปานงาม)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ศาสตราจารย์ ดร.ณชล ไชยรัตนะ)

..... กรรมการภายนอกมหาวิทยาลัย
(ดร.ศิษเฎศ ทองสิมา)

เจษฎา วีรเดชกำพล : ผลของการคัดเลือกสนิปตัวแทนต่อการวิเคราะห์การได้มากขึ้นจากเซตของยีนในบาทวิถีการให้สัญญาณจากฐานข้อมูล KEGG ในการศึกษาความสัมพันธ์ทั้งจีโนม. (Effects of tag SNP selection on gene set enrichment analysis of KEGG signalling pathways in genome-wide association studies) อ.ที่ปรึกษาหลัก : ศ. ดร.ประภาส จงสถิตย์วัฒนา, อ.ที่ปรึกษาร่วม : ศ. ดร.ณชล ไชยรัตนะ

วิทยานิพนธ์นี้นำเสนอการเปรียบเทียบระหว่างการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดและข้อมูลสนิปตัวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนม ชุดการวัดเปรียบเทียบสมรรถนะได้สร้างจากเจ็ดเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมจากการศึกษาความสัมพันธ์ทั้งจีโนมของเจ็ดโรคซับซ้อนโดย Wellcome Trust Case Control Consortium เจ็ดโรคซับซ้อนที่สนใจ ได้แก่ โรคอารมณ์สองขั้ว โรคหลอดเลือดแดงโคโรนารี โรคโครห์น ความดันเลือดสูง โรคข้ออักเสบรูมาตอยด์ เบาหวานชนิดที่ 1 และเบาหวานชนิดที่ 2 สนิปตัวแทนได้รับการคัดเลือกจากสนิปในตัวอย่างกลุ่มควบคุมโดยใช้ Tagger จากนั้นหนึ่งสนิปจะได้รับการคัดเลือกสำหรับใช้เป็นตัวแทนยีนโดยการหาค่าสูงสุดของค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจเป็นเงื่อนไขการคัดเลือก ถึงแม้ว่ามีการคำนวณค่าสถิติทดสอบสำหรับแต่ละสนิป ค่าสถิติทดสอบสำหรับสนิปตัวแทนจะใช้เป็นค่าสถิติทดสอบสำหรับสนิปที่มีตัวแทนด้วย ส่งผลให้ข้อมูลสนิปที่มีตัวแทนไม่จำเป็นสำหรับการวิเคราะห์บาทวิถี การวิเคราะห์บาทวิถีกระทำโดยใช้ GSEA-SNP ซึ่งเป็นเทคนิคที่ได้รับการพัฒนาต่อจากเทคนิคการวิเคราะห์การได้มากขึ้นจากเซตของยีนหรือ GSEA และสามารถระบุว่า เซตของยีนในบาทวิถีสัมพันธ์กับโรคซับซ้อนหรือไม่ การวิเคราะห์บาทวิถีสนใจเฉพาะบาทวิถีการให้สัญญาณจาก Kyoto Encyclopedia of Genes and Genomes (KEGG) ดังนั้นจุดประสงค์ของการวัดเปรียบเทียบสมรรถนะคือการเปรียบเทียบสมรรถนะการระบุบาทวิถีเป้าหมายที่สัมพันธ์กับแต่ละโรคซับซ้อนจากบาทวิถีการให้สัญญาณทั้งหมด โดยรวมการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดให้ผลการวิเคราะห์ไม่แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปตัวแทน ภายใต้เงื่อนไขการมีอยู่ของข้อมูลความไม่สัมพันธ์การเชื่องโยง ผลการศึกษาแสดงให้เห็นความเป็นไปได้ของการวิเคราะห์บาทวิถีโดยใช้เซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมซึ่งการเก็บข้อมูลจีโนไทป์จะอาศัยสนิปตัวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนม

สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่อนิสิต .....
ปีการศึกษา	2564	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6372020921 : MAJOR COMPUTER SCIENCE

KEYWORD: Pathway analysis, Genome-wide association study, Tag SNP, complex disease, gene set enrichment analysis

Jessada Weeradetkumpon : Effects of tag SNP selection on gene set enrichment analysis of KEGG signalling pathways in genome-wide association studies. Advisor: Prof. PRABHAS CHONGSTITVATANA, Ph.D. Co-advisor: Prof. Nachol Chaiyaratana, Ph.D.

This thesis 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. These complex diseases are bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. Tag SNPs were selected from SNPs in the controls using Tagger. Subsequently, 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 were also assigned to their tagged SNPs. As a result, tagged SNPs became redundant and were unnecessary in the pathway analysis. GSEA-SNP, which is an extension of gene set enrichment analysis (GSEA) and can identify whether gene sets in pathways are associated with a complex disease, was the chosen pathway analysis technique. Signalling pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) were the main focus. Therefore, the benchmarking aimed at comparing the ability to identify target pathways associated with each complex disease among all signalling pathways. Overall, the pathway analyses of all SNPs were similar to those of tag SNPs. Under the condition of linkage disequilibrium information availability, the results suggest the possibility of generalisation to pathway analysis of existing case-control datasets that exploit tag SNPs from genome-wide association studies.

Field of Study:	Computer Science	Student's Signature .....
Academic Year:	2021	Advisor's Signature .....
		Co-advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์ของ ศ.ดร.ประภาส จงสถิตวัฒนา อาจารย์ที่ปรึกษาวิทยานิพนธ์หลักและ ศ.ดร.ณชล ไชยรัตนะ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ซึ่งได้มอบโอกาส แนวคิด ตลอดจนไปถึงความรู้ในการทำวิทยานิพนธ์และสละเวลาคอยให้คำปรึกษาในด้านต่าง ๆ ตรวจสอบแก้ไข้จนทำให้การวิจัยครั้งนี้สามารถสำเร็จลุล่วงไปได้ด้วยดี

ขอขอบพระคุณกรรมการสอบวิทยานิพนธ์ รศ.ดร.เศรษฐา ปานงาม ประธานกรรมการ และ ดร.ศิษเฎศ ทองสิมา กรรมการภายนอก ที่กรุณาสละเวลามาตรวจสอบและให้คำแนะนำ ที่สามารถนำมาพัฒนาปรับปรุงวิทยานิพนธ์เล่มนี้ให้เกิดประโยชน์ต่อผู้อ่านได้เป็นอย่างดี

ขอขอบพระคุณพ่อแม่รวมไปถึงอาจารย์ทุกท่านที่ได้อบรม สั่งสอนให้ความรู้มากมายไม่ว่าจะทางวิชาการหรือการใช้ชีวิตตั้งแต่อดีตจนถึงปัจจุบัน

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

วิทยานิพนธ์ได้รับสนับสนุนทุนจาก ศ.ดร.ณชล ไชยรัตนะ

เจษฎา วีรเดชกำพล

สารบัญ

หน้า

..... ค

บทคัดย่อภาษาไทย ..... ค

..... ง

บทคัดย่อภาษาอังกฤษ ..... ง

กิตติกรรมประกาศ ..... จ

สารบัญ ..... ฉ

สารบัญตาราง ..... ซ

สารบัญรูปภาพ ..... ฌ

บทที่ 1 1

บทนำ 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของงานวิจัย 5

1.3 ขอบเขตการวิจัย 5

1.4 ขั้นตอนการดำเนินงาน 6

1.5 ประโยชน์ที่คาดว่าจะได้รับ 6

บทที่ 2 7

เซตข้อมูลและวิธีการวิจัย 7

2.2 การคัดเลือกสนิปตัวแทนโดยใช้ Tagger 8

2.3 การทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจ 9

2.4 การคัดเลือกสนิปสำหรับใช้เป็นตัวแทนยีน 10

2.5 GSEA-SNP 11

2.6 บาทวิถีการให้สัญญาณและบาทวิถีเป้าหมาย 14
บทที่ 3 16
ผลการวิจัยและอภิปรายผลการวิจัย 16
บทที่ 4 22
สรุปผลการวิจัยและข้อเสนอแนะ 22
4.1 สรุปผลการวิจัย 22
4.2 ข้อเสนอแนะ 22
บรรณานุกรม 23
ประวัติผู้เขียน 28

สารบัญตาราง

หน้า

ตารางที่ 1 จำนวนตัวอย่างของกลุ่มกรณีและกลุ่มควบคุม..... 7

ตารางที่ 2 การแจกแจงตัวอย่างกลุ่มกรณีและตัวอย่างกลุ่มควบคุมในเซตข้อมูลตามจีโนไทป์ที่สนิป .. 8

ตารางที่ 3 บาทวิถีการให้สัญญาณจาก KEGG ซึ่งเป็นบาทวิถีเป้าหมายสำหรับแต่ละโรคซับซ้อน..... 15

ตารางที่ 4 จำนวนสนิปทั้งหมดและจำนวนสนิปตัวแทนในแต่ละชุดการวัดเปรียบเทียบสมรรถนะ ... 17

ตารางที่ 5 จำนวนสนิปสำหรับใช้เป็นตัวแทนยีน จำนวนยีนที่มีสนิปสำหรับใช้เป็นตัวแทนยีน และจำนวนยีนที่ระบุในไฟล์บรรณนิทัศน์ของ NetAffx ในแต่ละชุดการวัดเปรียบเทียบสมรรถนะ 17

ตารางที่ 6 บาทวิถีเป้าหมายซึ่งผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 250K Nsp Array โดยใช้ GSEA-SNP ระบุว่า เซตของยีนในบาทวิถีสัมพันธ์กับแต่ละโรคซับซ้อน..... 18

ตารางที่ 7 บาทวิถีเป้าหมายซึ่งผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 250K Sty Array โดยใช้ GSEA-SNP ระบุว่า เซตของยีนในบาทวิถีสัมพันธ์กับแต่ละโรคซับซ้อน..... 18

ตารางที่ 8 บาทวิถีเป้าหมายซึ่งผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set โดยใช้ GSEA-SNP ระบุว่า เซตของยีนในบาทวิถีสัมพันธ์กับแต่ละโรคซับซ้อน..... 19

ตารางที่ 9 ผลการทดสอบสมมุติฐานว่า การคัดเลือกสนิปสำหรับใช้เป็นตัวแทนยีนไม่มีผลต่ออัตราการค้นพบเท็จสำหรับเซตของยีนในบาทวิถีโดยการทดสอบฟรีดแมน..... 20

สารบัญรูปภาพ

หน้า

รูปที่ 1 ตัวอย่างของจีโนไทป์ของพันธุ์ป่าโฮโมไซโกต 2

รูปที่ 2 ตัวอย่างของจีโนไทป์ของเฮเทอโรไซโกต 2

รูปที่ 3 ตัวอย่างจีโนไทป์ของพันธุ์กลายโฮโมไซโกต 3

รูปที่ 4 ขั้นตอนที่ใช้ในงานวิจัย 6

รูปที่ 5 หกสนิปและค่า r-squared สำหรับการอธิบายความไม่สมดุลการเชื่อมโยงระหว่างคู่สนิป 9

รูปที่ 6 สนิปตัวแทนที่ได้รับการคัดเลือกจากสนิปในรูปที่ 5 และการกำหนดค่าทดสอบสถิติแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับสนิปที่มีตัวแทน 11

รูปที่ 7 ผลการคำนวณคะแนนการได้มากขึ้นและคะแนนการได้มากขึ้นมีค่าบวก 13

รูปที่ 8 ผลการคำนวณคะแนนการได้มากขึ้นและคะแนนการได้มากขึ้นมีค่าลบ 14

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

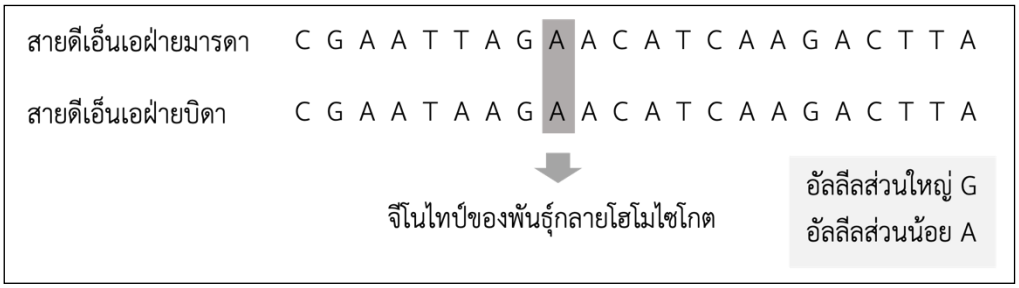
สนิป (Single Nucleotide Polymorphism หรือ SNP) เป็นเครื่องหมายพันธุกรรม (Genetic Marker) ที่พบได้ทั่วไปในจีโนม (Genome) ของสิ่งมีชีวิต แต่ละสนิปเป็นผลจากการแทนที่นิวคลีโอไทด์ (Nucleotide) หนึ่งตำแหน่งในจีโนม ข้อแตกต่างระหว่างสนิปและการกลายพันธุ์จุด (Point Mutation) คือการแปรผันของนิวคลีโอไทด์ในประชากรต้องมีความถี่อย่างน้อย 0.01 การแปรผันของนิวคลีโอไทด์นั้นจึงจะเรียกว่าสนิป [1] สำหรับมนุษย์ซึ่งมีดิพลอยด์จีโนม (Diploid Genome) สนิปส่วนใหญ่เป็นเครื่องหมายพันธุกรรมแบบสองอัลลีล (Allele) ส่งผลให้มีสามจีโนไทป์ (Genotype) ที่เป็นไปได้ที่ตำแหน่งที่ตั้ง (Locus) ของสนิป ได้แก่ จีโนไทป์ของพันธุ์ป่าโฮโมไซโกต (Homozygous Wild-type Genotype) จีโนไทป์ของเฮเทอโรไซโกต (Heterozygous Genotype) และจีโนไทป์ของพันธุ์กลายโฮโมไซโกต (Homozygous Variant Genotype) จีโนไทป์ของพันธุ์ป่าโฮโมไซโกต ประกอบด้วยสองอัลลีลส่วนใหญ่ (Major Allele หรือ Common Allele) ซึ่งเป็นอัลลีลที่มีความถี่ในประชากรสูงกว่าอัลลีลที่เหลือ จีโนไทป์ของเฮเทอโรไซโกตประกอบด้วยหนึ่งอัลลีลส่วนใหญ่และหนึ่งอัลลีลส่วนน้อย (Minor Allele หรือ Rare Allele) ซึ่งเป็นอัลลีลที่มีความถี่ในประชากรต่ำกว่าอัลลีลที่เหลือ จีโนไทป์ของพันธุ์กลายโฮโมไซโกตประกอบด้วยสองอัลลีลส่วนน้อย ตัวอย่างของจีโนไทป์ของพันธุ์ป่าโฮโมไซโกต จีโนไทป์ของเฮเทอโรไซโกต และจีโนไทป์ของพันธุ์กลายโฮโมไซโกตได้แสดงในรูปที่ 1, 2 และ 3 ตามลำดับ หลายการศึกษาพันธุกรรมมนุษย์อาศัยสนิปในการศึกษา เช่น การอธิบายโครงสร้างประชากร (Population Structure) [2] การระบุเครื่องหมายพันธุกรรมจำเพาะบรรพบุรุษ (Ancestry Informative Marker) [3] และการศึกษาความสัมพันธ์ทางพันธุกรรม (Genetic Association Study) [4-7]

การศึกษาความสัมพันธ์ทางพันธุกรรมสนใจการระบุเครื่องหมายพันธุกรรมที่อยู่ในหรือใกล้ยีน (Gene) ที่สามารถนำไปสู่การอธิบายภูมิไวรับ (Susceptibility) โรคทางพันธุกรรม (Genetic Disease) ภูมิไวรับโรคทางพันธุกรรมหลายโรค เช่น โรคหืด (Asthma) มะเร็ง (Cancer) เบาหวาน (Diabetes) และความดันเลือดสูง (Hypertension) ไม่สามารถอธิบายโดยทายกรรมแบบเมนเดล (Mendelian Inheritance) ดังนั้นโรคเหล่านี้จึงได้รับการเรียกว่าโรคซับซ้อน (Complex Disease) [8] ตามปกติแล้ว การศึกษาความสัมพันธ์ทางพันธุกรรมเป็นการศึกษากลุ่มกรณี-กลุ่มควบคุม (Case-Control Study) หรือการศึกษากลุ่มกรณี-กลุ่มร่วมรุ่น (Case-Cohort Study) โดยที่ตัวอย่างกลุ่มกรณี (Case Sample) คือตัวอย่างจากบุคคลเป็นโรค (Affected Individual) ในขณะที่ตัวอย่าง

กลุ่มควบคุม (Control Sample) และตัวอย่างกลุ่มร่วมรุ่น (Cohort Sample) คือตัวอย่างจากบุคคลไม่เป็นโรค (Unaffected Individual) ซึ่งได้จากการเก็บข้อมูลแบบตัดขวาง (Cross-Sectional Data Collection) และการเก็บข้อมูลแบบระยะยาว (Longitudinal Data Collection) ตามลำดับ [9] ในปัจจุบันการเก็บข้อมูลจีโนไทป์สามารถกระทำโดยพิจารณาจำนวนสนิปตั้งแต่หลักแสนถึงหลักล้าน โดยที่สนิปมีการกระจายในจีโนมของมนุษย์ซึ่งประกอบด้วยประมาณสามพันล้านนิวคลีโอไทด์และพิจารณาจำนวนตัวอย่างตั้งแต่หลักพันถึงหลักแสนในหนึ่งการศึกษาความสัมพันธ์ทางพันธุกรรม การศึกษาความสัมพันธ์ทางพันธุกรรมในลักษณะดังกล่าวเรียกว่าความสัมพันธ์ทั้งจีโนม (Genome-Wide Association Study หรือ GWAS) [10] ข้อมูลจากการศึกษาความสัมพันธ์ทั้งจีโนมมีจำนวนสนิปมากกว่าจำนวนตัวอย่างเสมอ ดังนั้นข้อมูลจากการศึกษาความสัมพันธ์ทั้งจีโนมจึงสอดคล้องกับบทนิยาม “P มาก N น้อย” ในสถิติวิเคราะห์ (Statistical Analysis) โดยที่ P คือจำนวนลักษณะประจำ (Attribute) (สนิป) และ N คือจำนวนตัวอย่าง [11]

รูปที่ 1 ตัวอย่างของจีโนไทป์ของพันธุ์ป่าโฮโมไซโกต

รูปที่ 2 ตัวอย่างของจีโนไทป์ของเฮเทอโรไซโกต



รูปที่ 3 ตัวอย่างจีโนไทป์ของพันธุ์กลายโฮโมไซโกต

ข้อมูลสนิปจาก International HapMap Project [12] ทำให้การออกแบบสนิปชิป (SNP Chip) สำหรับการศึกษาความสัมพันธ์ทั้งจีโนมเป็นไปได้ การออกแบบสนิปชิปสามารถแบ่งเป็นสองวิธี ได้แก่ การออกแบบสนิปชิปโดยอาศัยสนิปทั้งหมดและการออกแบบสนิปชิปโดยอาศัยสนิปตัวแทน (Tag SNP) การออกแบบสนิปชิปโดยอาศัยสนิปทั้งหมดใช้คุณภาพการเก็บข้อมูลจีโนไทป์ในการคัดเลือกสนิป ส่งผลให้ข้อมูลสนิปที่ได้จากสนิปชิปมีลักษณะกระจายในจีโนมอย่างสุ่ม ตัวอย่างของสนิปชิปที่ได้รับการออกแบบด้วยวิธีนี้คือสนิปชิปขนาด 111,000 และ 500,000 สนิปของ Affymetrix ในทางตรงกันข้าม การออกแบบสนิปชิปโดยอาศัยสนิปตัวแทนสนใจเฉพาะสนิปตัวแทนซึ่งเป็นสนิปที่มีสหสัมพันธ์(Correlation) หรือความไม่สมดุลการเชื่อมโยง (Linkage Disequilibrium) กับสนิปที่มีตัวแทน (Tagged SNP) ส่งผลให้ข้อมูลสนิปที่ได้จากสนิปชิปมีสหสัมพันธ์กับข้อมูลสนิปที่ไม่ได้จากสนิปชิป ตัวอย่างของสนิปชิปที่ได้รับการออกแบบด้วยวิธีนี้คือสนิปชิปขนาด 317,000 และ 555,000 สนิปของ Illumina [13]

การวิเคราะห์ข้อมูลสนิปจากการศึกษาความสัมพันธ์ทั้งจีโนมสามารถกระทำโดยการวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้ง (Single-Locus Analysis) และการวิเคราะห์ครั้งละหลายตำแหน่งที่ตั้ง (Multi-locus Analysis) [4-6] การวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้งเป็นการวิเคราะห์ที่ไม่ซับซ้อนและผลการวิเคราะห์ที่ได้ตีความง่าย อย่างไรก็ตาม การวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้งเหมาะสมสำหรับกรณีซึ่งสนิปที่สัมพันธ์กับโรคซับซ้อนมีผลหลัก (Main Effect) หรือผลหนึ่งตำแหน่งที่ตั้งแบบขอบ (Marginal Single-Locus Effect) เท่านั้น ข้อจำกัดดังกล่าวทำให้มีโอกาสการไม่ตรวจจับบางสนิปที่สัมพันธ์กับโรคซับซ้อน ในทางตรงกันข้าม การวิเคราะห์ครั้งละหลายตำแหน่งที่ตั้งไม่มีข้อจำกัดดังกล่าว อย่างไรก็ตาม การวิเคราะห์ครั้งละหลายตำแหน่งที่ตั้งต้องใช้ทรัพยากรในการคำนวณมากกว่าการวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้งและผลการวิเคราะห์ที่ได้ตีความยากกว่าผลการวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้ง

การคัดเลือกสนิปที่สัมพันธ์กับโรคซับซ้อนโดยการวิเคราะห์ครั้งละหนึ่งตำแหน่งที่ตั้งและหลายตำแหน่งที่ตั้งสามารถพิจารณาเป็นการคัดเลือกลักษณะประจำ (Attribute Selection) หรือการคัดเลือกตัวแปร (Variable Selection) จากมุมมองการรู้จำแบบ (Pattern Recognition) [14]

นอกจากการคัดเลือกสนิปที่สัมพันธ์กับโรคซับซ้อนโดยตรงแล้วการวิเคราะห์บาทวิถี (Pathway Analysis) [15] เป็นอีกการวิเคราะห์ซึ่งได้รับความสนใจในการศึกษาความสัมพันธ์ทั้งจีโนม การวิเคราะห์บาทวิถีใช้การจัดกลุ่มสนิปสำหรับใช้เป็นตัวแทนยีนตามบาทวิถีชีวภาพ (Biological Pathway) และมีเป้าหมายคือการตรวจจับบาทวิถีชีวภาพที่สัมพันธ์กับโรคซับซ้อน ดังนั้น การวิเคราะห์บาทวิถีจึงสามารถพิจารณาเป็นการวิเคราะห์ครั้งละหลายตำแหน่งที่ตั้งซึ่งสนใจกลุ่มเฉพาะของสนิปสำหรับใช้เป็นตัวแทนยีนเท่านั้นและเป็นการคัดเลือกลักษณะประจำเช่นกัน

หลายเทคนิคการวิเคราะห์บาทวิถีสำหรับการศึกษาความสัมพันธ์ทั้งจีโนมได้รับการพัฒนาจากเทคนิคการวิเคราะห์บาทวิถีสำหรับการวิเคราะห์การแสดงออกของยีน (Gene Expression Analysis) [15] GSEA-SNP เป็นหนึ่งในเทคนิคดังกล่าว [16] โดย GSEA-SNP ได้รับการพัฒนาจากเทคนิคการวิเคราะห์การได้มากขึ้นจากเซตของยีน (Gene Set Enrichment Analysis หรือ GSEA) [17] ตามปกติแล้ว ข้อมูลการแสดงออกของยีนที่ได้จากหนึ่งเซตของโพรบ (Probeset) พอเพียงสำหรับการใช้เป็นตัวแทนหนึ่งยีนในการวิเคราะห์โดยใช้ GSEA ดังนั้นข้อมูลสนิปที่ได้จากหนึ่งสนิปจึงเพียงพอสำหรับการใช้เป็นตัวแทนหนึ่งยีนในการวิเคราะห์โดยใช้ GSEA-SNP

ถึงแม้ว่าการศึกษาความสัมพันธ์ทั้งจีโนมต้องพิจารณาข้อมูลสนิปจำนวนมาก แต่การวิเคราะห์บาทวิถีจำเป็นต้องใช้ข้อมูลหนึ่งสนิปที่อยู่ในหรือใกล้ยีนเพื่อใช้เป็นตัวแทนแต่ละยีนเท่านั้น ส่งผลให้มีข้อมูลสนิปจำนวนมากที่ไม่ได้ใช้ในการวิเคราะห์บาทวิถี ดังนั้นจึงมีความเป็นไปได้ว่าการใช้ข้อมูลสนิปตัวแทนซึ่งได้รับการคัดเลือกจากสนิปทั้งหมดในการศึกษาความสัมพันธ์ทั้งจีโนมพอเพียงสำหรับการวิเคราะห์บาทวิถี นั่นคือการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนมให้ผลการวิเคราะห์ไม่แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปตัวแทนเท่านั้น ภายใต้เงื่อนไขการมีอยู่ของข้อมูลความไม่สมดุลการเชื่อมโยงระหว่างสนิปตัวแทนและสนิปที่มีตัวแทน การทดสอบแนวคิดนี้เป็นประโยชน์ต่อการวิเคราะห์ข้อมูลสนิปจากการศึกษาความสัมพันธ์ทั้งจีโนมในฐานข้อมูลสาธารณะ เช่น Database of Genotypes and Phenotypes (dbGaP) [18] โดยเฉพาะเมื่อใช้สนิปชิปขนาด 317,000 และ 555,000 สนิปของ Illumina ในการเก็บข้อมูลจีโนไทป์ เนื่องจากข้อมูลสนิปที่ได้จากสนิปชิปของ Illumina มีสหสัมพันธ์กับข้อมูลสนิปที่ไม่ได้จากสนิปชิปดังที่กล่าวข้างต้น

งานวิจัยนี้สนใจการเปรียบเทียบระหว่างการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดและข้อมูลสนิปตัวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนม ข้อมูลที่ใช้ในการเปรียบเทียบคือข้อมูลจากการศึกษาความสัมพันธ์ทั้งจีโนมโดย Wellcome Trust Case Control Consortium (WTCCC) ซึ่งการเก็บข้อมูลจีโนไทป์ใช้สนิปชิปขนาด 500,000 สนิปของ Affymetrix [19] ส่งผลให้ได้ข้อมูลสนิปซึ่งมีลักษณะกระจายในจีโนมอย่างสุ่ม ดังนั้นการคัดเลือกสนิปตัวแทนจากสนิปทั้งหมดซึ่งกระทำโดยใช้ Tagger [20] จึงมีลักษณะไม่แตกต่างจากการคัดเลือกสนิปสำหรับการออกแบบสนิปชิปของ

Illumina การวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดและข้อมูลสนิปตัวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนมกระทำโดยใช้ GSEA-SNP บาทวิถี (Pathway) ที่สนใจคือบาทวิถีการให้สัญญาณ (Signalling Pathway) จาก Kyoto Encyclopedia of Genes and Genomes (KEGG) [21] ขั้นตอนที่ใช้ในงานวิจัยได้สรุปในรูปที่ 4

1.2 วัตถุประสงค์ของงานวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อทดสอบว่า การวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนมให้ผลการวิเคราะห์แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปตัวแทนเท่านั้นหรือไม่

1.3 ขอบเขตการวิจัย

1. เซตข้อมูล (Dataset) ที่ใช้ในการวิจัยคือเซตข้อมูลจากการศึกษาความสัมพันธ์ทั้งจีโนมของเจ็ดโรคซับซ้อนโดย WTCCC ซึ่งการเก็บข้อมูลจีโนไทป์ใช้สนิปชิป Affymetrix GeneChip Human Mapping 500K Array Set

2. ชุดการวัดเปรียบเทียบสมรรถนะ (Benchmark Suite) มีสามชุด ได้แก่ ชุดการวัดเปรียบเทียบสมรรถนะ 250K Nsp Array ที่สร้างจากเซตข้อมูลซึ่งการเก็บข้อมูลจีโนไทป์ใช้สนิปชิป Affymetrix GeneChip Human Mapping 250K Nsp Array (ส่วนหนึ่งของสนิปชิป Affymetrix GeneChip Human Mapping 500K Array Set) ชุดการวัดเปรียบเทียบสมรรถนะ 250K Sty Array ที่สร้างจากเซตข้อมูลซึ่งการเก็บข้อมูลจีโนไทป์ใช้สนิปชิป Affymetrix GeneChip Human Mapping 250K Sty Array (ส่วนที่เหลือของสนิปชิป Affymetrix GeneChip Human Mapping 500K Array Set) และชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set

3. การคัดเลือกสนิปตัวแทนกระทำโดยใช้ Tagger และสนใจขีดเริ่มเปลี่ยน r^2 (r^2 Threshold) เท่ากับ 0.8 และ 0.9 สำหรับการอธิบายความไม่สมดุลการเชื่อมโยงระหว่างคู่สนิป

4. การคัดเลือกสนิปสำหรับใช้เป็นตัวแทนยีนกระทำโดยการหาค่าสูงสุดของค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจ (Cochran-Armitage Trend Test Statistic)

5. การวิเคราะห์บาทวิถีกระทำโดยใช้ GESA-SNP และบาทวิถีที่สนใจคือบาทวิถีการให้สัญญาณจาก KEGG เท่านั้น

1.4 ขั้นตอนการดำเนินงาน

1. ตั้งสมมุติฐานและออกแบบการทดลอง
2. เตรียมข้อมูลและจัดการข้อมูล
3. คัดเลือกสนิปตัวแทนโดยใช้ Tagger
4. คัดเลือกสนิปสำหรับใช้เป็นตัวแทนแต่ละยีน
5. วิเคราะห์บาทวิถีโดยใช้ GSEA-SNP และวิเคราะห์ผลการวิจัย
6. สรุปผลการวิจัย
7. จัดเตรียมบทความสำหรับการประชุมวิชาการ
8. เรียบเรียงวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ได้ข้อสรุปจากการทดสอบว่า การวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนมให้ผลการวิเคราะห์แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปตัวแทนเท่านั้นหรือไม่ ถ้าการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดให้ผลการวิเคราะห์ไม่แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปตัวแทนเท่านั้น แล้วระเบียบวิธีวิจัย (Methodology) ที่นำเสนอจะเป็นประโยชน์ต่อการวิเคราะห์ข้อมูลสนิปจากการศึกษาความสัมพันธ์ทั้งจีโนมในฐานข้อมูลสาธารณะโดยเฉพาะเมื่อข้อมูลสนิปที่ได้จากสนิปชิปมีสหสัมพันธ์กับข้อมูลสนิปที่ไม่ได้จากสนิปชิป

รูปที่ 4 ขั้นตอนที่ใช้ในงานวิจัย

บทที่ 2
เซตข้อมูลและวิธีการวิจัย

2.1 เซตข้อมูลและการจัดการข้อมูล

เซตข้อมูลที่ใช้คือเจ็ดเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุม (Case-Control Dataset) จากการศึกษาความสัมพันธ์ทั้งจีโนมของเจ็ดโรคซับซ้อนโดย WTCCC แต่ละเซตข้อมูลประกอบด้วยตัวอย่างกลุ่มกรณีจากบุคคลเป็นโรคในสหราชอาณาจักรซึ่งเป็นหนึ่งในเจ็ดโรคซับซ้อน ได้แก่ โรคอารมณ์สองขั้ว (Bipolar Disorder หรือ BD) โรคหลอดเลือดแดงโคโรนารี (Coronary Artery Disease หรือ CAD) โรคโครห์น (Crohn's Disease หรือ CD) ความดันเลือดสูง (Hypertension หรือ HT) โรคข้ออักเสบรูมาตอยด์ (Rheumatoid Arthritis หรือ RA) เบาหวานชนิดที่ 1 (Type 1 Diabetes หรือ T1D) และเบาหวานชนิดที่ 2 (Type 2 Diabetes หรือ T2D) นอกจากนี้ แต่ละเซตข้อมูลประกอบด้วยตัวอย่างกลุ่มควบคุมจากบุคคลไม่เป็นโรค ตัวอย่างกลุ่มควบคุมประกอบด้วยตัวอย่างจากหน่วยบริการเลือดสหราชอาณาจักร (UK Blood Services หรือ NBS) และตัวอย่างจากบุคคลที่เกิดในสหราชอาณาจักรในปี ค.ศ. 1958 (British Birth Cohort หรือ 58C) จำนวนตัวอย่างของทั้งสองกลุ่มได้สรุปในตารางที่ 1

ทุกข้อมูลมี 469,612 สนิป การเก็บข้อมูลจีโนไทป์ใช้สนิปชิป Affymetrix GeneChip Human Mapping 500K Array Set ข้อมูลจีโนไทป์ผ่านการควบคุมคุณภาพโดย WTCCC [19] งานวิจัยนี้สนใจเฉพาะสนิปซึ่งมีค่าความถี่ส่วนน้อย (Minor Allele Frequency หรือ MAF) ในตัวอย่างกลุ่มควบคุมมากกว่าหรือเท่ากับ 0.05 และสามารถระบุตำแหน่งในจีโนมได้เท่านั้น ซึ่งส่งผลให้สามารถคำนวณค่า r^2 [22] สำหรับการอธิบายความไม่สมดุลการเชื่อมโยงระหว่างคู่สนิปอย่างมีความเชื่อถือได้ หลังจากกำจัดสนิปในข้อมูลซึ่งไม่สอดคล้องกับเงื่อนไขแล้วเหลือสนิปสำหรับการทดลองทั้งหมด 367,623 สนิป

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

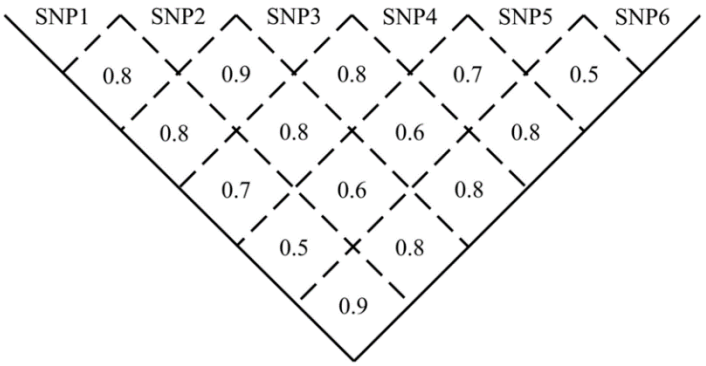
2.2 การคัดเลือกสนิปตัวแทนโดยใช้ Tagger

Tagger เป็นโปรแกรมสำหรับการคัดเลือกสนิปตัวแทนโดยไม่ใช้บล็อกของแฮ็ปโปลไทป์ (Haplotype Block-Free Approach) [20] Tagger สามารถคัดเลือกสนิปตัวแทนโดยการพิจารณาสหสัมพันธ์ระหว่างอัลลีล (Allele) ของคู่สนิป การคัดเลือกสนิปตัวแทนใช้ขั้นตอนวิธีละโมบ (Greedy Algorithm) ซึ่งอาศัยค่า r^2 สำหรับการอธิบายความไม่สมดุลการเชื่อมโยงระหว่างคู่สนิป ขั้นตอนวิธีละโมบเริ่มต้นด้วยการระบุสนิปตัวแทนซึ่งมีจำนวนสนิปเชื่อมโยง (Linked SNP) กับสนิปดังกล่าวสูงสุดโดยอิงจากขีดเริ่มเปลี่ยน r^2 สนิปตัวแทนนี้และสนิปเชื่อมโยงของสนิปตัวแทนนี้จะได้รับการรวมไว้ในหนึ่งผลแบ่งกั้น (Partition) ถ้ามีสนิปอื่นในผลแบ่งกั้นซึ่งเชื่อมโยงกับสนิปที่เหลือในผลแบ่งกั้นแล้วสนิปนี้จะเป็นสนิปตัวแทนเช่นกัน อย่างไรก็ตามหนึ่งสนิปตัวแทนเพียงพอสำหรับหนึ่งผลแบ่งกั้น จากนั้นขั้นตอนวิธีละโมบจะระบุสนิปตัวแทนจากสนิปที่เหลือในลักษณะเดียวกัน ถ้ามีสนิปซึ่งไม่เชื่อมโยงกับสนิปอื่น แล้วสนิปนี้จะเป็นสนิปตัวแทนซึ่งอยู่ในผลแบ่งกั้นของตัวเอง [23] ในงานวิจัยนี้ระยะทางสูงสุดระหว่างสนิปสำหรับการคำนวณค่า r^2 คือ 500 กิโลเบส (Kilobase)

พิจารณาตัวอย่างสำหรับแสดงการคัดเลือกสนิปตัวแทนโดยใช้ Tagger ในรูปที่ 5 ตัวอย่างนี้ประกอบด้วยหกสนิป ได้แก่ SNP1, SNP2, SNP3, SNP4, SNP5 และ SNP6 กำหนดให้ขีดเริ่มเปลี่ยน r^2 สำหรับการคัดเลือกสนิปตัวแทนเท่ากับ 0.8 ขั้นตอนวิธีละโมบคัดเลือก SNP2 เป็นสนิปตัวแทนแรกเพราะ SNP2 เชื่อมโยงกับ SNP1, SNP3, SNP4 และ SNP6 ส่งผลให้ผลแบ่งกั้นแรกประกอบด้วย SNP1, SNP2, SNP3, SNP4 และ SNP6 สนิปตัวแทนที่สองคือ SNP3 เพราะ SNP3 เชื่อมโยงกับ SNP1, SNP2, SNP4 และ SNP6 ดังนั้นจึงมีสนิปตัวแทนเดียวที่จำเป็นสำหรับผลแบ่งกั้นนี้ SNP5 เป็นสนิปตัวแทนสุดท้ายเพราะ SNP5 ไม่มีการเชื่อมโยงกับสนิปอื่น ดังนั้น SNP5 จึงเป็นสนิปตัวแทนเดียวในผลแบ่งกั้นที่สอง

ตารางที่ 2 การแจกแจงตัวอย่างกลุ่มกรณีและตัวอย่างกลุ่มควบคุมในเซตข้อมูลตามจีโนไทป์ที่สนิป

สถานะ	จีโนไทป์ที่สนิป			จำนวนตัวอย่าง
	จีโนไทป์ของพันธุ์ป่าโฮโมไซโกต	จีโนไทป์ของเฮเทอโรไซโกต	จีโนไทป์ของพันธุ์กลายโฮโมไซโกต	
กลุ่มกรณี	r_0	r_1	r_2	R
กลุ่มควบคุม	s_0	s_1	s_2	S
ทั้งหมด	n_0	n_1	n_2	N



รูปที่ 5 หกสนิปและค่า r-squared สำหรับการอธิบายความไม่สมดุลการเชื่อมโยงระหว่างคู่สนิป

2.3 การทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจ

การทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจ (Cochran-Armitage Trend Test หรือ CA Trend Test) เป็นหนึ่งในการทดสอบเชิงสถิติซึ่งได้รับความนิยมมากที่สุดในการศึกษาความสัมพันธ์ทางพันธุกรรม [24] พิจารณาเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมซึ่งมีจำนวนตัวอย่างตามจีโนไทป์ที่สนิปดังแสดงในตารางที่ 2 ค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจหรือ T_{CA} สามารถนิยามโดย

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

โดยที่ x_i เป็นตัวถ่วงน้ำหนักสำหรับจีโนไทป์ i สามแบบจำลองทางพันธุกรรม (Genetic Model) ที่สนใจในวิทยานิพนธ์นี้ ได้แก่ แบบจำลองลักษณะบวก (Additive Model) แบบจำลองลักษณะเด่น (Dominant Model) และแบบจำลองลักษณะด้อย (Recessive Model) สำหรับการทดสอบผลลักษณะบวก (Additive Effect) ตัวถ่วงน้ำหนัก $x_0 = 0, x_1 = 1$ และ $x_2 = 2$ ดังนั้นค่าสถิติทดสอบคือ

$$T_{CA}(add) = \frac{N}{R(N-R)} \frac{(N(r_1 + 2r_2) - R(n_1 + 2n_2))^2}{N(n_1 + 4n_2) - (n_1 + 2n_2)^2}$$

สำหรับการทดสอบผลลักษณะเด่น (Dominant Effect) ตัวถ่วงน้ำหนัก $x_0 = 0, x_1 = 1$ และ $x_2 = 1$ ดังนั้นค่าสถิติทดสอบคือ

$$T_{CA}(dom) = \frac{N}{R(N-R)} \frac{(N(r_1 + r_2) - R(n_1 + n_2))^2}{N(n_1 + n_2) - (n_1 + n_2)^2}$$

สำหรับการทดสอบผลลักษณะด้อย (Recessive Effect) ตัวถ่วงน้ำหนัก $x_0 = 0, x_1 = 0$ และ $x_2 = 1$ ดังนั้นค่าสถิติทดสอบคือ

$$T_{CA}(rec) = \frac{N}{R(N-R)} \frac{(Nr_2 - Rn_2)^2}{Nn_2 - {n_2}^2}$$

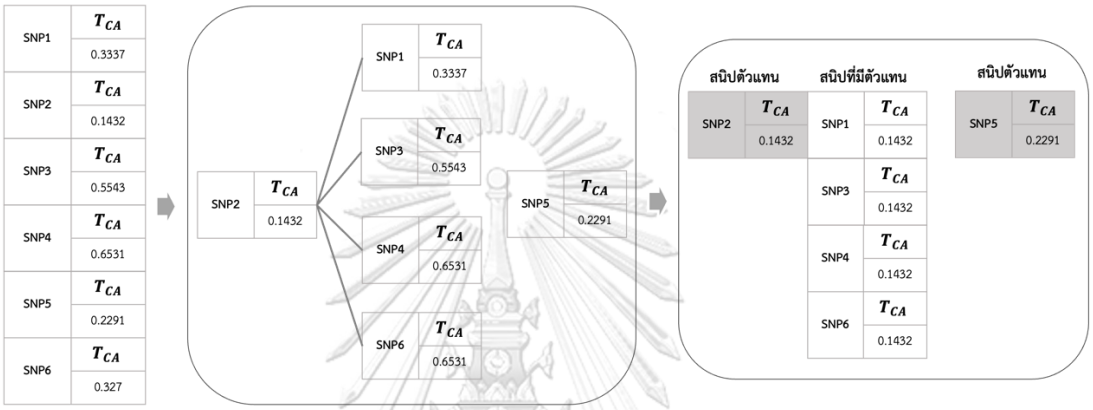
ค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจเป็นไปตามการแจกแจงไคกำลังสอง (χ^2 Distribution) ซึ่งมีหนึ่งระดับขั้นความเสรี (Degree of Freedom)

2.4 การคัดเลือกสนิปสำหรับใช้เป็นตัวแทนยีน

ตามปกติแล้วมีหลายสนิปที่อยู่ในหรือใกล้ยีน [25] แนะนำว่าสนิปซึ่งมีค่าสถิติทดสอบสุดขีด (Extreme) ที่สุดเมื่อเทียบกับสนิปที่อยู่ในหรือใกล้ยีนเดียวกันสามารถใช้เป็นตัวแทนยีนในการศึกษากลุ่มกรณี-กลุ่มควบคุม ค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจคือค่าสถิติทดสอบที่ใช้ในงานวิจัยนี้ สามค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับการทดสอบผลลักษณะบวก ผลลักษณะเด่นและผลลักษณะด้อยจะได้รับการคำนวณสำหรับแต่ละสนิปในเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุม แบบจำลองทางพันธุกรรมที่ได้รับการเลือกสำหรับแต่ละสนิปคือแบบจำลองทางพันธุกรรมที่ให้ค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสูงสุด กรณีที่สนใจข้อมูลสนิปทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนม ค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับแต่ละสนิปต้องได้รับการคำนวณ ในทางตรงกันข้ามกรณีที่สนใจข้อมูลสนิปตัวแทนเท่านั้นค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับสนิปที่มีตัวแทนซึ่งเชื่อมโยงกับสนิปตัวแทนและเปรียบเสมือนสนิปที่ไม่ได้จากสนิปชิปจะเท่ากับค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับสนิปตัวแทน อ้างอิงจากรูปที่ 5 รูปที่ 6 แสดงผลการคัดเลือกสนิปตัวแทนจากหกสนิป จะเห็นได้ว่าค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับสนิปตัวแทนจะใช้เป็นค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับสนิปที่มีตัวแทนโดยไม่สนใจว่าค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับสนิปตัวแทนมีค่ามากกว่าหรือน้อยกว่าค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับสนิปที่มีตัวแทน

สนิปสำหรับใช้เป็นตัวแทนยีนคือสนิปซึ่งมีค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจที่ได้รับการเลือกสูงสุดเมื่อเทียบกับสนิปที่อยู่ในหรือใกล้ยีนเดียวกัน สนิปที่อยู่ใกล้ยีนคือสนิปที่มีตำแหน่งไม่เกิน 500 กิโลเบสเมื่อนับย้อนหลังจากตำแหน่งเริ่มการถอดรหัส (Transcription Start Site) หรือเมื่อนับไปข้างหน้าจากตำแหน่งเลิกการถอดรหัส (Transcription Termination Site) [25, 26] การกำหนดขีดเริ่มเปลี่ยนระยะทางระหว่างตำแหน่งในจีโนมข้างต้นสอดคล้องกับข้อแนะนำการระบุ

ยีนให้กับสนิปในการวิเคราะห์บาทวิถีสำหรับการศึกษาความสัมพันธ์ทั้งจีโนม [27] 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 [28]



รูปที่ 6 สนิปตัวแทนที่ได้รับการคัดเลือกจากสนิปในรูปที่ 5 และการกำหนดค่าทดสอบสถิติแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับสนิปที่มีตัวแทน

2.5 GSEA-SNP

GSEA-SNP เป็นเทคนิคที่ได้รับการพัฒนาต่อจากเทคนิคการวิเคราะห์การได้มากขึ้นจากเซตของยีนหรือ GSEA สำหรับการวิเคราะห์การแสดงออกของยีน [17] การพัฒนาดังกล่าวส่งผลให้ GSEA-SNP เหมาะสมสำหรับการศึกษาความสัมพันธ์ทั้งจีโนม [16] GSEA-SNP สามารถระบุว่าเซตของยีน (Gene Set) ในบาทวิถีสัมพันธ์กับโรคซับซ้อนที่สนใจอย่างมีนัยสำคัญทางสถิติหรือไม่โดยใช้การคำนวณคะแนนการได้มากขึ้น (Enrichment Score) และการทดสอบการเรียงสับเปลี่ยน (Permutation Test) การทำงานของ GSEA-SNP สามารถอธิบายได้ดังนี้

พิจารณาเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมซึ่งประกอบด้วยหลายสนิปจาก N_G ยีน ถึงแม้ว่าการวิเคราะห์เซตข้อมูลซึ่งแต่ละยีนมีหลายสนิปสามารถกระทำได้โดยใช้ GSEA-SNP ในงานวิจัยนี้แต่ละยีนมีหนึ่งสนิปซึ่งได้รับการคัดเลือกสำหรับใช้เป็นตัวแทนยีนดังที่กล่าวข้างต้น [25] ยีนทั้งหมด N_G ยีนจะได้รับการเรียงลำดับตามค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจของสนิปสำหรับใช้เป็น

ตัวแทนยีนจากค่าสูงสุดไปค่าต่ำสุด สำหรับเซตของยีน L ซึ่งประกอบด้วย N_L ยีน คะแนนการได้มากขึ้นสำหรับเซตของยีนนี้หรือ $ES(L)$ สามารถนิยามโดย

$$ES(L) = \sum_{\substack{g_j \in L \\ j \le i^*}} \frac{c_j^\alpha}{N_C} - \sum_{\substack{g_j \notin L \\ j \le i^*}} \frac{1}{N_G - N_L}$$

โดยที่

$$i^* = \underset{1 \le i \le N_G}{\operatorname{argmax}} \left| \sum_{\substack{g_j \in L \\ j \le i}} \frac{c_j^\alpha}{N_C} - \sum_{\substack{g_j \notin L \\ j \le i}} \frac{1}{N_G - N_L} \right|$$

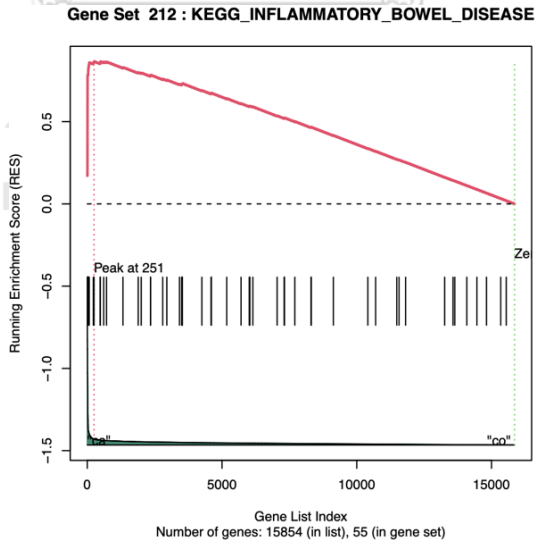
c_j คือค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจของสนิปสำหรับใช้เป็นตัวแทนยีน g_j, α คือ พารามิเตอร์ถ่วงน้ำหนักซึ่งได้รับการกำหนดให้เท่ากับ 1 [17, 25] และ $N_C = \sum_{g_j \in L} c_j^\alpha$ คะแนนการได้มากขึ้นสะท้อนถึงค่าเบี่ยงเบนสูงสุดจากศูนย์ของผลรวมค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสำหรับยีนในเซตของยีนที่ได้ระหว่างการแวะผ่าน (Traversal) ตามรายการยีน (Gene List) ซึ่งได้รับการเรียงลำดับตามค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจ รูปที่ 7 แสดงผลการคำนวณคะแนนการได้มากขึ้นและคะแนนการได้มากขึ้นมีค่าบวก ในขณะที่รูปที่ 8 แสดงผลการคำนวณคะแนนการได้มากขึ้นและคะแนนการได้มากขึ้นมีค่าลบ เส้นทึบแสดงการแวะผ่านตามรายการยีน ในขณะที่เส้นประแสดงดัชนีรายการยีน (Gene List Index) i^* สำหรับการคำนวณคะแนนการได้มากขึ้น

หลังจากการคำนวณคะแนนการได้มากขึ้น การทดสอบว่าเซตของยีนสัมพันธ์กับโรคซับซ้อนที่สนใจอย่างมีนัยสำคัญทางสถิติหรือไม่สามารถกระทำโดยใช้การทดสอบการเรียงสับเปลี่ยน การทดสอบการเรียงสับเปลี่ยนในงานวิจัยนี้ใช้ 1,000 เซตข้อมูลเรียงสับเปลี่ยน (Permutation Replicate) ซึ่งแต่ละเซตข้อมูลเรียงสับเปลี่ยนสร้างจากเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมที่ได้รับการเรียงสับเปลี่ยนเชิงสุ่มสถานะกรณีและสถานะควบคุมของตัวอย่างในเซตข้อมูลในขณะที่จำนวนตัวอย่างกลุ่มกรณีและจำนวนตัวอย่างกลุ่มควบคุมเป็นจำนวนเดิม จากนั้นคะแนนการได้มากขึ้นจะได้รับการคำนวณโดยใช้แต่ละเซตข้อมูลเรียงสับเปลี่ยน ค่าความน่าจะเป็นหรือค่าพี (p-value) ที่ได้จาก GSEA-SNP คือผลหารระหว่างจำนวนเซตข้อมูลเรียงสับเปลี่ยนซึ่งคะแนนการได้มากขึ้นสุดขีดกว่าหรือเท่ากับคะแนนการได้มากขึ้นซึ่งคำนวณโดยใช้เซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมและจำนวนเซตข้อมูลเรียงสับเปลี่ยนทั้งหมด

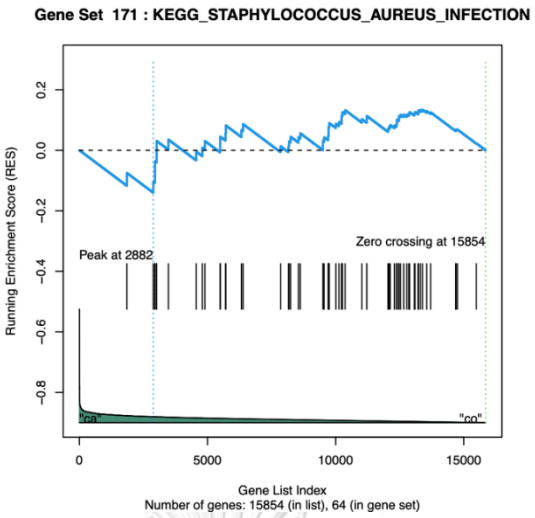
ตามปกติแล้ว หลายเซตของยีนจะได้รับการพิจารณาว่าแต่ละเซตของยีนสัมพันธ์กับโรคซับซ้อนที่สนใจอย่างมีนัยสำคัญทางสถิติหรือไม่ ดังนั้นการแก้สำหรับการทดสอบหลายสมมุติฐาน (Correction for Multiple Hypothesis Testing) จึงจำเป็นสำหรับ GSEA-SNP ในงานวิจัยนี้ อัตราการค้นพบเท็จ (False Discovery Rate หรือ FDR) เป็นค่าที่สนใจหลังการแก้สำหรับการทดสอบหลายสมมุติฐาน อัตราการค้นพบเท็จสามารถคำนวณโดยใช้เซตข้อมูลเรียงสับเปลี่ยนดังนี้ ค่าเฉลี่ยของคะแนนการได้มากขึ้นซึ่งคำนวณโดยใช้ทุกเซตข้อมูลเรียงสับเปลี่ยนจะได้รับการคำนวณสำหรับแต่ละเซตของยีน จากนั้นคะแนนการได้มากขึ้นซึ่งคำนวณโดยใช้เซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมและเซตข้อมูลเรียงสับเปลี่ยนจะได้รับการทำให้เป็นบรรทัดฐาน (Normalisation) โดยการหารด้วยค่าเฉลี่ยอัตราการค้นพบเท็จสำหรับเซตของยีน L^* ที่สนใจหรือ $FDR(L^*)$ สามารถคำนวณได้จาก

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

โดยที่ π คือตัวแปรที่ใช้ระบุเซตข้อมูลเรียงสับเปลี่ยน NES คือคะแนนการได้มากขึ้นที่ได้รับการทำให้เป็นบรรทัดฐาน (Normalized Enrichment Score) และ NES^* คือคะแนนการได้มากขึ้นที่ได้รับการทำให้เป็นบรรทัดฐานสำหรับเซตของยีน L^* [25]



รูปที่ 7 ผลการคำนวณคะแนนการได้มากขึ้นและคะแนนการได้มากขึ้นมีค่าบวก



รูปที่ 8 ผลการคำนวณคะแนนการได้มากขึ้นและคะแนนการได้มากขึ้นมีค่าลบ

2.6 บาทวิถีการให้สัญญาณและบาทวิถีเป้าหมาย

เซตของยีนในบาทวิถีที่สนใจในวิทยานิพนธ์นี้คือเซตของยีนในบาทวิถีการให้สัญญาณจาก KEGG มีบาทวิถีการให้สัญญาณทั้งหมด 223 บาทวิถี นอกจากนี้ หลักฐานการศึกษาความสัมพันธ์ทางพันธุกรรมแสดงให้เห็นว่า บางบาทวิถีการให้สัญญาณสัมพันธ์กับแต่ละโรคซับซ้อนที่สนใจ [21, 29] บาทวิถีการให้สัญญาณเหล่านี้คือบาทวิถีเป้าหมาย (Target Pathway) สำหรับการทดสอบสมรรถนะของ GSEA-SNP ในการระบุว่าเซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับแต่ละโรคซับซ้อนอย่างมีนัยสำคัญทางสถิติ บาทวิถีเป้าหมายสำหรับแต่ละโรคซับซ้อนได้แสดงในตารางที่ 3

สังเกตว่าไม่มีบาทวิถีการให้สัญญาณที่สัมพันธ์กับโรคอารมณ์สองขั้วใน KEGG [21] นอกจากนี้ไม่มียีนที่สัมพันธ์กับโรคอารมณ์สองขั้วใน KEGG เช่นกัน [21] อย่างไรก็ตาม การวิเคราะห์บาทวิถีโดยใช้เซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมจากการศึกษาความสัมพันธ์ทั้งจีโนมของโรคอารมณ์สองขั้วโดย WTCCC แสดงให้เห็นว่ามีสองบาทวิถีการให้สัญญา ได้แก่ บาทวิถี Cell Adhesion Molecules (CAMs) (hsa04514) และบาทวิถี Tight Junction (hsa04530) ที่สัมพันธ์โรคอารมณ์สองขั้ว [29] ดังนั้นสองบาทวิถีนี้จึงเป็นบาทวิถีเป้าหมายสำหรับโรคอารมณ์สองขั้ว

ตารางที่ 3 บาทวิถีการให้สัญญาณจาก KEGG ซึ่งเป็นบาทวิถีเป้าหมายสำหรับแต่ละโรคซับซ้อน

โรคซับซ้อน	บาทวิถีเป้าหมาย		
	KEGG ID	บาทวิถีการให้สัญญาณ	บรรณานุกรม
BD	hsa04514	Cell adhesion molecules (CAMs)	O' Dushlaine *et al.* [29]
	hsa04530	Tight junction	O' Dushlaine *et al.* [29]
CAD	hsa04022	cGMP-PKG signalling pathway	KEGG [21]
	hsa04310	Wnt signalling pathway	KEGG [21]
	hsa04928	Parathyroid hormone synthesis, secretion and action	KEGG [21]
CD	hsa04060	Cytokine-cytokine receptor interaction	KEGG [21]
	hsa04140	Regulation of autophagy	KEGG [21]
	hsa04621	NOD-like receptor signalling pathway	KEGG [21]
	hsa04630	Jak-STAT signalling pathway	KEGG [21]
	hsa05321	Inflammatory bowel disease	KEGG [21]
HT	hsa04925	Aldosterone synthesis and secretion	KEGG [21]
	hsa04960	Aldosterone-regulated sodium reabsorption	KEGG [21]
RA	hsa05323	Rheumatoid arthritis	KEGG [21]
T1D	hsa04060	Cytokine-cytokine receptor interaction	KEGG [21]
	hsa04151	PI3K-Akt signalling pathway	KEGG [21]
	hsa04612	Antigen processing and presentation	KEGG [21]
	hsa04630	Jak-STAT signalling pathway	KEGG [21]
	hsa04940	Type I diabetes mellitus	KEGG [21]
T2D	hsa03320	PPAR signalling pathway	KEGG [21]
	hsa04110	Cell cycle	KEGG [21]
	hsa04115	p53 signalling pathway	KEGG [21]
	hsa04141	Protein processing in endoplasmic reticulum	KEGG [21]
	hsa04310	Wnt signalling pathway	KEGG [21]
	hsa04330	Notch signalling pathway	KEGG [21]
	hsa04350	TGF-beta signalling pathway	KEGG [21]
	hsa04911	Insulin secretion	KEGG [21]
	hsa04930	Type II diabetes mellitus	KEGG [21]
	hsa04972	Pancreatic secretion	KEGG [21]

บทที่ 3
ผลการวิจัยและอภิปรายผลการวิจัย

ในงานวิจัยนี้ ชุดการวัดเปรียบเทียบสมรรถนะ (Benchmark Suite) ได้สร้างจากเจ็ดเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมจากการศึกษาความสัมพันธ์ทั้งจีโนมของเจ็ดโรคซับซ้อนโดย WTCCC เนื่องจากสนิปชิป Affymetrix GeneChip Human Mapping 500K Array Set ประกอบด้วยสนิปชิป Affymetrix GeneChip Human Mapping 250K Nsp Array และสนิปชิป Affymetrix GeneChip Human Mapping 250K Sty Array สนิปในแต่ละเซตข้อมูลจึงสามารถแบ่งเป็นสองส่วนไม่ซ้อนเหลื่อมส่งผลให้มีสามชุดการวัดเปรียบเทียบสมรรถนะ ได้แก่ ชุดการวัดเปรียบเทียบสมรรถนะ 250K Nsp Array ชุดการวัดเปรียบเทียบสมรรถนะ 250K Sty Array และชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set ขีดเริ่มเปลี่ยน r^2 สำหรับการคัดเลือกสนิปตัวแทนจากสนิปในตัวอย่างกลุ่มควบคุมโดยใช้ Tagger ที่สนใจคือ 0.8 และ 0.9 จำนวนสนิปทั้งหมดและจำนวนสนิปตัวแทนในแต่ละชุดการวัดเปรียบเทียบสมรรถนะได้แสดงในตารางที่ 4

ดังที่กล่าวข้างต้น สนิปซึ่งมีค่าสถิติทดสอบแนวโน้มเอียงคอคราน-อาร์มิเทจสูงสุดเมื่อเทียบกับสนิปที่อยู่ในหรือใกล้ยีนเดียวกันจะได้รับการคัดเลือกสำหรับใช้เป็นตัวแทนยีน เงื่อนไขดังกล่าวส่งผลให้ไม่สามารถระบุสนิปสำหรับใช้เป็นตัวแทนยีนได้ครบทุกยีน ดังนั้นจำนวนสนิปสำหรับใช้เป็นตัวแทนยีนรวมทั้งจำนวนยีนที่มีสนิปสำหรับใช้เป็นตัวแทนยีนจึงน้อยกว่าจำนวนยีนที่ระบุในไฟล์บรรณนิทัศน์ของ NetAffx จำนวนสนิปสำหรับใช้เป็นตัวแทนยีน จำนวนยีนที่มีสนิปสำหรับใช้เป็นตัวแทนยีน และจำนวนยีนที่ระบุในไฟล์บรรณนิทัศน์ของ NetAffx ในแต่ละชุดการวัดเปรียบเทียบสมรรถนะได้แสดงในตารางที่ 5 สังเกตว่า มากกว่าหนึ่งยีนมีสนิปสำหรับใช้เป็นตัวแทนยีนเป็นสนิปเดียวกัน ส่งผลให้จำนวนสนิปสำหรับใช้เป็นตัวแทนยีนน้อยกว่าจำนวนยีนที่มีสนิปสำหรับใช้เป็นตัวแทนยีน นอกจากนี้ เนื่องจากการคัดเลือกสนิปตัวแทนไม่มีผลต่อจำนวนสนิปสำหรับใช้เป็นตัวแทนยีน จำนวนสนิปสำหรับใช้เป็นตัวแทนยีนจึงมีจำนวนเท่ากันไม่ว่าจะสนใจสนิปทั้งหมดหรือสนิปตัวแทนในแต่ละชุดการวัดเปรียบเทียบสมรรถนะ

ตารางที่ 4 จำนวนสนิปทั้งหมดและจำนวนสนิปตัวแทนในแต่ละชุดการวัดเปรียบเทียบสมรรถนะ

ชุดการวัดเปรียบเทียบสมรรถนะ	จำนวนสนิปทั้งหมด	จำนวนสนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน r^2 = 0.9	จำนวนสนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน 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 จำนวนสนิปสำหรับใช้เป็นตัวแทนยีน จำนวนยีนที่มีสนิปสำหรับใช้เป็นตัวแทนยีน และจำนวนยีนที่ระบุในไฟล์บรรณนิทัศน์ของ NetAffx ในแต่ละชุดการวัดเปรียบเทียบสมรรถนะ

ชุดการวัดเปรียบเทียบสมรรถนะ	จำนวนสนิปสำหรับใช้เป็นตัวแทนยีน	จำนวนยีนที่มีสนิปสำหรับใช้เป็นตัวแทนยีน	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

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

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

โรคซับซ้อน	บาทวิถีเป้าหมายที่ระบุจากการใช้สนิปทั้งหมด	บาทวิถีเป้าหมายที่ระบุจากการใช้สนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน r^2 = 0.9	บาทวิถีเป้าหมายที่ระบุจากการใช้สนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน 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	บาทวิถีเป้าหมายที่ระบุจากการใช้สนิปทั้งหมด	บาทวิถีเป้าหมายที่ระบุจากการใช้สนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน r^2 = 0.9	บาทวิถีเป้าหมายที่ระบุจากการใช้สนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน 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	บาทวิถีเป้าหมายที่ระบุจากการใช้สนิปทั้งหมด	บาทวิถีเป้าหมายที่ระบุจากการใช้สนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน r^2 = 0.9	บาทวิถีเป้าหมายที่ระบุจากการใช้สนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน r^2 = 0.8
BD	-	-	-
CAD	-	-	-
CD	hsa05321	-	-
HT	-	-	-
RA	hsa05323	hsa05323	hsa05323
T1D	hsa04612, hsa04940	hsa04612, hsa04940	hsa04612, hsa04940
T2D	-	-	-

เนื่องจาก GSEA-SNP ใช้การทดสอบการเรียงสับเปลี่ยนในการประเมินค่าพีและอัตราการค้นพบเท็จ ดังนั้นการวิเคราะห์โดยใช้ GSEA-SNP จึงได้รับการทดลองซ้ำเพื่อทดสอบว่าการสุ่มในการทดสอบการเรียงสับเปลี่ยนไม่มีผลต่อการระบุเซตของยีนในบาทวิถีซึ่งสัมพันธ์กับโรคซับซ้อนที่สนใจอย่างมีนัยสำคัญทางสถิติ การทดลองซ้ำใช้สนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน 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

การวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะโดยใช้ GSEA-SNP แสดงให้เห็นว่าภายใต้เงื่อนไขการมีอยู่ของข้อมูลความไม่สมดุลการเชื่อมโยงระหว่างสนิปตัวแทนและสนิปที่มีตัวแทนอย่างสมบูรณ์ โดยรวมการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนมให้ผลการวิเคราะห์ไม่แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปตัวแทนเท่านั้น เพื่อยืนยันข้อสังเกตนี้ค่าเฉลี่ยของอัตราการค้นพบเท็จสำหรับ 223 เซตของยีนในบาทวิถีจากการวิเคราะห์แต่ละชุดการวัดเปรียบเทียบสมรรถนะโดยใช้ GSEA-SNP จะได้รับการใช้ในการทดสอบสมมุติฐานว่าการคัดเลือกสนิปสำหรับใช้เป็นตัวแทนยีนไม่มีผลต่ออัตราการค้นพบเท็จสำหรับเซตของยีนในบาทวิถีโดยการทดสอบฟรีดแมน (Friedman Test) กำหนดให้ตัวแปรกลุ่ม (Group Variable) คือสนิป

สำหรับใช้เป็นตัวแทนยีน นั่นคือสามกลุ่มที่สนใจ ได้แก่ สนิปสำหรับใช้เป็นตัวแทนยีนซึ่งได้รับการคัดเลือกจากสนิปทั้งหมด สนิปสำหรับใช้เป็นตัวแทนยีนซึ่งได้รับการคัดเลือกจากสนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน r^2 = 0.9 และสนิปสำหรับใช้เป็นตัวแทนยีนซึ่งได้รับการคัดเลือกจากสนิปตัวแทนเมื่อขีดเริ่มเปลี่ยน r^2 = 0.8 กำหนดให้ตัวแปรการจัดเป็นกลุ่มระเบียน (Blocking Variable) คือเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุม นั่นคือเจ็ดเซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมจากการศึกษาความสัมพันธ์ทั้งจีโนมของเจ็ดโรคซับซ้อน ผลการทดสอบฟรีดแมนในตารางที่ 9 แสดงให้เห็นว่าไม่สามารถปฏิเสธสมมุติฐานว่าง (Null Hypothesis) ที่ระดับนัยสำคัญ (Significance Level) 0.05

ตารางที่ 9 ผลการทดสอบสมมุติฐานว่า การคัดเลือกสนิปสำหรับใช้เป็นตัวแทนยีนไม่มีผลต่ออัตราการค้นพบเท็จสำหรับเซตของยีนในบาทวิถีโดยการทดสอบฟรีดแมน

ชุดการวัดเปรียบเทียบสมรรถนะ	ค่าสถิติทดสอบไคกำลังสอง (χ^2 Test Statistic)	ระดับขั้นความเสรี	ค่าพี
250K Nsp Array	3.7143	2	0.1561
250K Sty Array	1.0000	2	0.6065
500K Array Set	0.0741	2	0.9636

อย่างไรก็ตาม การวิเคราะห์เซตข้อมูลกลุ่มกรณี-กลุ่มควบคุมซึ่งกลุ่มกรณีคือกลุ่มบุคคลเป็นโรคโครห์นในชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set แสดงให้เห็นว่ากรณีที่สนใจข้อมูลสนิปตัวแทนเท่านั้น สมรรถนะในการระบุเซตของยีนในบาทวิถีเป้าหมายที่สัมพันธ์กับโรคซับซ้อนของ GSEA-SNP ลดลง ข้อสังเกตดังกล่าวสามารถอธิบายได้โดยพิจารณาสนิปชิป Affymetrix GeneChip Human Mapping 500K Array Set ซึ่งประกอบด้วยสนิปชิป Affymetrix GeneChip Human Mapping 250K Nsp Array และสนิปชิป Affymetrix GeneChip Human Mapping 250K Sty Array ดังนี้ GSEA-SNP สามารถระบุว่าเซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับโรคโครห์นจากผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 250K Sty Array ในทางตรงกันข้าม GSEA-SNP ไม่สามารถระบุว่าเซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับโรคโครห์นจากผลการวิเคราะห์ชุดการวัดเปรียบเทียบสมรรถนะ 250K Nsp Array นั่นคือสนิปที่ได้จากสนิปชิป Affymetrix GeneChip Human Mapping 250K Sty Array จำเป็นสำหรับการใช้เป็นตัวแทนยีนในการระบุว่าเซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับโรคโครห์น ดังนั้นกรณีที่สนใจข้อมูลสนิปทั้งหมดในชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set, GSEA-SNP สามารถระบุว่าเซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับโรคโครห์นเพราะมีสนิปสำหรับใช้เป็นตัวแทนยีนเป็นสนิปที่ได้จากสนิปชิป Affymetrix GeneChip

Human Mapping 250K Sty Array เพียงพอ ในทางตรงกันข้ามกรณีที่สนใจข้อมูลสนิปตัวแทนเท่านั้นในชุดการวัดเปรียบเทียบสมรรถนะ 500K Array Set, GSEA-SNP ไม่สามารถระบุว่าเซตของยีนในบาทวิถีเป้าหมายสัมพันธ์กับโรคโครห์นเพราะขาดสนิปสำหรับใช้เป็นตัวแทนยีนเป็นสนิปที่ได้จากสนิปชิป Affymetrix GeneChip Human Mapping 250K Sty Array ที่จำเป็น ดังนั้นสนิปชิปที่ใช้ในการเก็บข้อมูลจีโนไทป์มีผลต่อข้อสังเกตข้างต้น

บทที่ 4
สรุปผลการวิจัยและข้อเสนอแนะ

4.1 สรุปผลการวิจัย

ในงานวิจัยนี้ การวิเคราะห์การได้มากขึ้นจากเซตของยีนในบาทวิถีการให้สัญญาณจากฐานข้อมูล KEGG โดยใช้ข้อมูลสนิปทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนมให้ผลการวิเคราะห์ไม่แตกต่างจากการวิเคราะห์การได้มากขึ้นจากเซตของยีนในบาทวิถีการให้สัญญาณโดยใช้ข้อมูลสนิปตัวแทนเท่านั้น

4.2 ข้อเสนอแนะ

ในงานวิจัยนี้ การคัดเลือกสนิปตัวแทนใช้สนิปในตัวอย่างกลุ่มควบคุม ส่งผลให้มีข้อมูลความไม่สมดุลการเชื่อมโยงระหว่างสนิปตัวแทนและสนิปที่มีตัวแทนอย่างสมบูรณ์ ตามปกติแล้ว สนิปตัวแทนที่ใช้ในการออกแบบสนิปชิปโดยอาศัยสนิปตัวแทนจะได้จากการคัดเลือกสนิปในแผงสนิปอ้างอิง (Reference SNP Panel) ของประชากรเดียวกับหรือใกล้เคียงกับประชากรที่สนใจในการศึกษาความสัมพันธ์ทั้งจีโนม ข้อแตกต่างระหว่างการแจกแจงความถี่อัลลีล (Allele Frequency Distribution) ของสนิปในแผงสนิปอ้างอิงและสนิปในตัวอย่างกลุ่มควบคุมส่งผลต่อข้อมูลความไม่สมดุลการเชื่อมโยงระหว่าง สนิปตัวแทนและสนิปที่มีตัวแทน ดังนั้นการศึกษาผลของปัจจัยดังกล่าวต่อการทดสอบแนวคิดที่ว่า การวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปทั้งหมดจากการศึกษาความสัมพันธ์ทั้งจีโนมให้ผลการวิเคราะห์ไม่แตกต่างจากการวิเคราะห์บาทวิถีโดยใช้ข้อมูลสนิปตัวแทนจึงต้องได้รับการศึกษา

เนื่องจากการวิเคราะห์บาทวิถีในการศึกษาความสัมพันธ์ทั้งจีโนมจำเป็นต้องใช้สนิปตัวแทนเท่านั้น สนิปที่จำเป็นสำหรับการออกแบบสนิปชิปเพื่อการวิเคราะห์บาทวิถีคือสนิปตัวแทน สนิปตัวแทนที่ได้จากสนิปชิปสามารถใช้เป็นตัวแทนยีนโดยตรง ในทางตรงกันข้าม สนิปที่มีตัวแทนที่ไม่ได้จากสนิปชิปสามารถใช้เป็นตัวแทนยีนภายใต้เงื่อนไขการมีอยู่ของข้อมูลความไม่สมดุลการเชื่อมโยงระหว่างสนิปตัวแทนและสนิปที่มีตัวแทน นั่นคือสนิปตัวแทนสามารถใช้เป็นทั้งตัวแทนยีนโดยตรงและตัวแทนยีนโดยอ้อม ส่งผลให้มีข้อมูลสนิปสำหรับใช้เป็นตัวแทนยีนเพิ่มขึ้น นอกจากนี้ การออกแบบสนิปชิปโดยอาศัยสนิปตัวแทนเพื่อการวิเคราะห์บาทวิถีส่งผลให้จำนวนสนิปต่อหนึ่งตัวอย่างลดลง ดังนั้นการศึกษาความสัมพันธ์ทั้งจีโนมซึ่งสนใจเฉพาะการวิเคราะห์บาทวิถีจึงสามารถเพิ่มจำนวนตัวอย่างโดยไม่เพิ่มค่าใช้จ่ายการเก็บข้อมูลจีโนไทป์

บรรณานุกรม

[1] A. J. Brookes, "The essence of SNPs," *Gene,* vol. 234, no. 2, pp. 177-186, 1999/07/08/ 1999, doi: 10.1016/S0378-1119(99)00219-X.

[2] A. Raj, M. Stephens, and J. K. Pritchard, "fastSTRUCTURE: Variational inference of population structure in large SNP data sets," *Genetics,* vol. 197, no. 2, pp. 573-589, 2014, doi: 10.1534/genetics.114.164350.

[3] D. Setsirichok *et al.*, "Small ancestry informative marker panels for complete classification between the original four HapMap populations," *International Journal of Data Mining and Bioinformatics,* vol. 6, no. 6, pp. 651-674, 2012, doi: 10.1504/IJDMB.2012.050249.

[4] A. G. Heidema, J. M. A. Boer, N. Nagelkerke, E. C. M. Mariman, D. L. van der A, and E. J. M. Feskens, "The challenge for genetic epidemiologists: How to analyze large numbers of SNPs in relation to complex diseases," *BMC Genetics,* vol. 7, 2006, doi: 10.1186/1471-2156-7-23.

[5] C. M. Lewis, "Genetic association studies: Design, analysis and interpretation," *Briefings in Bioinformatics,* vol. 3, no. 2, pp. 146-153, 2002, doi: 10.1093/bib/3.2.146.

[6] G. Montana, "Statistical methods in genetics," *Briefings in Bioinformatics,* vol. 7, no. 3, pp. 297-308, 2006, doi: 10.1093/bib/bbl028.

[7] K. Van steen, "Travelling the world of gene-gene interactions," *Briefings in Bioinformatics,* vol. 13, no. 1, pp. 1-19, 2012, doi: 10.1093/bib/bbr012.

[8] N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science,* vol. 273, no. 5281, pp. 1516-1517, 1996, doi: 10.1126/science.273.5281.1516.

[9] A. Ziegler and Y. V. Sun, "Study designs and methods post genome-wide association studies," *Human Genetics,* vol. 131, no. 10, pp. 1525-1531, 2012/10/01 2012, doi: 10.1007/s00439-012-1209-8.

[10] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, "Benefits and limitations of genome-wide association studies," *Nature Reviews Genetics,* vol.

20, no. 8, pp. 467-484, 2019, doi: 10.1038/s41576-019-0127-1.

[11] G. Diao and A. N. Vidyashankar, "Assessing genome-wide statistical significance for large p small n problems," *Genetics,* vol. 194, no. 3, pp. 781-783, 2013, doi: 10.1534/genetics.113.150896.

[12] The International HapMap Consortium, "A haplotype map of the human genome," *Nature,* vol. 437, no. 7063, pp. 1299-320, Oct 27 2005, doi: 10.1038/nature04226.

[13] C. Wallace, R. J. Dobson, P. B. Munroe, and M. J. Caulfield, "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 Research,* vol. 17, no. 11, pp. 1596-1602, 2007, doi: 10.1101/gr.5996407.

[14] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics,* vol. 23, no. 19, pp. 2507-2517, 2007, doi: 10.1093/bioinformatics/btm344.

[15] K. Wang, M. Li, and H. Hakonarson, "Analysing biological pathways in genome-wide association studies," *Nature Reviews Genetics,* vol. 11, no. 12, pp. 843-854, 2010, doi: 10.1038/nrg2884.

[16] M. Holden, S. Deng, L. Wojnowski, and B. Kulle, "GSEA-SNP: Applying gene set enrichment analysis to SNP data from genome-wide association studies," *Bioinformatics,* vol. 24, no. 23, pp. 2784-2785, 2008, doi: 10.1093/bioinformatics/btn516.

[17] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 43, pp. 15545-15550, 2005, doi: 10.1073/pnas.0506580102.

[18] National Center for Biotechnology Information, dbGaP: Database of Genotypes and Phenotypes, 2021. Accessed on: Jun. 14, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/gap/.

[19] The Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature,*

vol. 447, no. 7145, pp. 661-78, Jun 7 2007, doi: 10.1038/nature05911.

[20] P. I. de Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler, "Efficiency and power in genetic association studies," *Nature Genetics,* vol. 37, no. 11, pp. 1217-23, Nov 2005, doi: 10.1038/ng1669.

[21] KEGG: Kyoto Encyclopedia of Genes and Genomes, KEGG Disease Database, 2021. Accessed on: Jun. 14, 2021. [Online]. Available: https://www.genome.jp/kegg/disease/,.

[22] W. G. Hill and A. Robertson, "Linkage disequilibrium in finite populations," *Theoretical and Applied Genetics,* vol. 38, no. 6, pp. 226-231, 1968, doi: 10.1007/BF01245622.

[23] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *American Journal of Human Genetics,* vol. 74, no. 1, pp. 106-120, 2004, doi: 10.1086/381000.

[24] P. D. Sasieni, "From genotypes to genes: Doubling the sample size," *Biometrics,* vol. 53, no. 4, pp. 1253-1261, 1997, doi: 10.2307/2533494.

[25] K. Wang, M. Li, and M. Bucan, "Pathway-based approaches for analysis of genomewide association studies," *American Journal of Human Genetics,* vol. 81, no. 6, pp. 1278-1283, 2007, doi: 10.1086/522374.

[26] S. Freytag *et al.*, "A network-based kernel machine test for the identification of risk pathways in genome-wide association studies," *Human Heredity,* vol. 76, no. 2, pp. 64-75, 2014, doi: 10.1159/000357567.

[27] A. Brodie, J. R. Azaria, and Y. Ofran, "How far from the SNP may the causative genes be?," *Nucleic Acids Research,* vol. 44, no. 13, pp. 6046-6054, 2016, doi: 10.1093/nar/gkw500.

[28] Affymetrix, Human Mapping 500K Array Set - Support Materials, 2017. Accessed on: Jun. 14, 2021. [Online]. Available: http://www.affymetrix.com/support/technical/byproduct.affx?product=500k.

[29] C. O'Dushlaine *et al.*, "Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility," *Molecular Psychiatry,* vol. 16, no. 3, pp. 286-292, 2011, doi:

10.1038/mp.2010.7.

ประวัติผู้เขียน

ชื่อ-สกุล	เจษฎา วีรเดชกำพล
วัน เดือน ปี เกิด	21 กุมภาพันธ์ 2540
สถานที่เกิด	นครปฐม
วุฒิการศึกษา	สาขาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ที่อยู่ปัจจุบัน	503 ต.ห้วยจรเข้ ถ.เพชรเกษม อ.เมือง จ.นครปฐม 73000
ผลงานตีพิมพ์	การประชุมวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาระดับชาติ ครั้งที่ 23 มหาวิทยาลัยขอนแก่น