

ผลกระทบของการควบคุมอัตราการค้ากับต่างประเทศที่มีต่อขั้นตอน
วิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการ
วิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง

นางสาวจิตตินันท์ ตั้งสุนันท์ธรรม

นางสาวเปรมิกา ชัยพรหม

นายวรากร ชินวรากร

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวិชากรรมศาสตรบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

ปีการศึกษา 2558

Effects of False Discovery Rate Control on an Omnibus
Permutation Test on Ensembles of Two-Locus Analyses

Ms. Jittinan Tangsununtham

Ms. Premmika Chaiprom

Mr. Varakorn Chinvarakorn

A PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF BACHELOR OF COMPUTER ENGINEERING
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER ENGINEERING
FACULTY OF ENGINEERING
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK
ACADEMIC YEAR 2015

ปริญญานิพนธ์เรื่อง : ผลกระทบของการควบคุมอัตราการค้าคนพม่าที่มีต่อขั้นตอน
 วิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการ
 วิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง
 ชื่อ : นางสาวจิตตินันท์ ตั้งสุนันท์ธรรม
 นางสาวเปรมมิกา ชัยพรหม
 นายวรกร ชินวรกร
 สาขาวิชา : วิศวกรรมคอมพิวเตอร์
 ภาควิชา : วิศวกรรมไฟฟ้าและคอมพิวเตอร์
 คณะ : วิศวกรรมศาสตร์
 อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.ณชล ไชยรัตน์
 ดร.ดำรงฤทธิ์ เศรษฐศิริโชค
 ปีการศึกษา : 2558

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ อนุมัติให้
 ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต
 สาขาวิชาวิศวกรรมคอมพิวเตอร์

..... หัวหน้าภาควิชาวิศวกรรมไฟฟ้า
 (ผู้ช่วยศาสตราจารย์ ดร.ณชล วิวัชรโกเศศ) และคอมพิวเตอร์

..... ประธานกรรมการ
 (รองศาสตราจารย์ ดร.ณชล ไชยรัตน์)

..... กรรมการ
 (รองศาสตราจารย์ ดร.วรา วราวิทย์)

..... วรรณ วัชร กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วรรณ วัชร)

..... ดิเรก วัฒนศิริ กรรมการ

(ดร.ดำรงศักดิ์ เศรษฐศิริโชค)

ลิขสิทธิ์ของภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

Projected Report Title : Effects of False Discovery Rate Control on an Omnibus
Permutation Test on Ensembles of Two-Locus Analyses

Name : Ms. Jittinan Tangsununtham
Ms. Premmika Chaiprom
Mr. Varakorn Chinvarakorn

Major Field : Computer Engineering


Department : Electrical and Computer Engineering

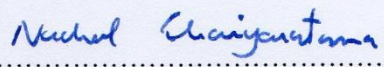
Faculty : Engineering

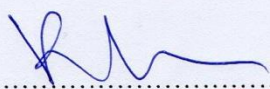
Project Advisor : Assoc. Prof. Dr. Nachol Chaiyaratana
Dr. Damrongrit Setsirichok

Academic Year : 2015

Accepted by the Faculty of Engineering, King Mongkut's University of Technology North
Bangkok in Partial Fulfillment of the Requirement for the Degree of Bachelor of Computer
Engineering.


..... Chairperson of Department of Electrical
(Asst. Prof. Dr. Nopadon Wiwatcharagoses) and Computer Engineering


..... Chairperson
(Assoc. Prof. Dr. Nachol Chaiyaratana)


..... Member
(Assoc. Prof. Dr. Vara Varavithya)

Waranyu Wongseree

Member

(Asst. Prof. Dr. Waranyu Wongseree)

Damrongrit

Member

(Dr. Damrongrit Setsirichok)

Copyright of the Department of Electrical and Computer Engineer, Faculty of Engineering

King Mongkut's University of Technology North Bangkok

บทคัดย่อ

ขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง (2LOmb) เป็นขั้นตอนวิธีที่ออกแบบมาสำหรับการระบุชนิดซึ่งมีความสัมพันธ์กับการเกิดโรคที่ซับซ้อน ขั้นตอนวิธีที่ได้รับการพิสูจน์ว่ามีประสิทธิภาพในการตรวจสอบฮิสเตซิสบริสุทธิ์ อย่างไรก็ตามขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง (2LOmb) ใช้การแก้ไขแบบ Bonferroni ซึ่งเป็นเทคนิคการแก้ไขที่นิยมกันสำหรับการทดสอบหลายสมมติฐาน ดังนั้นความเป็นไปได้ของการแทนที่การแก้ไขแบบ Bonferroni ด้วยการวิเคราะห์อัตราการค้นพบเท็จ (FDR) ซึ่งเป็นเทคนิคการแก้ไขที่นิยมน้อยกว่าจากการสำรวจผลการจำลองแสดงให้เห็นว่าขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง (2LOmb) และขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง โดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) มีความสามารถในการระบุชนิดที่ส่งผลให้เกิดโรคทั้งหมด เมื่ออันตรกิริยาฮิสเตซิสบริสุทธิ์ สองตำแหน่งที่ตั้ง และสามตำแหน่งที่ตั้งถูกพิจารณา อย่างไรก็ตามขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง โดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) รายงานชนิดที่ผิดพลาดของข้อมูลออกมามากกว่า

Abstract

An omnibus permutation test on ensembles of two-locus analyses (2LOmb) is an algorithm designed for identifying single nucleotide polymorphisms (SNPs) that are associated with a complex disease. The algorithm was proven to be efficient in pure epistasis detection. However, 2LOmb uses Bonferroni correction, which is a conservative correction technique for multiple hypothesis testing. Therefore, the possibility of replacing Bonferroni correction with false discovery rate analysis (FDR), which is a less conservative correction technique, was explored. The simulation results indicate that both 2LOmb and 2LOmb with embedded FDR analysis (2LOmbFDR) were capable of identifying all causative SNPs when purely epistatic two-locus and three-locus interactions were considered. Nonetheless, 2LOmbFDR reported more erroneous SNPs in its output.

กิตติกรรมประกาศ

ปริญญานิพนธ์เล่มนี้ข้าพเจ้าได้ทุ่มเททักษะ ความสามารถ เพื่อให้ผลงานนี้มีคุณภาพและมีประโยชน์สูงสุดต่อผู้ที่ได้ศึกษา ปริญญานิพนธ์นี้ไม่อาจสำเร็จได้โดยปราศจากบุคคลที่คอยให้ความช่วยเหลือ ต้องขอขอบพระคุณบุคคลเหล่านั้นมา ณ โอกาสนี้

กราบขอบพระคุณ รองศาสตราจารย์ ดร.ณชล ไชยรัตนะ และ ดร.ดำรงศักดิ์ เศรษฐศิริโชค อาจารย์ที่ปรึกษาปริญญานิพนธ์ที่ได้ให้ความรู้ คำปรึกษา คำแนะนำ ข้อคิดเห็น ข้อมูล และการสนับสนุนอย่างเต็มที่ ซึ่งเป็นประโยชน์อย่างยิ่งสำหรับปริญญานิพนธ์เล่มนี้

กราบขอบพระคุณอาจารย์ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ทุกท่านที่ได้ให้ความรู้ในวิชาชีพด้านวิศวกรรมและกำลังใจในการแก้ไขปัญหาในด้านการศึกษาตลอด 4 ปีที่ผ่านมา ทำให้ข้าพเจ้าสามารถนำความรู้ที่เรียนมาและทักษะการดำเนินชีวิตประจำวันมาประยุกต์ใช้ในปริญญานิพนธ์ได้เป็นอย่างดี

กราบขอบพระคุณ คุณพ่อและคุณแม่ที่ให้การสนับสนุนและความหวังใจต่อข้าพเจ้าตลอดมา ทำให้ข้าพเจ้าประสบความสำเร็จในการศึกษาไปได้ด้วยดี

ขอบคุณรุ่นพี่ รุ่นน้อง และเพื่อนทุกคนสำหรับคำแนะนำและความช่วยเหลือที่ทำให้ข้าพเจ้าทำงานได้อย่างมีประสิทธิภาพ

จิตตินันท์ ตั้งสุนันท์ธรรม

เปรมมิกา ชัยพรหม

วรกร ชินวรกร

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	๗
บทคัดย่อภาษาอังกฤษ	๘
กิตติกรรมประกาศ	๑๑
สารบัญตาราง	๑๒
สารบัญภาพ	๑๓
บทที่ 1. บทนำ	1
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์การวิจัย	1
1.3 ขอบเขตของการวิจัย	1
1.4 ขั้นตอนการวิจัย	2
1.5 ประโยชน์ที่คาดว่าจะได้รับ	2
บทที่ 2. วิธีและเทคนิคที่ใช้ในการทดลอง	3
2.1 การทดสอบการเรียงสับเปลี่ยนการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง โดยวิธีการแก้แบบบอนเฟอร์โรนี	3
2.2 อัตราการค้นพบเท็จ (False Discovery Rate)	8
2.3 การทดสอบการเรียงสับเปลี่ยนการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง โดยวิธีการแก้แบบอัตราการค้นพบเท็จ	9
บทที่ 3. ขั้นตอนและการดำเนินงาน	10
3.1 ข้อมูลที่ใช้ในการทดลอง	10

สารบัญ (ต่อ)

	หน้า
บทที่ 4. การทดลอง	20
4.1 การทดลอง	20
4.2 ผลการทดลอง	20
4.3 ตารางเวลาในการทดลอง	29
4.4 การทำการทดสอบความแตกต่างของค่าเฉลี่ยของสองประชากร (Paired t-Test)	30
บทที่ 5. สรุปผลการวิจัย	31
เอกสารอ้างอิง	32
ประวัติผู้แต่ง	33

สารบัญตาราง

ตารางที่	หน้า
2-1 ตัวอย่างตารางการจร	4
2-2 ตัวอย่างตารางการจร	4
3-1 แบบจำลองอันตรกิริยาสองตำแหน่งที่ตั้ง	10
3-2 แบบจำลองอันตรกิริยาสามตำแหน่งที่ตั้ง	11
3-3 ค่าพารามิเตอร์ใน genomeSIM file	12
3-4 ค่าพารามิเตอร์ใน Penetrance model file	13
3-5 ตัวอย่างข้อมูลที่ได้จากการรันโปรแกรมจีโนมซิม จำนวน 20 สนิป	16
3-6 ตัวอย่างข้อมูลที่ทำกรเรียงเปลี่ยนตำแหน่งใหม่จากข้อมูลในตารางที่ 3-5	17
4-1 จำนวนสนิปเฉลี่ยที่พบจากการทำ 2LOmb และ 2LOmbFDR โดยใช้ชุดข้อมูลสองตำแหน่งที่ตั้ง	25
4-2 จำนวนสนิปเฉลี่ยที่พบจากการทำ 2LOmb และ 2LOmbFDR โดยใช้ชุดข้อมูลสามตำแหน่งที่ตั้ง	29
4-3 เวลาเฉลี่ยที่ใช้ในการทดลองชุดข้อมูลสองตำแหน่งที่ตั้ง หน่วยเป็นวินาที (second)	29
4-4 เวลาเฉลี่ยที่ใช้ในการทดลองชุดข้อมูลสามตำแหน่งที่ตั้ง หน่วยเป็นวินาที (second)	30

สารบัญภาพ

ภาพที่	หน้า
3-1 ตัวอย่างข้อมูลเพนแทรกซ์สำหรับสองตำแหน่งที่ตั้ง ใน Penetrance table file	14
3-2 ตัวอย่างข้อมูลเพนแทรกซ์สำหรับสามตำแหน่งที่ตั้ง ใน Penetrance table file	15
4-1 กราฟแสดงจำนวนสนิปที่พบจากการทำขึ้น ตอนวิธีหาคข้อมูลสองตำแหน่งที่ตั้ง 20 สนิป	21
4-2 กราฟแสดงจำนวนสนิปที่พบจากการทำขึ้น ตอนวิธีหาคข้อมูลสองตำแหน่งที่ตั้ง 40 สนิป	22
4-3 กราฟแสดงจำนวนสนิปที่พบจากการทำขึ้น ตอนวิธีหาคข้อมูลสองตำแหน่งที่ตั้ง 80 สนิป	23
4-4 กราฟแสดงจำนวนสนิปที่พบจากการทำขึ้น ตอนวิธีหาคข้อมูลสองตำแหน่งที่ตั้ง 160 สนิป	24
4-5 กราฟแสดงจำนวนสนิปที่พบจากการทำขึ้น ตอนวิธีหาคข้อมูลสามตำแหน่งที่ตั้ง 20 สนิป	25
4-6 กราฟแสดงจำนวนสนิปที่พบจากการทำขึ้น ตอนวิธีหาคข้อมูลสามตำแหน่งที่ตั้ง 40 สนิป	26
4-7 กราฟแสดงจำนวนสนิปที่พบจากการทำขึ้น ตอนวิธีหาคข้อมูลสามตำแหน่งที่ตั้ง 80 สนิป	27
4-8 กราฟแสดงจำนวนสนิปที่พบจากการทำขึ้น ตอนวิธีหาคข้อมูลสามตำแหน่งที่ตั้ง 160 สนิป	28

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

1.1.1 ขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง (Omnibus Permutation Test on Ensembles of Two-Locus Analyses) ใช้การแก้แบบบอนเฟอร์โรนี (Bonferroni Correction) เมื่อมีการทดสอบหลายสมมติฐาน (Multiple Hypothesis Testing) อย่างไรก็ตาม การควบคุมอัตราการค้นพบเท็จ (False Discovery Rate Control) เป็นวิธีที่ได้รับการพิสูจน์ว่ามีประสิทธิภาพดีกว่าการแก้แบบบอนเฟอร์โรนี ดังนั้นจึงควรมีการศึกษาผลกระทบของการใช้การควบคุมอัตราการค้นพบเท็จแทนการแก้แบบบอนเฟอร์โรนีในขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง

1.2 วัตถุประสงค์การวิจัย

1.2.1 เพื่อศึกษาผลกระทบของการควบคุมอัตราการค้นพบเท็จที่มีต่อขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง

1.3 ขอบเขตของการวิจัย

1.3.1 ทำเฉพาะการใช้ควบคุมอัตราการค้นพบเท็จแทนการแก้แบบบอนเฟอร์โรนีในขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง

1.3.2 การศึกษาทำโดยอาศัยการจำลอง (Simulation) เท่านั้น

1.4 ขั้นตอนการวิจัย

1.4.1 ศึกษาอัตราการค้นพบเท็จมีต่อขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละสองตำแหน่งที่ตั้ง

1.4.2 ศึกษาอัตราการค้นพบเท็จแทนการแก้แบบบอนเฟอร์โรนี

1.4.3 เขียนโปรแกรมและทำการทดลอง

1.4.4 สรุปและวิเคราะห์ผลการทดลองพร้อมเสนอแนวทางแก้ไขปัญหา

1.4.5 จัดทำปริญญานิพนธ์และนำเสนอผลการทำปริญญานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ทราบผลกระทบของการควบคุมอัตราการค้นพบเท็จที่มีผลต่อขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวบรวมผลวิเคราะห์ครึ่งละสองตำแหน่ง

บทที่ 2

วิธีการและเทคนิคที่ใช้ในการทดลอง

2.1 การทดสอบการเรียงสับเปลี่ยนการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้งโดยวิธีการแก้แบบบอนเฟอร์โรนิ

2.1.1 การวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง (Two-Locus Analyses)

การวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง เริ่มจากการนำข้อมูลจากตารางเซตข้อมูลที่มีตัวอย่างกรณีและตัวอย่างควบคุมมาสร้างเป็นตารางการจร ซึ่งมีขนาดคอลัมน์เท่ากับ 2 (มีสถานะกรณีและสถานะควบคุม) และมีแถวเท่ากับ 9 (จาก 2locus genotype ที่แต่ละ locus มี genotype เป็น 0, 1 หรือ 2) จากนั้นจึงนำตารางดังกล่าวมาหาค่าสถิติไคกำลังสอง(Chi-Square Statistic) ดังสมการ

$$\chi^2 = \sum \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}} \quad (2-1)$$

ค่า observed_{ij} คือค่าของสมาชิกในแถวที่ i หลักที่ j และค่าคาดหวัง expected_{ij} หาได้ดัง

สมการ

$$\text{expected}_{ij} = \frac{(R_i)(C_j)}{N} \quad (2-2)$$

โดยที่ R_i คือ ผลรวมของสมาชิกในแถวที่ i

C_j คือ ผลรวมของสมาชิกในหลักที่ j

N คือ จำนวนสมาชิกทั้งหมด

จากนั้นคำนวณหาค่า χ^2 โดยใช้ค่าสถิติไคกำลังสอง(Chi-Square Statistic) และค่าระดับความเป็นอิสระ(degree of freedom) ซึ่งค่าระดับความเป็นอิสระ(degree of freedom) หาได้ดังสมการ

$$\text{degree of freedom} = (\text{row} - 1)(\text{column} - 1) \quad (2-3)$$

ตารางที่ 2-1 ตัวอย่างตารางการจร

	00	01	02	10	11	12	20	21	22
Case	4	16	24	12	34	16	32	27	35
Control	23	17	14	32	26	31	28	19	10

จากตารางที่ 2-1 เป็นตารางการจรที่มีขนาด 2×9 จะนับหาค่า Degree of Freedom ได้เท่ากับ $(2-1)(9-1)$ หรือได้ค่าเท่ากับ 8 แต่ถ้าหาก จำนวนข้อมูลในบางคอลัมน์เป็น 0 ทั้ง case และ control จะลดค่าคอลัมน์สำหรับการคำนวณหาค่า Degree of Freedom เช่น จากตารางที่ 2-2 คอลัมน์ที่ข้อมูลเท่ากับ 02 เป็น 0 ทั้ง case และ control จะหาค่า Degree of Freedom ได้เท่ากับ $(2-1)(8-1)$ หรือได้ค่าเท่ากับ 7

ตารางที่ 2-2 ตัวอย่างตารางการจร

	00	01	02	10	11	12	20	21	22
Case	15	16	0	19	34	22	32	27	35
Control	23	25	0	32	26	35	28	19	12

เมื่อได้ค่า Degree of Freedom มาแล้ว นำมาคำนวณค่า P-value โดยเข้าฟังก์ชัน

$$p_i = p\left(x^2, \text{DoF}\right) \quad (2-4)$$

โดยที่ p_i คือ ค่าพีที่ได้จากการวิเคราะห์ครั้งละสองตำแหน่งแบบบอนเฟอร์โรนินใน

ข้อมูลชุดที่ i

x^2 คือ ค่าสถิติไคกำลังสอง (Chi-Square Statistic)

DoF คือ ค่า Degree of Freedom

เมื่อได้ค่า p-value จากการวิเคราะห์ครั้งละสองตำแหน่งครบทุกชุดข้อมูล นำค่า p-value จากการวิเคราะห์ครั้งละสองตำแหน่งทุกค่ามาคูณกับ Bonferroni Correction Factor และหากค่า p-value จากการวิเคราะห์ครั้งละสองตำแหน่งที่คูณกับ Bonferroni Correction Factor แล้วมีค่าเกิน 1 ให้ปรับให้เท่ากับ 1

2.1.2 การทดสอบเรียงสับเปลี่ยน (Permutation test)

ในการทดสอบเรียงสับเปลี่ยนจะมีการหาค่าสถิติฟิชเชอร์ ซึ่งค่าสถิติฟิชเชอร์สามารถคิดได้ดังสมการ

$$T_i^e = -2 \sum_n \log(p_n) \quad (2-5)$$

โดยที่ T_i^e คือ ค่าสถิติฟิชเชอร์ของกลุ่มการรวมที่ e เซตข้อมูลที่ i

e คือ กลุ่มการรวมที่ e

i คือ เซตของข้อมูลดั้งเดิม หรือเซตของข้อมูลทำการเรียงสับเปลี่ยนที่ i

p_n คือ ค่าที่ได้จากการวิเคราะห์ครั้งละสองตำแหน่งแบบบอนเฟอร์โรนินในข้อมูลชุดที่ n

ในการเรียงสับเปลี่ยนข้อมูล จะนำข้อมูลที่มีมาทำการเรียงสับเปลี่ยน โดยที่จำนวนข้อมูลที่มีอยู่นั้น มีจำนวนเท่าเดิม เมื่อทำการเรียงสับเปลี่ยนข้อมูลแล้วจะนำมาหาค่าสถิติฟิชเชอร์ดังที่กล่าวมาข้างต้น

เมื่อคำนวณค่าสถิติฟิชเชอร์มาแล้ว สามารถคำนวณหา Raw p-value ของกลุ่มการรวมที่ e ของเซตข้อมูลดั้งเดิม โดยนำจำนวนของค่าสถิติฟิชเชอร์ของกลุ่มการรวมที่ e ของเซตข้อมูลทำการเรียงสับเปลี่ยนที่ i ที่มีค่ามากกว่าหรือเท่ากับค่าสถิติฟิชเชอร์ของกลุ่มการรวมที่ e ของเซตข้อมูลดั้งเดิม มาหารกับจำนวนที่ทำการเรียงสับเปลี่ยน ดังสมการ

$$p_o^e = \frac{\left| \left\{ i : 1 \leq i \leq t, T_i^e \geq T_o^e \right\} \right|}{t} \quad (2-6)$$

โดยที่ p_o^e คือ Raw p-value ของกลุ่มการรวมที่ e ของเซตข้อมูลดั้งเดิม

T_o^e คือ ค่าสถิติฟิชเชอร์ในกลุ่มการรวมที่ e ของเซตข้อมูลดั้งเดิม

T_i^e คือ ค่าสถิติฟิชเชอร์ในกลุ่มการรวมที่ e ของเซตข้อมูลที่ทำให้การเรียงสับเปลี่ยนที่ i

e คือ กลุ่มการรวมที่ e

i คือ เซตของข้อมูลที่ทำให้การเรียงสับเปลี่ยนที่ i

t คือ จำนวนที่ทำให้การเรียงสับเปลี่ยน

2.1.3 การกำหนด Global p-value (Global p-value determination)

การกำหนด Global p-value หาได้จาก Raw p-value ซึ่ง Raw p-value หาได้โดยนำจำนวนที่ค่าสถิติฟิชเชอร์ในกลุ่มการรวมที่ e ของเซตข้อมูลดั้งเดิม หรือเซตของข้อมูลที่ทำให้การเรียงสับเปลี่ยนที่ i ที่สนใจ ที่มีค่าน้อยกว่าหรือเท่ากับ ค่าสถิติฟิชเชอร์ในกลุ่มการรวมที่ e ของเซตข้อมูลที่ทำให้การเรียงสับเปลี่ยนที่ j ซึ่งเป็นเซตข้อมูลที่ไม่ใช่เซตข้อมูลที่ i มาหารกับจำนวนที่ทำให้การเรียงสับเปลี่ยน ดังสมการ

$$p_i^e = \frac{\left| \left\{ j: 0 \leq j \leq t, j \neq i, T_j^e \geq T_i^e \right\} \right|}{t} \quad (2-7)$$

โดยที่ p_i^e คือ Raw p-value ของกลุ่มการรวมที่ e ของเซตข้อมูลที่ i

T_i^e คือ ค่าสถิติฟิชเชอร์ในกลุ่มการรวมที่ e ของเซตข้อมูลดั้งเดิม หรือเซตของข้อมูลที่ทำให้การเรียงสับเปลี่ยนที่ i

T_j^e คือ ค่าสถิติฟิชเชอร์ในกลุ่มการรวมที่ e ของเซตข้อมูลที่ทำให้การเรียงสับเปลี่ยนที่ j

e คือ กลุ่มการรวมที่ e

i คือ เซตของข้อมูลดั้งเดิม หรือเซตของข้อมูลที่ทำให้การเรียงสับเปลี่ยนที่ i

j คือ เซตของข้อมูลที่ทำให้การเรียงสับเปลี่ยนที่ j ที่ไม่ใช่เซตข้อมูลที่ i

t คือ จำนวนที่ทำให้การเรียงสับเปลี่ยน

จากการคำนวณหา Raw p-value ดังที่กล่าวมาข้างต้น สามารถคำนวณหา Global p-value ในแต่ละกลุ่มการรวม โดยนำจำนวน Raw p-value ที่น้อยที่สุดของกลุ่มการรวมที่ e ของเซตข้อมูล เรียงสับเปลี่ยนที่ i ที่สนใจ (p_i^{\min})

$$p_i^{\min} = \min_e p_i^e \quad (2-8)$$

ที่น้อยกว่าหรือเท่ากับ Raw p-value ที่น้อยที่สุดของกลุ่มการรวมที่ e ของเซตข้อมูล ดั้งเดิม (p_o^{\min})

$$p_o^{\min} = \min_e p_o^e \quad (2-9)$$

จากนั้นนำจำนวนที่ได้มาหารกับจำนวนที่ได้ทำการเรียงสับเปลี่ยนข้อมูล ดังสมการ

$$p_{\text{global}} = \frac{\left| \left\{ i : 1 \leq i \leq t, p_i^{\min} \leq p_o^{\min} \right\} \right|}{t} \quad (2-10)$$

โดยที่ p_{global} คือ Global p-value

p_i^{\min} คือ Raw p-value ที่น้อยที่สุดของกลุ่มการรวมที่ e เซตข้อมูลที่ i

p_o^{\min} คือ Raw p-value ที่น้อยที่สุดของกลุ่มการรวมที่ e เซตข้อมูลดั้งเดิม

i คือ เซตของข้อมูลทำการเรียงสับเปลี่ยนที่ i

t คือ จำนวนทำการเรียงสับเปลี่ยน

2.1.4 การค้นหากลุ่มการรวมที่ดีที่สุด

การค้นหากลุ่มการรวมที่ดีที่สุด หาได้จากการวิเคราะห์ครึ่งละสองตำแหน่งที่ตั้ง จากนั้นทำการทดสอบเรียงสับเปลี่ยน จากนั้นกำหนด Global p-value และนำ Global p-value มาเปรียบเทียบกับค่า Global p-value ก่อนหน้าที่คำนวณได้และ Raw p-value มาเปรียบเทียบกับค่า Raw p-value ก่อนหน้าที่คำนวณได้เพื่อค้นหากลุ่มการรวมที่ดีที่สุด โดยถ้า Global p-value มีค่าเพิ่มขึ้น จะหยุดการค้นหากลุ่มการรวมที่ดีที่สุด และกลุ่มการรวมที่ดีที่สุดคือ กลุ่มการรวม ณ ตำแหน่งถัดไปและถ้า Raw p-value มีค่าเพิ่มขึ้น จะหยุดการค้นหากลุ่มการรวมที่ดีที่สุด และกลุ่มการรวมที่ดีที่สุดคือ กลุ่มการรวม ณ ตำแหน่งนั้น

2.2 อัตราการค้นพบเท็จ (False Discovery Rate)

อัตราการค้นพบเท็จ คือ อัตราของการเกิดความผิดพลาดประเภทที่ 1 ในการทดสอบสมมติฐานหลักเมื่อมีการทดสอบหลายสมมติฐาน ขั้นตอนการควบคุมอัตราการค้นพบเท็จถูกออกแบบมาเพื่อควบคุมสัดส่วนค่าคาดหวังของการปฏิเสธสมมติฐานหลักนั้นคือการค้นพบที่ไม่ถูกต้อง (การค้นพบเท็จ) ในกรณีที่สมมติฐานหลักเป็นจริง หมายความว่า การค้นพบทั้งหมดนั้นไม่เป็นจริง

ขั้นตอนการคำนวณหา q-value จากรายการ (list) ของ p-value

1. ให้เรียงค่า $p(1) \leq p(2) \leq p(3) \leq \dots \leq p(m)$ ตามจำนวน p-value
2. สำหรับค่า λ โดยให้ $R = \{0, 0.05, 0.1, \dots, 0.95\}$ คำนวณ $\hat{\pi}_0$ จากสูตร

$$\hat{\pi}_0 = \frac{\#\{p_j > \lambda\}}{m(1 - \lambda)} \quad (2-10)$$

โดยที่ m คือ จำนวนค่า p-value ที่มี

3. สำหรับแต่ละค่า $\lambda \in R$ ค่า B จากชุดแสดง $\hat{\pi}_0^{*b}$ ให้ $b = 1, 2, 3, \dots, B$

โดยมาจากการบูทสเตรปตัวอย่างจาก p-value จำนวน p ครั้ง

4. สำหรับแต่ละค่า $\lambda \in R$ ประมวลค่าของแต่ละ FDR เป็น

$$\hat{FDR}(\lambda) = \frac{1}{B} \sum_{b=1}^B \left[\hat{\pi}_0^{*b}(\lambda) - \min_{\lambda' \in R} \{\hat{\pi}_0(\lambda')\} \right]^2 \quad (2-11)$$

$$5. \text{ ให้ } \hat{\lambda} = \operatorname{argmin}_{\lambda \in R} \left\{ \hat{FDR}(\lambda) \right\} \quad (2-12)$$

$$6. \text{ คำนวณ } \hat{q}(p_{(m)}) = \hat{\pi}_0 \cdot p_{(m)} \quad (2-13)$$

$$\text{โดยที่ } \hat{\pi}_0 = \frac{\#\{p_j > \lambda\}}{m(1 - \lambda)} \quad (2-14)$$

7. สำหรับค่า $i = m-1, m-2, \dots, 1$

$$\text{คำนวณ } \hat{q}(p_{(i)}) = \min \left(\frac{\hat{\pi}_0 \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)}) \right) \quad (2-15)$$

8. ประเมินค่า q-value จากตำแหน่งที่ i ที่มีระดับนัยสำคัญสูงสุด คือ $\hat{q}(p_{(i)})$

2.3 การทดสอบการเรียงสับเปลี่ยนการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้งโดยวิธีการแก้แบบอัตรา

การค้นพบเท็จ

จากการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้งโดยวิธีการแก้แบบบอนเฟอร์โรนินในข้อ 2.1 ในขั้นตอนการทดสอบการเรียงสับเปลี่ยนในข้อ 2.1.2 จากสมการที่ 2-5

$$T_i^e = -2 \sum_n \log(p_n) \quad (2-5)$$

โดยที่ T_i^e คือ ค่าสถิติฟิชเชอร์ของกลุ่มการรวมที่ e เซตข้อมูลที่ i

e คือ กลุ่มการรวมที่ e

i คือ เซตของข้อมูลดั้งเดิม หรือเซตของข้อมูลที่ทำให้การเรียงสับเปลี่ยนที่ i

p_n คือ ค่าพีที่ได้จากการวิเคราะห์ครั้งละสองตำแหน่งแบบบอนเฟอร์โรนินในข้อมูลชุดที่ n จะเปลี่ยนจากค่า p_n เป็นค่า q_n โดยค่า q_n ได้มาจากระยะการคำนวณอัตราการค้นพบเท็จได้เป็นสมการ

$$T_i^e = -2 \sum_n \log(q_n) \quad (2-16)$$

โดยที่ T_i^e คือ ค่าสถิติฟิชเชอร์ของกลุ่มการรวมที่ e เซตข้อมูลที่ i

e คือ กลุ่มการรวมที่ e

i คือ เซตของข้อมูลดั้งเดิม หรือเซตของข้อมูลที่ทำให้การเรียงสับเปลี่ยนที่ i

q_n คือ ค่าคิวที่ได้จากการวิเคราะห์ครั้งละสองตำแหน่งแบบอัตราการค้นพบเท็จในข้อมูลชุดที่ n

จากนั้นทำตามขั้นตอนการกำหนด Global p-value (Global p-value determination) ในข้อ 2.1.3 และ ขั้นตอนการค้นหากลุ่มการรวมที่ดีที่สุดข้อ 2.1.4 ต่อไป

บทที่ 3

ขั้นตอนและการดำเนินงาน

3.1 ข้อมูลที่ใช้ในการทดลอง

3.1.1 แบบจำลองอันตรกิริยา (Interaction model)

แบบจำลองอันตรกิริยาเกิดจากการที่มี genotype มา interaction กัน โดยในการทดลองนี้ใช้ 2 แบบ ได้แก่ แบบจำลองอันตรกิริยาสองตำแหน่งที่ตั้ง (Two-locus interaction model) และแบบจำลองอันตรกิริยาสามตำแหน่งที่ตั้ง (Three-locus interaction model) ส่งผลให้ Penetrance of genotype ออกมาได้ดังตัวอย่าง แบบจำลองอันตรกิริยาสองตำแหน่งที่ตั้ง (Two-locus interaction model) และ แบบจำลองอันตรกิริยาสามตำแหน่งที่ตั้ง (Three-locus interaction model) (Culverhouse et al., 2002) แสดงในตารางที่ 3-1 และ ตารางที่ 3-2

ตารางที่ 3-1 แบบจำลองอันตรกิริยาสองตำแหน่งที่ตั้ง เมื่อ $0 \leq K \leq \frac{1}{4}$

GENOTYPE	PENETRANCE OF GENOTYPE		
	BB	Bb	bb
AA	0	0	4K
Aa	0	2K	0
aa	4K	0	0

โดยให้ $h^2 = 0.01$ ซึ่ง h^2 คือปัจจัยที่สนใจมีผลทำให้เกิดโรค ดังนั้น $h^2 = 0.01$ แปลว่าปัจจัยที่กำลังพิจารณาอยู่ใน genotype 2locus มีผล 0.01 ส่วนอีก 0.99 เป็นผลจากปัจจัยอื่น

$$\text{แทนค่าในสมการ} \quad 0.01 = \frac{2K}{1 - K} \quad (3-1)$$

$$\text{จะได้} \quad K = 0.004975$$

โดยที่ K คือค่าความชุกของโรค(prevalence) หรือความน่าจะเป็นที่ประชากรจะเป็นโรค
นำค่า K ไปแทนในตารางที่ 3-1 $2K = 0.00995$

$$4K = 0.0199$$

โดยจากตารางที่ 2-1 และตารางที่ 2-2 คู่ของ genotype แสดงเป็นตัวเลข 00, 01, 02, ..., 22
ซึ่ง 0 คือ Homozygous Wild-type ซึ่งในตารางที่ 3.1 คือ genotype AA และ BB, 1 คือ
Heterozygous Genotype ซึ่งในตารางที่ 3.1 คือ genotype Aa และ Bb และ 2 คือ Homozygous
Variant ซึ่งในตารางที่ 3-1 คือ genotype aa และ bb

ตารางที่ 3-2 แบบจำลองอันตรกิริยาสองตำแหน่งที่ตั้ง เมื่อ $0 \leq K \leq \frac{1}{16}$

GENOTYPE	PENETRANCE OF GENOTYPE								
	CC			Cc			cc		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	0	0	16K	0	0	0	0	0	0
Aa	0	0	0	0	4K	0	0	0	0
aa	0	0	0	0	0	0	16K	0	0

โดยให้ $h^2 = 0.01$ ซึ่ง h^2 คือปัจจัยที่สนใจมีผลทำให้เกิดโรค ดังนั้น $h^2 = 0.01$ แปลว่า
ปัจจัยที่กำลังพิจารณาอยู่ใน genotype 3locus มีผล 0.01 ส่วนอีก 0.99 เป็นผลจากปัจจัยอื่น

$$\text{แทนค่าในสมการ} \quad 0.01 = \frac{9K}{1 - K} \quad (3-2)$$

$$\text{จะได้} \quad K = 0.0011$$

โดยที่ K คือค่าความชุกของโรค (prevalence) หรือค่าที่แสดงประชากรที่เป็นโรคหารด้วย
ประชากรทั้งหมด

$$\text{นำค่า } K \text{ ไปแทนในตารางที่ 3-1} \quad 4K = 0.0044395$$

$$16K = 0.017758$$

จากนั้นนำค่าจากตารางที่ 3-1 ที่แทนค่า $2K = 0.00995$ และ $4K = 0.0199$ แล้วไปกำหนดค่าให้ PENTABLE ใน Penetrance table file ตามภาพที่ 3-1 และนำค่าจากตารางที่ 3-2 ที่แทนค่า $4K = 0.0044395$ และ $16K = 0.017758$ แล้วไปกำหนดค่าให้ PENTABLE ใน Penetrance table file ตามภาพที่ 3-2 ตามลำดับต่อไป

3.1.2 โปรแกรมจีโนมซิม (genomeSIM)

โปรแกรมจีโนมซิมเป็นโปรแกรมจำลองข้อมูลลักษณะทางพันธุกรรม โดยข้อมูลที่ได้จากการจำลองจะเป็นข้อมูล genotype ของประชากรที่จะมีสถานะของการเป็นโรคหรือไม่เป็นโรค ทำการจำลองสำหรับการทดลองทั้งหมด 800 ครั้งจากการกำหนดค่าพารามิเตอร์ดังตารางที่ 3-3 และตารางที่ 3-4 ดังนี้

ตารางที่ 3-3 ค่าพารามิเตอร์ใน genomeSIM file

ค่าพารามิเตอร์ (Parameter)	ค่าที่เปลี่ยนในการทดลอง	คำอธิบาย (Description)
RAND	1 – 100	กำหนด random seed ในการจำลอง
MODELFILES	Penetrance table file - sample.smod เป็น Penetrance table file สำหรับสองตำแหน่งที่ตั้ง ตัวอย่างดังภาพที่ 3-1 - sample3locus.smod เป็น Penetrance table file สำหรับสามตำแหน่งที่ตั้ง ตัวอย่างดังภาพที่ 3-2	ระบุ model file (.smod) ที่ใช้ในการทดลอง
ALLELEFREQS	สำหรับสองตำแหน่งที่ตั้ง 1 0.5 0.5 2 0.5 0.5 สำหรับสามตำแหน่งที่ตั้ง 1 0.5 0.5 2 0.5 0.5 3 0.5 0.5	ระบุความถี่อัลลีลสำหรับ SNPs ของแต่ละการจำลอง

ตารางที่ 3-3 (ต่อ) ค่าพารามิเตอร์ใน genomeSIM file

ค่าพารามิเตอร์ (Parameter)	ค่าที่เปลี่ยนในการทดลอง	คำอธิบาย (Description)
AFFECTED	200	จำนวนที่แสดงว่าเป็นโรค
UNAFFECTED	200	จำนวนที่แสดงว่าไม่เป็นโรค
SIMLOCI	20, 40, 80, 160 ตามลำดับ	จำนวนสลิปที่ใช้ทำการผลการทดลอง

ตารางที่ 3-4 ค่าพารามิเตอร์ใน Penetrance model file

ค่าพารามิเตอร์ (Parameter)	คำอธิบาย (Description)
DISEASELOCI	ระบุ loci ในการจำลอง
PENTABLE	ระบุค่าเพเนแทรนซ์สำหรับแต่ละจีโนไทป์จาก Interaction model

DISEASELOCI 1 2

PENTABLE

AABB 0.00000

AABb 0.00000

AAbb 0.01990

AaBB 0.00000

AaBb 0.00995

Aabb 0.00000

aaBB 0.01990

aaBb 0.00000

aabb 0.00000

ภาพที่ 3-1 ตัวอย่างข้อมูลเพเนทรานซ์สำหรับสองตำแหน่งที่ตั้ง ใน Penetrance table file

DISEASELOCI 1 2 3

PENTABLE

AABBCC 0.00000

AABbCC 0.00000

AAbbCC 0.01776

AaBBCC 0.00000

AaBbCC 0.00000

AabbCC 0.00000

aaBBCC 0.00000

aaBbCC 0.00000

aabbCC 0.00000

AABBCc 0.00000

AABbCc 0.00000

AAbbCc 0.00000

AaBBCc 0.00000

AaBbCc 0.00444

AabbCc 0.00000

aaBBCc 0.01776

aaBbCc 0.00000

aabbCc 0.00000

AABBcc 0.00000

AABbcc 0.00000

AAbbcc 0.00000

AaBBcc 0.00000

ภาพที่ 3-2 ตัวอย่างข้อมูลเพเนแทรนซ์สำหรับสามตำแหน่งที่ตั้ง ใน Penetrance table file

AaBbcc 0.00000

Aabbcc 0.00000

aaBBcc 0.00000

aaBbcc 0.00000

aabbcc 0.00000

ภาพที่ 3-2 (ต่อ) ตัวอย่างข้อมูลเพเนทรานซ์สำหรับสามตำแหน่งที่ตั้ง ใน Penetrance table file

หลังจากการรันโปรแกรมจีโนมซิมจะได้เอาต์พุตดังตารางที่ 3-5 และทำการแปลงข้อมูลเป็นดังตารางที่ 3-6 เพื่อเป็นข้อมูลเข้าสำหรับขั้นตอนวิธี

ตารางที่ 3-5 ตัวอย่างข้อมูลที่ได้จากการรัน โปรแกรมจีโนมซิม จำนวน 20 สนิป

Class	SNP1	SNP2	SNP3	...	SNP18	SNP19	SNP20
0	0	1	0		1	1	2
⋮							
0	⋮	⋮	⋮	...	⋮	⋮	⋮
1							
⋮							
1							

ตารางที่ 3-6 ตัวอย่างข้อมูลที่ใช้การเรียงเปลี่ยนตำแหน่งใหม่จากข้อมูลในตารางที่ 3-5

Sample	SNP1	SNP2	SNP3	...	SNP18	SNP19	SNP20	Class
1	0	1	0		1	1	2	0
⋮								⋮
200	⋮	⋮	⋮	...	⋮	⋮	⋮	0
201								1
⋮								⋮
400								1

หลังจากได้ข้อมูลที่ใช้การเรียงเปลี่ยนตำแหน่งใหม่จากข้อมูลที่ได้จากการรันโปรแกรม จีโนมซิมตามตารางที่ 3-6 แล้ว ซึ่งมีทั้งหมด 800 ชุดข้อมูลได้แก่

- ชุดข้อมูลสองตำแหน่งที่ตั้ง จำนวนสนิป 20 สนิป 100 ชุดข้อมูล (random seed 1 - 100)
- ชุดข้อมูลสองตำแหน่งที่ตั้ง จำนวนสนิป 40 สนิป 100 ชุดข้อมูล (random seed 1 - 100)
- ชุดข้อมูลสองตำแหน่งที่ตั้ง จำนวนสนิป 80 สนิป 100 ชุดข้อมูล (random seed 1 - 100)
- ชุดข้อมูลสองตำแหน่งที่ตั้ง จำนวนสนิป 160 สนิป 100 ชุดข้อมูล (random seed 1 - 100)
- ชุดข้อมูลสามตำแหน่งที่ตั้ง จำนวนสนิป 20 สนิป 100 ชุดข้อมูล (random seed 1 - 100)
- ชุดข้อมูลสามตำแหน่งที่ตั้ง จำนวนสนิป 40 สนิป 100 ชุดข้อมูล (random seed 1 - 100)
- ชุดข้อมูลสามตำแหน่งที่ตั้ง จำนวนสนิป 80 สนิป 100 ชุดข้อมูล (random seed 1 - 100)
- ชุดข้อมูลสามตำแหน่งที่ตั้ง จำนวนสนิป 160 สนิป 100 ชุดข้อมูล (random seed 1 - 100)

นำข้อมูลไปใช้สำหรับขั้นตอนวิธี 2 ขั้นตอนวิธี ได้แก่ ขั้นตอนวิธีการทดสอบการเรียง

สับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละสองตำแหน่งที่ตั้ง (2LOmb) และขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละสองตำแหน่งที่ตั้งโดยใช้การ

ควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) ซึ่งจากการทำแต่ละขั้นตอนวิธีจะได้จำนวนสนิปที่คัดเลือกมาสำหรับแต่ละชุดข้อมูล แล้วนำจำนวนสนิปที่ได้จากการทำขั้นตอนวิธีทั้งสองมาเปรียบเทียบกัน และเปรียบเทียบจำนวนสนิปที่ถูกต้องที่ได้จากการทำขั้นตอนวิธีทั้งสอง เมื่อสนิปที่ถูกต้องคือสนิปที่ถูกกำหนดให้ส่งผลต่อการเกิดโรคในค่า DISEASELOCI ตามภาพที่ 3-1 และภาพที่ 3-2 โดยเปรียบเทียบด้วยการทดสอบความแตกต่างของค่าเฉลี่ยของสองประชากร (Paired t-Test)

3.1.3 การทดสอบความแตกต่างของค่าเฉลี่ยของสองประชากร (Paired t-Test)

การใช้ T-TEST ใน Microsoft excel

T.TEST(array1,array2,tails,type)

ฟังก์ชัน T.TEST มีอาร์กิวเมนต์ดังนี้

- **Array1** (ต้องระบุ) ชุดข้อมูลชุดแรก ในการทดลองนี้ใช้จำนวนสนิปที่ได้จากการทำขั้นตอนวิธีทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง (2LOmb) 800 ครั้งเป็นชุดข้อมูลชุดแรก
- **Array2** (ต้องระบุ) ชุดข้อมูลชุดที่สอง ในการทดลองนี้ใช้จำนวนสนิปที่ได้จากการทำขั้นตอนวิธีทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้งโดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) 800 ครั้งเป็นชุดข้อมูลชุดที่สอง
- **Tails** (ต้องระบุ) ระบุจำนวน tail ของการแจกแจง ถ้า tails = 1 ฟังก์ชัน T.TEST จะส่งกลับการแจกแจงแบบด้านเดียว ถ้า tails = 2 ฟังก์ชัน T.TEST จะส่งกลับการแจกแจงแบบสองด้าน ในการทดสอบนี้ใช้ tails = 2
- **Type** (ต้องระบุ) ชนิดของ t-Test ที่จะใช้ ถ้า type = 1 จะใช้การทดสอบแบบคู่ ถ้า type = 2 จะใช้การทดสอบแบบค่าความแปรปรวนที่เท่ากันทั้งสองตัวอย่าง (homoscedastic) ถ้า type = 3 จะใช้การทดสอบแบบค่าความแปรปรวนที่ไม่เท่ากันในแต่ละตัวอย่าง (heteroscedastic) ในการทดสอบนี้ใช้ type = 1

เมื่อทำการเปรียบเทียบด้วยการทดสอบความแตกต่างของค่าเฉลี่ยของสองประชากร (Paired t-Test) แล้ว จะได้ค่า P-Value ออกมาอยู่ระหว่าง 0 ถึง 1 ซึ่งถ้าค่า P-Value ยิ่งน้อยแปลว่าชุดข้อมูลออกที่ได้มีความสอดคล้องกัน

บทที่ 4

ผลการทดลอง

4.1 การทดลอง

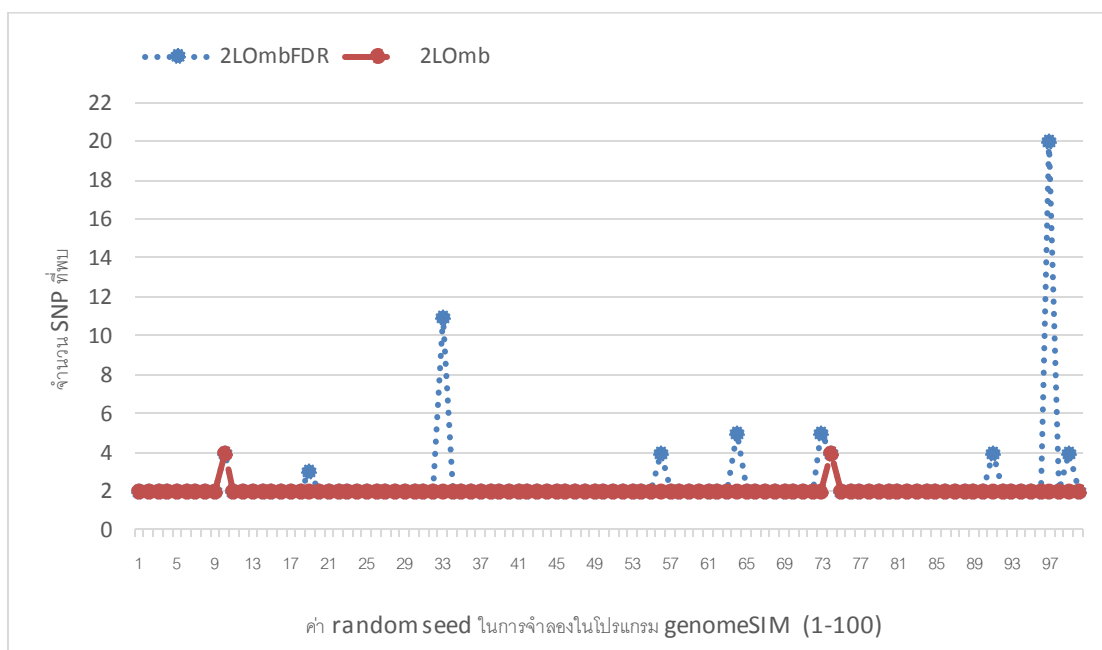
การทดลองเพื่อเปรียบเทียบผลการทดลองระหว่างขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละ สองตำแหน่งที่ตั้ง (2LOmb) และขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละ สองตำแหน่งที่ตั้ง โดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) สำหรับชุดข้อมูลสองตำแหน่งที่ตั้ง และชุดข้อมูลสามตำแหน่งที่ตั้ง โดยดัชนีที่นำมาเปรียบเทียบเป็นจำนวนสนิปจากการทำขั้นตอนวิธีทั้งสองและจำนวนสนิปที่ถูกต้องจากการทำขั้นตอนวิธีทั้งสอง ซึ่งเปรียบเทียบด้วยการทดสอบความแตกต่างของค่าเฉลี่ยของสองประชากร (Paired t-Test)

4.2 ผลการทดลอง

ผลการทดลองจากทั้งหมด 800 ชุดข้อมูล โดยชุดข้อมูลสองตำแหน่งที่ตั้งที่มีสนิปที่ถูกกำหนดให้ส่งผลต่อการเกิดโรคคือสนิปตัวที่ 1 กับ 2 และชุดข้อมูลสามตำแหน่งที่ตั้งที่มีสนิปที่ถูกกำหนดให้ส่งผลต่อการเกิดโรคคือสนิปตัวที่ 1, 2 กับ 3

4.2.1 ผลการทดสอบชุดข้อมูลสองตำแหน่งที่ตั้ง จำนวนสปี 20 สปี 100 ชุดข้อมูล

กราฟแสดงจำนวนสปีที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสองตำแหน่งที่ตั้ง 20 สปี
แสดงในภาพที่ 4-1

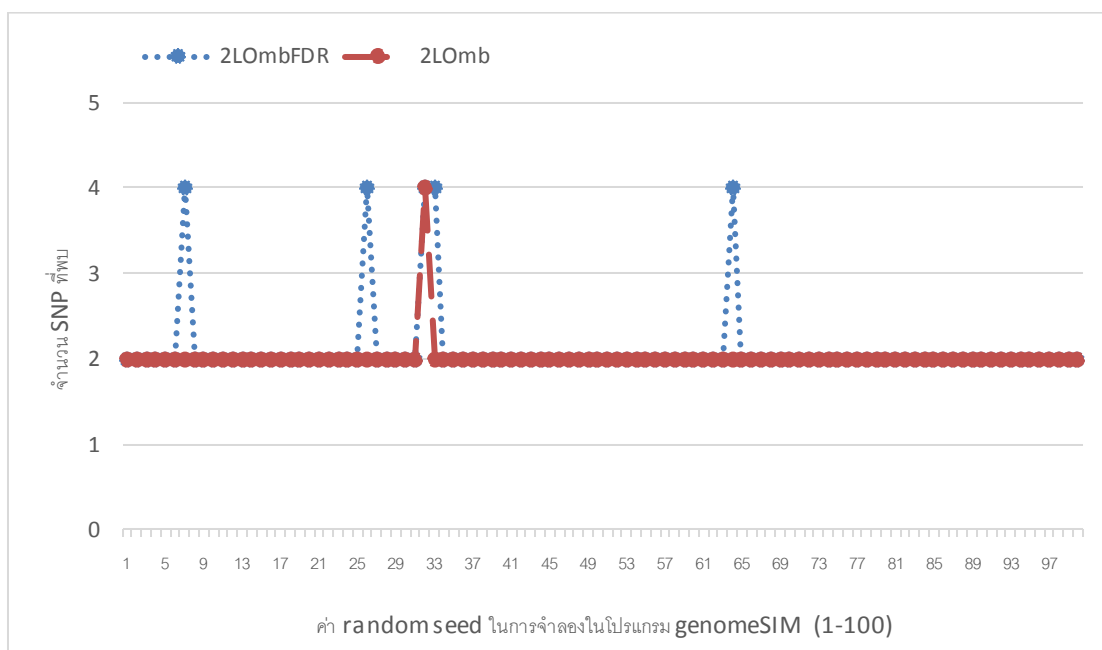


ภาพที่ 4-1 กราฟแสดงจำนวนสปีที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสองตำแหน่งที่ตั้ง 20 สปี

โดยข้อมูลทั้ง 100 ชุดข้อมูลมีจำนวนสปีที่ถูกต้องเท่ากับ 2 สปีทุกชุดข้อมูล

4.2.2 ผลการทดสอบชุดข้อมูลสองตำแหน่งที่ตั้ง จำนวนสนิป 40 สนิป 100 ชุดข้อมูล

กราฟแสดงจำนวนสนิปที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสองตำแหน่งที่ตั้ง 40 สนิป
แสดงในภาพที่ 4-2

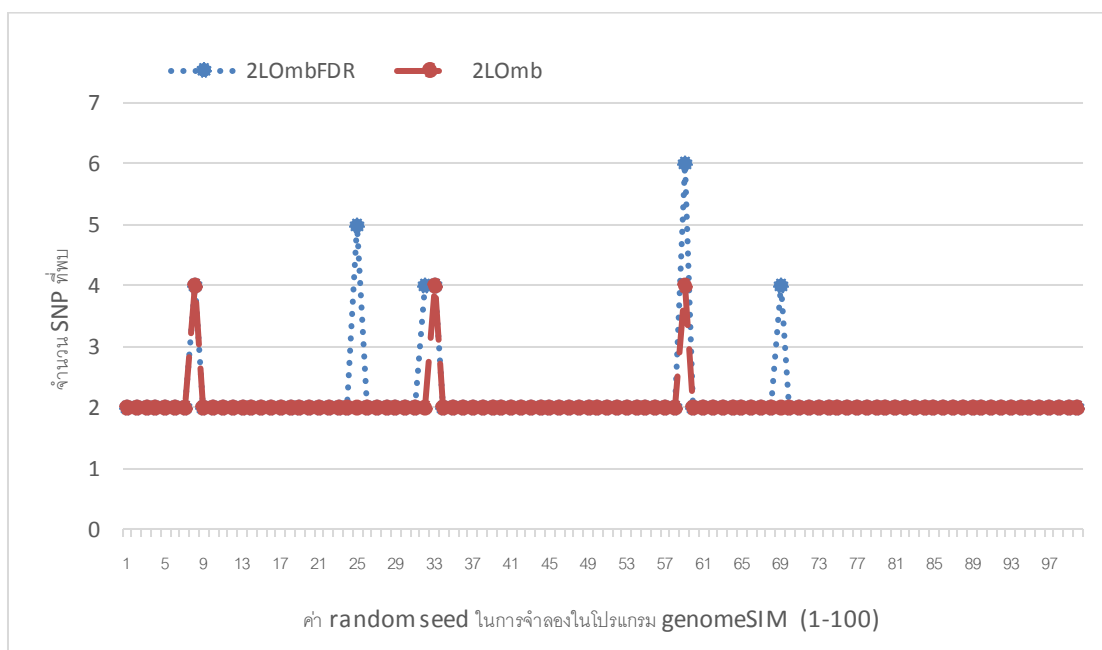


ภาพที่ 4-2 กราฟแสดงจำนวนสนิปที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสองตำแหน่งที่ตั้ง 40 สนิป

โดยข้อมูลทั้ง 100 ชุดข้อมูลมีจำนวนสนิปที่ถูกต้องเท่ากับ 2 สนิปทุกชุดข้อมูล

4.2.3 ผลการทดสอบชุดข้อมูลสองตำแหน่งที่ตั้ง จำนวนสปี 80 สปี 100 ชุดข้อมูล

กราฟแสดงจำนวนสปีที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสองตำแหน่งที่ตั้ง 80 สปี แสดงในภาพที่ 4-3

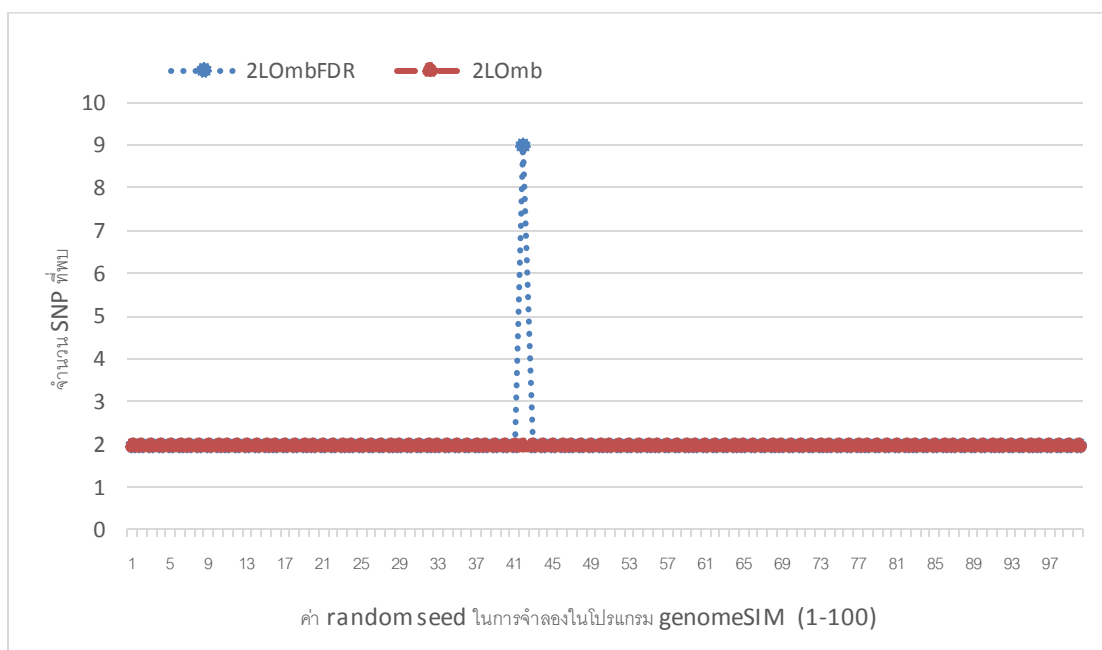


ภาพที่ 4-3 กราฟแสดงจำนวนสปีที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสองตำแหน่งที่ตั้ง 80 สปี

โดยข้อมูลทั้ง 100 ชุดข้อมูลมีจำนวนสปีที่ถูกต้องเท่ากับ 2 สปีทุกชุดข้อมูล

4.2.4 ผลการทดสอบชุดข้อมูลสองตำแหน่งที่ตั้ง จำนวนสปี 160 สปี 100 ชุดข้อมูล

กราฟแสดงจำนวนสปีที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสองตำแหน่งที่ตั้ง 160 สปี
แสดงในภาพที่ 4-4



ภาพที่ 4-4 กราฟแสดงจำนวนสปีที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสองตำแหน่งที่ตั้ง 160 สปี

โดยข้อมูลทั้ง 100 ชุดข้อมูลมีจำนวนสปีที่ถูกต้องเท่ากับ 2 สปีทุกชุดข้อมูล

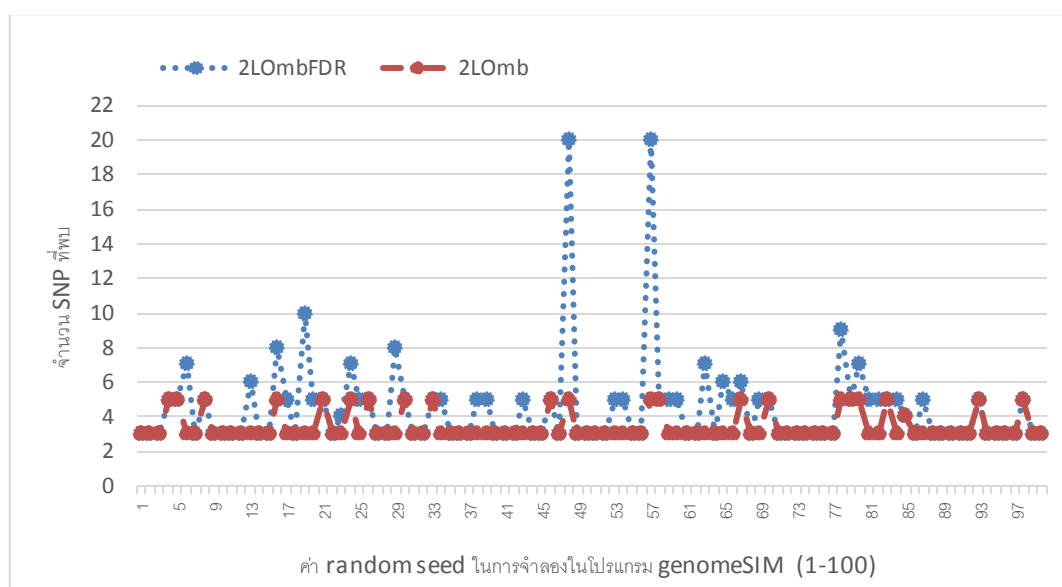
จำนวนสนิปเฉลี่ยที่พบจากการทำขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง (2LOmb) และขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง โดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) โดยใช้ชุดข้อมูลสองตำแหน่งที่ตั้ง ตามที่แสดงในตารางที่ 4-1

ตารางที่ 4-1 จำนวนสนิปเฉลี่ยที่พบจากการทำ 2LOmb และ 2LOmbFDR โดยใช้ชุดข้อมูลสองตำแหน่งที่ตั้ง

จำนวนสนิปของชุดข้อมูลสองตำแหน่งที่ตั้ง	2LOmbFDR	2LOmb
20	2.44	2.04
40	2.10	2.02
80	2.15	2.06
160	2.07	2.00

4.2.5 ผลการทดสอบชุดข้อมูลสามตำแหน่งที่ตั้ง จำนวนสนิป 20 สนิป 100 ชุดข้อมูล

กราฟแสดงจำนวนสนิปที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสามตำแหน่งที่ตั้ง 20 สนิป แสดงในภาพที่ 4-5

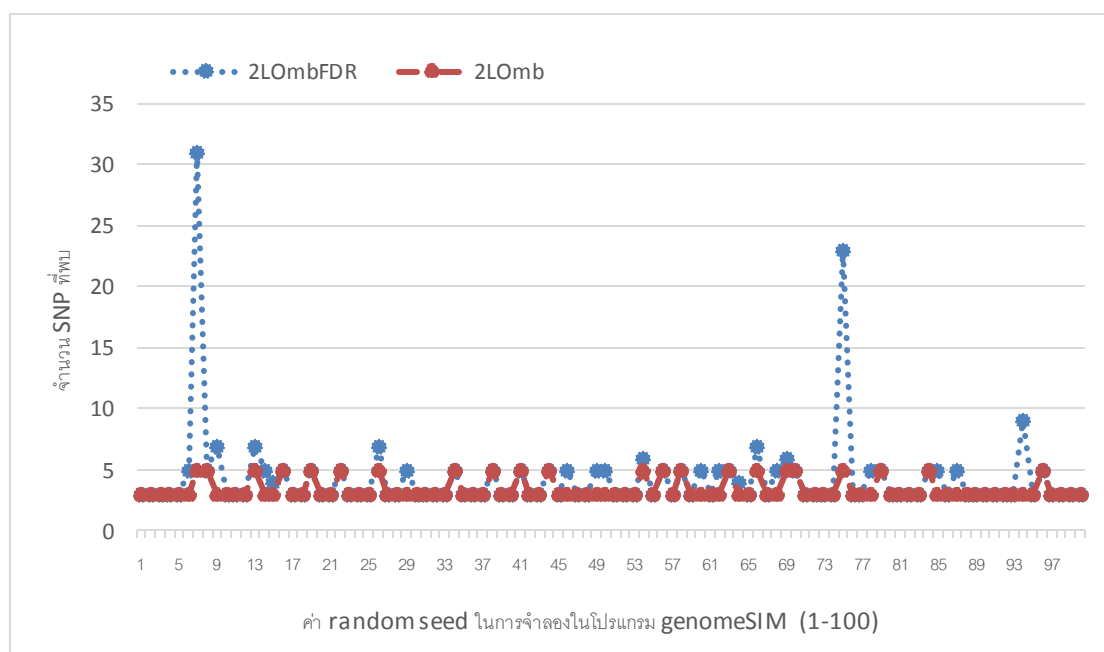


ภาพที่ 4-5 กราฟแสดงจำนวนสนิปที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสามตำแหน่งที่ตั้ง 20 สนิป

โดยข้อมูลทั้ง 100 ชุดข้อมูลมีจำนวนสปีที่ถูกต้องเท่ากับ 3 สปีทุกชุดข้อมูล

4.2.6 ผลการทดสอบชุดข้อมูลสามตำแหน่งที่ตั้ง จำนวนสปี 40 สปี 100 ชุดข้อมูล

กราฟแสดงจำนวนสปีที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสามตำแหน่งที่ตั้ง 40 สปี
แสดงในภาพที่ 4-6

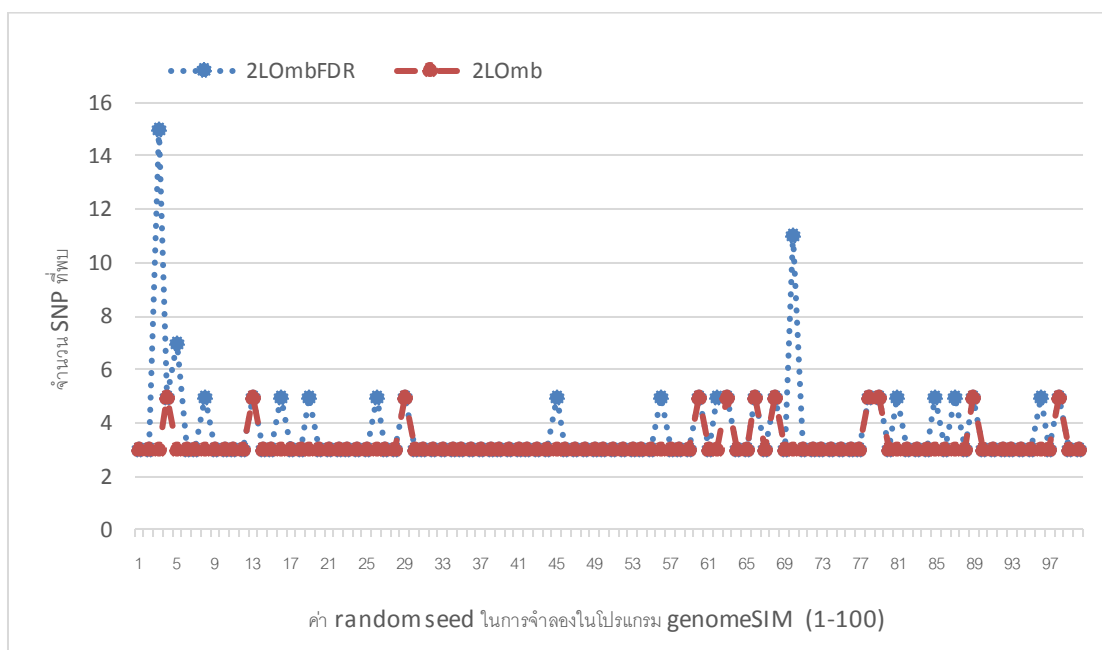


ภาพที่ 4-6 กราฟแสดงจำนวนสปีที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสามตำแหน่งที่ตั้ง 40 สปี

โดยข้อมูลทั้ง 100 ชุดข้อมูลมีจำนวนสปีที่ถูกต้องเท่ากับ 3 สปีทุกชุดข้อมูล

4.2.7 ผลการทดสอบชุดข้อมูลสามตำแหน่งที่ตั้ง จำนวนสปี 80 สปี 100 ชุดข้อมูล

กราฟแสดงจำนวนสปีที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสามตำแหน่งที่ตั้ง 80 สปี แสดงในภาพที่ 4-7

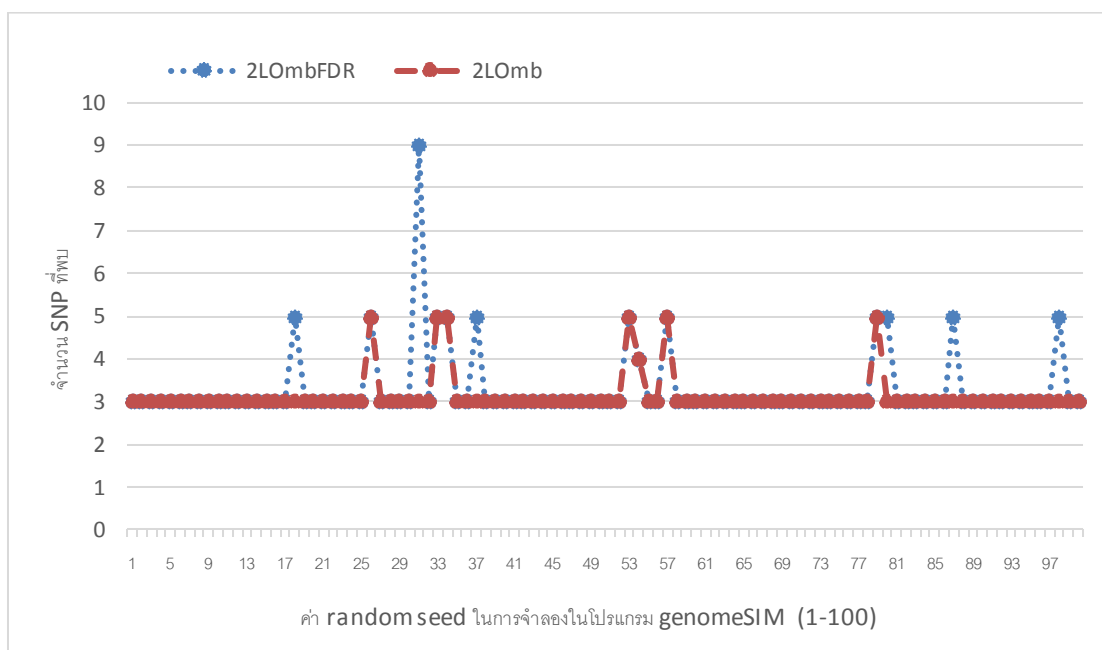


ภาพที่ 4-7 กราฟแสดงจำนวนสปีที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสามตำแหน่งที่ตั้ง 80 สปี

โดยข้อมูลทั้ง 100 ชุดข้อมูลมีจำนวนสปีที่ถูกต้องเท่ากับ 3 สปีทุกชุดข้อมูล

4.2.8 ผลการทดสอบชุดข้อมูลสามตำแหน่งที่ตั้ง จำนวนสนิป 160 สนิป 100 ชุดข้อมูล

กราฟแสดงจำนวนสนิปที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสามตำแหน่งที่ตั้ง 160 สนิป
แสดงในภาพที่ 4-8



ภาพที่ 4-8 กราฟแสดงจำนวนสนิปที่พบจากการทำขั้นตอนวิธีชุดข้อมูลสามตำแหน่งที่ตั้ง 160 สนิป

โดยข้อมูลทั้ง 100 ชุดข้อมูลมีจำนวนสนิปที่ถูกต้องเท่ากับ 3 สนิปทุกชุดข้อมูล

จำนวนสนิปเฉลี่ยที่พบจากการทำขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละสองตำแหน่งที่ตั้ง (2LOmb) และขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละสองตำแหน่งที่ตั้ง โดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) โดยใช้ชุดข้อมูลสามตำแหน่งที่ตั้ง ตามที่แสดงในตารางที่ 4-2

ตารางที่ 4-2 จำนวนสนิปเฉลี่ยที่พบจากการทำ 2LOmb และ 2LOmbFDR โดยใช้ชุดข้อมูลสามตำแหน่งที่ตั้ง

จำนวนสนิปของชุดข้อมูลสามตำแหน่งที่ตั้ง	2LOmbFDR	2LOmb
20	4.46	3.43
40	4.32	3.44
80	3.68	3.22
160	3.29	3.13

4.3 ตารางเวลาในการทดลอง

ตารางเวลาการทดลองของการทดลองทั้งหมดบน Server Intel(R) Xeon(R) X3330 2.66GHz quad-core processor and 4GB of RAM CentOS ได้ระยะเวลาเฉลี่ยที่ใช้ในการทดลองตามทีแสดงในตารางที่ 4-3 และตารางที่ 4-4

ตารางที่ 4-3 เวลาเฉลี่ยที่ใช้ในการทดลองชุดข้อมูลสองตำแหน่งที่ตั้ง หน่วยเป็นวินาที (second)

จำนวนสนิปของชุดข้อมูลสองตำแหน่งที่ตั้ง	2LOmbFDR	2LOmb
20	59.075	0.070
40	251.941	0.060
80	1101.095	0.100
160	4634.031	0.200

ตารางที่ 4-4 เวลาเฉลี่ยที่ใช้ในการทดลองชุดข้อมูลสามตำแหน่งที่ตั้ง หน่วยเป็นวินาที (second)

จำนวนสปีปของชุดข้อมูลสามตำแหน่งที่ตั้ง	2LOmbFDR	2LOmb
20	55.718	0.170
40	264.725	0.180
80	1103.677	0.190
160	4609.015	0.310

4.4 การทำการทดสอบความแตกต่างของค่าเฉลี่ยของสองประชากร (Paired t-Test)

จากการใช้ T-TEST ใน Microsoft excel โดยใช้คำสั่ง T.TEST (array1,array2,2,1)

ซึ่งทำการเปรียบเทียบชุดข้อมูลออกทั้งหมด 800 ชุดข้อมูลที่ผ่านขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง (2LOmb) และขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง โดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) มาเปรียบเทียบกัน ได้ค่า P-Value ออกมาเท่ากับ 4.36709E-10 ซึ่งแปลว่าค่าเฉลี่ยที่ได้จากการทดลองขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง (2LOmb) และขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง โดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ และขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครั้งละสองตำแหน่งที่ตั้ง โดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) มี false positive มากกว่าอย่างมีนัยสำคัญทางสถิติ

บทที่ 5

สรุปผลการวิจัย

ในปริณญานิพนธ์ฉบับนี้ได้นำเสนอผลการเปรียบเทียบขั้นตอนวิธีระหว่างขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละสองตำแหน่งที่ตั้ง (2LOmb) และขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละสองตำแหน่งที่ตั้งโดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) โดยทำการทดลองกับชุดข้อมูลทั้งหมด 800 ชุดข้อมูล

จากการทดลองแต่ละชุดข้อมูลพบว่าทั้งสองขั้นตอนวิธีสามารถตรวจพบสนิปที่ส่งผลให้เกิดโรคครบทั้งสองตำแหน่งทุกชุดข้อมูล ซึ่งทำให้การทำการทดสอบความแตกต่างของค่าเฉลี่ยของจำนวนสนิปที่ส่งผลให้เกิดโรคจากทั้งสองขั้นตอนวิธีไม่แตกต่างกัน และจากการทำการทดสอบความแตกต่างของค่าเฉลี่ยของจำนวนสนิปของข้อมูลออกจากการทำแต่ละขั้นตอนวิธีพบว่าค่าเฉลี่ยของจำนวนสนิปของข้อมูลออกจากการทำแต่ละขั้นตอนวิธีที่พบมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติคือมีค่า P-Value น้อยกว่า 0.05 และขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละสองตำแหน่งที่ตั้งโดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) มี false positive มากกว่าอย่างมีนัยสำคัญทางสถิติ

ประเด็นเรื่องเวลาเวลาของการทำขั้นตอนวิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการวิเคราะห์ครึ่งละสองตำแหน่งที่ตั้งโดยการใช้การควบคุมอัตราการค้นพบเท็จ (2LOmbFDR) ใช้เวลามากพอที่จะแสดงให้เห็นการเติบโตของเวลาว่าเป็นแบบ $O(mn^2)$ โดยที่ m คือจำนวน sample และ n คือจำนวนสนิป โดยในการทดลองไม่ได้เปลี่ยนจำนวน sample แต่เปลี่ยนแค่จำนวนสนิปอย่างเดียว และจากผลของเวลาที่ใช้ที่ได้จากการทดลองในตารางที่ 4-3 และตารางที่ 4-4 เนื่องจากจำนวนสนิปเพิ่มขึ้นทีละ 2 เท่า และเวลาเพิ่มขึ้นประมาณ 4 เท่า ทำให้เข้ารูปแบบการเติบโตของเวลาว่าเป็นแบบ $O(mn^2)$

เอกสารอ้างอิง

1. Heather J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. In : Human Molecular Genetics. Cambridge, UK : Oxford University Press ; 2002. Vol.20. No.11. p.2463-2468.
2. Am. J. Hum. Genet. A Perspective on Epistasis: Limits of Models Displaying No Main Effect. St. Louis ; 2002. p.461-471.
3. Scott M. Dudek, et al. Data Simulation Software for Whole-Genome Association and Other Studies in Human Genetics. Pacific Symposium on Biocomputing ; 2006. p.499-510.
4. William H. Press, et al. Numerical Recipes in C : the art of scientific computing. Cambridge : Cambridge University Press ; 1992.
5. John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. In : PNAS. Stanford ; 2003. Vol.100. No.16. p.9440-9445.
6. Waranyu Wongseree, et al. Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses. In : BMC Bioinformatics. BioMed Central Ltd. ; 2009.
7. Damrongrit Setsirichok, 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.
8. J.D.Storey, J.E.Taylor and D.Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Royal Statistical Society ; 2004. p.187-205.

ประวัติผู้แต่ง

ปริญญานิพนธ์เรื่อง : ผลกระทบของการควบคุมอัตราการค้าคนพม่าที่มีต่อขั้นตอน
วิธีการทดสอบการเรียงสับเปลี่ยนของกลุ่มการรวมผลการ
วิเคราะห์ครึ่งละสองตำแหน่งที่ตั้ง

สาขาวิชา : วิศวกรรมคอมพิวเตอร์

ภาควิชา : วิศวกรรมไฟฟ้าและคอมพิวเตอร์

คณะ : วิศวกรรมศาสตร์

ชื่อ : นางสาวจิตตินันท์ ตั้งสุนันท์ธรรม

ประวัติ

เกิดเมื่อวันที่ 15 เมษายน พ.ศ.2537 อยู่บ้านเลขที่ 38 ซอยเรวดี57แยก8 ตำบลตลาดขวัญ
อำเภอเมือง จังหวัดนนทบุรี สำเร็จการศึกษามัธยมศึกษาตอนปลายจากโรงเรียนสตรีนนทบุรี
จังหวัดนนทบุรี สาขาวิทยาศาสตร์- คณิตศาสตร์ ปีการศึกษา 2554 และสำเร็จการศึกษาในระดับ
ปริญญาตรี สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะ
วิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปีการศึกษา 2558

ชื่อ : นางสาวเปรมมิกา ชัยพรหม

ประวัติ

เกิดเมื่อวันที่ 7 ตุลาคม พ.ศ.2536 อยู่บ้านเลขที่ 30/32 ตำบลบ้านม้า อำเภอบางไทร จังหวัด
พระนครศรีอยุธยา สำเร็จการศึกษามัธยมศึกษาตอนปลายจากโรงเรียนคณะราษฎรบำรุงปทุมธานี
จังหวัดปทุมธานี สาขาวิทยาศาสตร์- คณิตศาสตร์ ปีการศึกษา 2554 และสำเร็จการศึกษาในระดับ
ปริญญาตรี สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะ
วิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปีการศึกษา 2558

ประวัติผู้แต่ง(ต่อ)

ชื่อ : นายวรากร ชินวรากร

ประวัติ

เกิดเมื่อวันที่ 14 สิงหาคม พ.ศ.2537 อยู่บ้านเลขที่ 864/23 ซอยพรานนก19 แขวงบ้านช่างหล่อ เขตบางกอกน้อย จังหวัดกรุงเทพฯ สำเร็จการศึกษามัธยมศึกษาตอนปลายจากโรงเรียนทิวธาภิเศก จังหวัดกรุงเทพฯ สาขาวิทยาศาสตร์-คณิตศาสตร์ ปีการศึกษา 2554 และสำเร็จการศึกษาในระดับปริญญาตรี สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปีการศึกษา 2558