

การคัดเลือกสปีตตัวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนม

นายธรรมบุญ กิจรสอนันต์

นางสาววรากร มณีแสง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร วิศวกรรมศาสตรบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

ปีการศึกษา 2563

Tag Single Nucleotide Polymorphism Selection from Genome-Wide
Association

Mr. Tummanoon Kitcharasanan

Ms. Warakorn Maneesang

FOR THE DEGREE OF BACHLOR ELECTRICAL ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

FACULTY OF ENGINEERING

KING MONGKUT'S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK

ACADEMIC YEAR 2020

ใบรับรองปริญญาานิพนธ์

ปริญญาานิพนธ์เรื่อง : การคัดเลือกสปีดวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนม

ชื่อ : นายธรรมนุญ กิจรสอนันต์ รหัสนักศึกษา 6001012630047

นางสาววรากร มณีแสง รหัสนักศึกษา 6001012630136

สาขาวิชา : วิศวกรรมคอมพิวเตอร์

ภาควิชา : วิศวกรรมไฟฟ้าและคอมพิวเตอร์

คณะ : วิศวกรรมศาสตร์

ที่ปรึกษา : รองศาสตราจารย์ ดร.ณชล ไชยรัตน์นะ

ผู้ช่วยศาสตราจารย์ ดร.วรัญญู วงษ์เสรี

ผู้ช่วยศาสตราจารย์ ดร.ดำรงค์ฤทธิ์ เศรษฐศิริโชค

ปีการศึกษา : 2563

ได้รับอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

..... หัวหน้าภาควิชาวิศวกรรมไฟฟ้า

(ผู้ช่วยศาสตราจารย์ ดร.ณชล ไชยรัตน์นะ) และคอมพิวเตอร์

..... ประธานกรรมการ

(ศาสตราจารย์ ดร.ณชล ไชยรัตน์นะ)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วรัญญู วงษ์เสรี)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ดำรงค์ฤทธิ์ เศรษฐศิริโชค)

Approval Project Certificate

Project : Tag Single Nucleotide Polymorphism Selection from Genome-Wide Association

Name : Mr. Tummanoon Kitcharasanan ID. 6001012630047

Ms. Warakorn Maneesange ID. 6001012630136

Major : Computer Engineering

Department : Electrical and Computer Engineering

Faculty : Engineering

Project Advisors : Prof. Dr. Nachol Chaiyaratana
Asst. Prof. Dr. Waranyu Wongseree
Asst. Prof. Dr. Damrongrit Setsirichok

Academic Year : 2020

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

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

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

..... Member
(Asst. Prof. Dr. Waranyu Wongseree)

..... Member
(Asst. Prof. Dr. Damrongrit Setsirichok)

บทคัดย่อ

โครงการนี้นำเสนอโปรโตคอลสำหรับการเชื่อมโยงข้อมูล Single Nucleotide Polymorphism หรือสเนป (SNP) จากศูนย์ข้อมูลเทคโนโลยีชีวภาพแห่งชาติ (NCBI) ที่สร้างครั้งที่ 36 ให้เป็นไปตามครั้งที่ 37 โปรโตคอลนี้ใช้ข้อมูลจากแหล่งต่าง ๆ ทั้งจาก NCBI, บทความ และข้อมูลจากแพลตฟอร์มสำหรับเก็บข้อมูลจีโนมใหญ่ การใช้โปรโตคอลนี้ถูกแสดงให้เห็นบนชุดข้อมูลจากการศึกษาความสัมพันธ์ทั้งจีโนมทั้งเจ็ดชุดข้อมูล ซึ่งดำเนินการโดย Wellcome Trust Case Control Consortium (WTCCC) การศึกษานี้มุ่งที่จะระบุถึงปัจจัยทางพันธุกรรมที่มีส่วนต่อโรคที่มีผลต่อการเกิดโรคซับซ้อนทั้งเจ็ดโรค ได้แก่ โรคอารมณ์สองขั้ว, โรคหลอดเลือดโคโรนารีหรือภาวะหัวใจขาดเลือด, โรคโครห์นหรือโรคที่เกิดการอักเสบเรื้อรังของระบบทางเดินอาหาร, โรคความดันโลหิตสูง, โรคข้ออักเสบรูมาตอยด์, โรคเบาหวานชนิดที่ 1 และโรคเบาหวานชนิดที่ 2 ชุดข้อมูลทั้งหมดมี 500,568 สเนป ซึ่งได้จากการเก็บข้อมูลจีโนมใหญ่โดย Affymetrix GeneChip Human Mapping 500K Array Set โปรโตคอลนี้สามารถนำไปใช้กับชุดข้อมูลทั้งหมดได้สำเร็จ ซึ่งทำให้เกิดการเชื่อมโยงข้อมูลสเนปที่สมบูรณ์ ด้วยข้อมูลสเนปที่พร้อมใช้งานตาม NCBI ที่สร้างครั้งที่ 37 จึงเป็นไปได้ที่จะคัดเลือกตัวแทนสเนปจาก 364,772 สเนปที่ผ่านการตรวจสอบคุณภาพที่กำหนดโดย WTCCC, Affymetrix และมีความถี่ของอัลลีลที่พบมากกว่า 0.05 วิธีการสำหรับการคัดเลือกตัวแทนสเนปคือ Tagger ซึ่งเป็นส่วนหนึ่งของซอฟต์แวร์ Haploview Tagger ได้เลือกสเนปประมาณ 55.45% เพื่อใช้เป็นตัวแทนสเนปจากแต่ละชุดข้อมูลควบคุมเมื่อกำหนดค่า r^2 ไว้ที่ 0.8 ซึ่งเป็นค่าขีดเริ่มเปลี่ยนทั่วไปสำหรับการกำหนดค่าภาวะความไม่สมดุลการเชื่อมโยงของโรคระหว่างสองตำแหน่ง เพอร์เซ็นต์ของตัวแทนสเนปที่ถูกเลือกนี้ มีความใกล้เคียงกับผลในรายงานการศึกษาก่อนหน้านี้ที่ใช้ชุดข้อมูล CEU (ชาวยุโรปที่มีเชื้อสายยุโรปเหนือและยุโรปตะวันตก) ซึ่งได้มาจากโครงการ International HapMap

Abstract

This project presents a protocol for mapping single nucleotide polymorphism (SNP) information from that according to the National Center for Biotechnology Information (NCBI) build 36 to that according to NCBI build 37. The protocol exploited information from various sources including NCBI, early literature and information provided by a genotyping platform manufacturer. The applicability of the protocol was demonstrated on the datasets from seven genome-wide association studies conducted by the Wellcome Trust Case Control Consortium (WTCCC). The studies attempt to identify genetic factors contributing to seven complex diseases: bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes. All datasets contain 500,568 SNPs, which were genotyped using the Affymetrix GeneChip Human Mapping 500K Array Set. The protocol was successfully applied to the datasets resulting in a complete SNP information mapping. With the availability of SNP information according to NCBI build 37, it is possible to extract tag SNPs from 364,772 SNPs that passed the quality control dictated by WTCCC, Affymetrix and a minor allele frequency threshold of 0.05. The chosen technique for tag SNP extraction was Tagger, which is a part of the Haploview software. Tagger chose approximately 55.45% of SNPs as tag SNPs from each case-control dataset when the r^2 threshold setting was 0.8, which is a common default threshold for defining linkage disequilibrium between two loci. The percentage of chosen tag SNPs was close to that reported in an early study involving the CEU (Utah residents with Northern and Western European ancestry) dataset obtained from the International HapMap Project.

กิตติกรรมประกาศ

ปริญญานิพนธ์เล่มนี้ไม่อาจเสร็จสมบูรณ์ได้หากปราศจากความช่วยเหลือจากศาสตราจารย์ ดร.ณชล ไชยรัตน์นะ ผู้ช่วยศาสตราจารย์ ดร.ดำรงศฤทธิ์ เศรษฐศิริโชค และผู้ช่วยศาสตราจารย์ ดร. วรัญญ วงษ์เสรี ที่คอยให้คำแนะนำและให้การสนับสนุน ตลอดทั้งการให้ความช่วยเหลือในทุก ๆ ด้าน จนทำให้ปริญญานิพนธ์เล่มนี้เสร็จสมบูรณ์ออกมาครบถ้วน ต้องขอขอบพระคุณอาจารย์ทุกท่านมา ณ โอกาสนี้

ข้าพเจ้าขอขอบคุณอาจารย์ท่านอื่น ๆ ในภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะ วิศวกรรมศาสตร์ ทุก ๆ ท่านที่คอยให้ความรู้ คำแนะนำ และคอยสั่งสอนข้าพเจ้าตลอดระยะเวลาที่ ศึกษาอยู่ ณ ที่แห่งนี้ จนข้าพเจ้าสามารถนำความรู้ที่ได้นำไปใช้ในการประกอบอาชีพในอนาคต

สุดท้ายนี้ต้องขอขอบคุณ นายอิสระ กุลอุดมชัยวัฒน์ รวมถึงเพื่อน ๆ ทุกคน รุ่นพี่ รุ่นน้อง และบุคลากรของสาขาวิชาวิศวกรรมคอมพิวเตอร์ทุกท่าน ที่คอยให้ความช่วยเหลือในตลอดระยะเวลา ที่ผ่านมา

ธรรมบุญ กิจรสอนันต์

วรากร มณีแสง

สารบัญ

	หน้า
ใบรับรองปริญญานิพนธ์	ก
Approval Project Certificate	ข
บทคัดย่อ	ค
Abstact	ง
กิตติกรรมประกาศ	จ
สารบัญภาพ	ซ
สารบัญตาราง	ฌ
บทที่ 1 บทนำ	1
1.1 ที่มาและความสำคัญ	1
1.2 วัตถุประสงค์	1
1.3 ขอบเขตของการทำโครงการ	1
1.4 ผลประโยชน์ที่คาดว่าจะได้รับ	1
1.5 แผนการดำเนินงาน	2
บทที่ 2 ทฤษฎี	3
2.1 ความหลากหลายทางพันธุกรรม	3
2.2 Tagger	3
2.3 ความถี่ของอัลลีลกลุ่มน้อย(Minor allele frequency : MAF)	4
2.4 โปรแกรม Haploview	7
2.4.1 การใช้งานโปรแกรม Haploview	10
2.4.2 ผลการศึกษาโปรแกรม Haploview	11
2.4.3 Haploview – Command line options	13
2.5 สเปคคอมพิวเตอรืที่ใช้	15
บทที่ 3 ขั้นตอนการดำเนินงาน	16
3.1 ข้อมูลที่ใช้	16
3.2 ขั้นตอนการเตรียมข้อมูล	17

สารบัญ (ต่อ)

	หน้า
3.2.1 ตรวจสอบข้อมูล WTCC กับ NSP และ STY annotations	18
3.2.2 นำสลิปมาค้นหาในเว็บ ncbi เพื่อหา RSID ตัวปัจจุบัน	20
3.2.3 นำสลิปมาค้นหาในเว็บ ncbi เพื่อหา RSID ตัวเก่า	22
3.2.4 ตรวจสอบสลิปกับไฟล์ GRCh37_hg19_AffylID2rsnumbers.txt	24
3.2.5 ตรวจสอบสลิปกับไฟล์ exclusion-list-snps-26_04_2007.txt	25
3.2.6 ตรวจสอบสลิปกับไฟล์ RsMergeArch.txt	26
3.2.7 ตรวจสอบสลิปโดยนำไปค้นหาบนเว็บ ncbi	27
3.2.8 ตรวจสอบข้อมูล WTCC ที่เหลือโดยใช้ String matching	28
3.2.9 นำสลิปเทียบเลขโครโมโซมกับ NSP และ STY annotations	29
3.3 จัดการข้อมูลให้อยู่ในรูปแบบ .ped และ .info	30
3.4 การใช้ Command line ในการคัดเลือกสลิปตัวแทน	33
บทที่ 4 ผลการดำเนินงาน	34
4.1 ผลลัพธ์การคัดเลือกสลิปตัวแทน	34
4.2 การศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรม Mapping250K_NSP.na32.annot รวมกับ Mapping250K_STY.na32.annot	35
4.3 การศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรม Mapping250K_NSP.na32.annot	43
4.4 การศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรม Mapping250K_STY.na32.annot	51
เอกสารอ้างอิง	59
ประวัติผู้แต่ง	61

สารบัญภาพ

ภาพที่	หน้า
2.1 แสดงข้อมูล (input) นามสกุล .ped ที่ใช้ในการศึกษาด้วยโปรแกรม Haploview	8
2.2 แสดงข้อมูล (input) นามสกุล .info ที่ใช้ในการศึกษาด้วยโปรแกรม Haploview	9
2.3 แสดงหน้าจอหลักของโปรแกรม Haploview	10
2.4 แสดงผลการวิเคราะห์จากการตรวจสอบคุณสมบัติของตำแหน่งสনিป	11
2.5 แสดงผลการวิเคราะห์ Linkage disequilibrium (LD Plot)	12
2.6 แสดงผลการวิเคราะห์บล็อก (Blocks) และแฮปโลไทป์ (Haplotype)	13
3.1 แสดงข้อมูลของโครโมโซม 22	16
3.2 แสดงข้อมูลที่ใช้ในการอ้างอิงของไฟล์ NSP Annotations	16
3.3 แสดงข้อมูลที่ใช้ในการอ้างอิงของไฟล์ STY Annotations	17
3.4 แสดงขั้นตอนการเตรียมข้อมูล	17
3.5 แสดงแผนภาพการคัดเลือก Prob Set ID และ RS ID	18
3.6 ผลลัพธ์ของการคัดเลือก Probe Set ID และ RS ID จากไฟล์ข้อมูล	18
3.7 แสดงแผนภาพการเปรียบเทียบข้อมูล WTCC กับข้อมูล annotations	19
3.8 ผลลัพธ์ของ Probe Set ID และ RS ID ที่ไม่ปรากฏกับ NSP annotations และ STY annotations	19
3.9 แสดงการค้นหาสนิปจากเว็บ ncbi	20
3.10 แสดงผลลัพธ์การค้นหาสนิปจากเว็บ ncbi	20
3.11 แสดงการใช้งาน BeautifulSoup package	21
3.12 แสดงการใช้งาน Regular Expression	21
3.13 แสดงผลลัพธ์การค้นหาสนิปจากเว็บ ncbi ในรูปแบบไฟล์ text	21
3.14 แสดงผลลัพธ์การค้นหาสนิปจากเว็บ ncbi	22
3.15 แสดงผลลัพธ์การค้นหาสนิปจากเว็บ ncbi แสดงผลลัพธ์การค้นหาสนิปจากเว็บ ncbi ในรูปแบบไฟล์ text	23
3.16 แสดงผลลัพธ์การเปรียบเทียบสนิปกับ NSP และ STY annotations	23

สารบัญภาพ (ต่อ)

ภาพที่	หน้า
3.17 ข้อมูลไฟล์ GRCh37_hg19_AffylID2rsnumbers.txt	24
3.18 แสดงผลลัพธ์การตรวจสอบสลิปว่าปรากฏในไฟล์ GRCh37_hg19_AffylID2rsnumbers.txt	25
3.19 ข้อมูลไฟล์ exclusion-list-snps-26_04_2007.txt	25
3.20 แสดงผลลัพธ์การตรวจสอบสลิปว่าปรากฏในไฟล์ exclusion-listsnps-26_04_2007.txt	25
3.21 ข้อมูลไฟล์ RsMergeArch.txt	26
3.22 แสดงผลลัพธ์การตรวจสอบการปรากฏของสลิปในไฟล์ RsMergeArch.txt	26
3.23 แสดงผลลัพธ์จากการค้นหาสลิปบนเว็บ ncbi	27
3.24 แสดงผลลัพธ์จากการค้นหาสลิปทั้งหมดบนเว็บ ncbi	27
3.25 แสดงการทำ flank	28
3.26 ผลลัพธ์การทำ string matching	28
3.27 แสดงการเปรียบเทียบสลิปกับไฟล์ annotations	29
3.28 แสดงข้อมูลนามสกุล .ped ที่ใช้ในการศึกษาด้วยโปรแกรม Haploview	30
3.29 แสดงข้อมูลนามสกุล .info ที่ใช้ในการศึกษาด้วยโปรแกรม Haploview	30
3.30 แสดงการใช้งาน Command line ในการคัดเลือกสลิปตัวแทน	33
3.31 แสดงผลการคัดเลือกสลิปตัวแทน	33

สารบัญตาราง

ตารางที่	หน้า
2.1 แสดงการคัดเลือกความถี่ของอัลลีลของ NSP และ STY annotations	4
2.2 แสดงการคัดเลือกความถี่ของอัลลีลของ NSP annotations	5
2.3 แสดงการคัดเลือกความถี่ของอัลลีลของ STY annotations	6
3.1 แสดงให้เห็นการผิดพลาดของเลขโครโมโซม	31
3.2 แสดงให้เห็นสลิปที่ไม่พบ position จากข้อมูล NSP และ STY Annotation	32
4.1 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค BD โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation	35
4.2 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค CAD โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation	36
4.3 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค CD โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation	37
4.4 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค HT โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation	38
4.5 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค RA โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation	39
4.6 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค T1D โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation	40
4.7 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค T2D โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation	41
4.8 แสดงค่าเฉลี่ยของจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับ โรคทั้ง 7 โรค โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation	42
4.9 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค BD โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation	43
4.10 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค CAD โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation	44

สารบัญตาราง (ต่อ)

ตารางที่	หน้า
4.11 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค CD โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation	45
4.12 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค HT โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation	46
4.13 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค RA โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation	47
4.14 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค T1D โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation	48
4.15 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค T1D โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation	49
4.16 แสดงค่าเฉลี่ยของจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์ กับโรคทั้ง 7 โรค โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation	50
4.17 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค BD โดยใช้ข้อมูลทางพันธุกรรม STY Annotation	51
4.18 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค CAD โดยใช้ข้อมูลทางพันธุกรรม STY Annotation	52
4.19 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค CD โดยใช้ข้อมูลทางพันธุกรรม STY Annotation	53
4.20 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค HT โดยใช้ข้อมูลทางพันธุกรรม STY Annotation	54
4.21 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค RA โดยใช้ข้อมูลทางพันธุกรรม STY Annotation	55

สารบัญตาราง (ต่อ)

ตารางที่	หน้า
4.22 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค T1D โดยใช้ข้อมูลทางพันธุกรรม STY Annotation	56
4.23 แสดงจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค T2D โดยใช้ข้อมูลทางพันธุกรรม STY Annotation	57
4.24 ตารางที่ 4.24 แสดงค่าเฉลี่ยของจำนวนการคัดเลือกสนิปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์ กับโรคทั้ง 7 โรค โดยใช้ข้อมูลทางพันธุกรรม STY Annotation	58

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญ

เป้าหมายของการศึกษาความสัมพันธ์ทั้งจีโนม (Genome-Wide Association Study) คือ การค้นหาสัณย (Single Nucleotide Polymorphism หรือ SNP) ในจีโนม (Genome) ที่สัมพันธ์กับโรคซับซ้อน (Complex Disease) ที่สนใจ เนื่องจากสัณยในข้อมูลมีจำนวนมาก ดังนั้นจึงคัดเลือกสัณยตัวแทน (Tag SNP) ซึ่งสามารถใช้เป็นตัวแทนของสัณยที่อยู่ในภาวะความไม่สมดุลการเชื่อมโยง (Linkage Disequilibrium) กับสัณยตัวแทนจึงเป็นเรื่องจำเป็นสำหรับการวิเคราะห์ข้อมูล WTCC โดยใช้เทคนิคทางชีวสารสนเทศศาสตร์ (Bioinformatics)

1.2 วัตถุประสงค์

เพื่อค้นหาวิธีที่เหมาะสมสำหรับคัดเลือกสัณยตัวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนม

1.3 ขอบเขตของการทำโครงการ

ทำการคัดเลือกสัณยตัวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรมเฉพาะจากฐานข้อมูลที่เข้าถึงได้เท่านั้น

1.4 ผลประโยชน์ที่คาดว่าจะได้รับ

ได้สัณยตัวแทนซึ่งสามารถใช้เป็นตัวแทนของสัณยที่อยู่ในภาวะความไม่สมดุลการเชื่อมโยงกับสัณยตัวแทนในการศึกษาความสัมพันธ์ทั้งจีโนม

1.5 แผนการดำเนินงาน

[illegible]

บทที่ 2

ทฤษฎี

2.1 ความหลากหลายทางพันธุกรรม (Single Nucleotide polymorphism : SNP) หรือ สนิป

สนิป หมายถึง ความแตกต่างหรือความหลากหลายทางพันธุกรรมระหว่างมนุษย์แต่ละคน ที่เกิดจากการเปลี่ยนแปลงลำดับเบสบนสายนิวคลีโอไทด์เพียงตำแหน่งเดียวที่ก่อให้เกิดผลที่แตกต่างกัน เช่น การเปลี่ยนลำดับเบสบนดีเอ็นเอจาก ATGGCTAA เป็น ATGGCTTA

สนิปจะถูกถ่ายทอดทางพันธุกรรมตามกฎของเมนเดล ความแตกต่างทางพันธุกรรมในแต่ละจุดนี้ทำให้มนุษย์แต่ละคนแตกต่างกันออกไป เช่น ความสูงต่างกัน ผิวสีต่างกัน เป็นโรคแตกต่างกัน เป็นต้น ความแตกต่างนั้นสามารถนำมาใช้เป็นเครื่องหมายชีวภาพสำหรับการสร้างแผนที่พันธุกรรม (genetic map) หรือนำมาศึกษาโรคที่ก่อให้เกิดโรคร้ายโรค ซึ่งจะเป็นประโยชน์ต่อการวินิจฉัยโรค

2.2 Tagger

Tagger คือ เครื่องมือสำหรับการทำ selection และ evaluation ของ tag SNPs จากข้อมูลจีโนมไทป์ ซึ่งเป็นการรวมความเรียบง่ายของวิธีการทำ pairwise tagging เข้ากับประโยชน์ด้านประสิทธิภาพของวิธี multimarker haplotype

อินพุตที่ผู้ใช้นำเข้าข้อมูลคือ ข้อมูลจีโนมไทป์ในรูปแบบ HapMap หรือ รูปแบบ pedigree ซึ่งจะคำนวณรูปแบบการเชื่อมโยงที่เป็นโรค หรืออีกวิธีหนึ่งคือผู้ใช้สามารถระบุ position ของโครโมโซม เพื่อระบุจีโนมที่สนใจ ซึ่งคุณลักษณะนี้จะเป็นประโยชน์อย่างยิ่งสำหรับการออกแบบ multiplex tag SNP ส่วนเอาต์พุตของ Tagger จะสร้างรายการ SNP ของแท็กและการทดสอบทางสถิติที่เกี่ยวข้องเพื่อรวบรวมตัวแปรที่สนใจทั้งหมดและรายงานความครอบคลุมโดยสรุปของ SNPs ของแท็กที่เลือก

Tagger ได้รับการพัฒนาโดย Paul de Bakker ในห้องทดลองของ David Altshuler และ Mark Daly ที่ศูนย์วิจัยพันธุกรรมมนุษย์ของโรงพยาบาล Massachusetts General Hospital และ Harvard Medical School และ Broad Institute

2.3 ความถี่ของอัลลีลกลุ่มน้อย (Minor allele frequency : MAF)

ความถี่ของอัลลีลกลุ่มน้อย คือ ความถี่ของอัลลีลที่พบบ่อยเป็นอันดับสองของประชากรที่สนใจ MAF ใช้กันอย่างแพร่หลายในการศึกษาพันธุศาสตร์ประชากร เนื่องจากใช้เพื่อแยกความแตกต่างระหว่างตัวแปรที่พบบ่อยและหายากในประชากร

จำนวนส니ปทั้งหมดที่จะใช้ในการวิเคราะห์จะต้องมีความถี่ของอัลลีลกลุ่มน้อย (Minor allele frequency) ต้องไม่ต่ำกว่า 0.05 เพื่อเป็นการขจัดสนิปที่อาจทำให้เกิดความคลื่อนออกไป

2.4 โปรแกรม Haploview

โปรแกรมแฮพลอวิว (Haploview) ใช้ในการวิเคราะห์ผลสถิติทางด้านพันธุศาสตร์จากข้อมูลจีโนมไทป์ของตัวอย่างแต่ละราย โปรแกรมสามารถแสดงค่าทางสถิติได้หลายรูปแบบ เช่น ค่าเฮเทอโรไซโกซิตี (Heterozygosity) ค่าฮาร์ดี-ไวเบิร์กทีสภาวะสมดุล (Hardy-Weinberg equilibrium: HWE) ร้อยละของจีโนมไทป์และความถี่ของอัลลีลกลุ่มน้อย (Minor allele frequency) อีกทั้งศึกษา Linkage disequilibrium (LD) ของตำแหน่งสนิปแต่ละตำแหน่งในแต่ละยีนส์และตำแหน่งสนิปว่ามีสนิปดังกล่าวอยู่ใน LD บล็อกเดียวกันหรือไม่ หรือมีการถ่ายทอดไปด้วยกันหรือไม่ ซึ่งเป็นโปรแกรมที่ใช้งานได้สะดวก นำเสนอข้อมูลในรูปแบบที่เข้าใจได้ง่าย

การนำเข้าข้อมูล (Input) สามารถนำเข้าข้อมูลได้หลายลักษณะ แต่ที่นิยมใช้แบบ standard linkage format โดยข้อมูลที่นำมาศึกษานั้นเป็น PED file และ INFO file

PED file ประกอบด้วย Family ID , Individual ID , Paternal ID , Maternal ID , Sex , Phenotype

Family ID	Individual ID	Paternal & Maternal ID		Sex	Phenotype	Genotype							
WTCCC125760	WTCCC125760	0	0	1	1	1	1	3	3	1	1	4	
WTCCC126352	WTCCC126352	0	0	2	1	1	3	3	3	1	1	4	
WTCCC126179	WTCCC126179	0	0	1	1	1	1	3	3	1	1	4	
WTCCC126013	WTCCC126013	0	0	2	1	1	1	3	3	1	1	4	
WTCCC126214	WTCCC126214	0	0	1	1	1	1	3	3	1	1	4	
WTCCC127641	WTCCC127641	0	0	1	1	3	3	3	3	1	1	4	
WTCCC126470	WTCCC126470	0	0	2	1	1	1	3	3	1	1	4	
WTCCC126474	WTCCC126474	0	0	1	1	1	1	3	3	1	1	4	
WTCCC126042	WTCCC126042	0	0	2	1	1	1	3	3	1	1	4	
WTCCC126864	WTCCC126864	0	0	1	1	1	3	3	3	1	1	4	
WTCCC126050	WTCCC126050	0	0	1	1	1	3	3	3	1	1	4	
WTCCC126693	WTCCC126693	0	0	1	1	1	3	3	3	1	1	4	
WTCCC127526	WTCCC127526	0	0	2	1	1	3	3	3	1	1	4	
WTCCC126795	WTCCC126795	0	0	2	1	1	1	3	3	1	1	4	
WTCCC126888	WTCCC126888	0	0	2	1	1	3	3	3	1	1	4	
WTCCC126978	WTCCC126978	0	0	1	1	1	1	3	3	1	1	4	
WTCCC126043	WTCCC126043	0	0	1	1	1	1	3	3	1	1	4	
WTCCC127116	WTCCC127116	0	0	2	1	1	1	3	3	1	1	4	

ภาพที่ 2.1 แสดงข้อมูล (input) นามสกุล .ped ที่ใช้ในการศึกษาด้วยโปรแกรม Haploview

Family ID	เป็นตัวเลขแสดง Individual family โดยตัวเลขจะเหมือนกันในกรณีที่เป็นข้อมูลที่มาจากรอบครัวเดียวกัน แต่ถ้าเป็นข้อมูลที่มาจากร่างครอบครัวตัวเลขจะไม่เหมือนกัน
Individual ID	เป็นหมายเลขประจำตัวเฉพาะของแต่ละบุคคล
Paternal ID	เป็นหมายเลขประจำตัวเฉพาะของแต่ละบุคคลที่เป็นบิดาของบุคคลนั้น ในกรณีวิเคราะห์แบบ Association analysis ไม่ใช่ข้อมูลครอบครัวจึงกำหนดค่าบิดาเป็น 0
Maternal ID	เป็นหมายเลขประจำตัวเฉพาะของแต่ละบุคคลที่เป็นมารดาของบุคคลนั้น ในกรณีวิเคราะห์แบบ Association analysis ไม่ใช่ข้อมูลครอบครัวจึงกำหนดค่ามารดาเป็น 0
Sex	เป็นการระบุเพศ โดยใช้เลข 1 แทนเพศชาย และใช้เลข 2 แทนเพศหญิง และให้ค่าเป็น 0 เมื่อไม่ทราบเพศ
Affectation status	เป็นการระบุสถานะการเป็นโรค โดยเลข 1 แทนสถานะที่ไม่เป็นโรค เลข 2 แทนสถานะที่เป็นโรค และเลข 0 แทนสถานะที่ไม่ทราบ
Marker genotype	แต่ละ marker แสดงจีโนไทป์ของแต่ละสปี โดยกำหนดจีโนไทป์เป็นลำดับเบส A, T, C, G หรือหมายเลข 1 – 4 (1 = A, 2 = C, 3 = G, 4 = T)

INFO file รูปแบบไฟล์จะเป็น Marker information ภายในไฟล์จะประกอบด้วย 2 คอลัมน์ คือ Marker name และ position

Marker name	Position
rs2471469	12238077
rs17036071	12244544
rs12490159	12266404
rs7614818	12270285
rs17036088	12271056
rs17671592	12273414
rs310751	12273621
rs310749	12273768
rs167467	12281183
rs11929414	12285628
rs1562041	12285734
rs17036126	12287863

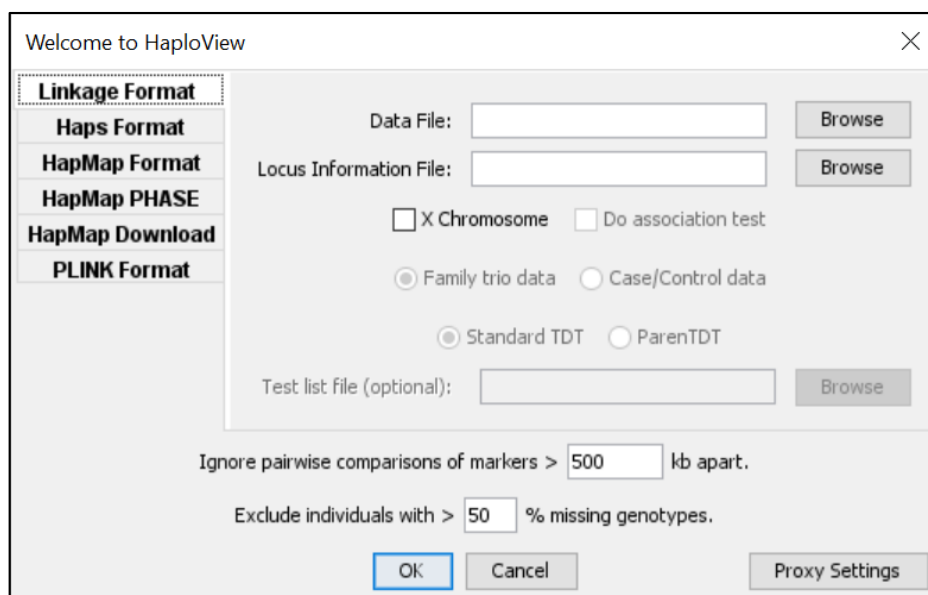
ภาพที่ 2.2 แสดงข้อมูล (input) นามสกุล .info ที่ใช้ในการศึกษาด้วยโปรแกรม Haploview

Marker name ชื่อของตำแหน่งสลับ

Position ระยะห่างของลำดับคู่เบส

2.4.1 การใช้งานโปรแกรม Haploview

1. เปิดโปรแกรม Haploview ขึ้นมา โดยสามารถเรียกใช้งานผ่าน shortcut ที่หน้า Desktop หรือจาก Start Menu > Programes > Haploview > Haploview

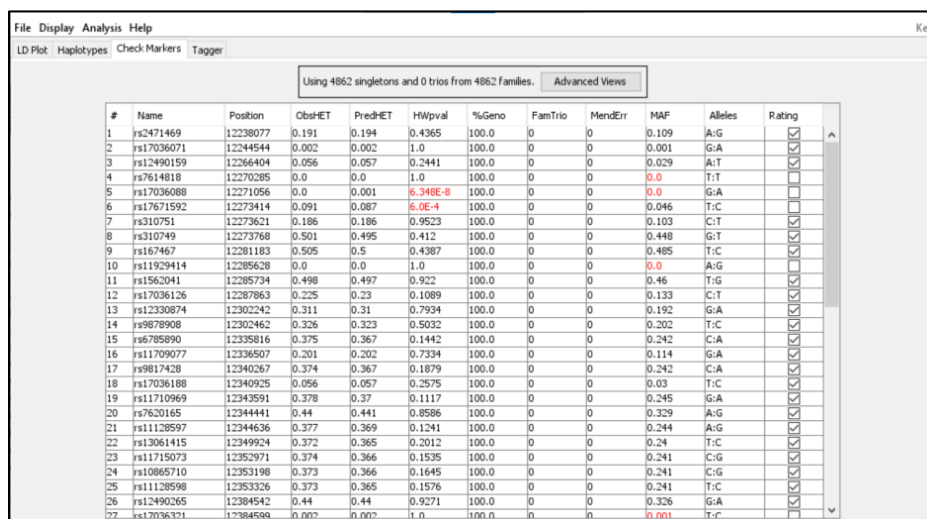


ภาพที่ 2.3 แสดงหน้าจอหลักของโปรแกรม Haploview

2. เลือกไฟล์ข้อมูลที่ต้องการนำมาศึกษาโดยใน Data file ระบุนามสกุลไฟล์เป็น .data เช่น Sample.data และใน Locus Information File ระบุนามสกุลไฟล์เป็น .info เช่น Sample.info โดยโปรแกรมจะทำการนำเข้าไฟล์ทั้งสองแบบอัตโนมัติ

2.4.2 ผลการศึกษาโปรแกรม Haploview

1. การตรวจสอบคุณสมบัติของตำแหน่งสแนป (Check Markers) โปรแกรมจะทำการคำนวณข้อมูลพื้นฐานของแต่ละสแนป และรายงานผลสำหรับการตรวจสอบคุณสมบัติต่าง ๆ



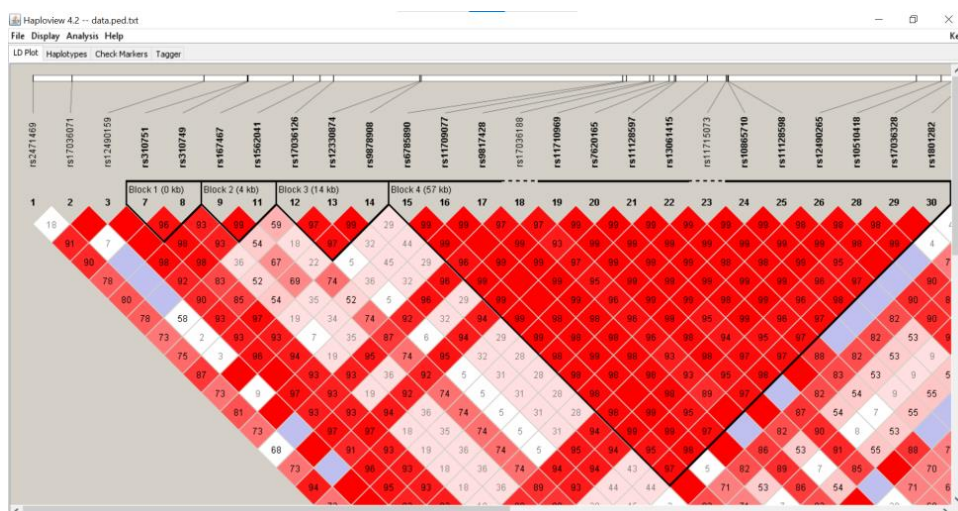
#	Name	Position	ObsHET	PredHET	HWpval	%Gene	FamTrio	MendErr	MAF	Alleles	Rating
1	rs2471469	12238077	0.191	0.194	0.4365	100.0	0	0	0.109	A:G	<input checked="" type="checkbox"/>
2	rs17036071	12244544	0.002	0.002	1.0	100.0	0	0	0.001	G:A	<input checked="" type="checkbox"/>
3	rs12490159	12266404	0.056	0.057	0.2441	100.0	0	0	0.029	A:T	<input checked="" type="checkbox"/>
4	rs7614818	12270285	0.0	0.0	1.0	100.0	0	0	0.0	T:T	<input checked="" type="checkbox"/>
5	rs17036088	12271056	0.0	0.001	6.348E-8	100.0	0	0	0.0	G:A	<input checked="" type="checkbox"/>
6	rs17671592	12273414	0.091	0.087	6.0E-4	100.0	0	0	0.046	T:C	<input checked="" type="checkbox"/>
7	rs310751	12273621	0.186	0.186	0.9523	100.0	0	0	0.103	C:T	<input checked="" type="checkbox"/>
8	rs310749	12273768	0.501	0.495	0.412	100.0	0	0	0.448	G:T	<input checked="" type="checkbox"/>
9	rs167467	12281183	0.505	0.5	0.4387	100.0	0	0	0.485	T:C	<input checked="" type="checkbox"/>
10	rs11929414	12285628	0.0	0.0	1.0	100.0	0	0	0.0	A:G	<input checked="" type="checkbox"/>
11	rs1562041	12285734	0.498	0.497	0.922	100.0	0	0	0.46	T:G	<input checked="" type="checkbox"/>
12	rs17036126	12287863	0.225	0.23	0.1089	100.0	0	0	0.133	C:T	<input checked="" type="checkbox"/>
13	rs12330874	12302242	0.311	0.31	0.7934	100.0	0	0	0.192	G:A	<input checked="" type="checkbox"/>
14	rs9878908	12302462	0.326	0.323	0.5032	100.0	0	0	0.202	T:C	<input checked="" type="checkbox"/>
15	rs6785890	12335816	0.375	0.367	0.1442	100.0	0	0	0.242	C:A	<input checked="" type="checkbox"/>
16	rs11709077	12336507	0.201	0.202	0.7334	100.0	0	0	0.114	G:A	<input checked="" type="checkbox"/>
17	rs9817428	12340267	0.374	0.367	0.1879	100.0	0	0	0.242	C:A	<input checked="" type="checkbox"/>
18	rs17036188	12340925	0.056	0.057	0.2575	100.0	0	0	0.03	T:C	<input checked="" type="checkbox"/>
19	rs11710969	12343591	0.378	0.37	0.1117	100.0	0	0	0.245	G:A	<input checked="" type="checkbox"/>
20	rs7620165	12344441	0.44	0.441	0.8586	100.0	0	0	0.329	A:G	<input checked="" type="checkbox"/>
21	rs11128597	12344636	0.377	0.369	0.1241	100.0	0	0	0.244	A:G	<input checked="" type="checkbox"/>
22	rs13061415	12349924	0.372	0.365	0.2012	100.0	0	0	0.24	T:C	<input checked="" type="checkbox"/>
23	rs11715073	12352971	0.374	0.366	0.1535	100.0	0	0	0.241	C:G	<input checked="" type="checkbox"/>
24	rs10665710	12353198	0.373	0.366	0.1645	100.0	0	0	0.241	C:G	<input checked="" type="checkbox"/>
25	rs11128598	12353326	0.373	0.365	0.1576	100.0	0	0	0.241	T:C	<input checked="" type="checkbox"/>
26	rs12490265	12384542	0.44	0.44	0.9271	100.0	0	0	0.326	G:A	<input checked="" type="checkbox"/>
27	rs17036321	12384598	0.002	0.002	1.0	100.0	0	0	0.001	T:C	<input checked="" type="checkbox"/>

ภาพที่ 2.4 แสดงผลการวิเคราะห์จากการตรวจสอบคุณสมบัติของตำแหน่งสแนป

#	เลขตำแหน่งสแนป
Name	ชื่อของนิปส์จากไฟล์ Marker Information
Position	ตำแหน่งของสแนปจากไฟล์ Marker Information
ObsHET	ค่าที่ได้จากการนับ Observed heterozygosity
PredHET	ค่าที่ได้จากการคำนวณ Predicted heterozygosity
HWpval	ค่า P-value ของ Hardy-Weinberg equilibrium คือโอกาสความน่าจะเป็นที่ข้อมูลจีโนไทป์นี้มีการกระจายแบบสมดุล
%Gene	ค่าเปอร์เซ็นต์ของการศึกษาจีโนไทป์ที่ได้ผลของแต่ละตำแหน่งสแนป
FamTrio	จำนวนของครอบครัวที่มีผลจีโนไทป์ครบ
MendErr	จำนวนครอบครัวที่มีผลจีโนไทป์ไม่ครบไม่เป็นไปตามกฎของเมนเดล

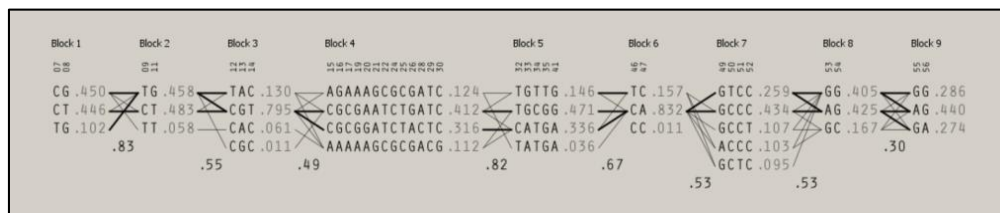
MAF	ค่า minor allele frequency ของแต่ละแอลลีล
Allele	แสดง Major และ Minor ของสลิปแต่ละตำแหน่ง
Rating	ค่าที่บอกให้โปรแกรมวิเคราะห์ผล ถ้ามีเครื่องหมายถูกอยู่แสดงว่าตำแหน่งสลิปนั้นผ่านการทดสอบ แต่ถ้าไม่มีแสดงว่าตำแหน่งสลิปนั้นไม่ผ่านการทดสอบ

- การวิเคราะห์ Linkage disequilibrium (LD Plot) การศึกษาว่าตำแหน่งสลิปที่อยู่ใกล้กันมีโอกาสถ่ายทอดไปด้วยกันมากกว่าหรือน้อยกว่าค่าคาดหวัง ซึ่งพิจารณาจากค่า Lewontin's coefficient (D') ที่คำนวณจากระยะทางและความถี่แอลลีลของแต่ละคู่สลิป หากค่า D' มากกว่า 0.8 หมายความว่า มีการถ่ายทอดไปด้วยกันของคู่เครื่องหมายพันธุกรรมได้บ่อยจนถือว่ามี Linkage disequilibrium ที่จะวิเคราะห์ว่าเป็น LD Blok เดียวกัน



ภาพที่ 2.5 แสดงผลการวิเคราะห์ Linkage disequilibrium (LD Plot)

- การวิเคราะห์ (Blocks) และแฮพลไทป์ (Haplotype) โปรแกรมสามารถสร้าง Haplotype block ได้หลายแบบ การศึกษานี้เลือกรูปแบบ Confidence intervals เป็นรูปแบบเริ่มต้น โปรแกรมจะตัดตำแหน่งสลิปที่ศึกษาที่มีค่า $MAF < 0.05$ ออก ส่วนการแสดงผลแฮพลไทป์ในแต่ละบล็อกจะประกอบด้วยความถี่ของแฮพลไทป์ในแต่ละแฮพลไทป์และเส้นเชื่อมโยงจากบล็อกหนึ่งไปอีกบล็อกหนึ่งที่อยู่ติดกันหากมีหลายบล็อกของสลิปหลายตำแหน่ง โปรแกรมจะแสดงเฉพาะแฮพลไทป์หรือแอลลีลที่มีการแสดงออกมาที่หน้าจอเท่านั้น



ภาพที่ 2.6 แสดงผลการวิเคราะห์บล็อก (Blocks) และแฮพลোটป์ (Haplotype)

2.4.3 Haploview - Command line options

Haploview สามารถรันจาก command line โดยมีหรือไม่มีหน้าต่างแสดงผลได้ทั้งคู่ บางครั้ง ด้วยข้อจำกัดทางหน่วยความจำหรือความคุ้นชินของผู้ใช้งาน อินเทอร์เน็ตแบบปกติมักจะช้า และมีความยุ่งยากมากกว่า command line เพื่อที่จะทำการประมวลผลชุดข้อมูลหลายชุดหรือ ต้องการการคำนวณอย่างรวดเร็วบนชุดข้อมูลที่ใหญ่มาก ๆ จึงขอแนะนำให้สร้างความคุ้นชินกับ command line ที่ Haploview รองรับ

Haploview สามารถเริ่มได้จาก terminal โดยใช้

```
java -jar Haploview.jar
```

หากต้องการที่จะรัน Haploview โดยตรงจาก terminal จะเป็นต้องปิดการใช้งาน GUI หรือ graphical user interface ใช้คำสั่งดังต่อไปนี้

```
java -jar Haploview.jar -nogui
```

หากต้องการจะแสดงคำสั่งหลัก ๆ บน command line ใช้คำสั่งต่อไปนี้

```
java -jar Haploview.jar -nogui -help
```

คำสั่งที่ต้องการข้อมูลอินพุตเพิ่มเติม จะเห็นได้ในลักษณะต่อไปนี้

```
-chromosome <chrom#>; -startpos <start# in kb>; -panel  
<PanelName>
```

โดย <> จะเป็นตัวระบุที่จำเป็นต้องใส่พารามิเตอร์เพิ่มเข้าไป หากเว้นว่างไว้จะทำให้โปรแกรมไม่ทำงานตามที่ระบุไว้

General options

คำสั่งต่อไปนี้สามารถใช้เป็นอาร์กิวเมนต์เมื่อเริ่มใช้งาน Haploview ได้

-h, -help

แสดงข้อมูล help information

-n, -nogui

Command line mode - ไม่แสดงหน้าจอ

-q, quiet

Quiet mode - ย่อขนาดของข้อมูลเอาต์พุตไปที่ command line

-log <filename>

สร้าง logfile information ไปยัง filename ที่ระบุ (ค่าเริ่มต้นคือ Haploview.log หากไม่มีการระบุ)

-out <fileroot>

ระบุตำแหน่ง fileroot ที่จะใช้สำหรับไฟล์ output ทั้งหมด

-memory <memsize>

จัดสรรหน่วยความจำ <memsize> megabytes ของหน่วยความจำไปที่ process ของ Haploview (ค่าเริ่มต้นคือ 512MB)

2.5 สเปคคอมพิวเตอร์ที่ใช้

Specifications

Operating System:

Windows 10 Education 64-bit (10.0, Build 19041)

Processor:

Name: AMD Ryzen 5 3600

Number of CPU cores: 6

Number of Threads: 12

Base Clock: 3.6GHz

Technology: 7nm

Caches:

Total L1 Data-Cache: 192KB

Total L1 Instruction-Cache: 192KB

Total L2 Cache: 3MB

Total L3 Cache: 32MB

Mainboard:

Model: AM4 GIGABYTE B450M S2H

Memory:

Type: DDR4

Channel: Dual

Size: 16GB

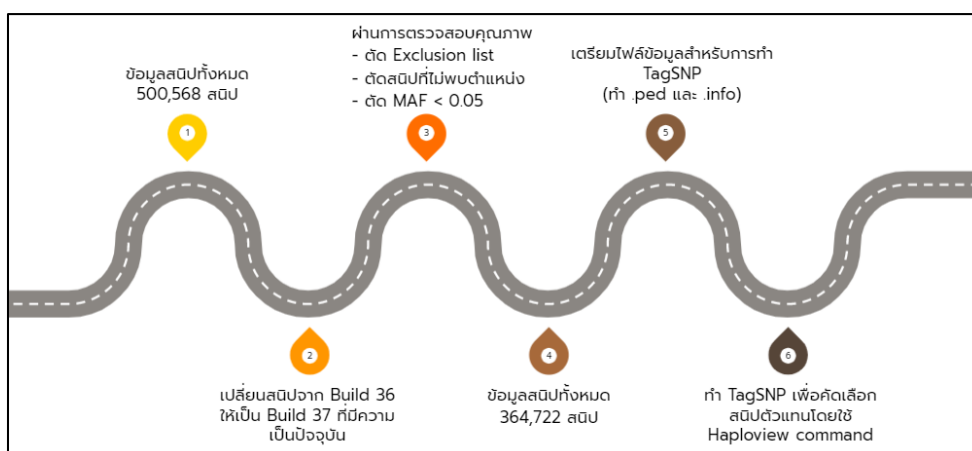
Note : ข้อมูลชุดนี้ได้มาจาก CPU-Z ซึ่งเป็น software สำหรับรวบรวมข้อมูลต่าง ๆ ของอุปกรณ์บนเครื่องคอมพิวเตอร์ (System information software) รวมกับข้อมูลบนเว็บไซต์อย่างเป็นทางการของ AMD

บทที่ 3

ขั้นตอนการดำเนินงาน

3.1 ภาพรวมการดำเนินงาน

การทำงานจะเริ่มต้นจาก การนำข้อมูลสลิปทั้งหมด 500,568 สลิปซึ่งมาจากข้อมูล WTCCC ซึ่งเป็นข้อมูลของกลุ่มตัวอย่าง นำมาเชื่อมโยงสลิปให้มีความเป็นปัจจุบันโดยเปลี่ยนสลิปจาก NCBI ที่สร้างครั้งที่ 36 ให้เป็นข้อมูลสลิปที่พร้อมใช้งานตาม NCBI ที่สร้างครั้งที่ 37 จากนั้นนำสลิปมาตรวจสอบคุณภาพที่กำหนดโดย ตัดสลิปที่ไม่พบตำแหน่ง , ตัด Exclusion list และตัดสลิปที่มีความถี่อัลลีลน้อยกว่า 0.05 จะเหลือสลิปสำหรับคัดเลือกตัวแทนสลิป 363,028 ที่ผ่านการตรวจสอบคุณภาพ



ภาพที่ 3.1 ภาพรวมของโปรเจก

3.1 ข้อมูลที่ใช้

ข้อมูล WTCCC ข้อมูลของกลุ่มตัวอย่างถูกจัดเก็บมาเป็นชุดของสลิปตั้งแต่สลิปของโครโมโซม 1 ถึง โครโมโซม X ซึ่งในแต่ละสลิปประกอบด้วย ชื่อสลิป , ข้อมูลจีโนไทป์ เป็นต้น ซึ่งไฟล์นี้มาจากบริษัท Affymetrix ซึ่งเป็นแบรนด์ของผลิตภัณฑ์ DNA microarray ที่ขายโดย Thermo Fisher Scientific ซึ่งมีต้นกำเนิดจากบริษัท วิจัยและพัฒนาเทคโนโลยีชีวภาพแห่งสหรัฐอเมริกา

Mapping250K_NSP Annotations ข้อมูลที่ใช้จะอ้างอิงจากไฟล์ NSP Annotations ไฟล์ จากปี 2532 เป็นข้อมูลจาก ศูนย์ข้อมูลเทคโนโลยีชีวภาพแห่งชาติหอสมุดแพทยศาสตร์แห่งชาติ สหรัฐอเมริกา หลังจากนั้นจะเรียกชุดข้อมูลนี้โดยย่อว่า NSP Annotations

Mapping20K_Nip.nip.22.2mont																					
File	Home	Insert	Page Layout	Formulas	Data	Review	View	Help	Tell me what you want to do												Share
Cut									General												Autosave
Paste									Font												Find & Replace
Copy									Conditional Formatting												Sort & Filter
Format									Table												Filter - Select
Alignment									Style												Editing
E41																					
	Probe Set	ID	Chromosome	Physical Position (kb)	Strand	ChIP-seq signal (CytoBand)	Flank	Allele A	Allele B	K	L	J	K	L	M	N	O	P	Q	R	
11	Probe Set 1	DBS10012	1	150661138																	
22	SNP A	17061919	17	105069113		0.p31	gatattatg	g	gatattatg	GC	ENST000002.72.00022	AT125299E	GC	GC	ATC.010204	0.2023	1	49.0	0	CG	
23	SNP A	17061818	17	105069118		0.q12	gatattatg	g	gatattatg	GC	ENST000001.08.00867	DB15134	GC	GC	ATC.03.30	0.417	1	50.0	0	CG	
24	SNP A	17061701	17	105069112		0.q12	gatattatg	g	gatattatg	GC	ENST000002.52.25224	AT10887	GC	GC	ATC.01.21	0.3318	1	50.0	0	CG	
25	SNP A	17061707	17	105069113		0.q12.3	gatattatg	g	gatattatg	GC	ENST000001.78.30360	AT10276	GC	GC	ATC.03.27	0.2796	1	50.0	0	CG	
26	SNP A	17061616	17	105069116		0.q22	gatattatg	A	gatattatg	GC	ENST000002.1002.0047	AT125145	GC	GC	ATC.03.88686	0.242875	46.0	0	CG		
27	SNP A	17061613	17	105069115		0.p12.1	gatattatg	g	gatattatg	GC	ENST000001.47.21277	AT10121	GC	GC	ATC.02.2408	1	50.0	0	CG		
28	SNP A	17061419	17	105069100		0.p11.1	gatattatg	A	gatattatg	GC	ENST000001.18.40234	AT00318	GC	GC	ATC.03.30632	0.408126	49.0	0	CG		
29	SNP A	17061412	17	105069103		0.p11.2	gatattatg	A	gatattatg	GC	ENST000001.190.0007	AT125242	GC	GC	ATC.04.14	0.317	1	50.0	0	CG	
30	SNP A	17061418	17	105069109		0.p33.3	gatattatg	A	gatattatg	GC	ENST000001.16.12131	AT0730	GC	GC	ATC.11.0	0.370	1	50.0	0	CG	
31	SNP A	17061611	17	105069121		0.q13.3	gatattatg	A	gatattatg	GC	ENST000001.82.9843	AT04264	GC	GC	ATC.01.0	0.2	1	50.0	0	CG	
32	SNP A	17061610	17	105069120		0.q13.3	gatattatg	A	gatattatg	GC	ENST000001.198.0001	AT04264	GC	GC	ATC.01.0	0.2	1	50.0	0	CG	
33	SNP A	17061415	17	105069117		0.p33.3	gatattatg	A	gatattatg	GC	ENST000001.162.0006	AT02582	GC	GC	ATC.01.0	0.3432	1	50.0	0	CG	
34	SNP A	17061576	17	105069131		0.p21.3	gatattatg	A	gatattatg	GC	ENST000001.119.00281	AT02170	GC	GC	ATC.01.0	0.05768	1	50.0	0	CG	
35	SNP A	17061413	17	105069142		0.p14.2	gatattatg	A	gatattatg	GC	ENST000001.43.2804	AT10116	GC	GC	ATC.03.8757	0.273219	49.0	0	CG		
36	SNP A	17061412	17	105069145		0.q23.2	gatattatg	A	gatattatg	GC	ENST000001.127.2890	AT10200	GC	GC	ATC.01.16325	0.273219	49.0	0	CG		
37	SNP A	17061704	17	105069124		0.q24.1	gatattatg	A	gatattatg	GC	ENST000001.150.6662	AT02378	GC	GC	ATC.01.0	0.2112	1	50.0	0	CG	
38	SNP A	17061739	17	105069187		0.q22.3	gatattatg	G	gatattatg	GC	ENST000001.72.40571	AT13584	GC	GC	ATC.02.3	0.375	1	4			
39	SNP A	17061702	17	105091169		0.q22	gatattatg	G	gatattatg	GC	ENST000001.23.1919	AT081309	GC	GC	ATC.03.1972	0.31291	48.0	0	CG		
40	SNP A	17061703	17	105091164		0.q24.12	gatattatg	A	gatattatg	GC	ENST000001.119.7753	AT081309	GC	GC	ATC.01.1	0.301726	48.0	0	CG		
41	SNP A	17061737	17	105100883		0.p15.4	gatattatg	A	gatattatg	GC	ENST000001.10.0690	AT1145426	GC	GC	ATC.01.0	0.2282	1	50.0	0	CG	
42	SNP A	17061738	17	105100882		0.p13.3	gatattatg	A	gatattatg	GC	ENST000001.58.00006	AT13585	GC	GC	ATC.02.16325	0.284285	47.0	0	CG		

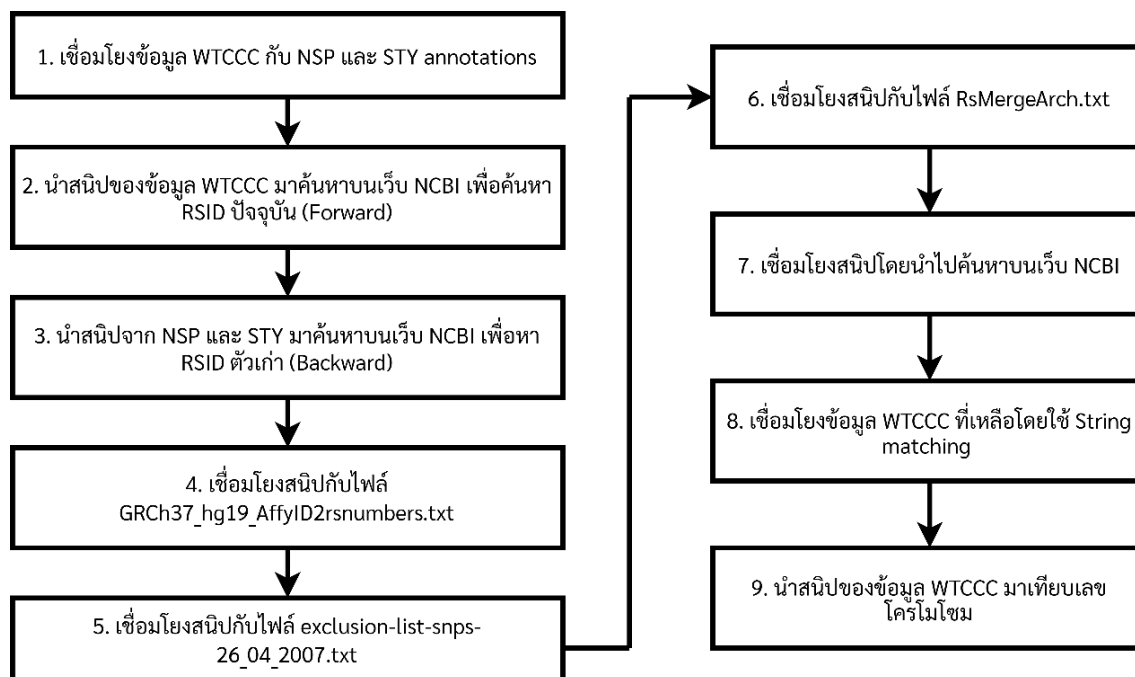
ภาพที่ 3.3 ข้อมูลที่ใช้ในการอ้างอิงของไฟล์ NSP Annotations

Mapping250K_STY Annotations ข้อมูลที่ใช้จะอ้างอิงจากไฟล์ STY Annotations ไฟล์ จากปี 2532 เป็นข้อมูลจาก ศูนย์ข้อมูลเทคโนโลยีชีวภาพแห่งชาติหอสมุดแพทยศาสตร์แห่งชาติ สหรัฐอเมริกา หลังจากนั้นจะเรียกชุดข้อมูลนี้โดยย่อว่า STY Annotations

ภาพที่ 3.4 ข้อมูลที่ใช้ในการอ้างอิงของไฟล์ STY Annotations

3.2 ขั้นตอนการเตรียมข้อมูล

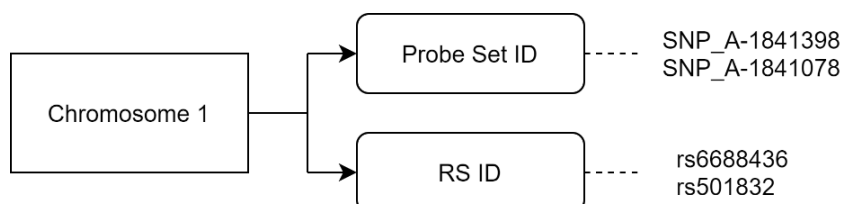
จุดประสงค์ในการเตรียมข้อมูลคือ ต้องการเชื่อมโยงสลับของข้อมูล WTCCC ตั้งแต่ โครโมโซม 1 ถึง โครโมโซม X นำมาเชื่อมโยงสลับให้มีความเป็นปัจจุบันโดยเปลี่ยนสลับจาก NCBI ที่สร้างครั้งที่ 36 ให้เป็นข้อมูลสลับที่พร้อมใช้งานตาม NCBI ที่สร้างครั้งที่ 37 ซึ่งมีขั้นตอนการเตรียมข้อมูลดังนี้



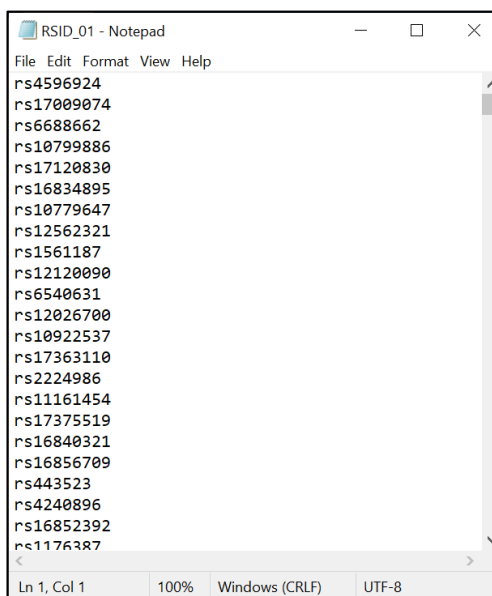
ภาพที่ 3.5 ขั้นตอนการเตรียมข้อมูล

3.2.1 เชื่อมโยงข้อมูล WTCCC กับ NSP และ STY annotations

- นำข้อมูล WTCCC ตั้งแต่ โครโมโซมที่ 1 ถึง โครโมโซม X มาทำการคัดเลือกสลับ โดยแต่ละสลับจะมีซ้ำกันอยู่ 1,504 สลับ นำมาคัดแยก Probe Set ID และ RS ID ที่ไม่ซ้ำกันออกจากกัน โดยเก็บเป็น text ไฟล์



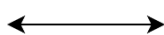
ภาพที่ 3.6 แผนภาพการคัดเลือก Prob Set ID และ RS ID



ภาพที่ 3.7 ผลลัพธ์ของการคัดเลือก Probe Set ID และ RS ID จากไฟล์ข้อมูล

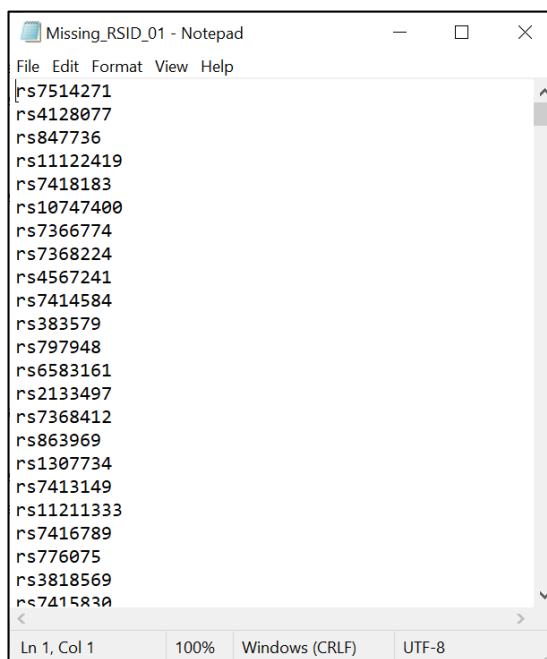
- ตรวจสอบข้อมูล WTCCC ที่ผ่านการคัดเลือก Probe Set ID และ RS ID ซึ่งได้มาเป็นไฟล์ text ดังแสดงในภาพที่ 4 มาทำการตรวจสอบกับ NSP annotations และ STY annotations ว่ามีสลับที่ตรงกันหรือไม่ ซึ่งจะใช้โค้ดในการตรวจสอบ และเก็บผลลัพธ์เป็นไฟล์ text

SNP_A-1841398
SNP_A-1841078
SNP_A-1780619
SNP_A-1780419
SNP_A-1780612
rs6688436
rs501832



Probe Set ID	RS ID	Chromosome
SNP_A-1780358	rs17325399	5
SNP-A-1780551	rs12454921	18

ภาพที่ 3.8 แสดงแผนภาพการเปรียบเทียบข้อมูล WTCC กับข้อมูล annotations



ภาพที่ 3.9 ผลลัพธ์ของ Probe Set ID และ RS ID ที่ไม่ปรากฏกับ NSP และ STY annotations

สรุป เมื่อทำการตรวจสอบ Probe Set ID และ RS ID ของข้อมูล WTCCC ตั้งแต่โครโมโซมที่ 1 ถึงโครโมโซม X จากข้อมูลทั้งหมด 500,568 ตัว ซึ่งมีข้อมูลที่ตรงกับ NSP และ STY annotations ทั้งหมด 486,176 ตัว ดังนั้นข้อมูลที่ไม่ตรงกับ NSP และ STY annotations มีทั้งหมด 6,379 ตัว

3.2.2 นำสนิปของข้อมูล WTCCC มาค้นหาบนเว็บ NCBI เพื่อค้นหา RSID ปัจจุบัน (Forward)

- เข้าที่เว็บ <https://www.ncbi.nlm.nih.gov/snp/?term=> ใช้ RS ID ในการค้นหาว่าสนิปที่ทำการค้นหานั้นเปลี่ยนแปลงไปเป็น RS ID ไດ
- ใช้การค้นหาสนิปแบบ Forward ซึ่งเป็นการนำสนิปจากข้อมูล WTCCC ไปทำการค้นหาบนเว็บ ncbi และเลือกผลลัพธ์สนิปปัจจุบันมาใช้เพื่อตรวจสอบกับ NSP และ STY annotations ซึ่งมีขั้นตอนการทำได้ดังนี้
 - นำสนิปมาค้นหาในเว็บ ncbi โดยเลือกตัวเลือกในช่องแรกเป็น SNP และใส่ RS ID ลงในช่องค้นหา



ภาพที่ 3.10 การใช้สนิปในการค้นหาสนิปปัจจุบันจากเว็บ ncbi

- ผลลัพธ์ที่เราต้องการคือ RS ID ปัจจุบันของ RS ID เดิม

Search results

Items: 3

☐ rs58282973 has merged into **rs2923297** [Homo sapiens]

1.

Variant type:	SNV
Alleles:	T>A,C,G [Show Flanks]
Chromosome:	1:105026862 (GRCh38) 1:105569484 (GRCh37)

Canonical SPDI:
NC_000001.11:105026861:T:A,NC_000001.11:105026861:T:C,NC_000001.11:105026861:T:G

Gene: LOC105378880 (Varview)

Functional Consequence: upstream_transcript_variant

Validated: by frequency,by alfa,by cluster

MAF: T=0.039491/879 (ALFA)
T=0./0 (GENOME_DK)
G=0./0 (KOREAN)

...more

HGVS: NC_000001.11:g.105026862T>A, NC_000001.11:g.105026862T>C,
NC_000001.11:g.105026862T>G, NC_000001.10:g.105569484T>A,
NC_000001.10:g.105569484T>C, NC_000001.10:g.105569484T>G

ภาพที่ 3.11 ผลลัพธ์การค้นหาลีนจากเว็บ ncbi

- ใช้แพ็คเกจ BeautifulSoup ของ python ในการดึงข้อมูลจากหน้าเว็บเพื่อค้นหาข้อมูลที่เราต้องการ

```

1 import re
2 import requests
3 from bs4 import BeautifulSoup
4
5 page = requests.get("https://www.ncbi.nlm.nih.gov/snp/?term=rs3818569")
6 soup = BeautifulSoup(page.content, 'html.parser')
7 print(soup)

```

```

<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

<html lang="en" xml:lang="en" xmlns="http://www.w3.org/1999/xhtml">
<head xmlns:xi="http://www.w3.org/2001/XInclude"><meta content="text/html; charset=utf-8" http-equiv="Content-Type"/>
<!-- meta -->
<meta content="dbSNP is a public-domain archive for human single nucleotide variations, microsatellites, and small-scale ins
ertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping in
formation for both common variations and clinical mutations." name="description"/><meta content="noindex,nofollow,noarchive"
name="robots"/>
<meta content="entrez" name="ncbi_app"/><meta content="snp" name="ncbi_db"/><meta content="rs3818569" name="ncbi_term"/><met
a content="snpsum" name="ncbi_report"/><meta content="html" name="ncbi_format"/><meta content="20" name="ncbi_pagesize"/><me
ta content="snp_id" name="ncbi_sortorder"/><meta content="1" name="ncbi_pageno"/><meta content="3" name="ncbi_resultcount"/>
<meta content="search" name="ncbi_op"/><meta content="snpsum" name="ncbi_pdid"/><meta content="CE885E7A012EDEC1_1851SID" nam
e="ncbi_sessionid"/><meta content="all" name="ncbi_filter"/><meta content="false" name="ncbi_stat"/><meta content="false" na
me="ncbi_hitstat"/>
<!-- title -->
<title>rs3818569 - SNP - NCBI</title>
<!-- Common JS and CSS -->

```

ภาพที่ 3.12 การใช้งาน BeautifulSoup package

- ใช้ Regular Expression ในการค้นหาข้อความที่มีลักษณะตรงกับที่ต้องการ

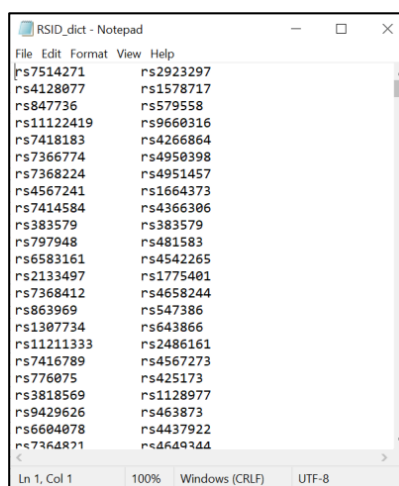
```

1 get_rs = soup.find('a', attrs={'href': re.compile("^/snp/rs")})
2 get_rs = str(get_rs)
3 print(get_rs)
4 rs = re.search("rs[0-9]+", get_rs)
5 print(rs.group())
6
<a href="/snp/rs1128977">rs1128977</a>
rs1128977

```

ภาพที่ 3.13 การใช้งาน Regular Expression

- ผลลัพธ์การค้นหาสนิปจากเว็บ ncbi



ภาพที่ 3.14 ผลลัพธ์การค้นหาสนิปจากเว็บ ncbi ในรูปแบบไฟล์ text

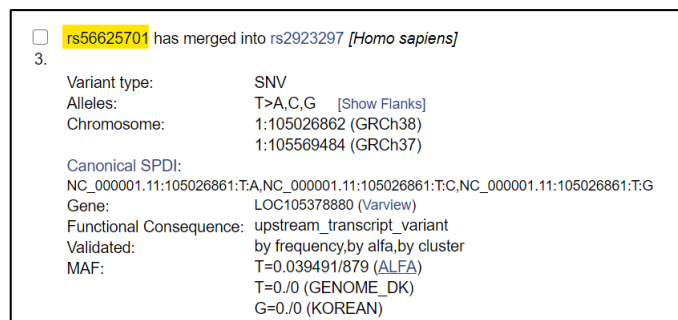
สรุป เมื่อทำการค้นหา RS ID เดิมของข้อมูล WTCC ทั้งหมด 6,379 ตัว เมื่อนำไปค้นหาที่เว็บ ncbi จะได้ผลลัพธ์เป็น RS ID ปัจจุบัน ซึ่งเมื่อทำการค้นหาจะได้ผลลัพธ์ดังนี้ พบ RS ID ปัจจุบันทั้งหมด 5,085 ตัว และไม่พบ RS ID ปัจจุบันทั้งหมด 1,294 ตัว

3.2.3 นำสนิปจาก NSP และ STY มาค้นหบนเว็บ NCBI เพื่อหา RSID ตัวเก่า (Backward)

- ใช้ <https://www.ncbi.nlm.nih.gov/snp/?term=> เพื่อการใช้ RS ID ในการค้นหาว่าสนิปที่ทำการค้นหานั้นมี RS ID ตัวก่อนหน้าเป็น RS ID ไດ
- ใช้การค้นหาสนิปแบบ Backward ซึ่งเป็นการนำสนิปจาก annotations ที่ยังไม่ปรากฏทั้งหมดไปทำการค้นหบนเว็บ ncbi และเลือกผลลัพธ์เป็นสนิปตัวก่อนหน้าทั้งหมดมาใช้เพื่อตรวจสอบกับข้อมูล WTCC ซึ่งมีขั้นตอนการทำดังนี้
 - นำสนิปของข้อมูล WTCC ที่ตรงกับ NSP annotations และ STY annotations ทั้งหมด มาตัดสนิปที่พบแล้วออกจาก NSP และ STY annotations ซึ่งเมื่อ

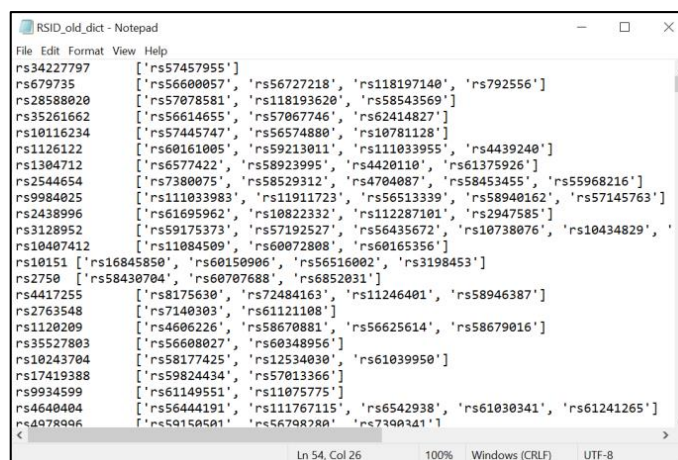
ตัดสนิปแล้วจะเหลือจำนวนสนิปใน NSP และ STY annotations ทั้งหมด 13,797 ตัว

- นำสนิปจำนวน 13,797 ตัว มาทำการค้นหาในเว็บ ncbi เพื่อทำการค้นหา RS ID ตัวเก่าของสนิปนั้น
- ค้นหา RS ID ที่ต้องการและนำผลลัพธ์คือ RS ID เก่าทั้งหมดนำมาใช้



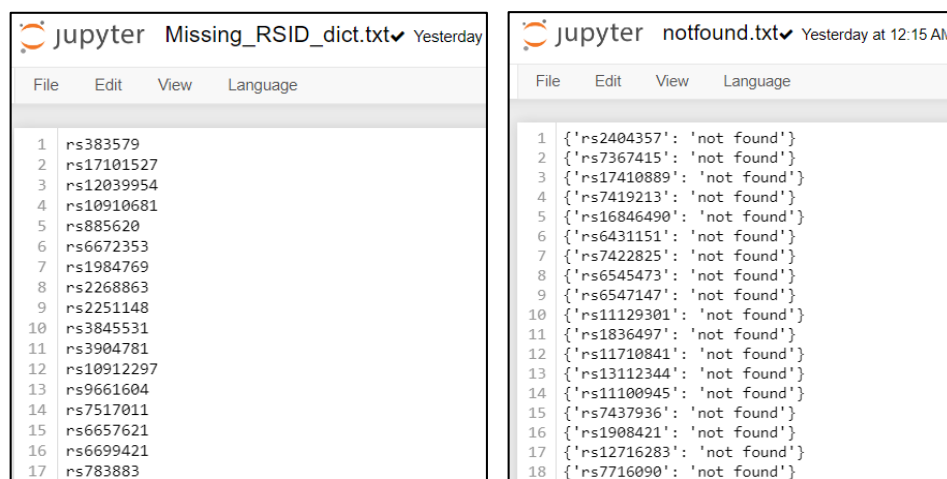
ภาพที่ 3.15 แสดงผลลัพธ์การค้นหาสนิปจากเว็บ ncbi

- ผลลัพธ์การค้นหาสนิปจากเว็บ ncbi



ภาพที่ 3.16 แสดงผลลัพธ์การค้นหาสนิปจากเว็บ ncbi ในรูปแบบไฟล์ text

- นำสนิปที่ไม่พบจากการค้นหา ncbi แบบค้นหา RS ID ปัจจุบัน ทั้งหมด 1,294 ตัว มาทำการตรวจสอบกับผลลัพธ์ของการค้นหาสนิปตัวเก่า
- ผลลัพธ์การเปรียบเทียบสนิปกับ NSP และ STY annotations

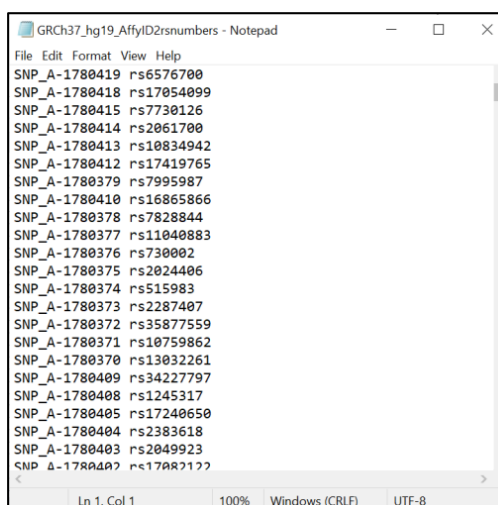


ภาพที่ 3.17 แสดงผลลัพธ์การเปรียบเทียบสปีกับ NSP และ STY annotations

สรุป นำสปีของ NSP และ STY annotations โดยตัดตัวที่ปรากฏแล้วว่าตรงกับสปีในข้อมูล WTCCC ออก จากนั้นนำมาค้นหบนเว็บ ncbi เพื่อค้นหา RS ID ตัวก่อนหน้า หรือ ตัวเก่า นำผลลัพธ์ที่ได้นำมาเช้คกับสปีของข้อมูล WTCCC ที่ยังไม่ปรากฏใน annotations โดยจะได้ว่าเมื่อเปรียบเทียบข้อมูลแล้วมีสปีที่ตรงกันทั้งหมด 1,239 ตัว ทำให้เหลือข้อมูลของโครโมโซมที่ยังไม่พบทั้งหมด 509 ตัว

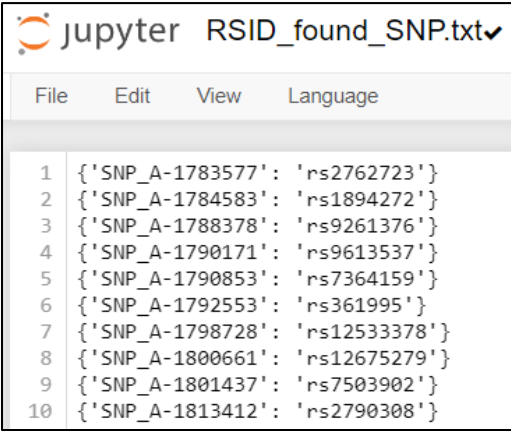
3.2.4 เชื่อมโยงสปีกับไฟล์ GRCh37_hg19_AffyID2rsnumbers.txt

นำสปีที่ยังไม่ปรากฏกับข้อมูล annotations ทั้งหมด 509 ตัว มาตรวจสอบกับไฟล์ GRCh37_hg19_AffyID2rsnumbers.txt โดยใช้ RS ID ในการตรวจสอบ ถ้า RS ID ตรงกัน เก็บ Probset ID นำมาใช้เพื่อนำไปตรวจสอบว่าปรากฏในไฟล์ annotations หรือไม่ โดยมีขั้นตอนการดำเนินงานนี้



ภาพที่ 3.18 ข้อมูลไฟล์ GRCh37_hg19_AffyID2rsnumbers.txt

- ผลลัพธ์ตรวจสอบสนิปว่าปรากฏในไฟล์ GRCh37_hg19_AffyID2rsnumbers



1	{'SNP_A-1783577': 'rs2762723'}
2	{'SNP_A-1784583': 'rs1894272'}
3	{'SNP_A-1788378': 'rs9261376'}
4	{'SNP_A-1790171': 'rs9613537'}
5	{'SNP_A-1790853': 'rs7364159'}
6	{'SNP_A-1792553': 'rs361995'}
7	{'SNP_A-1798728': 'rs12533378'}
8	{'SNP_A-1800661': 'rs12675279'}
9	{'SNP_A-1801437': 'rs7503902'}
10	{'SNP_A-1813412': 'rs2790308'}

ภาพที่ 3.19 แสดงผลลัพธ์การเชื่อมโยงสนิปว่าปรากฏในไฟล์

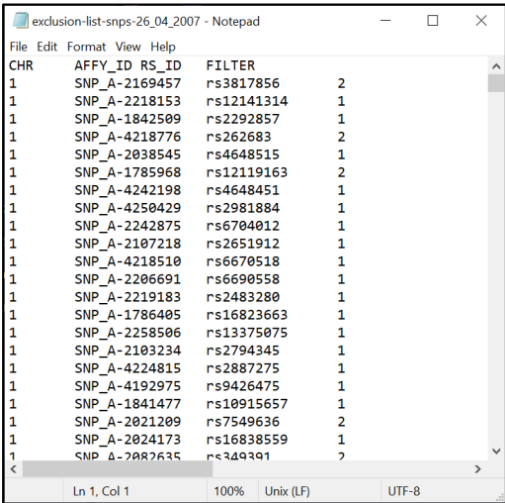
GRCh37_hg19_AffyID2rsnumbers

- นำผลลัพธ์มาตรวจสอบในไฟล์ NSP และ STY annotations โดยนำ Prob set ID ไปตรวจสอบว่าปรากฏอยู่ใน annotations หรือไม่

สรุป จากสนิปทั้งหมดที่ยังไม่ปรากฏใน annotations ทั้งหมด 509 ตัว พบ RS ID ที่ปรากฏในไฟล์ GRCh37_hg19_AffyID2rsnumbers.txt ทั้งหมด 371 ตัว และยังไม่พบทั้งหมด 138 ตัว

3.2.5 เชื่อมโยงสนิปกับไฟล์ exclusion-list-snps-26_04_2007.txt

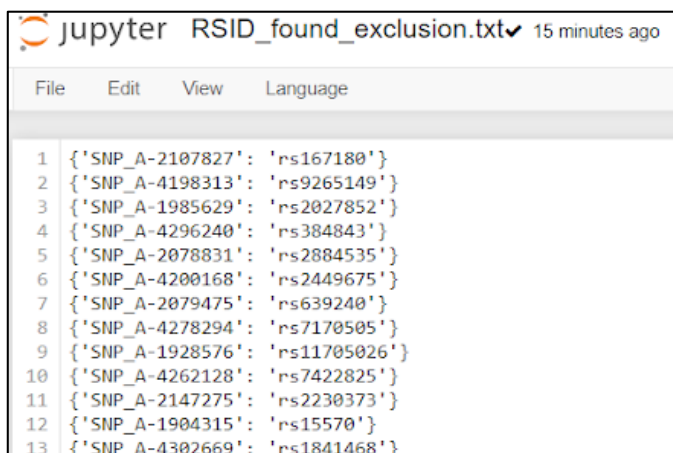
นำสนิปที่ยังไม่ปรากฏในไฟล์ annotation ทั้งหมด 138 ตัว มาตรวจสอบกับไฟล์ exclusion-list-snps-26_04_2007.txt โดยใช้ RS ID ในการตรวจสอบ ถ้า RS ID ตรงกัน เก็บ Probset ID นำมาใช้เพื่อนำไปตรวจสอบว่าปรากฏในไฟล์ annotations หรือไม่ โดยมีขั้นตอนการดำเนินงานดังนี้



CHR	AFFY_ID	RS_ID	FILTER
1	SNP_A-2169457	rs3817856	2
1	SNP_A-2218153	rs12141314	1
1	SNP_A-1842509	rs2292857	1
1	SNP_A-4218776	rs262683	2
1	SNP_A-2038545	rs4648515	1
1	SNP_A-1785968	rs12119163	2
1	SNP_A-4242198	rs4648451	1
1	SNP_A-4250429	rs2981884	1
1	SNP_A-2242875	rs6704012	1
1	SNP_A-2107218	rs2651912	1
1	SNP_A-4218510	rs6670518	1
1	SNP_A-2206691	rs6690558	1
1	SNP_A-2219183	rs2483280	1
1	SNP_A-1786405	rs16823663	1
1	SNP_A-2258506	rs13375075	1
1	SNP_A-2103234	rs2794345	1
1	SNP_A-4224815	rs2887275	1
1	SNP_A-4192975	rs9426475	1
1	SNP_A-1841477	rs10915657	1
1	SNP_A-2021209	rs7549636	2
1	SNP_A-2024173	rs16838559	1
1	SNP_A-7087635	rs349391	?

ภาพที่ 3.20 ข้อมูลไฟล์ exclusion-list-snps-26_04_2007.txt

- ผลลัพธ์ตรวจสอบสนิปว่าปรากฏในไฟล์ exclusion-list-snps-26_04_2007.txt



```

jupyter RSID_found_exclusion.txt 15 minutes ago
File Edit View Language

1 {'SNP_A-2107827': 'rs167180'}
2 {'SNP_A-4198313': 'rs9265149'}
3 {'SNP_A-1985629': 'rs2027852'}
4 {'SNP_A-4296240': 'rs384843'}
5 {'SNP_A-2078831': 'rs2884535'}
6 {'SNP_A-4200168': 'rs2449675'}
7 {'SNP_A-2079475': 'rs639240'}
8 {'SNP_A-4278294': 'rs7170505'}
9 {'SNP_A-1928576': 'rs11705026'}
10 {'SNP_A-4262128': 'rs7422825'}
11 {'SNP_A-2147275': 'rs2230373'}
12 {'SNP_A-1904315': 'rs15570'}
13 {'SNP_A-4302669': 'rs1841468'}

```

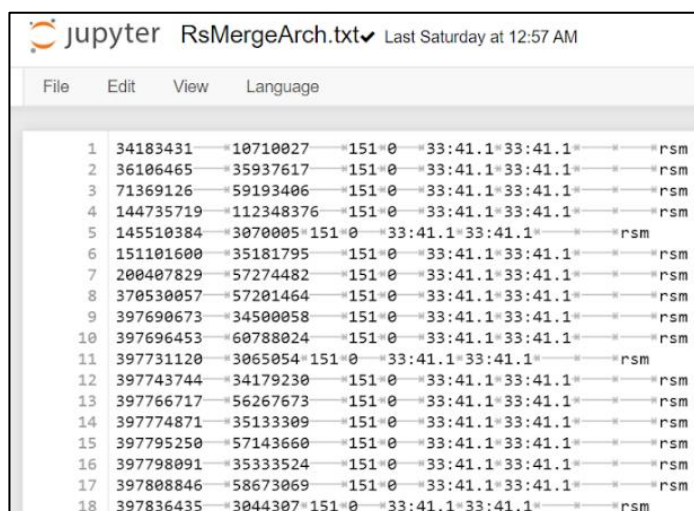
ภาพที่ 3.21 แสดงผลลัพธ์การตรวจสอบสนิปว่าปรากฏในไฟล์ exclusion-list-snps-26_04_2007.txt

- นำผลลัพธ์มาตรวจสอบในไฟล์ NSP annotations และ STY annotations โดยนำ Prob set ID ไปตรวจสอบว่าปรากฏอยู่ใน annotations หรือไม่

สรุป จากสนิปทั้งหมดที่ยังไม่ปรากฏใน annotations ทั้งหมด 138 ตัว พบ RS ID ที่ปรากฏในไฟล์ exclusion-list-snps-26_04_2007.txt ทั้งหมด 38 ตัว และยังไม่พบทั้งหมด 100 ตัว

3.2.6 เชื่อมโยงสนิปกับไฟล์ RsMergeArch.txt

นำสนิปที่ยังไม่ปรากฏกับข้อมูล annotations ทั้งหมด 509 ตัว มาตรวจสอบกับไฟล์ RsMergeArch.txt โดยใช้ RS ID ในการตรวจสอบ ถ้า RS ID ตรงกัน เก็บ RS ID ปัจจุบันนำมาใช้เพื่อนำไปตรวจสอบว่าปรากฏในไฟล์ annotations หรือไม่ โดยมีขั้นตอนการดำเนินงาน



```

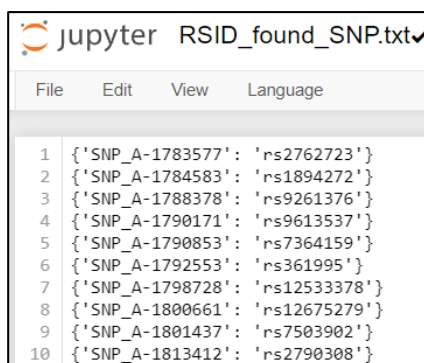
jupyter RsMergeArch.txt Last Saturday at 12:57 AM
File Edit View Language

1 34183431 → 10710027 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
2 36106465 → 35937617 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
3 71369126 → 59193406 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
4 144735719 → 112348376 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
5 145510384 → 3070005 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
6 151101600 → 35181795 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
7 200407829 → 57274482 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
8 370530057 → 57201464 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
9 397690673 → 34500058 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
10 397696453 → 60788024 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
11 397731120 → 3065054 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
12 397743744 → 34179230 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
13 397766717 → 56267673 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
14 397774871 → 35133309 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
15 397795250 → 57143660 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
16 397798091 → 35333524 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
17 397808846 → 58673069 → 151 → 0 → 33:41.1 → 33:41.1 → rsm
18 397836435 → 3044307 → 151 → 0 → 33:41.1 → 33:41.1 → rsm

```

ภาพที่ 3.22 ข้อมูลไฟล์ RsMergeArch.txt

- ผลลัพธ์การตรวจสอบการปรากฏของสลิปในไฟล์ RsMergeArch.txt



```

1 {'SNP_A-1783577': 'rs2762723'}
2 {'SNP_A-1784583': 'rs1894272'}
3 {'SNP_A-1788378': 'rs9261376'}
4 {'SNP_A-1790171': 'rs9613537'}
5 {'SNP_A-1790853': 'rs7364159'}
6 {'SNP_A-1792553': 'rs361995'}
7 {'SNP_A-1798728': 'rs12533378'}
8 {'SNP_A-1800661': 'rs12675279'}
9 {'SNP_A-1801437': 'rs7503902'}
10 {'SNP_A-1813412': 'rs2790308'}

```

ภาพที่ 3.23 แสดงผลลัพธ์การตรวจสอบการปรากฏของสลิปในไฟล์ RsMergeArch.txt

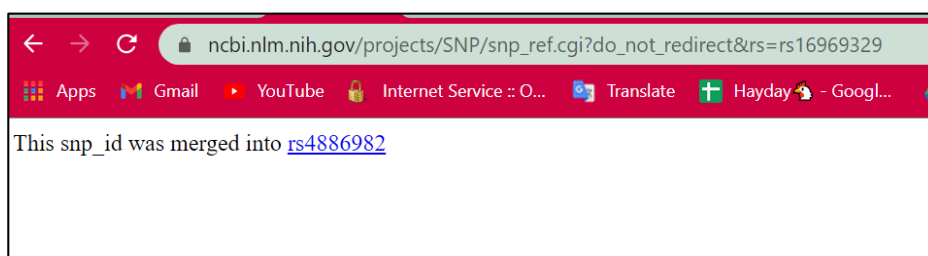
- นำผลลัพธ์มาตรวจสอบในไฟล์ NSP annotations และ STY annotations โดยนำ RS ID ไปตรวจสอบว่าปรากฏอยู่ใน annotations หรือไม่

สรุป จากสลิปทั้งหมดที่ยังไม่ปรากฏใน annotations ทั้งหมด 509 ตัว พบ RS ID ที่ปรากฏในไฟล์ RsMergeArch.txt ทั้งหมด 10 ตัว และยังไม่พบทั้งหมด 90 ตัว

3.2.7 ตรวจสอบสลิปโดยนำไปค้นหบนเว็บ ncbi

นำสลิปที่ยังไม่ปรากฏข้อมูลในไฟล์ annotations มาทั้งหมด 90 ตัว มาทำการค้นหบนเว็บ https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?do_not_redirect&rs= โดยสามารถทำได้ดังนี้

- ตัวอย่างทำการค้นหาสลิป rs16969329 ในเว็บ ncbi จะได้



ภาพที่ 3.24 แสดงผลลัพธ์จากการค้นหาสลิปบนเว็บ ncbi

rs2099870	SNP_A-4193165		
rs2611045			
rs7987360	SNP_A-2030054		
rs6974319			
rs6008833	SNP_A-1863986		
rs13126656	SNP_A-1931068		
rs1405371	SNP_A-1963269		
rs4887245	SNP_A-1870909		
rs17109001	SNP_A-2217864		
rs2645549			
rs237407	SNP_A-4276865		
rs12716283		rs6869312	was withdrawn on June 16, 2015.
rs16917285	SNP_A-1825047	rs3178681	rs1134429
rs10805714		rs9686984	rs4975915

ภาพที่ 3.25 แสดงผลลัพธ์จากการค้นหาสลิปทั้งหมดบนเว็บ ncbi

- นำผลลัพธ์มาตรวจสอบในไฟล์ NSP annotations และ STY annotations โดยนำ Probe set ID หรือ RS ID ไปตรวจสอบว่าปรากฏอยู่ใน annotations หรือไม่
- สรุป จากสลับทั้งหมดที่ยังไม่ปรากฏใน annotations ทั้งหมด 90 ตัว พบสลับที่ปรากฏในไฟล์ annotations ทั้งหมด 70 ตัว และยังไม่พบทั้งหมด 20 ตัว

3.2.8 ตรวจสอบข้อมูล WTCC ที่เหลือโดยใช้ String matching

นำสลับจากข้อมูล WTCC ที่ยังไม่ปรากฏในข้อมูล annotations ไปค้นหาลำดับใน https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?do_not_redirect&rs= จากนั้นเลื่อนลงไป Fasta sequence เพื่อหา flank ของสลับ โดยต้องหาทั้ง 5' Near Seq 30 bp และ 3' Near Seq 30 bp ดังภาพที่ 3-32 จากนั้นนำไปตรวจสอบกับ flank ของข้อมูลอ้างอิงที่ยังไม่ตรงกับข้อมูลโครโมโซมโดยต้องเช็คทั้ง forward และ reverse ถ้าตรงกัน แสดงว่าสลับนั้นคือสลับของข้อมูลอ้างอิง

```

Fasta sequence (Legend)
>gnl|dbSNP|rs2611045|allelePos=501|totalLen=1001|taxid=9606|snpc1
TACTCTTTT TTCTCTAAAC TTCTCTCTT GTTCATTTT ATTCAATTGA TCTCAATCA
CTGATACCTT TTCTCCACT TGATCAATTA GCTAGTGAA GCTTGTGAT GCATCAGTA
GTTCCTATGC TACGGTTTT AGCTCCATCA GTCATTAA GCTCTCTCT ACATGTTTA
TTCTAGTTAG CCATTAATCT AATCTTTTT CAAGGTTTT AGCTCTCTT CAATGGTTA
GAACATGCTC CTTAGCTCA GAGAAGTGT TATTCTGAC CTCTGAAGC CTAATCTGT
CAACTCTCA AAGTATTCT CTGTCCAGCT TTGTCTATT GCTGCAAGG AGCTGTGATC
CTTTGGAGGA GAAAAGGCAT TCTGGTTTT AGAATTTAA GTTTTTCTG TCTGTTTCT
CCCCATCTT GTGGTTTTAT CTACCTTGG TCTTGATGA TGGTGACCTA CAGATGAGGT
TTTGGTGTG ATGTGCTTTT
y
GTTGATGTTG ATGCTATTCC TTCTGTTTG TTAGTTTTCT TTCTAACAGC CAGGTCCTC
AGCTGCAGGT CTGTTGGAGT TTGCTGGAGG TCTACTCCAG ACCTGATTG CCGTGTGTC
ACCATGGGAG GCTGCGGAC AGCAAAATAT GCAGAACAGC AAATATTGCT GCCTGATCCT
TCCTCTGGA GCTTTGTCTC AGATGAGCAC CCAGCTGTAT GAGGTGTCAG TCAATCCCTA
CTGCGAGATG TCTCCAGTT AGCTACAGC GGGGTCAAGG ACCCACTTGA GAGGCAAGT
TGCTCATTCT CAGAGCTCAA ACACATGCT GGGAGAACCA CTGCTCTCT CAGAGCTGTC
AGACAGGAC GTTAAATCT GCACAAGTT CTGCTGCTT TTGTTAGCT ATGTCTGCC
CCGAGAGGT GAGTCTACAG AGGACAGGAG GCTTCTTAA GCTGCGGTG GCTCCATCA
GTTTGAAGCT TCTGGTGGCT

```

ภาพที่ 3.26 แสดงการหา flank

- แสดงการแปลง flank โดยใช้ string matching

	rsid	flank	flankinverse
0	rs2611045	gtgtgatgtgtctttgtgtgatgtgtgcta	tagcatcaacatcaacaaagcacatcaaac
1	rs6974319	aagaggatcaaatatttaactttatgcatc	gatggcataaagttaaatattgtatctctt
2	rs2645549	ctgtgatcaattttaaccataaagcaaat	attttgctttatgttaaaatgatcacaaag
3	rs6869312	tggtgtggttggaacaacattttacactgt	acagtgtaaaagtgtgttcaaccaccaaca
4	rs4975915	caaaccagaagagagtggttcaatttaac	gttgaatattgaaccactctctctgtgttg
5	rs16992891	tacagctacatacacctcacacacacac	gtgtaggtgtgtgtggtgtgtgtgtgtgt
6	rs12677255	agcaaaccttttcttgcatccagtgtttcca	tggaaacacttggtgaggaagaaagtgtgt
7	rs10949768	gttcagtcagttgggacttaggtttatttt	aaatgaaatcctaagtcctcaactgaagac
8	rs6608401	gctctctatctgtcctaatgagaccagggccc	gggctgtgtgtctcattaggacagatagagac
9	rs10996500	ctggagaggataggcaaataggaaacatttt	aaagtgttctatttgccatctctctcag
10	rs4300971	atataggccaatggaaagacggatgctcag	ctgaggcatccgtcttctcattggcctatat
11	rs10141046	atcaaaaccacagtgtcatgcatcaccacca	tgggtgaaatggcatgcatgtgtgtgtgt
12	rs4859065	tggaaacaaaatgagccacattgtcaagtca	tgacttgacaatgtgtgtcattttgtgtcca
13	rs4917351	gtgtgtatgagctgtgtgaagcctctgtgtg	cacacagaggcttacaacagactcatacac
14	rs10017381	caaccattgtagaacaagtgtgtgtgtgtgt	caggaaatcacacactgtgtctacatgtgtg
15	rs9498958	aaaagagcttccacgttttacaagttaaga	tcttaactgtagaacgtggaagcctttt
16	rs4362999	cctgaggaatgctactgtctgcacaaatgg	ccattgtggcagacaaagtgcattctcagg
17	rs2364212	atagtaccagtagacattttcagctctctct	agagaagactgaaaaatgtctactgtgtact

ภาพที่ 3.27 ผลลัพธ์การทำ string matching

- นำ flank ของสนิป มาตรวจสอบกับข้อมูล flank ของไฟล์ annotation

```
rs2611045 tagcatcaacatcaacaaaagcacatcaacac fw SNP_A-1980085
rs6974319 gatggcataaagttaaataatttgatcctctt fw SNP_A-2286738
rs2645549 attttgctttatggttaaaaattgatcacaag fw SNP_A-2162858
rs6869312 tgttggtgggtgaacaacacttttacactgt rev SNP_A-1916230
rs4975915 gttgaatattgaaccactctcttctggttg fw SNP_A-1818663
rs12677255 tggaaacactggatgccaggaaaaagtgtgct fw SNP_A-4202345
rs10949768 gttcagtcagttgggacttaggatttcatttt rev SNP_A-4245737
rs6608401 gctctctatctgtcctaattgagaccaggccc rev SNP_A-4261601
rs10996500 ctggagaggatatggcaaataggaacactttt rev SNP_A-4296772
rs4300971 atataggccaatggaaagaacggatgcctcag rev SNP_A-2133172
rs10141046 atcaaaaccacagtgcattgccatttcacacca rev SNP_A-2135664
rs4859065 tggaaacaaaaatgagccacattgtcaagtca rev SNP_A-1899331
rs4917351 cacacagaggcttacaacagactcatcacac fw SNP_A-2173868
rs10017381 caaccattgtagaacaagtgtggtgattcctg rev SNP_A-4287601
rs4362999 ccattgtggcagacaagtagcgattcctcagg fw SNP_A-2207466
rs2364212 agagaagactgaaaaatgtctactggtactat fw SNP_A-2097515
```

ภาพที่ 3.28 แสดงการเปรียบเทียบสนิปกับไฟล์ annotation

สรุป จากสนิปทั้งหมดที่ยังไม่ปรากฏใน annotations ทั้งหมด 18 ตัว พบสนิปที่ปรากฏในไฟล์ annotations ทั้งหมด 70 ตัว และยังไม่พบทั้งหมด 2 ตัว

3.2.9 นำสนิปของข้อมูล WTCCC มาเทียบเลขโครโมโซม

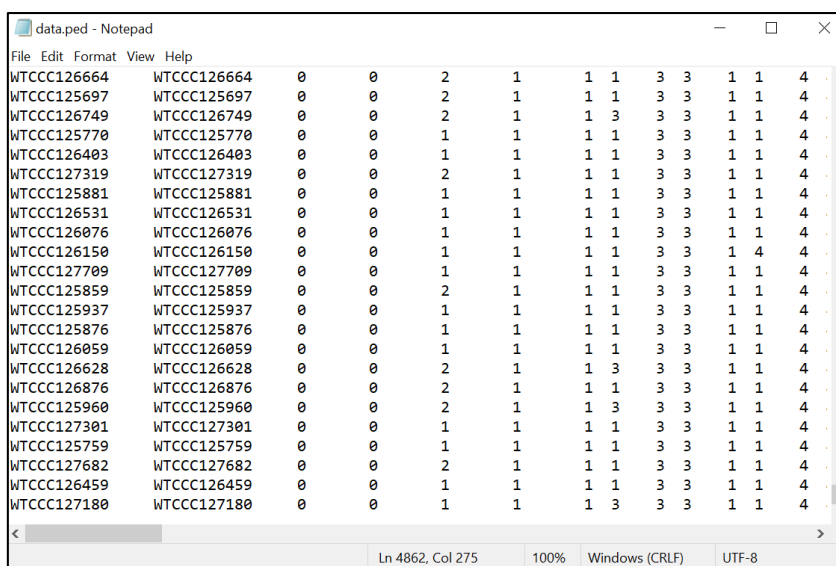
นำสนิปจากไฟล์ annotations มาเทียบกับสนิปที่ยังไม่ปรากฏจะใช้เลขโครโมโซมในการเปรียบเทียบ เนื่องจากข้อมูลเลขโครโมโซมไม่เหมือนกันสามารถทำให้เทียบได้ว่าสนิปนี้ตรงกับสนิปจากข้อมูล annotations ตัวใด

สรุป จากจำนวนสนิปทั้งหมด 500,568 ตัว สามารถเทียบกับสนิปกับไฟล์ annotations ได้ทั้งหมดทุกตัว

เนื่องจากข้อมูล WTCC ที่เป็นกลุ่มตัวอย่างมีการเปลี่ยนแปลงชื่อ เราจึงต้องทำการเทียบข้อมูล WTCC กับไฟล์ annotations ให้เป็นสนิปปัจจุบัน เพื่อเตรียมข้อมูลสำหรับการคัดเลือกสนิปตัวแทนต่อไป

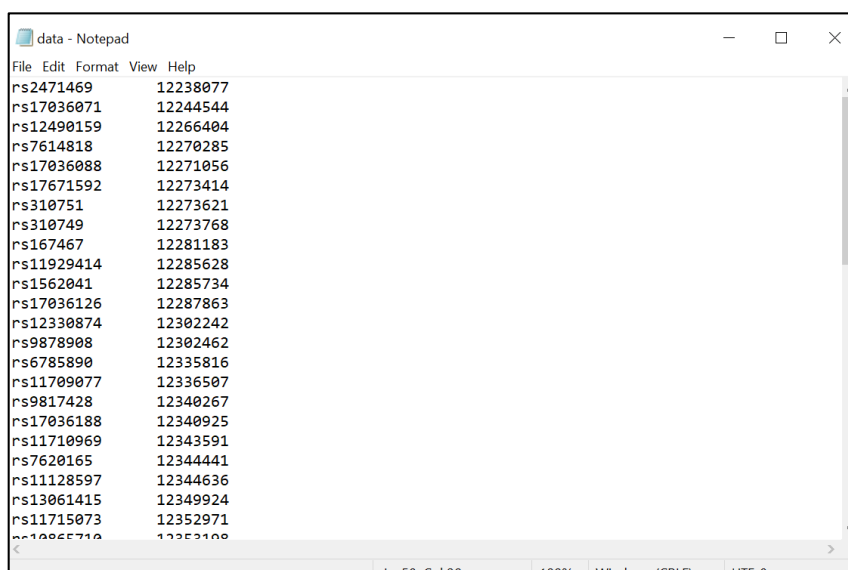
3.3 จัดการข้อมูลให้อยู่ในรูปแบบ pedigree

นำข้อมูลแต่ละโครโมโซมที่ผ่านกระบวนการเตรียมข้อมูลมาทำการตัดสลิปที่ไม่ผ่านการคัดกรองโดยใช้เปรียบเทียบกับไฟล์ exclusion list โดยจำนวนสลิปทั้งหมดของแต่ละโครโมโซมที่จะใช้ในคัดเลือกสลิปตัวแทนจะต้องมีความถี่ของอัลลีลกลุ่มน้อยไม่ต่ำกว่า 0.05 ซึ่งต้องผ่านกระบวนการคัดเลือก Minor allele frequency จากนั้นจัดการข้อมูลให้อยู่ในรูปแบบ pedigree เพื่อนำไปคัดเลือกสลิปตัวแทนผ่านการใช้โปรแกรม Haploview



ID	ID	0	0	2	1	1	1	3	3	1	1	4
WTCCC126664	WTCCC126664	0	0	2	1	1	1	3	3	1	1	4
WTCCC125697	WTCCC125697	0	0	2	1	1	1	3	3	1	1	4
WTCCC126749	WTCCC126749	0	0	2	1	1	3	3	3	1	1	4
WTCCC125770	WTCCC125770	0	0	1	1	1	1	3	3	1	1	4
WTCCC126403	WTCCC126403	0	0	1	1	1	1	3	3	1	1	4
WTCCC127319	WTCCC127319	0	0	2	1	1	1	3	3	1	1	4
WTCCC125881	WTCCC125881	0	0	1	1	1	1	3	3	1	1	4
WTCCC126531	WTCCC126531	0	0	1	1	1	1	3	3	1	1	4
WTCCC126076	WTCCC126076	0	0	1	1	1	1	3	3	1	1	4
WTCCC126150	WTCCC126150	0	0	1	1	1	1	3	3	1	4	4
WTCCC127709	WTCCC127709	0	0	1	1	1	1	3	3	1	1	4
WTCCC125859	WTCCC125859	0	0	2	1	1	1	3	3	1	1	4
WTCCC125937	WTCCC125937	0	0	1	1	1	1	3	3	1	1	4
WTCCC125876	WTCCC125876	0	0	1	1	1	1	3	3	1	1	4
WTCCC126059	WTCCC126059	0	0	1	1	1	1	3	3	1	1	4
WTCCC126628	WTCCC126628	0	0	2	1	1	3	3	3	1	1	4
WTCCC126876	WTCCC126876	0	0	2	1	1	1	3	3	1	1	4
WTCCC125960	WTCCC125960	0	0	2	1	1	3	3	3	1	1	4
WTCCC127301	WTCCC127301	0	0	1	1	1	1	3	3	1	1	4
WTCCC125759	WTCCC125759	0	0	1	1	1	1	3	3	1	1	4
WTCCC127682	WTCCC127682	0	0	2	1	1	1	3	3	1	1	4
WTCCC126459	WTCCC126459	0	0	1	1	1	1	3	3	1	1	4
WTCCC127180	WTCCC127180	0	0	1	1	1	3	3	3	1	1	4

ภาพที่ 3.28 แสดงข้อมูลนามสกุล .ped ที่ใช้ในการศึกษาด้วยโปรแกรม Haploview



ID	ID
rs2471469	12238077
rs17036071	12244544
rs12490159	12266404
rs7614818	12270285
rs17036088	12271056
rs17671592	12273414
rs310751	12273621
rs310749	12273768
rs167467	12281183
rs11929414	12285628
rs1562041	12285734
rs17036126	12287863
rs12330874	12302242
rs9878908	12302462
rs6785890	12335816
rs11709077	12336507
rs9817428	12340267
rs17036188	12340925
rs11710969	12343591
rs7620165	12344441
rs11128597	12344636
rs13061415	12349924
rs11715073	12352971
rs10065710	12353100

ภาพที่ 3.29 แสดงข้อมูลนามสกุล .info ที่ใช้ในการศึกษาด้วยโปรแกรม Haploview

จากการเปรียบเทียบข้อมูล WTCC กับข้อมูล NSP และ STY Annotation โดยสังเกตเห็นข้อมูลที่ผิดพลาดจากเลขของโครโมโซมที่ไม่ตรงกัน ซึ่งจะแสดงให้เห็นดังตารางที่ โดยจากการตรวจสอบสืบจากเว็บ NCBI สังเกตได้ว่าข้อมูลตัวเลขโครโมโซมของ WTCC นั้นถูกต้อง

ตารางที่ 3.1 แสดงให้เห็นการผิดพลาดของเลขโครโมโซม โดยเปรียบเทียบจากข้อมูล WTCC กับข้อมูล NSP และ STY Annotation

SNP ID from NSP-STY Annotation	SNP ID from WTCC data	Chromosome from NSP-STY Annotation	Chromosome from WTCC data
rs10029864	rs12511652	X	04
rs10059910	rs12697429	04	05
rs10176963	rs12467570	10	02
rs4640846	rs9327509	12	05
rs9686472	rs12653450	X	05
rs9798668	rs13373187	20	21
SNP_A-1831731	rs7774545	01	06
SNP_A-1899331	rs11710841	01	03
SNP_A-1916230	rs12716283	X	05
SNP_A-2078697	rs11975333	04	07
SNP_A-2135664	rs11159743	08	14
SNP_A-2207466	rs12213858	14	06
SNP_A-4287601	rs11100945	12	04
SNP_A-4296772	rs10996500	06	10

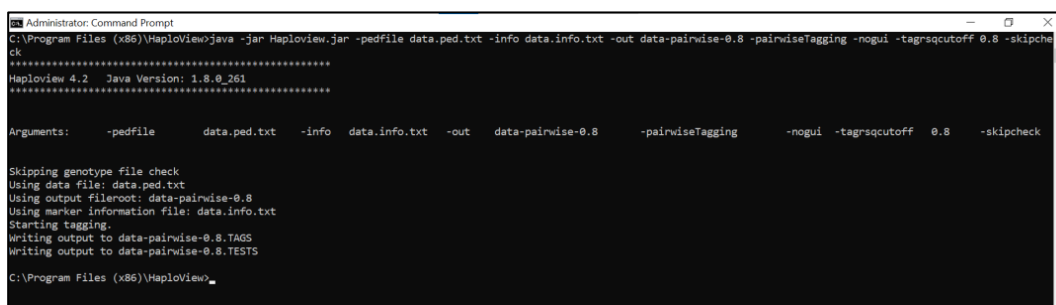
จากข้อมูล NSP และ STY Annotation สังเกตได้ว่า มีสลิปบางส่วนที่ไม่สามารถค้นหา position ของสลิปได้ โดยจะแสดงดังตารางที่ ซึ่งสลิปถูกตัดโดย exclusion list ทั้งหมด

ตารางที่ 3.2 แสดงให้เห็นสลิปที่ไม่พบ position จากข้อมูล NSP และ STY Annotation

Chromosome	SNP
1	SNP_A-1876089 , SNP_A-4255977 , SNP_A-2128780 , SNP_A-1845230 , SNP_A-2000157 , SNP_A-1781105
7	SNP_A-1782317 , SNP_A-4273084 , SNP_A-2196787 , SNP_A-1813412 , SNP_A-2049537 , SNP_A-2024620
13	SNP_A-4237637 , SNP_A-1782274 , SNP_A-1895472

3.4 การใช้ Command line ในการคัดเลือกสปีดัวแทน

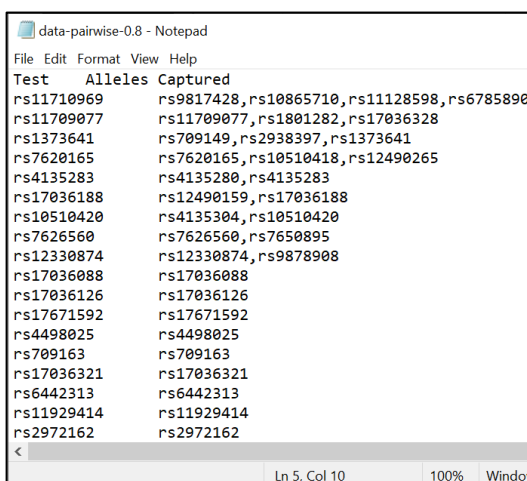
1. เปิดหน้าต่าง Command Prompt โดยเปลี่ยนตำแหน่งการทำงานโดยใช้คำสั่ง `cd` ไปที่ "C:\Program Files\Haploview"
2. เลือกวิธีในการวิเคราะห์ผล เลือกตัวอย่างไฟล์ input และกำหนดไฟล์ output โดยใช้คำสั่งแตกต่างกัน เช่น `java -jar Haploview.jar -pedfile data.ped -info data.info -out data-pairwise-0.8 -pairwiseTagging -nogui -tagsqcutoff 0.8 -skipcheck` ดังแสดงในภาพที่ 36



```
Administrator: Command Prompt
C:\Program Files (x86)\Haploview>java -jar Haploview.jar -pedfile data.ped.txt -info data.info.txt -out data-pairwise-0.8 -pairwiseTagging -nogui -tagsqcutoff 0.8 -skipcheck
Haploview 4.2  Java Version: 1.8.0_261
Arguments:      -pedfile      data.ped.txt      -info      data.info.txt      -out      data-pairwise-0.8      -pairwiseTagging      -nogui      -tagsqcutoff      0.8      -skipcheck
Skipping genotype file check
Using data file: data.ped.txt
Using output fileroot: data-pairwise-0.8
Using marker information file: data.info.txt
Starting tagging.
Writing output to data-pairwise-0.8.TAGS
Writing output to data-pairwise-0.8.TESTS
C:\Program Files (x86)\Haploview>
```

ภาพที่ 3.29 แสดงการใช้งาน Command line ในการคัดเลือกสปีดัวแทน

การคัดเลือกสปีดัวแทน โปรแกรมจะเริ่มสร้างไฟล์ output ไว้สำหรับเก็บข้อมูลที่ได้จากการวิเคราะห์ข้อมูล input ตามชื่อที่ผู้ใช้งานทำการกำหนดไว้ เช่น ตั้งชื่อ output เป็น data-pairwise-0.8.out ผลจากการศึกษาจะแสดงเป็นไฟล์ Notepad ที่แสดงผลการศึกษาการจับคู่สปีที่สามารถเป็นตัวแทนของกันได้ ดังแสดงในภาพที่ 37



Test	Alleles Captured
rs11710969	rs9817428,rs10865710,rs11128598,rs6785890,
rs11709077	rs11709077,rs1801282,rs17036328
rs1373641	rs709149,rs2938397,rs1373641
rs7620165	rs7620165,rs10510418,rs12490265
rs4135283	rs4135280,rs4135283
rs17036188	rs12490159,rs17036188
rs10510420	rs4135304,rs10510420
rs7626560	rs7626560,rs7650895
rs12330874	rs12330874,rs9878908
rs17036088	rs17036088
rs17036126	rs17036126
rs17671592	rs17671592
rs4498025	rs4498025
rs709163	rs709163
rs17036321	rs17036321
rs6442313	rs6442313
rs11929414	rs11929414
rs2972162	rs2972162

ภาพที่ 3.30 แสดงผลการคัดเลือกสปีดัวแทน

บทที่ 4

สรุปผลการวิจัย

4.1 ผลลัพธ์การคัดเลือกสnpตัวแทน

การศึกษาความสัมพันธ์ของจีโนมกับเครื่องหมาย SNP ได้กลายเป็นเครื่องมือมาตรฐานในการค้นหายีนที่มีพื้นฐานของโรคที่ซับซ้อน โดยใช้ tagSNP เพื่อให้ได้สnpตัวแทนซึ่งสามารถใช้เป็นตัวแทนของสnpที่อยู่ในภาวะความไม่สมดุลการเชื่อมโยงกับสnpตัวแทนในการศึกษาความสัมพันธ์ ซึ่งจะแบ่งการศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรมแตกต่างกัน โดยมีโรคทั้งหมด 7 ประเภท ได้แก่

1. Bipolar Disorder (BD) โรคอารมณ์สองขั้ว
2. Coronary Artery Disease (CAD) โรคหลอดเลือดโคโรนารี หรือ ภาวะหัวใจขาดเลือด
3. Crohn's Disease (CD) โรคโครห์น หรือ โรคที่เกิดการอักเสบเรื้อรังของระบบทางเดินอาหาร
4. Hypertension (HT) โรคความดันโลหิตสูง
5. Rheumatoid Arthritis (RA) โรคข้ออักเสบรูมาตอยด์
6. Type 1 diabetes (T1D) โรคเบาหวานชนิดที่ 1
7. Type 2 diabetes (T2D) โรคเบาหวานชนิดที่ 2

โดยการคัดเลือกสnpตัวแทนจะแบ่งออกเป็น 3 ส่วนคือ

- การศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรม Mapping250K_NSP.na32.annot รวมกับ Mapping250K_STY.na32.annot ซึ่งเวลาในการคัดเลือกสnpตัวแทนใช้เวลาไปทั้งหมด **3 ชั่วโมง 20 นาที**
- การศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรม Mapping250K_NSP.na32.annot ซึ่งเวลาในการคัดเลือกสnpตัวแทนใช้เวลาไปทั้งหมด **50 นาที**
- การศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรม Mapping250K_STY.na32.annot ซึ่งเวลาในการคัดเลือกสnpตัวแทนใช้เวลาไปทั้งหมด **50 นาที**

4.2 การศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรม

Mapping250K_NSP.na32.annot รวมกับ Mapping250K_STY.na32.annot

ตารางที่ 4.1 แสดงจำนวนการคัดเลือกส니ปตัวแทนที่ใช้เป็นตัวแทนของส니ปที่สัมพันธ์กับโรค BD โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	28393	14924	0.5256
2	29853	15590	0.5222
3	24790	13078	0.5276
4	22921	11847	0.5169
5	23774	12380	0.5207
6	23826	12238	0.5136
7	19207	10419	0.5425
8	20112	10553	0.5247
9	16995	9288	0.5465
10	20778	10632	0.5117
11	19320	9693	0.5017
12	18244	9663	0.5297
13	13910	7399	0.5319
14	11184	6237	0.5577
15	10183	5959	0.5852
16	10909	6357	0.5827
17	8158	4795	0.5878
18	10662	5890	0.5524
19	4635	2921	0.6302
20	9054	4949	0.5466
21	5285	2893	0.5474
22	4442	2700	0.6078
X	7293	2908	0.3987
Total	363928	193313	0.5396

ตารางที่ 4.2 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค CAD โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	28393	14895	0.5246
2	29853	15576	0.5218
3	24790	13058	0.5267
4	22921	11829	0.5161
5	23774	12383	0.5209
6	23826	12218	0.5128
7	19207	10417	0.5424
8	20112	10544	0.5243
9	16995	9279	0.5460
10	20778	10637	0.5119
11	19320	9687	0.5014
12	18244	9661	0.5295
13	13910	7391	0.5313
14	11184	6223	0.5564
15	10183	5950	0.5843
16	10909	6351	0.5822
17	8158	4789	0.5870
18	10662	5887	0.5521
19	4635	2923	0.6306
20	9054	4962	0.5480
21	5285	2887	0.5463
22	4442	2707	0.6094
X	7293	3904	0.5353
Total	363928	194158	0.5453

ตารางที่ 4.3 แสดงจำนวนการคัดเลือกสปีทัวแทนที่ใช้เป็นตัวแทนของสปีทัวที่สัมพันธ์กับโรค CD โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	28393	14927	0.5257
2	29853	15592	0.5223
3	24790	13079	0.5276
4	22921	11845	0.5168
5	23774	12394	0.5213
6	23826	12234	0.5135
7	19207	10421	0.5426
8	20112	10545	0.5243
9	16995	9298	0.5471
10	20778	10644	0.5123
11	19320	9694	0.5018
12	18244	9669	0.5300
13	13910	7402	0.5321
14	11184	6234	0.5574
15	10183	5954	0.5847
16	10909	6361	0.5831
17	8158	4798	0.5881
18	10662	5895	0.5529
19	4635	2920	0.6300
20	9054	4966	0.5485
21	5285	2888	0.5465
22	4442	2708	0.6096
X	7293	3909	0.5360
Total	363928	194377	0.5458

ตารางที่ 4.4 แสดงจำนวนการคัดเลือกส니ปตัวแทนที่ใช้เป็นตัวแทนของส니ปที่สัมพันธ์กับโรค HT โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	28393	14895	0.5246
2	29853	15586	0.5221
3	24790	13066	0.5271
4	22921	11821	0.5157
5	23774	12378	0.5207
6	23826	12237	0.5136
7	19207	10418	0.5424
8	20112	10543	0.5242
9	16995	9283	0.5462
10	20778	10634	0.5118
11	19320	9692	0.5017
12	18244	9661	0.5295
13	13910	7394	0.5316
14	11184	6224	0.5565
15	10183	5956	0.5849
16	10909	6356	0.5826
17	8158	4792	0.5874
18	10662	5890	0.5524
19	4635	2925	0.6311
20	9054	4957	0.5475
21	5285	2891	0.547
22	4442	2710	0.6101
X	7293	3909	0.536
Total	363928	194218	0.5455

ตารางที่ 4.5 แสดงจำนวนการคัดเลือกสปีทัวแทนที่ใช้เป็นตัวแทนของสปีทัวที่สัมพันธ์กับโรค RA โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	28393	14921	0.5255
2	29853	15590	0.5222
3	24790	13066	0.5271
4	22921	11839	0.5165
5	23774	12377	0.5206
6	23826	12232	0.5134
7	19207	10405	0.5417
8	20112	10540	0.5241
9	16995	9279	0.546
10	20778	10632	0.5117
11	19320	9694	0.5018
12	18244	9672	0.5301
13	13910	7385	0.5309
14	11184	6228	0.5569
15	10183	5940	0.5833
16	10909	6338	0.581
17	8158	4786	0.5867
18	10662	5891	0.5525
19	4635	2920	0.63
20	9054	4958	0.5476
21	5285	2886	0.5461
22	4442	2708	0.6096
X	7293	3906	0.5356
Total	363928	194193	0.5453

ตารางที่ 4.6 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค T1D โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	28393	14910	0.5251
2	29853	15593	0.5223
3	24790	13061	0.5269
4	22921	11841	0.5166
5	23774	12390	0.5212
6	23826	12243	0.5139
7	19207	10419	0.5425
8	20112	10536	0.5239
9	16995	9282	0.5462
10	20778	10648	0.5125
11	19320	9694	0.5018
12	18244	9666	0.5298
13	13910	7395	0.5316
14	11184	6230	0.557
15	10183	5955	0.5848
16	10909	6354	0.5825
17	8158	4797	0.588
18	10662	5894	0.5528
19	4635	2921	0.6302
20	9054	4960	0.5478
21	5285	2891	0.547
22	4442	2709	0.6099
X	7293	3906	0.5356
Total	363928	194295	0.5456

ตารางที่ 4.7 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค T2D โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	28393	14899	0.5247
2	29853	15578	0.5218
3	24790	13058	0.5267
4	22921	11835	0.5163
5	23774	12383	0.5209
6	23826	12232	0.5134
7	19207	10418	0.5424
8	20112	10548	0.5245
9	16995	9302	0.5473
10	20778	10635	0.5118
11	19320	9702	0.5022
12	18244	9669	0.53
13	13910	7402	0.5321
14	11184	6232	0.5572
15	10183	5958	0.5851
16	10909	6351	0.5822
17	8158	4788	0.5869
18	10662	5897	0.5531
19	4635	2923	0.6306
20	9054	4957	0.5475
21	5285	2893	0.5474
22	4442	2708	0.6096
X	7293	3912	0.5364
Total	363928	194280	0.5457

ตารางที่ 4.8 แสดงค่าเฉลี่ยของจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรคทั้ง 7 โรค โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation

Disease	Using Tag SNPs	remain
BD	193313	0.5396
CAD	194158	0.5453
CD	194377	0.5458
HT	194218	0.5455
RA	194193	0.5453
T1D	194295	0.5456
T2D	194280	0.5457
Mean		0.5445

พิจารณาจากตารางที่ โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation จะเห็นได้ว่า ค่าเฉลี่ยของจำนวนสลิปที่ใช้เป็นตัวแทนสลิปที่สัมพันธ์กับโรคทั้ง 7 โรคนี้ขึ้นอยู่กับประมาณ 0.5445

4.3 การศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรม

Mapping250K_NSP.na32.annot

ตารางที่ 4.9 แสดงจำนวนการคัดเลือกส니ปตัวแทนที่ใช้เป็นตัวแทนของส니ปที่สัมพันธ์กับโรค BD โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	14437	9051	0.6269
2	16404	10013	0.6104
3	13854	8526	0.6154
4	13849	8232	0.5944
5	12999	7926	0.6097
6	13339	7976	0.5979
7	10662	6673	0.6259
8	11153	6836	0.6129
9	9119	5800	0.636
10	10737	6527	0.6079
11	10082	5929	0.5881
12	9777	6081	0.622
13	8206	4994	0.6086
14	6006	3886	0.647
15	5097	3456	0.678
16	5216	3563	0.6831
17	3631	2471	0.6805
18	5955	3796	0.6374
19	1975	1435	0.7266
20	4370	2859	0.6542
21	2956	1829	0.6187
22	1848	1326	0.7175
X	4024	2548	0.6332
Total	195696	121733	0.6362

ตารางที่ 4.10 แสดงจำนวนการคัดเลือกสปีตัวแทนที่ใช้เป็นตัวแทนของสปีที่สัมพันธ์กับโรค CAD โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	14437	9038	0.626
2	16404	10014	0.6105
3	13854	8521	0.6151
4	13849	8223	0.5938
5	12999	7926	0.6097
6	13339	7981	0.5983
7	10662	6672	0.6258
8	11153	6829	0.6123
9	9119	5799	0.6359
10	10737	6527	0.6079
11	10082	5933	0.5885
12	9777	6080	0.6219
13	8206	4985	0.6075
14	6006	3884	0.6467
15	5097	3458	0.6784
16	5216	3561	0.6827
17	3631	2475	0.6816
18	5955	3798	0.6378
19	1975	1434	0.7261
20	4370	2864	0.6554
21	2956	1826	0.6177
22	1848	1326	0.7175
X	4024	2555	0.6349
Total	195696	121709	0.6362

ตารางที่ 4.11 แสดงจำนวนการคัดเลือกสปีปตัวแทนที่ใช้เป็นตัวแทนของสปีปที่สัมพันธ์กับโรค CD โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	14437	9054	0.6271
2	16404	10031	0.6115
3	13854	8532	0.6159
4	13849	8235	0.5946
5	12999	7927	0.6098
6	13339	7977	0.598
7	10662	6673	0.6259
8	11153	6836	0.6129
9	9119	5800	0.636
10	10737	6527	0.6079
11	10082	5929	0.5881
12	9777	6081	0.622
13	8206	4994	0.6086
14	6006	3886	0.647
15	5097	3456	0.678
16	5216	3563	0.6831
17	3631	2471	0.6805
18	5955	3796	0.6374
19	1975	1435	0.7266
20	4370	2859	0.6542
21	2956	1829	0.6187
22	1848	1326	0.7175
X	4024	2548	0.6332
Total	195696	121765	0.6363

ตารางที่ 4.12 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค HT โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation

Chromosome	Number of SNPs	Number of Tag SNPs	ratio
1	14437	9042	0.6263
2	16404	10023	0.611
3	13854	8519	0.6149
4	13849	8220	0.5935
5	12999	7925	0.6097
6	13339	7980	0.5982
7	10662	6673	0.6259
8	11153	6834	0.6127
9	9119	5791	0.635
10	10737	6520	0.6072
11	10082	5935	0.5887
12	9777	6082	0.6221
13	8206	4987	0.6077
14	6006	3882	0.6464
15	5097	3459	0.6786
16	5216	3554	0.6814
17	3631	2472	0.6808
18	5955	3799	0.638
19	1975	1436	0.7271
20	4370	2866	0.6558
21	2956	1827	0.6181
22	1848	1327	0.7181
X	4024	2554	0.6347
Total	195696	121707	0.6362

ตารางที่ 4.13 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค RA โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	14437	9051	0.6269
2	16404	10015	0.6105
3	13854	8518	0.6148
4	13849	8233	0.5945
5	12999	7937	0.6106
6	13339	7976	0.5979
7	10662	6670	0.6256
8	11153	6831	0.6125
9	9119	5797	0.6357
10	10737	6529	0.6081
11	10082	5937	0.5889
12	9777	6087	0.6226
13	8206	4985	0.6075
14	6006	3890	0.6477
15	5097	3453	0.6775
16	5216	3554	0.6814
17	3631	2472	0.6808
18	5955	3800	0.6381
19	1975	1436	0.7271
20	4370	2863	0.6551
21	2956	1825	0.6174
22	1848	1331	0.7202
X	4024	2549	0.6334
Total	195696	121739	0.6363

ตารางที่ 4.14 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค T1D โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	14437	9041	0.6262
2	16404	10019	0.6108
3	13854	8519	0.6149
4	13849	8222	0.5937
5	12999	7935	0.6104
6	13339	7988	0.5988
7	10662	6677	0.6262
8	11153	6835	0.6128
9	9119	5802	0.6363
10	10737	6535	0.6086
11	10082	5936	0.5888
12	9777	6079	0.6218
13	8206	4991	0.6082
14	6006	3888	0.6474
15	5097	3465	0.6798
16	5216	3565	0.6835
17	3631	2473	0.6811
18	5955	3800	0.6381
19	1975	1435	0.7266
20	4370	2864	0.6554
21	2956	1828	0.6184
22	1848	1327	0.7181
X	4024	2549	0.6334
Total	195696	121773	0.6365

ตารางที่ 4.15 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค T2D โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	14437	9045	0.6265
2	16404	10016	0.6106
3	13854	8515	0.6146
4	13849	8221	0.5936
5	12999	7935	0.6104
6	13339	7979	0.5982
7	10662	6675	0.6261
8	11153	6839	0.6132
9	9119	5803	0.6364
10	10737	6530	0.6082
11	10082	5938	0.589
12	9777	6082	0.6221
13	8206	4994	0.6086
14	6006	3889	0.6475
15	5097	3458	0.6784
16	5216	3551	0.6808
17	3631	2472	0.6808
18	5955	3794	0.6371
19	1975	1435	0.7266
20	4370	2865	0.6556
21	2956	1825	0.6174
22	1848	1330	0.7197
X	4024	2558	0.6357
Total	195696	121749	0.6364

ตารางที่ 4.16 แสดงค่าเฉลี่ยของจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรคทั้ง 7 โรค โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation

Disease	Using Tag SNPs	remain
BD	121733	0.6362
CAD	121709	0.6362
CD	121765	0.6363
HT	121707	0.6362
RA	121739	0.6363
T1D	121773	0.6365
T2D	121749	0.6364
Mean		0.6363

พิจารณาจากตารางที่ โดยใช้ข้อมูลทางพันธุกรรม NSP Annotation จะเห็นได้ว่าค่าเฉลี่ยของจำนวนสลิปที่ใช้เป็นตัวแทนสลิปที่สัมพันธ์กับโรคทั้ง 7 โรคนั้นอยู่ที่ประมาณ 0.6363

4.3 การศึกษาความสัมพันธ์ทั้งจีโนมโดยใช้ข้อมูลทางพันธุกรรม

Mapping250K_STY.na32.annot

ตารางที่ 4.17 แสดงจำนวนการคัดเลือกส니ปตัวแทนที่ใช้เป็นตัวแทนของสนิปที่สัมพันธ์กับโรค BD โดยใช้ข้อมูลทางพันธุกรรม STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	13956	9238	0.6619
2	13449	9084	0.6754
3	10936	7426	0.679
4	9072	6194	0.6828
5	10775	7229	0.6709
6	10487	7031	0.6704
7	8545	5906	0.6912
8	8959	6050	0.6753
9	7876	5407	0.6865
10	10041	6497	0.647
11	10082	5929	0.5881
12	8467	5687	0.6717
13	5704	3985	0.6986
14	5178	3599	0.6951
15	5086	3577	0.7033
16	5693	3945	0.693
17	4527	3152	0.6963
18	4707	3252	0.6909
19	2660	1933	0.7267
20	4684	3181	0.6791
21	2329	1610	0.6913
22	2594	1835	0.7074
X	3269	2192	0.6705
Total	169076	113939	0.6805

ตารางที่ 4.18 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค CAD โดยใช้ข้อมูลทางพันธุกรรม STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	13956	9227	0.6611
2	13449	9082	0.6753
3	10936	7427	0.6791
4	9072	6197	0.6831
5	10775	7227	0.6707
6	10487	7023	0.6697
7	8545	5903	0.6908
8	8959	6055	0.6759
9	7876	5411	0.687
10	10041	6492	0.6465
11	10082	5934	0.5886
12	8467	5688	0.6718
13	5704	3981	0.6979
14	5178	3590	0.6933
15	5086	3573	0.7025
16	5693	3950	0.6938
17	4527	3148	0.6954
18	4707	3260	0.6926
19	2660	1932	0.7263
20	4684	3184	0.6798
21	2329	1609	0.6909
22	2594	1838	0.7086
X	3269	2190	0.6699
Total	169076	113921	0.6805

ตารางที่ 4.19 แสดงจำนวนการคัดเลือกสลับตัวแทนที่ใช้เป็นตัวแทนของสลับที่สัมพันธ์กับโรค CD โดยใช้ข้อมูลทางพันธุกรรม STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	13956	9230	0.6614
2	13449	9071	0.6745
3	10936	7437	0.68
4	9072	6194	0.6828
5	10775	7226	0.6706
6	10487	7036	0.6709
7	8545	5904	0.6909
8	8959	6049	0.6752
9	7876	5412	0.6872
10	10041	6497	0.647
11	10082	5937	0.5889
12	8467	5688	0.6718
13	5704	3985	0.6986
14	5178	3599	0.6951
15	5086	3572	0.7023
16	5693	3946	0.6931
17	4527	3150	0.6958
18	4707	3259	0.6924
19	2660	1932	0.7263
20	4684	3184	0.6798
21	2329	1612	0.6921
22	2594	1838	0.7086
X	3269	2195	0.6715
Total	169076	113953	0.6807

ตารางที่ 4.20 แสดงจำนวนการคัดเลือกสปีดัวแทนที่ใช้เป็นตัวแทนของสปีดัวที่สัมพันธ์กับโรค HT โดยใช้ข้อมูลทางพันธุกรรม STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	13956	9223	0.6609
2	13449	9082	0.6753
3	10936	7428	0.6792
4	9072	6188	0.6821
5	10775	7230	0.671
6	10487	7027	0.6701
7	8545	5903	0.6908
8	8959	6052	0.6755
9	7876	5409	0.6868
10	10041	6499	0.6472
11	10082	5935	0.5887
12	8467	5689	0.6719
13	5704	3983	0.6983
14	5178	3591	0.6935
15	5086	3572	0.7023
16	5693	3947	0.6933
17	4527	3153	0.6965
18	4707	3254	0.6913
19	2660	1930	0.7256
20	4684	3182	0.6793
21	2329	1611	0.6917
22	2594	1837	0.7082
X	3269	2197	0.6721
Total	169076	113922	0.6805

ตารางที่ 4.21 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค RA โดยใช้ข้อมูลทางพันธุกรรม STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	13956	9235	0.6617
2	13449	9080	0.6751
3	10936	7425	0.679
4	9072	6191	0.6824
5	10775	7224	0.6704
6	10487	7026	0.67
7	8545	5895	0.6899
8	8959	6045	0.6747
9	7876	5410	0.6869
10	10041	6494	0.6467
11	10082	5937	0.5889
12	8467	5683	0.6712
13	5704	3980	0.6978
14	5178	3595	0.6943
15	5086	3569	0.7017
16	5693	3942	0.6924
17	4527	3150	0.6958
18	4707	3253	0.6911
19	2660	1929	0.7252
20	4684	3179	0.6787
21	2329	1610	0.6913
22	2594	1833	0.7066
X	3269	2196	0.6718
Total	169076	113881	0.6802

ตารางที่ 4.22 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค T1D โดยใช้ข้อมูลทางพันธุกรรม STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	13956	9222	0.6608
2	13449	9081	0.6752
3	10936	7428	0.6792
4	9072	6203	0.6838
5	10775	7223	0.6703
6	10487	7028	0.6702
7	8545	5901	0.6906
8	8959	6045	0.6747
9	7876	5411	0.687
10	10041	6497	0.647
11	10082	5936	0.5888
12	8467	5689	0.6719
13	5704	3986	0.6988
14	5178	3594	0.6941
15	5086	3575	0.7029
16	5693	3950	0.6938
17	4527	3153	0.6965
18	4707	3254	0.6913
19	2660	1928	0.7248
20	4684	3187	0.6804
21	2329	1610	0.6913
22	2594	1838	0.7086
X	3269	2201	0.6733
Total	169076	113940	0.6807

ตารางที่ 4.23 แสดงจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรค T2D โดยใช้ข้อมูลทางพันธุกรรม STY Annotation

Chromosome	Number of SNPs	Using Tag SNPs	Remain
1	13956	9230	0.6614
2	13449	9078	0.675
3	10936	7424	0.6789
4	9072	6192	0.6825
5	10775	7222	0.6703
6	10487	7025	0.6699
7	8545	5910	0.6916
8	8959	6042	0.6744
9	7876	5413	0.6873
10	10041	6500	0.6473
11	10082	5938	0.589
12	8467	5693	0.6724
13	5704	3983	0.6983
14	5178	3594	0.6941
15	5086	3575	0.7029
16	5693	3950	0.6938
17	4527	3146	0.6949
18	4707	3261	0.6928
19	2660	1930	0.7256
20	4684	3181	0.6791
21	2329	1609	0.6909
22	2594	1836	0.7078
X	3269	2195	0.6715
Total	169076	113927	0.6805

ตารางที่ 4.24 แสดงค่าเฉลี่ยของจำนวนการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรคทั้ง 7 โรค โดยใช้ข้อมูลทางพันธุกรรม STY Annotation

Disease	Using Tag SNPs	remain
BD	113939	0.6805
CAD	113921	0.6805
CD	113953	0.6807
HT	113922	0.6805
RA	113881	0.6802
T1D	113940	0.6807
T2D	113927	0.6805
Mean		0.6805

พิจารณาจากตารางที่ โดยใช้ข้อมูลทางพันธุกรรม STY Annotation จะเห็นได้ว่าค่าเฉลี่ยของจำนวนสลิปที่ใช้เป็นตัวแทนสลิปที่สัมพันธ์กับโรคทั้ง 7 โรคนั้นอยู่ที่ประมาณ 0.6805

เมื่อพิจารณาการคัดเลือกสลิปตัวแทนที่ใช้เป็นตัวแทนของสลิปที่สัมพันธ์กับโรคทั้ง 7 โรค โดยใช้ข้อมูลทางพันธุกรรม NSP และ STY Annotation นั้นจะได้ค่าเฉลี่ยอยู่ที่ 0.5457 ซึ่งมีจำนวนสลิปที่ใช้เป็นตัวแทนสลิปนั้นน้อยกว่าการใช้ข้อมูลทางพันธุกรรม NSP Annotation และ ข้อมูลทางพันธุกรรม STY Annotation โดยจะมีค่าเฉลี่ยอยู่ที่ประมาณ 0.6 ซึ่งเกิดจากเมื่อทำการรวมข้อมูลทางพันธุกรรมของไฟล์ NSP และ STY Annotation เข้าด้วยกัน จะสังเกตได้ว่ามีสลิปบางส่วนที่สามารถเป็นตัวแทนของทั้งข้อมูลของ NSP และ STY Annotation ได้ ดังนั้นเมื่อแยกข้อมูลของ NSP และ STY Annotation ออกจากกัน จึงทำให้จำเป็นต้องคัดเลือกสลิปตัวแทนเพิ่มในกลุ่มของข้อมูลนั้น เป็นผลทำให้จำนวนสลิปตัวแทนที่ได้จากการคัดเลือกตัวแทนสลิปของข้อมูล NSP Annotation รวมเข้ากับ STY Annotation มีจำนวนสลิปตัวแทนน้อยกว่า

เอกสารอ้างอิง

- [1] D. O. Stram et al., “Choosing Haplotype-Tagging SNPS Based on Unphased Genotype Data Using a Preliminary Sample of Unrelated Subjects with an Example from the Multiethnic Cohort Study,” *Hum Hered*, vol. 55, no. 1, pp. 27–36, 2003.
- [2] D. O. Stram, “Software for tag single nucleotide polymorphism selection,” *Hum Genomics*, vol. 2, no. 2, p. 144, 2005.
- [3] D. O. Stram, “Tag SNP selection for association studies,” *Genet. Epidemiol.*, vol. 27, no. 4, pp. 365–374, Dec. 2004.
- [4] P. I. W. de Bakker, R. Yelensky, I. Pe’er, S. B. Gabriel, M. J. Daly, and D. Altshuler, “Efficiency and power in genetic association studies,” *Nat Genet*, vol. 37, no. 11, pp. 1217–1223, Nov. 2005.
- [5] Broad Institute, “Haploview,” May 16, 2008. <https://www.broadinstitute.org/Haploview/Haploview> (accessed Mar. 25, 2021).
- [6] J. M. VanLiere and N. A. Rosenberg, “Mathematical properties of the measure of linkage disequilibrium,” *Theoretical Population Biology*, vol. 74, no. 1, pp. 130–137, Aug. 2008.
- [7] 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. 661–678, Jun. 2007.
- [8] Thermo Fisher, “GeneChip Array Annotation Files - Mapping250K_NSP Annotations, CSV format, Release 32 (83 MB, 7/15/11) and Mapping250K_STY Annotations, CSV format, Release 32 (76 MB, 7/15/11)” www.thermofisher.com/id/en/home/life-science/microarray-analysis/microarray-data-analysis/genechip-array-annotation-files.html (accessed Mar. 25, 2021).
- [9] Enhancing Neuro Imaging Genetics Through Meta Analysis, “GRCh37_hg19_AffyID2rsnumbers.txt,” http://enigma.ini.usc.edu/wp-content/uploads/2012/04/GRCh37_hg19_AffyID2rsnumbers.txt (accessed Mar. 25, 2021).

เอกสารอ้างอิง (ต่อ)

- [10] National Center for Biotechnology Information, “RsMergeArch.bep.gz,” https://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/database/organism_data/RsMergeArch.bcp.gz (accessed Mar. 25, 2021).
- [11] E. Candès, Y. Fan, L. Janson, and J. Lv, “Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 3, pp. 551–577, 2018.

ประวัติผู้แต่ง

ปริญญาบัตรเรื่อง : การคัดเลือกสปีดวแทนจากการศึกษาความสัมพันธ์ทั้งจีโนม
 สาขาวิชา : วิศวกรรมคอมพิวเตอร์
 ภาควิชา : วิศวกรรมไฟฟ้าและคอมพิวเตอร์
 คณะ : วิศวกรรมศาสตร์
 ชื่อ : นายธรรมบุญ กิจรสอนันต์
 ประวัติ :

เกิดเมื่อวันที่ 10 ธันวาคม พ.ศ. 2541 อยู่บ้านเลขที่ 1931 ซอยกาญจนาภิเษก 008 แขวงบาง
 แคน เขตบางแค จังหวัดกรุงเทพฯ ฯ สำเร็จการศึกษาระดับมัธยมศึกษาตอนปลายจากโรงเรียนมัธยมวัด
 สิมห์ สาขาวิทยาศาสตร์-คณิตศาสตร์ ปีการศึกษา 2559 และสำเร็จการศึกษาในระดับปริญญาตรี
 สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
 มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปีการศึกษา 2562

ชื่อ : นางสาวรากร มณีแสง
 ประวัติ :

เกิดเมื่อวันที่ 17 ธันวาคม พ.ศ. 2541 อยู่บ้านเลขที่ 51/11 หมู่ 1 ตำบลท่าเสา อำเภอกระทุ่ม
 แบน จังหวัดสมุทรสาคร สำเร็จการศึกษาระดับมัธยมศึกษาตอนปลาย จากโรงเรียนนวมินทราชินูทิศ
 สตรีวิทยา พุทธมณฑล สาขาวิทยาศาสตร์-คณิตศาสตร์ ปีการศึกษา 2559 และสำเร็จการศึกษาใน
 ระดับปริญญาตรี สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะ
 วิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปีการศึกษา 2562