เครื่องมือวิเคราะห์แบบรวมสำหรับการแปรผันของจำนวนชุดดีเอ็นเอบนลำดับเอ็กโซม

น.ส.เสาวภาค จันทร์วิกูล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมซอฟต์แวร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

An Integrated Analysis Tool for Copy Number Variation on Whole Exome Sequencing

Miss Saowwapark Chanwigoon

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Software Engineering
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2019

หัวข้อวิทยานิพนธ์	เครื่องมือวิเคราะห์แบบรวมสำหรับการแปรผันของจำนวนชุดดีเอ็นเอบนลำดับเอ็กโซม
โดย	น.ส.เสาวภาค จันทร์วิกูล
สาขาวิชา	วิศวกรรมซอฟต์แวร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร.ดวงดาว วิชาดากุล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

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

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.วิวัฒน์ วัฒนาวุฒิ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร.ดวงดาว วิชาดากุล)

..... กรรมการ
(รองศาสตราจารย์ ดร.ทวิตีย์ เสนีวงศ์ ณ อยุธยา)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.นุวีย์ วิวัฒนวัฒนา)

เสาวภาค จันทร์วิกูล : เครื่องมือวิเคราะห์แบบรวมสำหรับการแปรผันของจำนวนชุดดีเอ็นเอบนลำดับเอ็กโซม. (An Integrated Analysis Tool for Copy Number Variation on Whole Exome Sequencing) อ.ที่ปรึกษาหลัก : อ. ดร.ดวงดาว วิชาดากุล

การศึกษาการแปรผันของจำนวนชุดดีเอ็นเอบนลำดับเอ็กโซม หรือที่เรียกว่าการศึกษาซีเอ็นวีบนเอ็กโซม เป็นหนึ่งในการศึกษาการแปรผันเชิงโครงสร้างของสารพันธุกรรมที่ได้รับความนิยมในปัจจุบัน การศึกษานี้สามารถช่วยนักวิจัยให้เข้าใจวิวัฒนาการของมนุษย์ การวินิจฉัยโรค และการตอบสนองต่อยาที่ใช้รักษาโรคได้ แต่ในขณะเดียวกันการศึกษานี้ก็เป็นเรื่องที่ท้าทาย เนื่องจากกระบวนการเพื่อให้ได้มาซึ่งเอ็กโซมมาพร้อมกับค่าผลบวกเท็จสูงมาก และเมื่อนำเอ็กโซมที่มีค่าผลบวกเท็จสูงนี้มาตรวจจับหาการแปรผันของจำนวนชุดดีเอ็นเอบนลำดับเอ็กโซมก็ยิ่งก่อให้เกิดค่าผลบวกเท็จสูงยิ่งขึ้น อีกทั้งการแปรผันนี้ยังมีคุณลักษณะที่หลากหลายทำให้นักวิจัยไม่สามารถสร้างเครื่องมือตรวจจับที่ครอบคลุมคุณลักษณะทั้งหมดได้ นักวิจัยได้พยายามจัดการกับปัญหานี้ด้วยการสร้างเครื่องมือตรวจจับการแปรผันจำนวนมากที่มีลักษณะการทำงานที่แตกต่างกัน อย่างไรก็ตามยังไม่มีเครื่องมือตรวจจับการแปรผันเครื่องมือใดสามารถแก้ปัญหานี้ได้ อีกทั้งเครื่องมือตรวจจับการแปรผันส่วนใหญ่ขาดความสะดวก และความยืดหยุ่นในการใช้งาน เช่น ผู้ใช้ต้องติดตั้งเครื่องมือตรวจจับการแปรผันผ่านทางคอมมานไลน์ เครื่องมือตรวจจับไม่ให้คำอธิบายประกอบให้กับการแปรผันที่ตรวจจับได้ และไม่เปิดเผยข้อมูลที่ใช้ในการประมวลผล เป็นต้น จากปัญหาดังกล่าวผู้วิจัยจึงสร้างเครื่องมือตรวจจับการแปรผันของจำนวนชุดดีเอ็นเอบนลำดับเอ็กโซมแบบบูรณาการในรูปแบบเว็บแอปพลิเคชันซึ่งง่ายต่อการติดตั้งที่ชื่อว่า “อินซีเอ็นวี” เครื่องมือนี้สามารถรวมผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวีหลายเครื่องมือเพื่อเพิ่มความแม่นยำในการตรวจจับการแปรผัน หาความสัมพันธ์ของการแปรผันจากหลายตัวอย่างเพื่อช่วยในการวินิจฉัยโรค และจำกัดความเป็นไปได้ในการค้นพบการแปรผันตำแหน่งใหม่ที่ยังไม่เคยถูกรายงานในฐานข้อมูลการแปรผันทางพันธุกรรม เป็นต้น

สาขาวิชา	วิศวกรรมซอฟต์แวร์	ลายมือชื่อนิสิต .....
ปีการศึกษา	2562	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

5970349821 : MAJOR SOFTWARE ENGINEERING
KEYWORD: copy number variation, CNV, whole-exome sequencing, visualization

Saowwapark Chanwigoon : An Integrated Analysis Tool for Copy Number Variation on Whole Exome Sequencing. Advisor: Duangdao Wichadakul, Ph.D.

The study of copy number variations (CNVs) on whole-exome sequencing (WES) helps researchers gain insight into human genome diversity and predisposition to diseases. On the contrary, this study poses a major challenge of high false-positive rates from extracting exome. When researchers apply this kind of exome to detect CNVs, the false-positive rates become much higher. Moreover, CNVs have many characteristics making pre-built CNV tools unsuccessful in detecting all types of CNVs with full coverage. Researchers have tried to deal with those problems by creating a lot of CNV detection tools having various characteristics; however, those tools have still failed. Besides, numerous CNV detection tools lack the ease and flexibility of use. For example, users have to install CNV detection tools from command line; moreover, CNV detection tools do not have genome annotation, cannot prioritize the results including presenting it graphically, and users are unable to access the genome data of CNV tools in public. To solve those obstacles, we present inCNV, the web application that is installed easily, can integrate the CNV results from multiple CNV detection tools to improve the precision of detecting CNVs, find the relationships between CNVs to predict diseases, and limit the scope of potential novel CNVs.

Field of Study: Software Engineering Student's Signature .....
Academic Year: 2019 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์ของ อ.ดร.ดวงดาว วิชาดากุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งสละเวลาให้คำปรึกษา ช่วยตรวจสอบแก้ไขข้อบกพร่องต่าง ๆ จนทำให้การวิจัยครั้งนี้สำเร็จลุล่วงไปได้ด้วยดี

ขอขอบพระคุณกรรมการสอบวิทยานิพนธ์ รศ.ดร.วิวัฒน์ วัฒนาวุฒิ รศ.ดร.ทวิตีย์ เสนีวงศ์ ณ อยุธยา และ ผศ.ดร.นุวีย์ วิวัฒนวัฒนา ที่กรุณาสละเวลาให้คำแนะนำ ตรวจสอบ และแก้ไขวิทยานิพนธ์ ซึ่งเป็นประโยชน์ในการทาวิทยานิพนธ์ฉบับนี้อย่างยิ่ง

ขอขอบพระคุณ คุณศักยภาพ ผิวเหลือง ที่กรุณาสละเวลาในออกแบบซอฟต์แวร์ที่นำเสนอในวิทยานิพนธ์ฉบับนี้ให้สามารถอัพเดทคำอธิบายจีโนมได้อย่างอัตโนมัติเพื่อให้ซอฟต์แวร์นี้มีความน่าเชื่อถือ และสามารถใช้งานได้ในระยะยาว รวมถึงขอขอบพระคุณในความช่วยเหลือสำหรับการเขียนโปรแกรมบรรจุซอฟต์แวร์นี้ลงในสภาพแวดล้อมด็อกเกอร์เพื่อให้ผู้ใช้สามารถติดตั้งซอฟต์แวร์ได้โดยง่าย

ขอขอบพระคุณ คุณณัฐสุดา นวมะชิติ ที่กรุณาสละเวลาให้คำแนะนำในการนำเสนอผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจจับซีเอ็นวีหลายเครื่องมือในมุมมองรูปภาพที่ง่ายต่อการทำความเข้าใจของนักวิจัยทางชีวสารสนเทศ

สุดท้ายนี้ ข้าพเจ้าหวังเป็นอย่างยิ่งว่า เนื้อหาในวิทยานิพนธ์ฉบับนี้จะเป็นประโยชน์แก่ผู้อื่นไม่มากก็น้อย

เสาวภาค จันทร์วิกูล

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ด้วยการเกิดขึ้นมาของเทคโนโลยีเอ็นจีเอส (next generation sequencing: NGS) ทำให้การศึกษาการแปรผันเชิงโครงสร้างของลำดับเบสทั้งหมดบนจีโนม (whole genome sequencing: WGS) และลำดับเบสทั้งหมดบนเอ็กโซม (whole exome sequencing: WES) ได้รับความนิยมอย่างกว้างขวาง โดยหนึ่งในการแปรผันที่เป็นที่นิยมในการศึกษาคือ การศึกษาการแปรผันเชิงโครงสร้างที่เรียกว่าซีเอ็นวี (copy Number of variation: CNV)

ซีเอ็นวีเป็นสาเหตุหลักของการเปลี่ยนแปลงแก้ไขสำเนาบนชิ้นส่วนจีโนมจำนวนมากทำให้คนแต่ละคนมีลักษณะที่แตกต่างกัน และเกิดวิวัฒนาการของมนุษย์ [1-3] รวมถึงเป็นสาเหตุของการเกิดโรค [4] อีกด้วย ดังนั้นการศึกษาซีเอ็นวีสามารถช่วยในเรื่องของการศึกษาความหลากหลายของมนุษย์ ความเสี่ยงต่อการเกิดโรคได้ และการตอบสนองต่อยาที่ใช้ในการรักษาโรค ตัวอย่างโรคที่สัมพันธ์กับซีเอ็นวี เช่น ออทิสติก (Autism Spectrum Disorder) [5], โรคจิตเภท (schizophrenia) [6], เบาหวานชนิดที่ 2 (type-2 diabetes) [7], หัวใจพิการแต่กำเนิด (Congenital heart) [8], การสูญเสียการได้ยินชนิดไม่เป็นกลุ่มอาการ (non-syndromic hearing loss) [9, 10], โรคจอประสาทตาเสื่อมที่สืบทอดมา (the inherited retinal degeneration) [11] และอื่น ๆ ความสัมพันธ์นี้สามารถช่วยในการวินิจฉัย ทำนาย และหาแนวทางในการรักษาโรค อย่างไรก็ตามการศึกษาซีเอ็นวียังคงประสบปัญหาหลายประการ เช่น ความแม่นยำในการตรวจจับซีเอ็นวี ความซับซ้อนของซีเอ็นวีตัวนั้น ๆ และการตีความหมายของซีเอ็นวีที่ตรวจจับได้

การศึกษาซีเอ็นวีสามารถทำได้ทั้งบนเอ็กโซม และจีโนม ถึงแม้ว่าเอ็กโซมจะมีปริมาณข้อมูลน้อยกว่าจีโนมมาก แต่การศึกษาซีเอ็นวีบนเอ็กโซมก็เป็นที่นิยมมากกว่าอันเป็นผลมาจาก (1) โปรเจคส่วนใหญ่ที่เริ่มต้นในการศึกษาสารพันธุกรรมยังเริ่มต้นมาจากการศึกษาเอ็กโซม ทำให้การศึกษาจีโนมดูเป็นเรื่องที่ท้าทายเกินไป [12] และ (2) ค่าใช้จ่าย และเวลาในการให้ได้มาซึ่งจีโนมสูงกว่าของเอ็กโซมมาก [13] (3) ข้อมูลเอ็กโซมเป็นข้อมูลที่ถอดรหัสมาเฉพาะบริเวณที่เป็นยีนซึ่งสามารถแปลรหัสไปเป็นโปรตีน ทำให้สามารถตีความในเชิงฟังก์ชันและผลกระทบได้ง่ายกว่าโดยเฉพาะในเชิงคลินิก ด้วยเหตุผลดังกล่าวจึงทำให้ที่ผ่านมาการตรวจหาความผิดปกติของร่างกายด้วยการวิเคราะห์ซีเอ็นวีบนเอ็กโซมเป็นทางเลือกที่น่าสนใจมากกว่าการศึกษาซีเอ็นวีบนจีโนม

อย่างไรก็ตาม ปัจจุบันยังไม่มีเครื่องตรวจจับซีเอ็นวีบนเอ็กโซมเครื่องมือใดสามารถตรวจจับซีเอ็นวีได้ครอบคลุม และแม่นยำเพียงพอแก่ความต้องการ [14-16] เนื่องจากกระบวนการให้ได้มาซึ่งเอ็กโซมจะมาพร้อมกับค่าอคติ (bias) จากสารเคมีที่ใช้สำหรับคัดเลือกเอ็กซอน (exome sequencing capture kit) และจีซีคอนเทนต์ (GC content) จากการจับคู่ระหว่างคู่เบสกวานีน (guanine) และไซโตซีน (cytosine) บนสายดีเอ็นเอในส่วนที่กำลังพิจารณา ส่งผลให้การวิเคราะห์ซีเอ็นวีเกิดความผิดพลาดมีค่าผลบวกเท็จ (false positive: FP) สูง และแม้ว่ามีการพัฒนาเครื่องมือการตรวจจับซีเอ็นวี (CNV detection tools) อย่างต่อเนื่อง แต่เครื่องตรวจจับซีเอ็นวีแต่ละเครื่องมือมีความจำเพาะต่อซีเอ็นวีลักษณะที่แตกต่างกัน เช่น CONTRA [17] เหมาะกับการหาซีเอ็นวีในบริเวณแคบ [14], CNVnator [18] เหมาะกับการหาซีเอ็นวีทุกขนาด และได้อัตราการทำนายที่ถูกต้องสูง (true positive rate: TPR) แต่ก็ได้ค่าอัตราการทำนายที่เป็นเท็จหลากหลาย (false discovery rate: FDR) [19], CODEX [20] เหมาะกับซีเอ็นวีที่พบเจอได้ยาก (rare CNVs), EXCAVATOR [21] เหมาะกับการตรวจจับซีเอ็นวีที่เกิดในโรคมะเร็งทั้งซีเอ็นวีที่มีขนาดสั้น และยาว รวมถึงใช้ตัวอย่างข้อมูลเอ็กโซมของผู้ป่วยจำนวนไม่มาก ขณะที่ XHMM [22] และ CoNIFER [23] เหมาะกับการตรวจจับซีเอ็นวีบริเวณที่มีซีเอ็นวีอยู่น้อย และใช้ตัวอย่างจำนวนมาก [15] เป็นต้น

นอกจากนี้ เครื่องมือตรวจจับซีเอ็นวียังขาดความสะดวกในการใช้งาน เช่น ความลำบากในการแปลผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวี เครื่องมือส่วนใหญ่มักแสดงผลอยู่ในรูปแบบของไฟล์ข้อความ (plain text format) และขาดคำอธิบายประกอบ (annotation) เช่น CONTRA, CoNIFER และ CNVnator เป็นต้น ทำให้ผู้ใช้ต้องหาความหมายของซีเอ็นวีที่ได้โดยการเทียบกับฐานข้อมูลต่าง ๆ ด้วยตนเอง และแม้ว่าเครื่องมือบางเครื่องมือจะมีการแสดงผลลัพธ์ด้วยภาพ (visualization) และมีคำอธิบายประกอบ แต่ก็ยังมีข้อจำกัดให้ผู้ใช้ต้องเขียนสคริปต์ (script) เพื่อเรียกใช้งาน เช่น GenVisR [24], CNView [25] และ iCopyDAV [26] เป็นต้น นอกจากนี้ ถึงแม้ว่าเครื่องมือบางตัวจะมีการแสดงผลลัพธ์ผ่านการใช้ภาพเป็นตัวปฏิสัมพันธ์กับผู้ใช้ หรือที่เรียกว่า “จียูไอ” (graphical user interface: GUI) แต่ก็ยังจำกัดการใช้งานให้สามารถใช้จียูไอได้ก็ต่อเมื่อใช้อัลกอริทึมของตัวเครื่องมือเอง ทำให้ขาดความยืดหยุ่นในการใช้งาน ผู้ใช้ไม่สามารถนำผลลัพธ์จากเครื่องมืออื่นมาแสดงผลด้วยจียูไอผ่านเครื่องมือเหล่านี้ เช่น DeAnnCNV [27] และ Ginkgo [28] ท้ายสุด ผู้ใช้ไม่สามารถคัดเลือกซีเอ็นวีที่ตรวจจับได้ผ่านทางการใช้คำสำคัญ หรือ คีย์เวิร์ด (keywords) ทำให้ยากต่อการจัดลำดับความสำคัญของซีเอ็นวีที่สนใจ

นอกจากข้อจำกัดที่มักพบข้างต้น ยังพบปัญหาในการติดตั้ง และดูแลรักษาเครื่องมือตรวจจับซีเอ็นวี เนื่องจากเครื่องมือส่วนใหญ่ที่มีในปัจจุบันจะติดตั้งผ่านทางคอมมานไลน์ (command line) ผู้ใช้ต้องดาวน์โหลดไลบรารี (libraries) ที่ใช้ในเครื่องมือ และตรวจสอบความเข้ากันได้ของไลบรารี

เหล่านั้นทำให้ยากต่อการติดตั้ง เช่น GenVisR และ CNView ส่งผลให้ผู้ที่ขาดทักษะด้านการเขียนโปรแกรมไม่สามารถติดตั้งโปรแกรมเหล่านี้ด้วยตนเองได้ บางเครื่องมือต้องใช้งานผ่านทางเว็บไซต์ที่ผู้พัฒนาเตรียมไว้ให้ หากผู้พัฒนาหยุดให้บริการผู้ใช้ก็จะไม่สามารถใช้งานเครื่องมือนั้นต่อไป นอกจากนี้ เครื่องมือส่วนใหญ่ไม่ได้บอกที่มาของคำอธิบายประกอบ (annotation) และบางเครื่องมือไม่สามารถอัพเดตคำอธิบายประกอบได้ ทำให้ผู้ใช้งานเครื่องมือตรวจจับซีเอ็นวีขาดความมั่นใจในความเข้ากันได้ของคำอธิบายประกอบกับข้อมูลที่ลำดับเบสเอ็กโซม (WES) ที่ผู้ใช้เลือกใช้

จากเหตุผลที่กล่าวมาข้างต้น จึงได้เกิดแนวคิดการพัฒนาเครื่องมือตรวจจับซีเอ็นวีในลักษณะบูรณาการ “อินซีเอ็นวี (inCNV)” ซึ่งเป็นเว็บแอปพลิเคชัน (standalone web-based application) ที่ผู้ใช้สามารถเลือกติดตั้งบนเครื่องคอมพิวเตอร์ของตนเอง หรือว่าติดตั้งบนเครื่องเซิร์ฟเวอร์ได้ ตัวโปรแกรมมีความยืดหยุ่นสามารถรวมผลลัพธ์ของเครื่องมือตรวจจับซีเอ็นวีหลายตัวที่มีลักษณะต่างกันเพื่อทำให้เกิดความครอบคลุมในการศึกษาซีเอ็นวีบนเอ็กโซมมากขึ้น หรือสามารถรวมผลลัพธ์ซีเอ็นวีจากกลุ่มตัวอย่างเพื่อหาความพันธ์ของกลุ่มคนกับซีเอ็นวีที่สนใจ ซอฟต์แวร์อินซีเอ็นวีให้ความสะดวกในการใช้งานโดยใช้ภาพเป็นตัวปฏิสัมพันธ์กับผู้ใช้ เชื่อมโยงคำอธิบายประกอบ (annotation) กับฐานข้อมูลสาธารณะ (public databases) ที่เกี่ยวข้องกับซีเอ็นวี และจัดลำดับความสำคัญของซีเอ็นวีได้โดยการฟิลเตอร์ (filter) ซีเอ็นวีที่ตรวจจับได้ผ่านคีย์เวิร์ด และจำนวนเครื่องมือ หรือตัวอย่างที่ตรวจจับซีเอ็นวีไว้ในบริเวณเดียวกัน นอกจากนี้ผู้ใช้ยังสามารถติดตั้งซอฟต์แวร์อินซีเอ็นวีในหน่วยงานของตนเองผ่านดอกเกอร์ได้โดยง่าย โดยวิทยานิพนธ์นี้อยู่บนพื้นฐานแนวคิดการสร้างประโยชน์เพื่องานวิจัยในเชิงพันธุศาสตร์มนุษย์ และเพื่อการวินิจฉัยโรคในระดับคลินิกผ่านการแพทย์จีโนมิกส์ในอนาคต

1.2 วัตถุประสงค์ของงานวิจัย

งานวิจัยนี้ได้ออกแบบ และพัฒนาซอฟต์แวร์โดยมีวัตถุประสงค์ดังต่อไปนี้

1) เพื่อสร้างระบบที่สามารถบูรณาการผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวีหลายเครื่องมือของหนึ่งตัวอย่าง (sample) ให้ตรวจจับซีเอ็นวีได้ครอบคลุม และแม่นยำมากขึ้น หมายเหตุความแม่นยำนี้ขึ้นอยู่กับอัลกอริทึมที่เครื่องมือตรวจจับซีเอ็นวีแต่ละเครื่องมือใช้ด้วย
2) เพื่อสร้างระบบที่สามารถบูรณาการผลลัพธ์ของหลายตัวอย่างจากเครื่องมือตรวจจับซีเอ็นวีหนึ่ง ๆ เพื่อหาความสัมพันธ์ระหว่างตัวอย่างที่สนใจ กลุ่มตัวอย่าง และโรคทางพันธุกรรม
3) เพื่อสร้างเครื่องมือตรวจจับซีเอ็นวีที่ง่ายต่อการใช้งาน สามารถให้คำอธิบายประกอบ คัดกรอง และจัดลำดับความสำคัญของซีเอ็นวีตามความต้องการของผู้ใช้ได้

4) เพื่อสร้างเครื่องมือตรวจจับซีเอ็นวีที่ง่ายต่อการติดตั้ง และดูแลรักษา

1.3 **ขอบเขตการวิจัย**

1) ผู้ใช้สามารถเลือกผลลัพธ์ของเครื่องมือตรวจจับซีเอ็นวีเครื่องมือใด ๆ ก็ได้เพื่อเป็นข้อมูลนำเข้าของซอฟต์แวร์อินซีเอ็นวีที่นำเสนอ โดยมีเงื่อนไขว่า ผลลัพธ์ที่ได้จากเครื่องมือตรวจจับซีเอ็นวีเครื่องใด ๆ จะต้องเป็นไฟล์ข้อความนามสกุล ".txt" ซึ่งมีข้อมูลดังต่อไปนี้ ชื่อตัวอย่าง, ชื่อโครโมโซมที่ตรวจพบซีเอ็นวี, ตำแหน่งเบสเริ่มต้นบนโครโมโซมที่ตรวจพบซีเอ็นวี, ตำแหน่งเบสสิ้นสุดบนโครโมโซมที่ตรวจพบซีเอ็นวี และประเภทของซีเอ็นวี

2) ก่อนการวิเคราะห์หาซีเอ็นวี ผู้ใช้จำเป็นต้องกำหนดค่าตัวแปรต่าง ๆ ในไฟล์ผลลัพธ์ของเครื่องมือตรวจจับซีเอ็นวีแต่ละเครื่องมือเพื่อให้ซอฟต์แวร์อินซีเอ็นวีสามารถเข้าใจไฟล์ผลลัพธ์เหล่านั้นได้ โดยผู้ใช้ต้องกำหนดค่าตัวแปรผ่านทางจียูไอ (graphical user interface: GUI) ที่ระบบเตรียมไว้สำหรับทำการแมปปิ้ง (mapping) คำหลัก (keywords) ว่าคำใด หมายถึง ชื่อตัวอย่าง, ชื่อโครโมโซม, ตำแหน่งเบสเริ่มต้นบนโครโมโซมที่ตรวจพบซีเอ็นวี, ตำแหน่งเบสสิ้นสุดที่ตรวจพบซีเอ็นวี และประเภทของซีเอ็นวี

3) ซอฟต์แวร์อินซีเอ็นวีทำงานผ่านทางเว็บแอปพลิเคชัน โดยรองรับการทำงานผ่านบราวเซอร์อย่างน้อย 3 ตัว ประกอบด้วย โครม (Chrome), ซาฟารี (Safari) และโอเปรา (Opera)

4) ซอฟต์แวร์อินซีเอ็นวีเป็นซอฟต์แวร์แบบครอสแพลตฟอร์ม (cross-platform software) สามารถติดตั้งได้บนทุกระบบปฏิบัติการที่มีด็อกเกอร์เอ็นจิน (docker engine) ติดตั้งอยู่

1.4 **ขั้นตอนการดำเนินงาน**

1) ศึกษาความรู้พื้นฐานทางชีวสารสนเทศ (bioinformatics) อันได้แก่ จีโนมอ้างอิงมนุษย์ (human reference genome), เอ็กโซม, เทคโนโลยีการหาลำดับเบสแบบเอ็นจีเอส (Next Generation Sequencing: NGS) และซีเอ็นวี

2) ศึกษาลักษณะและการทำงานของเครื่องมือตรวจจับซีเอ็นวีจำนวนมาก เพื่อใช้เป็นข้อมูลในการออกแบบซอฟต์แวร์อินซีเอ็นวี เช่น เครื่องมือตรวจจับซีเอ็นวีที่ดีควรแสดงผลลัพธ์อะไรบ้าง และควรนำเสนอในรูปแบบใด เพื่อให้สามารถพัฒนาซอฟต์แวร์อินซีเอ็นวีให้ออกมาในรูปแบบที่เหมาะสม

3) ศึกษาลักษณะผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจจับซีเอ็นวีจำนวนมากเพื่อหาแนวทางการรวมผลลัพธ์ และเพิ่มความน่าเชื่อถือในการตรวจจับซีเอ็นวี

4) ศึกษาคำอธิบายจีโนม (genome annotation) จากฐานข้อมูลสาธารณะหลายแห่ง จากนั้นเลือก และผนวกข้อมูลเหล่านั้นลงในซอฟต์แวร์อินซีเอ็นวีเพื่อระบุความหมาย และความสำคัญให้กับซีเอ็นวีที่ตรวจจับได้

5) ออกแบบจียูไอให้สามารถแสดงผลลัพธ์ของ “การรวมซีเอ็นวีจากเครื่องมือตรวจจับซีเอ็นวีหลายเครื่องมือ” หรือ “การรวมผลลัพธ์ของหลายตัวอย่าง” ให้ง่ายต่อการใช้งาน โดยใช้ความรู้เรื่องเอ็กโซมและการแสดงผลกราฟฟิกของซีเอ็นวีบนจีโนมที่ได้ทำการศึกษามาก่อนหน้านี้

6) ออกแบบวิธีการจัดลำดับความสำคัญของซีเอ็นวีที่ตรวจจับได้

7) ศึกษาแพทเทิร์นการเขียนโปรแกรม เทคโนโลยี และเครื่องมือในการพัฒนาซอฟต์แวร์อินซีเอ็นวี

8) เขียนโปรแกรมพัฒนาซอฟต์แวร์อินซีเอ็นวี

9) สืบค้นเพิ่มเติมข้อมูลตัวอย่างจากแหล่งข้อมูลที่น่าเชื่อถือ เพื่อใช้เป็นข้อมูลนำเข้าในการทดสอบประสิทธิภาพการทำงานของซอฟต์แวร์อินซีเอ็นวี

10) ทดสอบการทำงานของซอฟต์แวร์อินซีเอ็นวีด้วยข้อมูลนำเข้าจากข้อ 9)

11) สรุปผลการวิจัย

12) ตีพิมพ์ผลงานวิจัย

13) สรุปผลและเรียบเรียงวิทยานิพนธ์

1.5 **ประโยชน์ที่คาดว่าจะได้รับ**

1) สร้างเครื่องมือรูปแบบใหม่ที่สามารถรวมผลลัพธ์ของซีเอ็นวีจากเครื่องมือตรวจจับซีเอ็นวีหลายเครื่องมือ หรือสามารถรวมผลลัพธ์ของซีเอ็นจากหลายตัวอย่าง แล้วแสดงผลในรูปแบบกราฟฟิกซึ่งยังไม่เคยมีใครทำมาก่อน

2) ได้ซอฟต์แวร์ที่สามารถเพิ่มความครอบคลุม และแม่นยำในการตรวจจับซีเอ็นวี

3) ได้ซอฟต์แวร์ที่ช่วยจำกัดขอบเขตในการหาซีเอ็นวีที่มีแนมโน้มจะเป็นตำแหน่งใหม่ที่ยังไม่ถูกบันทึกในฐานข้อมูลซีเอ็นวีได้โดยง่าย

4) ได้ซอฟต์แวร์ที่ช่วยจำกัดขอบเขตในการหา *de novo* CNVs ของตัวอย่างที่อยู่ในครอบครัวเดียวกัน [5, 29]

5) ได้ซอฟต์แวร์ที่ช่วยสนับสนุนการตัดสินใจในการวินิจฉัยโรคของคน หรือตัวอย่างที่สนใจ

6) ได้ซอฟต์แวร์ที่ช่วยนักวิจัยทางการแพทย์สามารถคัดเลือกซีเอ็นวีที่น่าสนใจไปทดสอบในห้องปฏิบัติการเพิ่มเติม

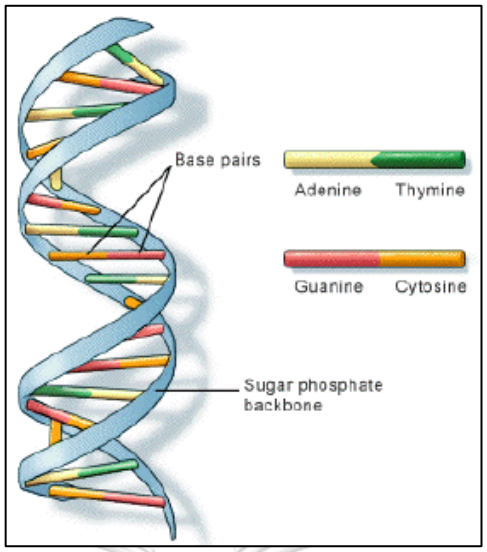
บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

เนื่องจากงานวิจัยนี้เป็นการผลิตซอฟต์แวร์ในรูปแบบเว็บแอปพลิเคชัน เพื่อการวิเคราะห์แบบรวมของการแปรผันจำนวนชุดดีเอ็นเอบนลำดับเบสเอ็กโซม ทฤษฎีที่เกี่ยวข้องจึงได้ถูกแบ่งออกเป็นสองส่วนหลักประกอบด้วย ทฤษฎีที่เกี่ยวข้องทางด้านชีวสารสนเทศ อันได้แก่ จีโนมอ้างอิงของมนุษย์ (human reference genome) เอ็กโซม (exome) เทคโนโลยีการหาลำดับเบสแบบเอ็นจีเอส (next generation sequencing: NGS) ซีเอ็นวี (copy number variants: CNVs) ไฟล์วีซีเอฟ (VCF file) และฐานข้อมูลสาธารณะที่เกี่ยวข้อง เช่น คำอธิบายจีโนมจากบราวเซอร์จีโนมของมหาวิทยาลัยแห่งแคลิฟอเนียร์ซานตาครูส (UCSC Genome Browser) ฐานข้อมูลอองซอมเบลอ (Ensembl) ฐานข้อมูลดีจีวี (DGV) และฐานข้อมูลคลินวาร์ (ClinVar) ซึ่งเป็นข้อมูลหลักที่ระบบต้องการหรือนำมาใช้เป็นส่วนประกอบ สำหรับทฤษฎีที่เกี่ยวข้องในส่วนของการออกแบบและพัฒนาซอฟต์แวร์ประกอบด้วย แองกูลาร์เฟรมเวิร์ค (Angular framework) โหนดเจเอส (NodeJS) โมเดล-วิว-คอนโทรลเลอร์ (Model-View-Controller: MVC) และวัตถุการเข้าถึงข้อมูล (Data access object: DAO)

2.1 ทฤษฎีที่เกี่ยวข้องทางชีวสารสนเทศ

2.1.1 จีโนมอ้างอิงของมนุษย์ (Human reference genome)

จีโนมอ้างอิงของมนุษย์ คือ ข้อมูลรหัสพันธุกรรมทั้งหมดของมนุษย์ ร่างกายมนุษย์มีโครโมโซมร่างกาย 22 คู่ และโครโมโซมเพศ 1 คู่ (โครโมโซม X และโครโมโซม Y) ภายในโครโมโซมจะประกอบด้วย ดีเอ็นเอ (DNA) 2 เส้น ซึ่งจับกันด้วยเบส หรือ นิวคลีโอไทด์ (nucleotide) ดังนี้ อะดีนีน (Adenine: A) จับกับ ไทมีน (Thymine: T) และกวานีน (Guanine: G) จับกับ ไซโตซีน (Cytosine: C) ได้เป็นดีเอ็นเอเกลียวคู่ (double-stranded DNA) ดังรูปที่ 1 โดยลำดับเบสเหล่านี้เป็นตัวกำหนดลักษณะที่แสดงออก หรือปรากฏให้เห็น (phenotype) ของแต่ละคน เช่น สีผิว สีตา การเกิดโรค และความเสี่ยงต่อการเกิดโรค เป็นต้น ดังนั้นการหาลำดับเบสของรหัสพันธุกรรมมนุษย์จึงสามารถช่วยในการวิเคราะห์การแพทย์ได้ โดยสามารถช่วยทำนายแนวโน้มการเกิดโรคในอนาคต ช่วยวินิจฉัยโรคที่เป็นอยู่ และช่วยในการผลิตยารักษาให้ตรงกับโรคได้



รูปที่ 1 โครงสร้างโมเลกุลของดีเอ็นเอ

(ที่มา: รูปที่ 1 ของ[30])

2.1.2 เอ็กโซม (Exome)

ดีเอ็นเอในจีโนมจะถูกแบ่งออกเป็น 2 ส่วน คือ ส่วนที่เป็นยีน และส่วนที่ไม่ใช่ยีน โดยบริเวณที่เป็นยีนนั้นมีหน้าที่รับผิดชอบในการสังเคราะห์โปรตีน ซึ่งมีขั้นตอนดังนี้

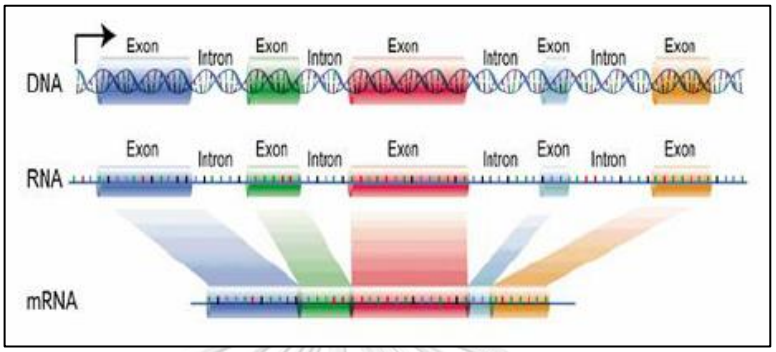
1) การถอดรหัส (transcription) ข้อมูลเข้ารหัสที่อยู่ในยีนจะถูกสำเนาลงในโมเลกุลของเมสเซนเจอร์อาร์เอ็นเอ หรือ เอ็มอาร์เอ็นเอ (messenger RNA: mRNA)

2) การแปลรหัส (translation) แปลรหัสจากเอ็มอาร์เอ็นเอ ไปเป็นสายของกรดอะมิโน หรือที่เรียกว่า สายพอลิเพปไทด์ (polypeptides) ซึ่งเป็นส่วนประกอบของโปรตีน

แต่ละยีนประกอบด้วยเอ็กซอน (exon) และ อินทรอน (intron) ดังรูปที่ 2 ในกระบวนการถอดรหัสของสิ่งมีชีวิตในกลุ่มยูคาริโอต (Eukaryote) อินทรอนจะถูกคัดออก เหลือไว้เพียงบริเวณที่เป็นเอ็กซอนที่จะถูกแปลรหัสไปเป็นโปรตีน ดังนั้น เอ็กซอน จะถูกเรียกว่าเป็นส่วนที่เข้ารหัสโปรตีน (protein-coding) [31-33] ซึ่งส่วนที่เข้ารหัสโปรตีนเหล่านี้เป็นส่วนที่สำคัญมากในการทำความเข้าใจการทำงานในกระบวนการต่าง ๆ ทางชีววิทยา และเราเรียกส่วนของเอ็กซอนทั้งหมดในจีโนมที่ถูกเข้ารหัสพันธุกรรมว่า “เอ็กโซม (exome)”

ในการให้ได้มาซึ่งเอ็กโซมนั้นมาพร้อมกับค่าอคติจำนวนมาก ในปัจจุบันชุดของสารเคมีที่ใช้สำหรับคัดเลือกเอ็กซอนทั้งหมดจากสายดีเอ็นเอ (exome sequencing capture kit) ที่เป็นที่

นิยมมี 3 ชุด คือ Illumina TrueSeq, Agilent SureSelect และ NimbleGen SeqCap EZ [34] แต่เนื่องจากสารเคมีเหล่านี้มีค่าประสิทธิผลโดยเฉลี่ยอยู่ที่ 40 - 70% และสารเคมีแต่ละชุดก็มีความสามารถในคัดเลือกเอ็กซอนทั้งหมดจากสายดีเอ็นเอได้ดีไม่เท่ากัน ดังนั้นจึงก่อให้เกิดค่าอคติในการเรียงลำดับเบสทั้งหมดบนเอ็กโซมจำนวนมาก

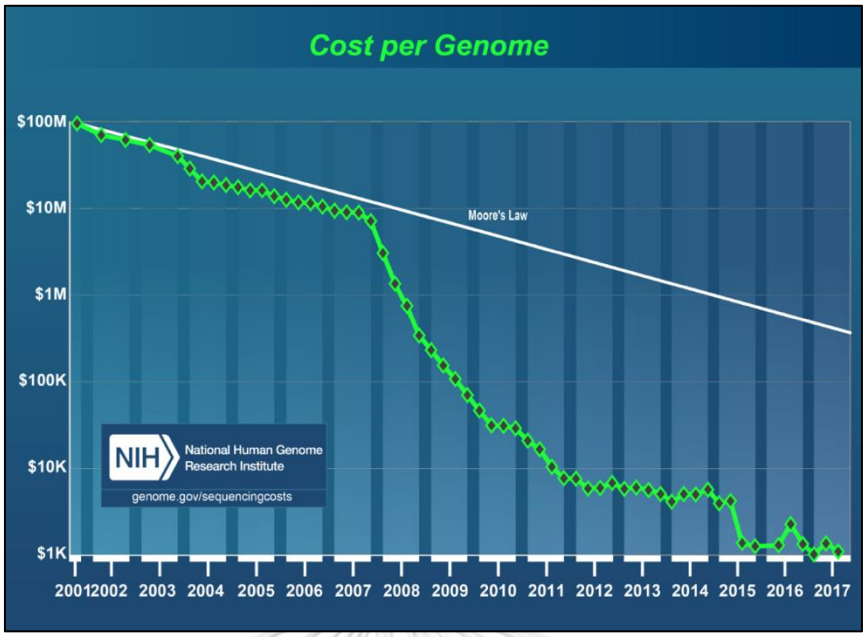


รูปที่ 2 โครงสร้างโมเลกุลของดีเอ็นเอ

(ที่มา: รูปที่ 2 ของ [35])

2.1.3 เทคโนโลยีการหาลำดับเบสแบบเอ็นจีเอส (Next Generation Sequencing: NGS)

โครงการจีโนมของมนุษย์ (human genome project) ได้ใช้วิธีการหาลำดับเบสของเฟรด แซงเกอร์ (Fred Sanger) ที่เรียกว่าการหาลำดับเบสแบบแซงเกอร์ (Sanger Sequencing) [36] หาจีโนมมนุษย์อันแรกได้สำเร็จโดยในปี ค.ศ.2001 ใช้เวลาทั้งหมด 13 ปี ใช้เงินประมาณ 2.7 พันล้านเหรียญสหรัฐ [37-39] และต่อมาในปี ค.ศ.2005 ได้มีการใช้วิธีใหม่ที่เรียกว่าเอ็นจีเอส (Next Generation Sequencing: NGS) ในเชิงพาณิชย์เป็นครั้งแรก ซึ่งวิธีการนี้ได้ช่วยลดต้นทุนในการให้ได้มาซึ่งจีโนมมนุษย์จำนวนมาก โดยในปี ค.ศ.2008 เทคโนโลยีการหาลำดับเบสแบบเอ็นจีเอสสามารถลดค่าใช้จ่ายในการหาจีโนมมนุษย์คนหนึ่งเหลือประมาณ 1.5 ล้านเหรียญสหรัฐ ใช้เวลาเพียง 5 เดือน [12] และในปลายปี ค.ศ. 2015 ค่าใช้จ่ายสำหรับลำดับเบสของจีโนมมนุษย์คนหนึ่งอยู่ที่ประมาณ 1,500 เหรียญสหรัฐ และใช้เวลา 2 – 3 สัปดาห์ ส่วนค่าใช้จ่ายสำหรับลำดับเบสทั้งหมดบนเอ็กโซมมนุษย์คนหนึ่งจะต่ำกว่า 1,000 เหรียญสหรัฐ และใช้เวลา 2 – 3 วัน [40] และค่าใช้จ่ายเหล่านี้ก็มีมีแนวโน้มจะลดลงเรื่อย ๆ และลงมากกว่ากฎของมัวร์ (Moore's law) ดังรูปที่ 3



รูปที่ 3 กราฟค่าใช้จ่ายในการหาจีโนมมนุษย์หนึ่งคน

(ที่มา: รูปที่ 1 ของ [40])

อิลูมินา (Illumina) เป็นค่ายผู้ผลิตเครื่องมือสำหรับการถอดรหัสพันธุกรรมจีโนมมนุษย์ค่ายหนึ่งที่พัฒนาการหาจีโนมแบบเอ็นจีเอสด้วยวิธีเฉพาะของตน คือ การหาลำดับเบสด้วยการต่อเบส และการถ่ายภาพแสงที่แสดงออกของแต่ละเบสที่ต่อเพิ่มเข้าไป (sequencing-by-synthesis: SBS) ซึ่งเป็นวิธีที่ได้รับความนิยมมากจนค่ายอิลูมินาได้มีอิทธิพลครอบงำตลาดด้านการทำจีโนมของมนุษย์เหนือค่ายอื่น ๆ ดังนั้น ในที่นี้จะพูดถึงการทำเอ็นจีเอสสำหรับการหาลำดับเบสของเอ็กโซม และจีโนม ด้วยแพลตฟอร์มอิลูมินา (Illumina platform) เท่านั้น โดยขั้นตอนการทำเอ็นจีเอสด้วยแพลตฟอร์มอิลูมินา มีดังต่อไปนี้

2.13.1 การเตรียมถังข้อมูลลำดับเบส (Library Preparation)

มีขั้นตอนดังต่อไปนี้ (รูปที่ 4)

1) ตัดสายดีเอ็นเอเป็นเส้นเล็ก ๆ เท่า ๆ กัน (fragmentation)

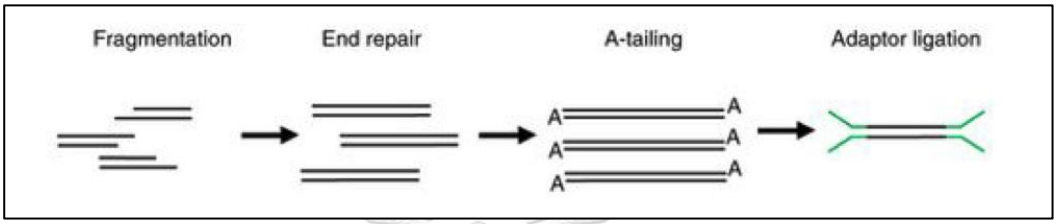
2) ซ่อมปลายสายดีเอ็นเอที่ถูกตัด (end repair)

3) ใส่เบสอะดีนีนไว้ปลายสายดีเอ็นเอ (A-tailing)

4) ผูกสายดีเอ็นเอกับอะแดปเตอร์ (adapter ligation)

5) เพิ่มจำนวน และคุณสมบัติของสายดีเอ็นเอด้วยการทำพีซีอาร์ (polymerase chain reaction enrichment: PCR enrichment)

6) เลือกดีเอ็นเอที่ผูกกับอะแดปเตอร์ที่มีคุณภาพ มีขนาดเท่ากัน และมีจำนวนที่มากพอเพื่อทำขั้นตอนต่อไป



รูปที่ 4 ขั้นตอนการผูกอะแดปเตอร์ไว้ที่ปลายสายดีเอ็นเอ

(ที่มา: รูปที่ 1 ของ [41])

2.1.3.2 การเตรียมทำโฟลว์เซลล์ (Flow-Cell Preparation)

โฟลว์เซลล์ คือ แผ่นกระจกใสที่มีหนึ่ง หรือ หลายช่องก็ได้ หลังจากทำโฟลว์เซลล์แล้วจะได้สายเบสจำนวนมากซึ่งถูกสำเนามาจากต้นฉบับ และถูกเรียงเป็นแนวฉากกับระนาบโฟลว์เซลล์

2.1.3.3 การถอดรหัสโดยการสังเคราะห์และต่อเบส (Sequencing by synthesis: SBS)

เป็นกระบวนการสำเนาสายดีเอ็นเอด้วยเบสที่ใส่สารสารเรืองแสงลงไป แล้วอ่านค่าลำดับเบสจากสารเรืองแสง เช่น สีเขียว หมายถึง อะดีนีน, สีฟ้า หมายถึง ไซโตซีน, สีเหลือง หมายถึง กัวนีน และสีแดง หมายถึง ไธมีน เป็นต้น

การเรียงลำดับเบสมีทั้งแบบเรียงจากด้านเดียว (single-end sequencing) และจากสองด้าน (paired-end sequencing) โดยลำดับเบสทั้งหมดบนสายดีเอ็นเอที่อ่านได้ยาวต่อเนื่องกัน 1 เส้น จะเรียกว่า 1 รีด (read)

2.1.3.4 การบันทึกข้อมูลลงบนไฟล์

เส้นรีดที่อ่านได้จากเครื่องถอดรหัสจะถูกบันทึกไว้ในไฟล์รูปแบบฟาสคิว (FASTQ) รีดเหล่านี้จะถูกนำไปเทียบกับจีโนมอ้างอิง ได้ผลลัพธ์ของการเทียบกับจีโนมอ้างอิง (reference genome) เป็นไฟล์รูปแบบแซม (SAM) แล้วถูกบีบอัดเป็นไฟล์แบม (BAM) โดยยิ่งมีเส้น

รีดจำนวนมาก ข้อมูลก็ยิ่งมาก เมื่อนำมาประมวลผลข้อมูลก็จะได้ข้อมูลที่มีความแม่นยำมากขึ้นตามไปด้วย

2.1.4 ซีเอ็นวี (Copy Number Variants: CNVs)

ซีเอ็นวีเป็นรูปแบบหนึ่งของการแปรผันเชิงโครงสร้าง หรือที่เรียกว่า “เอสวี (structural variation: SV)” ซึ่งเป็นสาเหตุหลักของการเปลี่ยนแปลงแก้ไขสำเนาบนชิ้นส่วนจีโนมจำนวนมากทำให้คนแต่ละคนมีลักษณะที่แตกต่างกันหรือเป็นสาเหตุของโรคบางโรค โดยแต่เดิมบริเวณที่ถือว่าเป็นซีเอ็นวีต้องมีขนาดการแปรผันเชิงโครงสร้างของจีโนมอย่างน้อย 1 กิโลเบส (kilobase: kb) [42, 43] แต่ด้วยเทคโนโลยีที่ก้าวหน้าขึ้นทำให้ปัจจุบันสามารถหาการแปรผันเชิงโครงสร้างนี้ในบริเวณที่แคบลงได้ ทำให้ในทางปฏิบัติซีเอ็นวีมีขนาดการแปรผันเชิงโครงสร้างอย่างน้อย 50 เบส [44]

เนื่องจากลำดับเบสของดีเอ็นเอบนโครโมโซมมีโอกาสปรับเปลี่ยนเป็นปกติ และกระบวนการนี้เองที่ทำให้มนุษย์มีวิวัฒนาการ [1-3] ดังนั้นตำแหน่งที่เกิดซีเอ็นวีขึ้นจึงมีทั้งตำแหน่งที่ปรับเปลี่ยนโดยปกติ ซึ่งไม่ได้มีปัจจัยก่อให้เกิดโรค และแบบที่เป็นปัจจัยต่อการเกิดโรค [4] ตัวอย่างโรคที่มีความสัมพันธ์กับซีเอ็นวี เช่น โรคออทิสซึม (autism) [45, 46] โรคจิตเภท (schizophrenia) [47] โรคโครห์น (Crohn’s disease) [48, 49] โรคข้อต่อรูมาตอยด์ (Rheumatoid arthritis) [48] โรคเบาหวานชนิดที่ 1 (type 1 diabetes) [48] โรคอ้วน (obesity) [50] และโรคอื่น ๆ ที่สามารถพบได้อีกจำนวนมาก [51-54]

2.1.4.1 ประเภทของซีเอ็นวี

ซีเอ็นวีมีหลายประเภท และสามารถถูกจัดประเภทโดยเกณฑ์การแบ่งดังนี้ คือ แบ่งตามอัตราการพบ และแบ่งตามรูปร่างของซีเอ็นวี

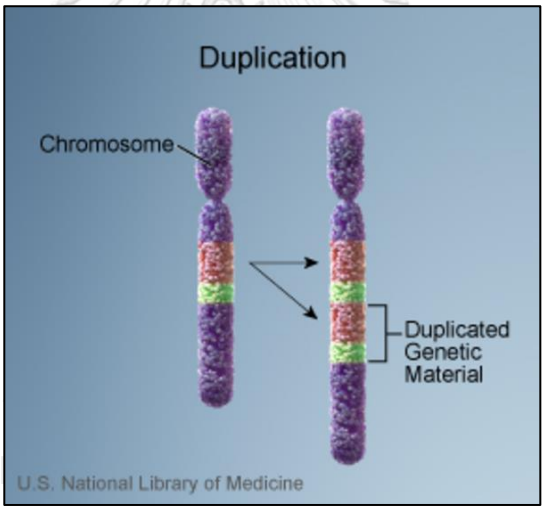
2.1.4.1.1 แบ่งตามอัตราการพบ

ซีเอ็นวีแบ่งตามอัตราการพบได้เป็น 2 ประเภท คือ ซีเอ็นวีที่พบได้โดยทั่วไป (common CNVs) และซีเอ็นวีที่พบได้ยาก (rare CNVs) [55] โดยซีเอ็นวีที่พบได้โดยทั่วไปจะถูกพบมากกว่า 1% ของประชากร มักมีขนาดสั้น มีความยาวน้อยกว่า 10 กิโลเบส และมักเกี่ยวข้องกับโปรตีนที่มีความสำคัญกับยา และระบบภูมิคุ้มกัน เช่น โรคสะเก็ดเงิน (Psoriasis) [56] โรคโครห์น (Crohn’s disease) [57] และโรคไตอักเสบ (glomerulonephritis) [58] เป็นต้น ในขณะที่ซีเอ็นวีที่พบได้ยากจะถูกพบน้อยกว่า 1% ของประชากร มักมีขนาดยาวกว่าซีเอ็นวีที่พบได้โดยทั่วไปมาก มีความยาวตั้งแต่หลายแสนถึงหนึ่งล้านเบส ซีเอ็นวีประเภทนี้มักเกิดในครอบครัวโดยการปฏิสนธิระหว่างไข่ และสเปิร์ม หรือเกิดจากการส่งผ่านเพียงไม่กี่รุ่นในครอบครัว เกี่ยวข้องกับโรค

ปัญญาอ่อน (mental retardation) [54] โรคพัฒนาการล่าช้า (developmental delay) [59] โรคจิตเภท (schizophrenia) [60] และโรคออทิสซึม (autism) [61] เป็นต้น โดยมีการคาดการณ์ไว้ว่าซีเอ็นวีที่พบได้ยากนี้จะเกี่ยวข้องกับโรคความผิดปกติทางระบบประสาท(neurocognitive diseases) มากกว่ารูปแบบความแปรผันทางพันธุกรรมแบบอื่น

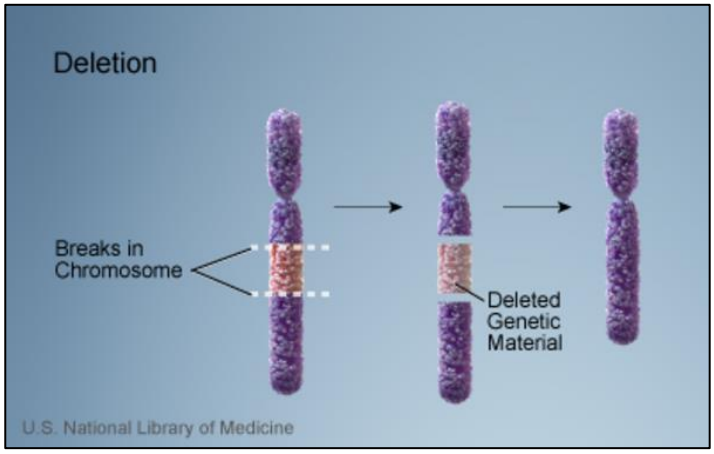
2.4.1.1.2 แบ่งตามรูปร่างของซีเอ็นวี

ซีเอ็นวีแบ่งตามการจัดเรียงลำดับเบสได้เป็น 2 ประเภทหลัก คือ duplication (รูปที่ 5) และ deletion (รูปที่ 6) ซึ่งโดยเฉลี่ยแล้วคนแต่ละคนจะมีซีเอ็นวีเฉลี่ย 1,000 การแปรผัน ต่อ 1 ล้านคู่เบส [62] คิดเป็น 12% - 16% ของจีโนมมนุษย์ [44, 63] โดยซีเอ็นวีประเภท duplication คือ ซีเอ็นวีที่มีรูปแบบการแปรผันเชิงโครงสร้างบนจีโนมที่ไม่สมมาตร มีการเพิ่มขึ้นของชิ้นส่วนจีโนมบางส่วน และ ซีเอ็นวีประเภท deletion คือ ซีเอ็นวีที่มีรูปแบบการแปรผันเชิงโครงสร้างบนจีโนมที่ไม่สมมาตร มีการหายไปของชิ้นส่วนจีโนมบางส่วน



รูปที่ 5 การแปรผันแบบ duplication บนโครโมโซม

(ที่มา: [64])



รูปที่ 6 การแปรผันแบบ deletion บนโครโมโซม

(ที่มา: [65])

2.1.4.2 กระบวนการตรวจจับซีเอ็นวี

กระบวนการตรวจจับซีเอ็นวีซึ่งเป็นที่นิยมมากในตอนนี้คือ วิธีการนับรีด (read count) [66] หรือที่เรียกว่า การหาความลึกของรีด (read depth) เป็นกระบวนการนับจำนวนเส้นรีดในบริเวณใด ๆ ของลำดับเบสบนโครโมโซม ซึ่งมีขั้นตอนดังนี้

2.1.4.2.1 การทำแมปปิ้ง หรือการเตรียมข้อมูล

1) กำจัดรีดที่มีคุณสมบัติไม่ดีออกไป เช่น ซ้ำ หรือ มีคะแนนคุณภาพของเบสต่ำ[1] เป็นต้น

2) แมปปิ้งโดยเอารีดไปเทียบกับจีโนมอ้างอิง (reference genome) ดังรูปที่ 7 - 8 ซึ่งเป็นภาพที่ได้จากโปรแกรม NGB[2] โดยที่เส้นสีเทาในแนวนอน 1 เส้น คือ 1 รีด และจำนวนรีดที่ตกลงอยู่ในบริเวณหนึ่ง คือ ความลึกของรีดในบริเวณหนึ่ง (coverage)

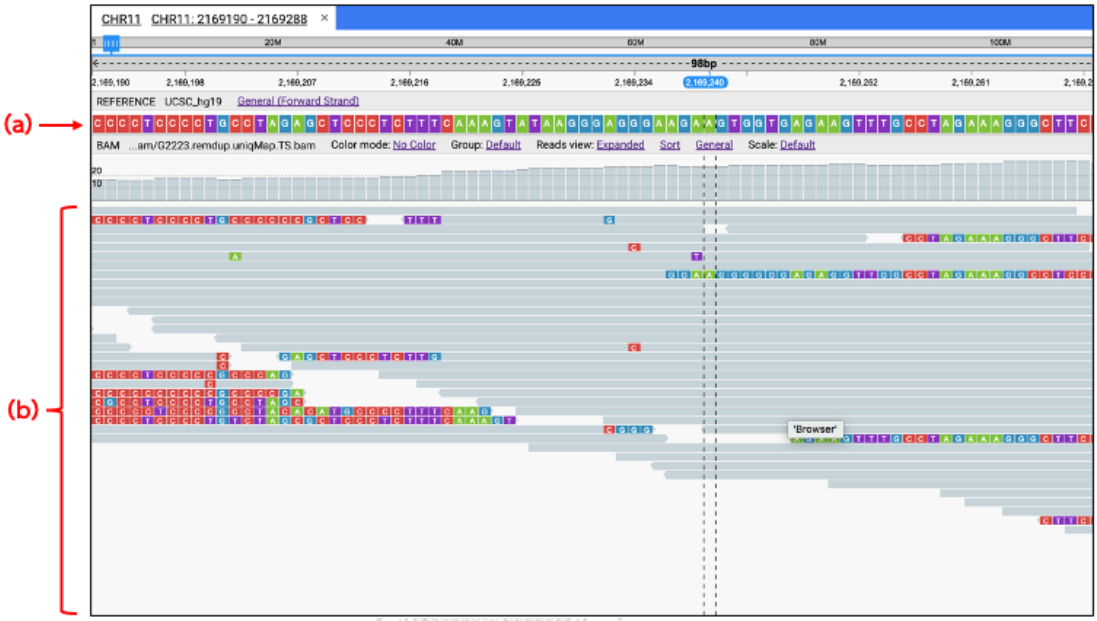
2.1.4.2.2 นับรีดที่แมปได้ในแต่ละบิน

บิน (bin) คือหน้าต่างหรือขอบเขตบริเวณในจีโนม ที่ใช้คำนวณหาความลึกของรีด หากขนาดของบินเล็กเกินไปก็จะแมปกับเส้นรีดได้น้อยเกินไป เกิดเป็นผลลบเท็จ (false negative: FN) และหากขนาดของบินใหญ่เกินไปก็จะแมปกับเส้นรีดได้จำนวนมากเกินไปทำ

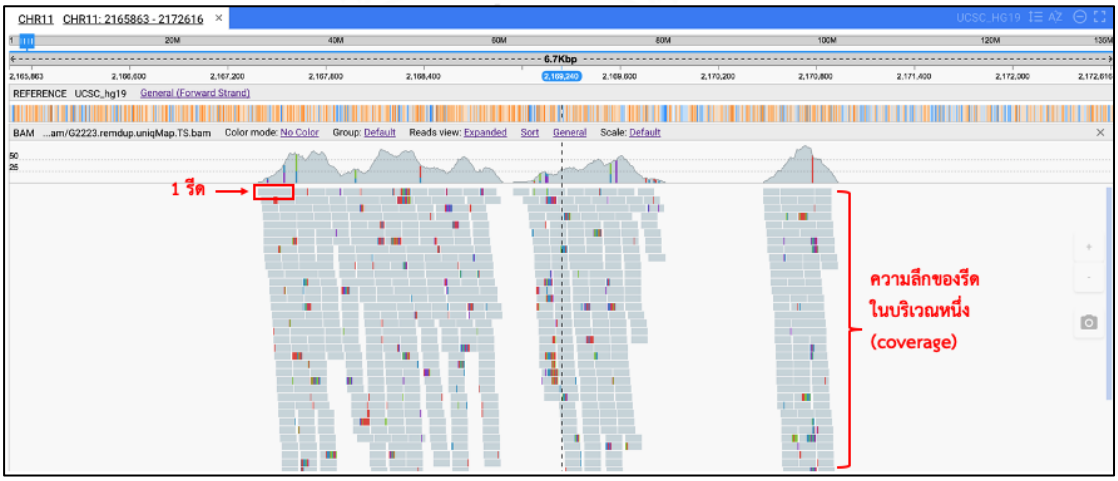
[1] รีดที่ไม่ดี มักหมายถึงรีดที่มีคะแนนคุณภาพของเบสจากอัลกอริทึม Phred score ต่ำกว่า 20 คะแนน และรีดที่มีคุณภาพดีมักหมายถึงรีดที่มีเบสซึ่งมีคะแนนคุณภาพสูงกว่า 30 คะแนน (ในทางปฏิบัติเครื่องอะลูมินามักให้คะแนนเบสอยู่ที่ 0 – 60 คะแนน)

[2]New Genome Browser (NGB) is a Web client-server tool, available at https://github.com/epam/NGB

ให้ขาดความถูกต้องเกิดเป็นผลบวกเท็จ (false positive: FP) ดังนั้นขนาดของบินที่เหมาะสมสำหรับลำดับเบสในแต่ละตัวอย่างจะแตกต่างกันไปตามค่าความครอบคลุมของรีดที่ได้ตอนทำรีดจากกระบวนการหาลำดับเบสซึ่งมักมีค่าอยู่ที่ 50-1000 คู่เบส [34]



รูปที่ 7 ตัวอย่างผลการแมปปิ้งของรีดจำนวนมากกับบริเวณในจีโนมอ้างอิง แสดงผลด้วยโปรแกรม NGB แบบซูมออก



รูปที่ 8 ตัวอย่างผลการแมปปิ้งของรีดจำนวนมากกับบริเวณในจีโนมอ้างอิง แสดงผลด้วยโปรแกรม NGB แบบซูมเข้า โดย (a) แสดงจีโนมอ้างอิง และ (b) แสดงข้อมูลจากไฟล์แบม

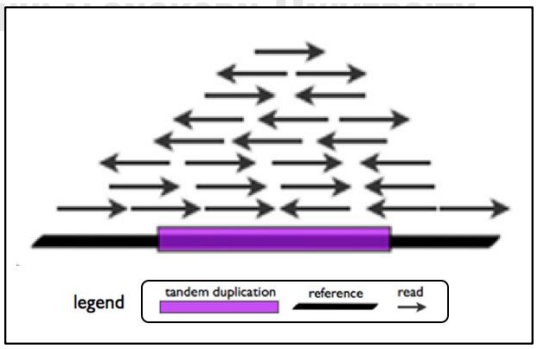
2.1.4.2.3 การทำนอร์มัลไลเซชัน (Normalization)

การทำนอร์มัลไลเซชัน คือ การนำค่ารีดที่ได้มาทำให้เป็นค่ามาตรฐาน เนื่องจากการได้มาซึ่งลำดับเบสจีโนม และบนเอ็กโซมมาพร้อมกับค่าอคติจำนวนมาก ได้แก่ ค่าอคติที่เกิดจากค่าจีซีคอนเทนต์ (guanine-cytosine content: GC content)[3] และค่าอคติที่เกิดจากค่าความสามารถในการแมปปิ้งรีดไปยังจีโนมอ้างอิง (genomic mappability) [67] จึงจำเป็นต้องมีการจัดการค่ารีดเหล่านั้นให้อยู่ในรูปแบบมาตรฐานก่อนการคำนวณหาตำแหน่งที่เกิดซีเอ็นวี

นอกจากนี้สำหรับเอ็กโซมก็ยังมีค่าอคติจากสารเคมีที่ใช้สำหรับคัดเลือกเอ็กซอนทั้งหมดจากสายดีเอ็นเอ (exome sequencing capture kit) อีกด้วย จึงทำให้ค่าอคติบนเอ็กโซมสูงกว่าค่าอคติบนจีโนมมาก และยังไม่มีเครื่องมือจับซีเอ็นวีตัวใดทำนอร์มอไลเซชันซีเอ็นวีบนเอ็กโซมได้สมบูรณ์ ดังนั้น การทำนอร์มอไลเซชันจึงถือเป็นส่วนที่ท้าทายสำหรับการพัฒนาเครื่องมือหาซีเอ็นวีบนเอ็กโซม

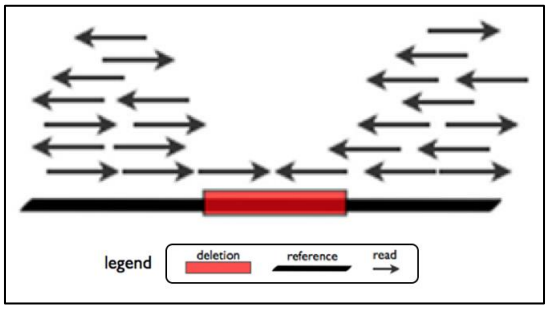
2.1.4.2.4 การประมาณจำนวนสำเนา (Estimation of copy number)

การประมาณผลลัพธ์ในเส้นจีโนมเพื่อพิจารณาว่าบริเวณใดเป็นซีเอ็นวีและเป็นซีเอ็นวีประเภทใด (duplication หรือ deletion) โดยพิจารณาจากจำนวนสำเนาในแต่ละบิน (bin) และอาศัยสมมติฐานที่ว่าลำดับเบสบนโครโมโซมบริเวณใดที่มีค่าความลึกของรีดมากก็น่าจะเป็นซีเอ็นวีประเภท duplication (รูปที่ 9) และเป็น deletion ในทางตรงกันข้าม (รูปที่ 10) ซึ่งมีเครื่องมือตรวจจับซีเอ็นวีจำนวนมากใช้การกระจายตัวแบบพัวซ์ซอง (Poisson distribution) เข้ามาหาความเป็นไปได้ของจำนวนรีดที่เกิดขึ้นในบริเวณที่นับจำนวนเส้นรีดได้ยาก



รูปที่ 9 การหาซีเอ็นวีแบบ duplication ด้วยความลึกของรีด
(ที่มา: รูปที่ 1 (A) ของ [68])

[3] จำนวนคู่ของเบสกวานีน (Guanine: G) ที่อยู่ติดกันกับ ไซโตซีน (Cytosine: C)



รูปที่ 10 การหาซีเอ็นวีแบบ deletion ด้วยความลึกของรีด
(ที่มา: รูปที่ 1 (D) ของ [68])

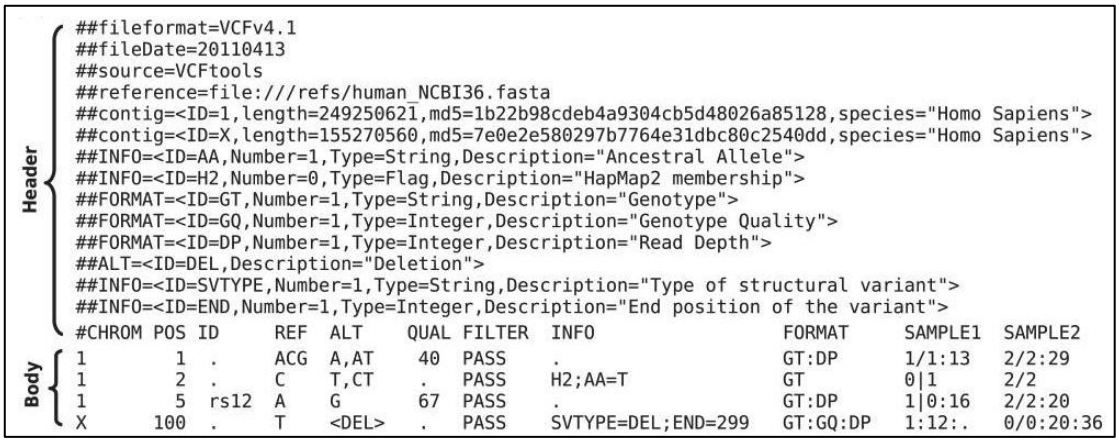
2.1.4.2.6 การแบ่งส่วน (Segmentation)

หลังจากที่ตรวจพบตำแหน่งที่เกิดซีเอ็นวีแล้วสิ่งต้องที่พิจารณาต่อไปก็คือ การแบ่งส่วน ซึ่งเป็นการหาขนาดความยาวโดยประมาณของซีเอ็นวีแต่ละตำแหน่งโดยพิจารณาว่าซีเอ็นวีที่ตรวจจับได้มีความยาวกี่คู่เบส โดยพื้นฐานแล้วจะใช้หลักการความน่าจะเป็นพิจารณาบินที่อยู่ขนาบข้างตำแหน่งที่ตรวจพบว่าเป็นซีเอ็นวีว่ามีความต่อเนื่องของเส้นรีดเป็นความยาวเท่าไหร่

2.1.5 ไฟล์วีซีเอฟ (VCF File)

ไฟล์วีซีเอฟเป็นรูปแบบไฟล์ที่ใช้สำหรับบันทึกข้อมูลการแปรผันเชิงโครงสร้างของจีโนม มีลักษณะเป็นไฟล์ข้อความ และมักอยู่ในรูปแบบบีบอัด โดยจะบอกตำแหน่ง ช่วงการแปรผัน และรูปแบบของการแปรผัน ตัวไฟล์จะแบ่งออกเป็น 2 ส่วนคือ ส่วนหัวข้อมูล (header) และส่วนข้อมูล (data) ดังรูปที่ 11

ส่วนหัวข้อมูลจะประกอบด้วย ส่วนเมตะดาต้า (metadata) ซึ่งเป็นคำอธิบายข้อมูลต่าง ๆ ในไฟล์ ขึ้นต้นด้วยเครื่องหมายแฮช (hash) 2 อัน และมีลักษณะเป็น key=value และส่วนชื่อคอลัมน์ซึ่งขึ้นต้นด้วยเครื่องหมายแอช 1 อัน ในขณะที่ส่วนข้อมูลจะแสดงข้อมูลของการแปรผันในรูปแบบที่ระบุไว้ในส่วนหัวของข้อมูล โดยไฟล์วีซีเอฟมีข้อมูลบังคับ 8 คอลัมน์ซึ่งแสดงรายละเอียดดังตารางที่ 1 และสามารถมีข้อมูลอื่น ๆ นอกเหนือจากนี้ได้ และข้อมูลที่นิยมเพิ่มเติมคือ ชื่อตัวอย่าง (sample) ที่ตรวจพบการแปรผันทางพันธุกรรม



รูปที่ 11 ตัวอย่างไฟล์วีซีเอฟ

(ที่มา: รูปที่ 1 (A) ของ [69])

ตารางที่ 1 ฟิลด์หลักในไฟล์วีซีเอฟ

คอลัมน์	ฟิลด์	ประเภทของข้อมูล	คำอธิบาย
1	CHROM	String	ชื่อของโครโมโซม
2	POS	Int	ตำแหน่งของเบสแรก
3	ID	String	ระบุชื่อของการแปรผันนี้
4	REF	String	ระบุเบสบนจีโนมอ้างอิง
5	ALT	String	ระบุเบสที่มีการแปรผันไป
6	QUAL	Int	ระบุคุณภาพของเบสด้วย Phred score
7	FILTER	String	ระบุการผ่านการกรองผลลัพธ์
8	INFO	String	ระบุข้อมูลเพิ่มเติม

2.1.6 บราวเซอร์จีโนมของมหาวิทยาลัยแห่งแคลิฟอเนียร์ซานตาครูส (UCSC Genome Browser)

บราวเซอร์จีโนมของมหาวิทยาลัยแห่งแคลิฟอเนียร์ซานตาครูส เป็นเว็บเซิร์ฟเวอร์ที่รวบรวมข้อมูลจีโนมของสิ่งมีชีวิตหลากหลายสายพันธุ์ เช่น มนุษย์ หนู สัตว์เลี้ยงลูกด้วยนมอื่น ๆ และแมลง เป็นต้น โดยเว็บเซิร์ฟเวอร์นี้เป็นหนึ่งในงานวิจัยของมหาวิทยาลัยแห่งแคลิฟอเนียร์ ซานตาครูส (University of California, Santa Cruz: UCSC) ร่วมมือในเชิงวิจัยกับมหาวิทยาลัย และหน่วยงานต่าง ๆ เช่น สำหรับการรวบรวมข้อมูลจีโนมมนุษย์จะเป็นการร่วมมือกับโครงการจีโนมมนุษย์ในระดับนานาชาติ (International Human Genome Project) ซึ่งได้รับทุนจากสถาบันวิจัยจีโนมมนุษย์แห่งชาติ (National Human Genome Research Institute) ในส่วนวิธีการใช้ข้อมูลจากเว็บ

เซิร์ฟเวอร์จีโนมนี้ ผู้ใช้สามารถทำโดยสืบค้นข้อมูลจีโนมที่ต้องการผ่านทางเว็บบราวเซอร์ และสามารถดาวน์โหลดไฟล์ข้อมูลที่ต้องการ

2.1.7 ฐานข้อมูลอองซอมเบลอ (Ensembl)

ฐานข้อมูลอองซอมเบลอ (Ensembl) คือ ฐานข้อมูลสาธารณะที่รวบรวมคำอธิบายชื่อยีนจำนวนมากบนจีโนมอ้างอิง ซึ่งรายละเอียดของคำอธิบายชื่อยีนนี้ประกอบด้วย ตำแหน่งยีนบนโครโมโซม ชื่อยีน และลักษณะเฉพาะของยีนนั้น ๆ ที่เกี่ยวข้องกับลักษณะที่แสดงออก หรือปรากฏให้เห็น (phenotype) เช่น สีผิว สีตา และการเกิดโรค [70] รูปที่ 12 แสดงตัวอย่างข้อมูลจากฐานข้อมูลอองซอมเบลอซึ่งระบุข้อมูลในบรรทัดที่ 30 ว่า โครโมโซมที่ 10 ตำแหน่งเบสที่ 4,6892 – 7,4163 มียีนที่ชื่อ TUBB8 และระบุข้อมูลในบรรทัดที่ 33 ว่าที่โครโมโซม และยีนตัวเดียวกันนี้ ตำแหน่งเบสที่ 4,6892 – 4,8114 มีเอ็กซอนชื่อ ENSE00002607146

	A	B	C	D	E	F	G	H	I	J	K	L
1	##gff-version 3											
2	##sequence-region 10 1 133797422											
3	#!genome-build Ensembl GRCh38.p12											
4	#!genome-version GRCh38											
5	#!genome-date 2013-12											
6	#!genome-build-accession NCBI:GCA_000001405.27											
7	#!genebuild-last-updated 2018-01											
8	10	Ensembl	chromosome	1	133797422	.	.	.	ID=chromosome:10;Alias=CM000672.2,chr10,NC_000010.11			
9	###											
10	10	.	biological_region	10494	12018	690	.	.	external_name=oe %3D 0.65;logic_name=cpg			
11	10	.	biological_region	11624	11698	1	-	.	external_name=rank %3D 1;logic_name=firstef			
12	10	.	biological_region	11774	12029	1	+	.	external_name=rank %3D 1;logic_name=firstef			
13	10	havana	ncRNA_gene	14061	16544	.	-	.	ID=gene:ENSG00000260370;Name=AC215217.1;biotype=lincRNA;			
14	10	havana	lnc_RNA	14061	14604	.	-	.	ID=transcript:ENST00000562162;Parent=gene:ENSG00000260370			
15	10	havana	exon	14061	14299	.	-	.	Parent=transcript:ENST00000562162;Name=ENSE00002584618;c			
16	10	havana	exon	14497	14604	.	-	.	Parent=transcript:ENST00000562162;Name=ENSE00002606019;c			
17	10	havana	lnc_RNA	14138	16544	.	-	.	ID=transcript:ENST00000566940;Parent=gene:ENSG00000260370			
18	10	havana	exon	14138	14299	.	-	.	Parent=transcript:ENST00000566940;Name=ENSE00002615256;c			
19	10	havana	exon	16502	16544	.	-	.	Parent=transcript:ENST00000566940;Name=ENSE00002578035;c			
20	###											
21	10	havana	pseudogene	44712	46884	.	+	.	ID=gene:ENSG00000237297;Name=AL713922.1;biotype=unproces			
22	10	havana	pseudogenic_trans	44712	46884	.	+	.	ID=transcript:ENST00000416477;Parent=gene:ENSG00000237297			
23	10	havana	exon	44712	44901	.	+	.	Parent=transcript:ENST00000416477;Name=ENSE00001738331;c			
24	10	havana	exon	44952	45204	.	+	.	Parent=transcript:ENST00000416477;Name=ENSE00001680617;c			
25	10	havana	exon	45309	45405	.	+	.	Parent=transcript:ENST00000416477;Name=ENSE00001782515;c			
26	10	havana	exon	45834	45882	.	+	.	Parent=transcript:ENST00000416477;Name=ENSE00001802091;c			
27	10	havana	exon	46245	46359	.	+	.	Parent=transcript:ENST00000416477;Name=ENSE00001630430;c			
28	10	havana	exon	46842	46884	.	+	.	Parent=transcript:ENST00000416477;Name=ENSE00001620962;c			
29	###											
30	10	ensembl_havana	gene	46892	74163	.	-	.	ID=gene:ENSG00000261456;Name=TUBB8;biotype=protein_codin			
31	10	havana	mRNA	46892	74163	.	-	.	ID=transcript:ENST00000564130;Parent=gene:ENSG00000261456			
32	10	havana	three_prime_UTR	46892	47056	.	-	.	Parent=transcript:ENST00000564130			
33	10	havana	exon	46892	48114	.	-	.	Parent=transcript:ENST00000564130;Name=ENSE00002607146;c			
34	10	havana	CDS	47057	48114	.	-	2	ID=CDS:ENSP00000457610;Parent=transcript:ENST00000564130;p			

รูปที่ 12 ตัวอย่างข้อมูลคำอธิบายจีโนมจากฐานข้อมูลอองซอมเบลอ

2.1.8 ฐานข้อมูลดีจีวี (Database of Genomic Variants: DGV)

ฐานข้อมูลดีจีวี คือ ฐานข้อมูลสาธารณะที่เก็บรวบการแปรผันทางพันธุกรรมของคนปกติ พร้อมคำอธิบาย ซึ่งเป็นคำอธิบายที่เกิดจากการศึกษาหาความแปรผันเชิงโครงสร้างของสารพันธุกรรมในบุคคลที่มีร่างกายปกติ (control individuals) และไม่มีอาการเจ็บป่วยของโรคมาเก็บรวบรวมไว้เพื่อใช้ในการศึกษาพันธุศาสตร์ [71] รูปที่ 13 กลุ่มของข้อความที่เน้นด้วยสีส้มแสดงถึงตัวอย่างข้อมูลที่ซอฟต์แวร์อินซีเอ็นวีเลือกใช้จากฐานข้อมูลดีจีวี ซึ่งได้แก่ ชื่อโครโมโซม (chr) ตำแหน่งเบสเริ่มต้นบนโครโมโซม (start) ตำแหน่งเบสสิ้นสุดบนโครโมโซม (end) และประเภทของซีเอ็นวี (type)

chr	start	end	state	id	type	num_variant	num_sample	num_sample
chr1	10001	177368	CNV	CNVR1	Gain	54	50	4
chr1	317770	471368	CNV	CNVR2	Gain	26	22	1
chr1	521413	1708649	CNV	CNVR3	Loss	1409	821	402
chr1	1866297	1867172	CNV	CNVR4	Loss	2	2	0
chr1	1912935	1913930	CNV	CNVR5	Loss	4	4	0
chr1	2024490	2027497	CNV	CNVR6	Loss	43	41	2
chr1	2037656	2038273	CNV	CNVR7	Loss	3	2	1
chr1	2052961	2056112	CNV	CNVR8	Gain	30	30	0
chr1	2073716	2074039	CNV	CNVR9	Loss	2	2	0
chr1	2566002	2634182	CNV	CNVR10	Loss	150	147	3
chr1	2684255	2695595	CNV	CNVR11	Gain	12	12	0
chr1	2876448	2876895	CNV	CNVR12	Loss	2	2	0
chr1	2911489	2911934	CNV	CNVR13	Loss	383	383	0
chr1	3209858	3212147	CNV	CNVR14	Loss	8	7	1
chr1	3215213	3217688	CNV	CNVR15	Loss	13	13	0
chr1	3716941	3717153	CNV	CNVR16	Loss	88	82	6
chr1	4123545	4127968	CNV	CNVR17	Loss	172	163	9
chr1	4154970	4155665	CNV	CNVR18	Loss	500	499	1
chr1	4284551	4286005	CNV	CNVR19	Loss	5	4	1

รูปที่ 13 ตัวอย่างข้อมูลคำอธิบายจีโนมที่ได้จากฐานข้อมูลดีจีวี

2.1.9 ฐานข้อมูลคลินวาร์ (ClinVar)

ฐานข้อมูลคลินวาร์ (ClinVar) เป็นฐานข้อมูลสาธารณะที่รายงานข้อมูลการแปรผันของจีโนมมนุษย์ และลักษณะที่แสดงออก หรือปรากฏให้เห็น (phenotype) โดยมีหลักฐานสนับสนุน [72] รูปที่ 14 แสดงตัวอย่างข้อมูลที่ซอฟต์แวร์อินซีเอ็นวีนำมาจากฐานข้อมูลคลินวาร์ ประกอบด้วย ชื่อโครโมโซม (chromosome) ตำแหน่งเบสเริ่มต้นบนโครโมโซม (start) ตำแหน่งเบสสิ้นสุดบนโครโมโซม (stop) ชื่อของยีนตามฐานข้อมูลโอมิม (OMIM) ลิสต์ชื่อของโรค หรือ ลักษณะฟีโนไทป์ตามฐานข้อมูลคลินวาร์ และเวอร์ชันตามจีโนมอ้างอิง

Type	PhenotypeIDS	PhenotypeList	Assembly	Chromosome	Start	Stop
indel	MedGen:C3150901,OMIM:613647	Spastic paraplegia 48, autosomal	GRCh37	7	4820844	4820847
indel	MedGen:C3150901,OMIM:613647	Spastic paraplegia 48, autosomal	GRCh38	7	4781213	4781216
deletion	MedGen:C3150901,OMIM:613647	Spastic paraplegia 48, autosomal	GRCh37	7	4827366	4827379
deletion	MedGen:C3150901,OMIM:613647	Spastic paraplegia 48, autosomal	GRCh38	7	4787735	4787748
single nucleotide	MedGen:C4551772,OMIM:251300	Galloway-Mowat syndrome 1	GRCh37	15	85342440	85342440
single nucleotide	MedGen:C4551772,OMIM:251300	Galloway-Mowat syndrome 1	GRCh38	15	84799209	84799209
single nucleotide	MedGen:C4748791,OMIM:618241;M	MITOCHONDRIAL COMPLEX I DE	GRCh37	11	126145284	126145284
single nucleotide	MedGen:C4748791,OMIM:618241;M	MITOCHONDRIAL COMPLEX I DE	GRCh38	11	126275389	126275389
single nucleotide	MedGen:C4748791,OMIM:618241	MITOCHONDRIAL COMPLEX I DE	GRCh37	11	126147412	126147412
single nucleotide	MedGen:C4748791,OMIM:618241	MITOCHONDRIAL COMPLEX I DE	GRCh38	11	126277517	126277517
single nucleotide	MeSH:D030342,MedGen:C0950123;N	Inborn genetic diseases;MITOCH	GRCh37	14	32031331	32031331
single nucleotide	MeSH:D030342,MedGen:C0950123;N	Inborn genetic diseases;MITOCH	GRCh38	14	31562125	31562125
deletion	MedGen:C4748792,OMIM:618242	MITOCHONDRIAL COMPLEX I DE	NCBI36	14	30932976	31194846
deletion	MedGen:C4748792,OMIM:618242	MITOCHONDRIAL COMPLEX I DE	GRCh37	14	31863225	32125095
deletion	MedGen:C4748792,OMIM:618242	MITOCHONDRIAL COMPLEX I DE	GRCh38	14	31394019	31655889
single nucleotide	MedGen:C3150874,OMIM:613610	Cranioectodermal dysplasia 2	GRCh37	2	20189045	20189045
single nucleotide	na;Human Phenotype Ontology:HP:0(	Alzheimer disease, susceptibility	GRCh37	6	26093141	26093141
single nucleotide	na;Human Phenotype Ontology:HP:0(	Alzheimer disease, susceptibility	GRCh38	6	26092913	26092913

รูปที่ 14 ตัวอย่างข้อมูลคำอธิบายจีโนมที่ได้จากฐานข้อมูลคลินวาร์

2.2 ทฤษฎีที่เกี่ยวข้องทางซอฟต์แวร์

2.2.1 แองกูลาร์เฟรมเวิร์ค (Angular framework)

แองกูลาร์ คือ เฟรมเวิร์คส่วนหน้า (frontend framework) ที่พัฒนาโดยกูเกิล (Google) สำหรับพัฒนาเว็บแอปพลิเคชันในฝั่งของไคลแอนต์ (client) แบบ single-page applications (SPA) เขียนด้วยภาษาไทป์สคริป (typescript) ตัวอย่างแอปพลิเคชันที่พัฒนาด้วยแองกูลาร์ เช่น Gmail, Youtube TV, Microsoft Office Online และ Xbox Live เป็นต้น

2.2.1.1 ข้อดี

แองกูลาร์มีข้อดีหลายประการดังต่อไปนี้

1. มีการพัฒนาต่อเนื่องโดยทีมงานของกูเกิลซึ่งเป็นบริษัทซอฟต์แวร์ค่ายใหญ่ของโลก จึงมีแนวโน้มว่าแองกูลาร์จะมีการพัฒนาต่อไปในอนาคต ส่งผลให้โปรแกรมที่พัฒนาด้วยแองกูลาร์มีแนวโน้มที่จะได้รับการสนับสนุนในระยะยาว

2. มีโครงสร้างการเขียนโปรแกรมที่ดี โดยมีการกำหนดให้ผู้พัฒนาแบ่งโปรแกรมเขียนเป็นคอมโพเน้นต์ (component) ย่อย ๆ และใช้แพทเทิร์น dependency injection (DI) ทำให้เขียนโปรแกรมได้ยืดหยุ่น และเป็นระเบียบ (กล่าวถึงรายละเอียดต่อไปในด้านล่าง) ทำให้สามารถใช้คนจำนวนมากพัฒนาระบบร่วมกันได้อย่างมีประสิทธิภาพมากขึ้น จึงเหมาะสำหรับ

การพัฒนาระบบที่มีขนาดใหญ่ที่มีความซับซ้อน และต้องการบำรุงรักษาในระยะยาว

3. แองกูลาร์มีเครื่องมือที่จำเป็นในการพัฒนาเว็บแอปพลิเคชันอยู่จำนวนมาก ทำให้ไม่จำเป็นต้องดาวน์โหลดไลบรารี (library) อื่น ๆ มาใช้งานร่วมด้วย ส่งผลให้ผู้พัฒนาคนใหม่สามารถพัฒนาระบบเดิมต่อได้สะดวก หากพัฒนาระบบด้วย React หรือ Vue.js ผู้พัฒนาต้องเลือกไลบรารีที่สนับสนุน routing, dependency injection, forms ฯลฯ มาใช้งานร่วมด้วย ซึ่งไลบรารีเหล่านี้จะถูกเลือกโดยผู้ออกแบบระบบ หากผู้ออกแบบระบบเลิกพัฒนาโปรเจคเดิมแล้ว การหาผู้พัฒนาคนใหม่จะเป็นไปได้ยาก เพราะผู้พัฒนาคนใหม่ต้องมีความเข้าใจในไลบรารีต่าง ๆ ที่ผู้ออกแบบระบบได้เลือกไว้

4. สนับสนุนการทำ responsive web design (RWD) ทำให้ user interface (UI) ปรับเปลี่ยนขนาดได้ตามหน้าจอที่แสดงผล จึงสามารถพัฒนาโค้ดหน้าเว็บชุดเดียวแล้วใช้ได้ทั้งบนเครื่องคอมพิวเตอร์แบบตั้งโต๊ะ (desktop) และมือถือ (mobile)

2.2.1.2 ข้อเสีย

แองกูลาร์มีข้อเสียบางประการ ดังต่อไปนี้

1. โปรแกรมที่พัฒนาด้วยแองกูลาร์มีขนาดใหญ่เมื่อเทียบกับการพัฒนาด้วย React และ Vue.js ทำให้ใช้เวลาในการ compile นานกว่า
2. เนื่องจากแองกูลาร์มีโครงสร้างการเขียนโปรแกรมที่ดี มีรูปแบบการเขียนและแพทเทิร์นที่ค่อนข้างตายตัว ส่งผลให้ผู้สนใจใช้เวลาในการเรียนรู้นาน โดยเฉพาะเมื่อเทียบกับการพัฒนาด้วย React และ Vue.js

2.2.1.3 แพทเทิร์น dependency Injection (DI pattern)

Dependency injection เป็นหนึ่งในแพทเทิร์นของการออกแบบโปรแกรมซึ่งเป็นจุดเด่นของแองกูลาร์ โดยแองกูลาร์นำแพทเทิร์นนี้มาใช้ในองค์ประกอบหลักหลายส่วนในการพัฒนาโปรแกรม เช่น การทำการพิสูจน์ตัวตน (authentication) การทำการทดสอบโปรแกรม (testing) การสร้างคอมโพเน้นต์ของตนเอง (custom component) ที่ต้องการการนำกลับมาใช้ใหม่ ฯลฯ แพทเทิร์นนี้ช่วยเพิ่มความยืดหยุ่นให้กับคอมโพเน้นทำให้สามารถนำคอมโพเน้นต์กลับมาใช้ได้

ใหม่ ผู้พัฒนาไม่ต้องแก้ไขโค้ดหลักทั้งหมด พัฒนาเพิ่มเฉพาะคลาสใหม่ที่ต้องการเรียกใช้งาน และเรียกใช้งานโดยปรับเปลี่ยนเพียงชื่อของคลาส ดังตัวอย่างด้านล่าง

2.2.1.3.1 ตัวอย่างการใช้แพทเทิร์น DI ในโมดูลการพิสูจน์ตัวตน

การพิสูจน์ตัวตน (authentication) จะต้องมีการใส่ข้อมูลในส่วนหัว (header) ของ http request เพื่อระบุสิทธิ์การเข้าถึงของผู้ใช้งาน แล้วส่ง http request นี้ไปยังระบบส่วนหลัง (backend) เพื่อให้ระบบส่วนหลังดำเนินการตามที่ผู้ใช้งานร้องขอ แต่ก่อนที่ระบบส่วนหลังจะดำเนินการตามนั้น ระบบส่วนหลังจะขอตรวจสอบสิทธิ์การเข้าถึงข้อมูลของผู้ร้องขอก่อน หากตรวจสอบแล้วพบว่าผู้ใช้งานมีสิทธิ์ในการเข้าถึง ระบบส่วนหลังก็จะดำเนินการตามที่ผู้ใช้งานร้องขอ ขั้นตอนนี้จะเกิดขึ้นทุกครั้งที่ผู้ใช้งานส่ง http request ไปยังระบบส่วนหลัง ดังนั้น แองกูลาร์จึงมีบิลต์อินอินเตอร์เฟส (built-in interface) ชื่อ HttpInterceptor เพื่อให้ผู้ใช้สามารถอิมพลีเมนต์ (implement) อินเตอร์เฟสนี้ตามความต้องการ และการใช้งานอินเตอร์เฟสนี้ใช้หลักการของ DI ซึ่งแสดงในรูปที่ 15-16

รูปที่ 15 แสดงคลาส AuthenInterceptor ซึ่งมาจากการอิมพลีเมนต์อินเตอร์เฟสชื่อ HttpInterceptor เพื่อแทรกข้อมูลการพิสูจน์ตัวตน (authentication) ในส่วนหัวของ http request ของผู้ใช้งาน และรูปที่ 16 แสดงโค้ดตัวอย่างของโมดูล authentication ซึ่งเป็นโมดูลที่ผู้พัฒนาสร้างขึ้นเพื่อให้แองกูลาร์ใช้แพทเทิร์น DI สร้างอินสแตนซ์ (instance) ของคลาส AuthenInterceptor แล้วอินเจค (inject) ลงในโมดูลนี้ ทำให้ผู้พัฒนาไม่จำเป็นต้องแก้ไขโปรแกรมทั้งหมด เพียงแค่เปลี่ยนค่าของแอตทริบิวต์ชื่อ useClass เป็นชื่อคลาสที่ต้องการให้แองกูลาร์ทำแพทเทิร์น DI ก็เพียงพอ


```
@Injectable()
export class AuthenInterceptor implements HttpInterceptor {
  constructor(private authService: AuthenService) {}

  intercept(req: HttpRequest<any>, next: HttpHandler) {
    const authToken = this.authService.getToken();
    const authRequest = req.clone({
      headers: req.headers.set('Authorization', 'Bearer ' + authToken)
    });
    return next.handle(authRequest);
  }
}
```

รูปที่ 15 โค้ดตัวอย่างของคลาส AuthenInterceptor

```
@NgModule({
  declarations: [SigninComponent, SignupComponent],
  imports: [SharedModule, AuthenRoutingModule],
  providers: [
    { provide: HTTP_INTERCEPTORS, useClass: AuthenInterceptor, multi: true }
  ]
})
export class AuthenModule {
  constructor() {}
}
```

รูปที่ 16 โค้ดตัวอย่างของโมดูล authentication (หมายเหตุ “HTTP_INTERCEPTORS” เป็นชื่อของบิลต์อินอินเตอร์เฟสอันหนึ่งของแองกูลาร์)

2.2.1.3.2 ตัวอย่างการใช้แพทเทิร์น DI ในการทดสอบโปรแกรม

แองกูลาร์สะดวกต่อการทดสอบโปรแกรมโดยให้ผู้พัฒนาสามารถแก้ไขปรับเปลี่ยนสภาพแวดล้อมของโปรแกรมได้โดยการเปลี่ยนชื่อคลาสที่ต้องการสร้างอินสแตนซ์แล้ว อินเจคลงในโค้ดชุดเดิม ดังรูปที่ 17 - 18

```
@NgModule ({
  ...
  providers: [{provide: ProductService, useClass: ProductService}]
})
```

รูปที่ 17 โมดูล product ในสภาพแวดล้อม production โดยโมดูลนี้จะสร้างอินสแตนซ์จากคลาส ProductService

```
@NgModule ({
  ...
  providers: [{provide: ProductService, useClass: MockProductService}]
})
```

รูปที่ 18 โมดูล product ในสภาพแวดล้อม test โดยโมดูลนี้จะสร้างอินสแตนซ์จากคลาส MockProductService ซึ่งมีลักษณะคล้ายคลึงกับคลาส ProductService

จากเหตุผลดังกล่าวข้างต้นผู้วิจัยจึงนำแองกูลาร์เฟรมเวิร์คมาใช้ในการพัฒนาส่วนหน้า (frontend) ของซอฟต์แวร์อินซีเอ็นวีเพื่อให้ซอฟต์แวร์มีความยืดหยุ่นในการออกแบบ และสามารถพัฒนาต่อได้ง่ายในระยะยาว

2.2.2 โหนดเจเอส (NodeJS)

โหนดเจเอส คือเทคโนโลยีในการรันภาษาจาวาสคริปต์ให้สามารถทำงานเป็นเซิร์ฟเวอร์ได้ ทำงานในลักษณะซิงเกิลเธรด (single thread) ยกเว้นส่วนที่ทำงานร่วมกับอินพุต และเอ้าพุตของโปรแกรม (I/O) เช่น การอ่านไฟล์ การเขียนไฟล์ และการส่งการร้องขอเพื่อเข้าถึงฐานข้อมูล เป็นต้น รูปแบบการเขียนโปรแกรมในโหนดเจเอสไม่ได้มีข้อตกลงตายตัว ขึ้นอยู่กับผู้พัฒนาว่าต้องการออกแบบโปรแกรมอย่างไร มีเว็บแอปพลิเคชันเฟรมเวิร์ค (web application framework) ที่นิยมใช้งานหลายอัน และหนึ่งในนั้นคือ เอ็กเพรสเฟรมเวิร์ค (express framework) ซึ่งมีหน้าที่หลักในการจัดการ http request ที่เข้ามายังโหนดเจเอส เร้าเส้นทาง (route) ให้กับ request เหล่านั้น และส่ง http response กลับไปยังเครื่องผู้ใช้งานฝั่งไคลเอนต์ (client)

สำหรับงานวิจัยนี้ ผู้วิจัยได้ใช้โหนดเจเอสร่วมกับเอ็กเพรสเฟรมเวิร์คมาพัฒนาส่วนหลัง (backend) ของซอฟต์แวร์อินซีเอ็นวีเพื่อให้ผู้วิจัยสามารถกำหนดรูปแบบการเขียนโปรแกรมตามความเหมาะสมของซอฟต์แวร์ได้สะดวก

2.2.3 โมเดล-วิว-คอนโทรลเลอร์ (Model-View-Controller: MVC)

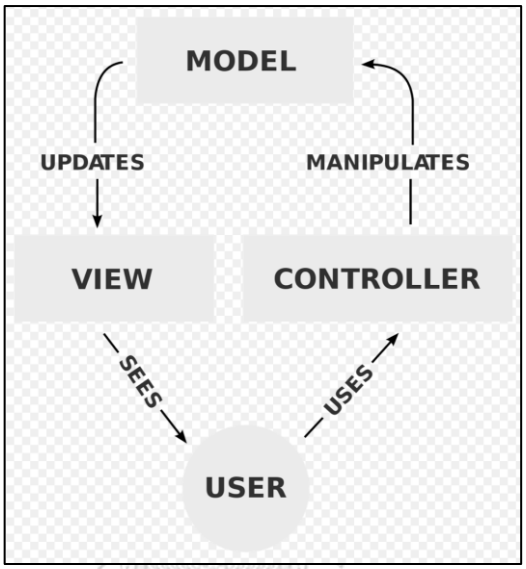
โมเดล-วิว-คอนโทรลเลอร์เป็นแพทเทิร์นการออกแบบโปรแกรมแบบหนึ่งที่แบ่งการทำงานของโปรแกรมเป็น 3 องค์ประกอบ อันได้แก่ Model (M) View (V) และ Controller (C)

โมเดล (model: M) คือ ส่วนของการเก็บรวบรวมข้อมูล ส่วนนี้จะติดต่อกับแหล่งข้อมูลที่โปรแกรมต้องการนำมาประมวลผล หรือว่านำมาแสดงผล เช่น ฐานข้อมูล (database) ไฟล์ข้อความ และไฟล์ภาพ เป็นต้น

วิว (view: V) คือ ส่วนของการแสดงผลซึ่งเป็นการแสดงข้อมูลต่าง ๆ ทั้งจากโมเดล (model) โดยตรง และจากการผ่านการประมวลในคอนโทรลเลอร์ก่อนนำมาแสดง เช่น การแสดงข้อมูลบนหน้าเว็บผ่านทางข้อความ ตาราง หรือ แผนภาพ เป็นต้น

คอนโทรลเลอร์ (controller: C) คือ ส่วนของการทำงาน ซึ่งจะเป็นส่วน business logic ของโปรแกรม จะทำงานตามที่ได้รับคำสั่งจากผู้ใช้ผ่านทางส่วนวิว (view)

ยกตัวอย่างกรณีการทำงานบนเว็บแอปพลิเคชันดังรูปที่ 19 เมื่อผู้ใช้เห็นหน้าเว็บซึ่งก็คือส่วนวิว (view) ผู้ใช้ก็จะมีปฏิสัมพันธ์แล้วสั่งงานโปรแกรม ส่วนคอนโทรลเลอร์ (controller) ก็จะประมวลผลตามคำสั่ง รวมถึงปรับเปลี่ยนข้อมูลที่เก็บไว้ที่ส่วนโมเดล (model) เมื่อข้อมูลในโมเดลถูกเปลี่ยนแปลงแล้ว หน้าเว็บหรือส่วนวิวก็จะถูกปรับเปลี่ยนตามไปด้วย



รูปที่ 19 แผนภาพแสดงกระบวนการทำงานของ MVC

(ที่มา: [73])

ข้อดี คือ

1. เนื่องจากโปรแกรมมีการแบ่งส่วนการทำงานที่ชัดเจน คือ ส่วนโมเดล ส่วนวิว และส่วนคอนโทลเลอร์ ทำให้คนในทีมสามารถแยกพัฒนาโปรแกรมได้พร้อมกัน 3 ส่วนได้
2. ง่ายต่อการปรับเปลี่ยนแก้ไขโปรแกรม เมื่อแก้ไขโปรแกรมส่วนหนึ่ง จะไม่กระทบกับส่วนที่เหลือ
3. ง่ายต่อการทดสอบโปรแกรม สามารถแยกทดสอบโปรแกรมเป็นส่วน ๆ ได้ ไม่ต้องเสียเวลาในการทดสอบทั้งหมดทุกครั้งที่มีการแก้ไขเกิดขึ้น
4. สามารถพัฒนาส่วนวิว (view) ได้หลากหลายโดยใช้โค้ดในส่วนของโมเดล และคอนโทรลเลอร์ชุดเดิม

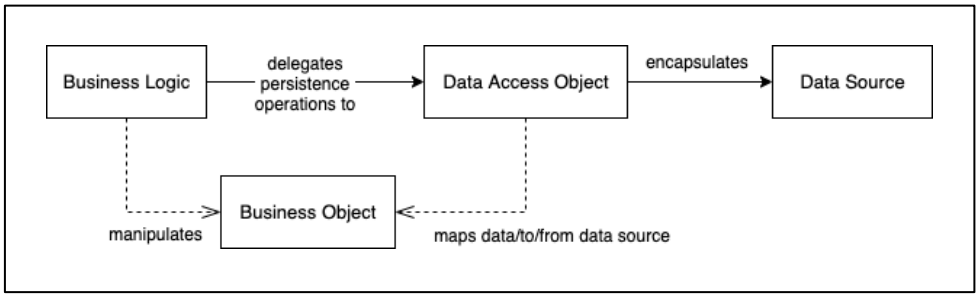
ข้อเสีย คือ

1. เพิ่มความซับซ้อนให้กับการออกแบบซอฟต์แวร์ เนื่องจากต้องแยกส่วนของซอฟต์แวร์ออกเป็น 3 ส่วน รวมถึงต้องมีการกำหนดวิธี และรูปแบบการเร้าเส้นทางในการรับ หรือส่ง ข้อมูลไปยังแต่ละส่วนที่ชัดเจน
2. สำหรับระบบที่มีขนาดเล็กจะเป็นการเพิ่มงานเกินความจำเป็น กล่าวคือ โค้ดในการทำงานหลักตาม business logic อาจมีเพียงเล็กน้อย แต่ผู้พัฒนาต้องเขียนโปรแกรมเพิ่มโค้ดเพื่อแบ่งระบบให้เป็น 3 ส่วนตามแพทเทิร์นโมเดล-วิว-คอนโทรลเลอร์
3. ผู้พัฒนาระบบต้องมีความรู้ และความเข้าใจในการทำงานของโมเดล-วิว-คอนโทรลเลอร์อย่างดีจึงจะสามารถพัฒนาระบบได้

สำหรับงานวิจัยนี้ผู้วิจัยได้นำโมเดล-วิว-คอนโทรลเลอร์มาประยุกต์ใช้กับซอฟต์แวร์อินซีเอ็นวีโดยการนำรูปแบบของโมเดล (model) และคอนโทรลเลอร์ (controller) มาใช้ในส่วนของโปรแกรมส่วนหลัง (backend) ซึ่งพัฒนาด้วยโหนดเจเอส และนำรูปแบบของวิว (view) มาใช้ในส่วนของการแสดงผลลัพธ์ที่โปรแกรมส่วนหน้า (frontend) ซึ่งพัฒนาด้วยแองกูลาร์ เพื่อให้สามารถแบ่งการพัฒนาซอฟต์แวร์ออกจากกันให้เป็นสัดส่วน

2.2.4 วัตถุการเข้าถึงข้อมูล (Data access object: DAO)

เป็นแพทเทิร์นการเขียนโปรแกรมรูปแบบหนึ่งที่เก็บชุดคำสั่งการเข้าถึงฐานข้อมูล ซึ่งมีประโยชน์คือ ช่วยแยกชั้นข้อมูลออกจาก business logic ออกจากแหล่งข้อมูล เช่น ฐานข้อมูล (database) หรือไฟล์ ทำให้ลดความซับซ้อนของโปรแกรม สร้างระเบียบในการเข้าถึงข้อมูลทำให้ทั้งโปรแกรมสามารถเข้าถึงข้อมูลได้ในรูปแบบเดียวกัน และช่วยให้ผู้พัฒนานำโค้ดเก่ากลับมาใช้ใหม่ได้ ดังรูปที่ 20

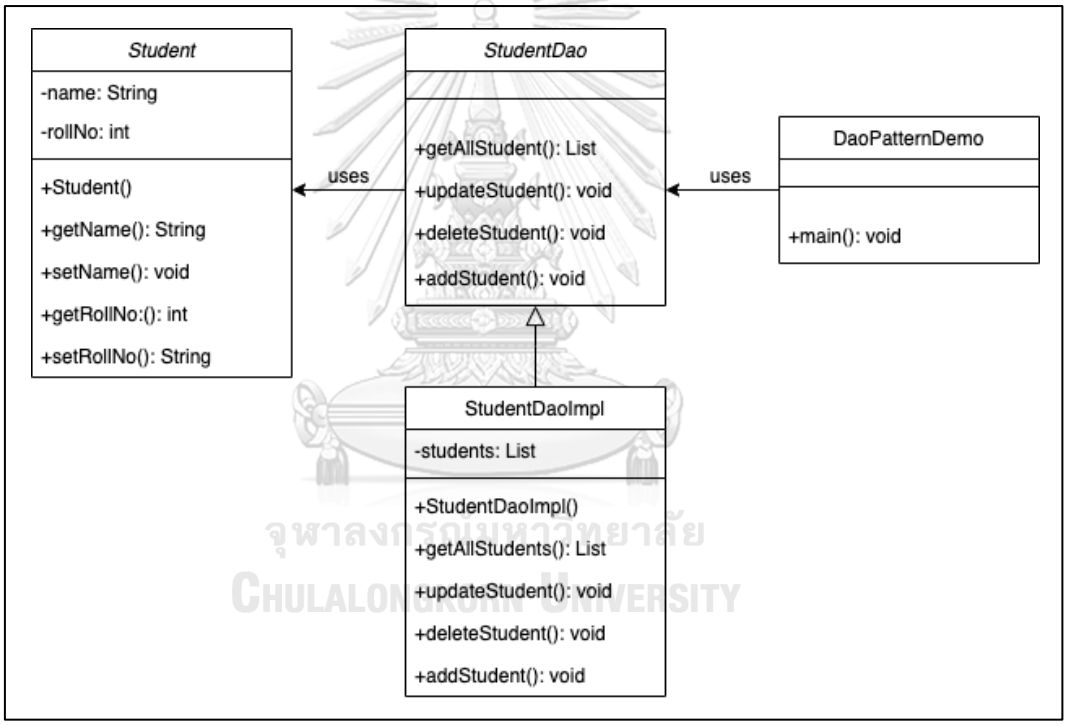


รูปที่ 20 แผนภาพแสดงบทบาทของวัตถุการเข้าถึงข้อมูลในระบบ

(ที่มา: รูปที่ 32.1 ของ [74])

วัตถุการเข้าถึงข้อมูลจะเขียนในรูปแบบของอินเตอร์เฟสดังรูปที่ 21 คลาส DaoPatternDemo เป็นคลาสหลักเพื่อเรียกใช้ดูข้อมูลนักศึกษา StudentDao ซึ่งเป็นอินเตอร์เฟสของคลาส StudentDaoImpl ซึ่งจะสามารถเข้าถึงฐานข้อมูลเพื่อแก้ไข ปรับเปลี่ยน เรียกดูข้อมูลจากฐานข้อมูล แล้วแมปข้อมูลมาใส่ไว้ในอินสแตนซ์ของ คลาส Student เพื่อนำไปใช้แสดงผลต่อไป

สำหรับงานวิจัยนี้ผู้วิจัยได้ประยุกต์วัตถุการเข้าถึงข้อมูลมาใช้กับซอฟต์แวร์อินซีเอ็นวี โดยการสร้างคลาสซึ่งเก็บรวมรวมคำสั่ง SQL ที่ใช้ในการเข้าถึงข้อมูลในแต่ละตารางของฐานข้อมูลเชิงสัมพันธ์ (relational database) หรือ สร้างคลาสสำหรับเก็บรวบรวมคำสั่งการเข้าถึงไฟล์แต่ละไฟล์ขึ้นมา เพื่อแยกโค้ดในส่วนของการเข้าถึงแหล่งข้อมูลออกจากโค้ดส่วนของ business logic



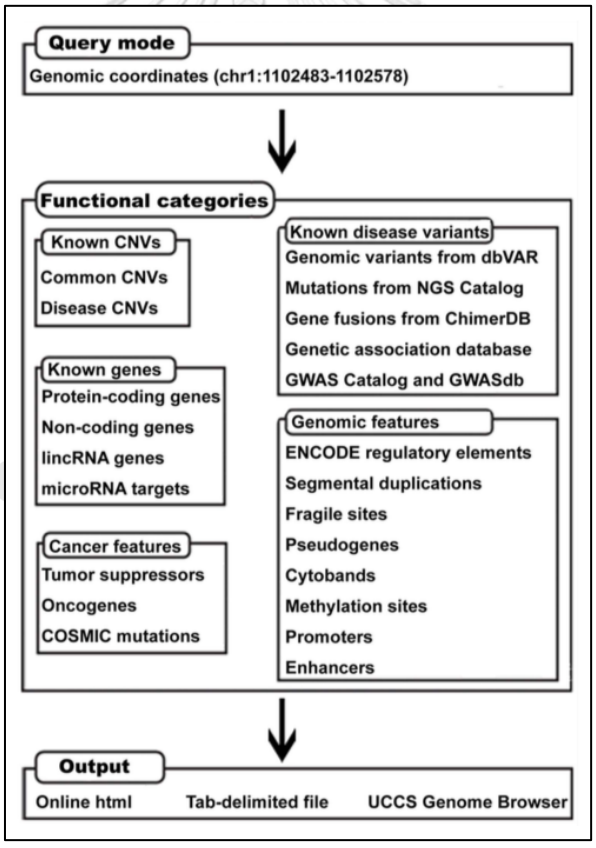
รูปที่ 21 คลาสไดอะแกรมแสดงตัวอย่างการใช้งานวัตถุการเข้าถึงข้อมูล

(ที่มา: [75])

บทที่ 3
งานวิจัยที่เกี่ยวข้อง

3.1 CNVannotator: A Comprehensive Annotation Server for Copy Number Variation in the Human Genome

CNVannotator [76] เป็นเว็บเซิร์ฟเวอร์ที่ให้คำอธิบายจีโนมแก่ซีเอ็นวีโดยอนุญาตผู้ใช้อัพโหลดข้อมูลซีเอ็นวีที่สนใจเข้าระบบ แล้ว CNVannotator จะส่งผลลัพธ์คำอธิบายจีโนมที่มีความสัมพันธ์กับข้อมูลซีเอ็นวีนั้นกลับให้ผู้ใช้ทางหน้าเว็บ นอกจากนี้ CNVannotator กำหนดแบบฟอร์มการนำเข้าข้อมูลซีเอ็นวีให้อยู่ในรูปแบบตารางที่ง่ายต่อการใช้งาน และใช้ข้อมูลคำอธิบายจีโนมที่เกี่ยวเนื่องกันจากฐานข้อมูลจำนวนมาก ผู้ใช้สามารถดาวน์โหลดผลลัพธ์เป็นคำอธิบายเหล่านั้นจากระบบในรูปแบบของข้อความธรรมดา (plain text) การทำงานโดยภาพรวมมีลักษณะดังแสดงในรูปที่ 22



รูปที่ 22 แผนภาพการทำงานโดยภาพรวมของ CNVannotator
(ที่มา: รูปที่ 1 ของ [76])

คำอธิบายจีโนมที่ CNVannotator เลือกใช้ถูกรวบรวมมาจากฐานข้อมูล 18 แหล่ง และแบ่งประเภทข้อมูลที่สนใจเป็น ซีเอ็นวีที่มีการรายงานมาก่อน (known CNVs) ยีนที่มีการรายงานมาก่อน (known genes) คุณลักษณะของมะเร็ง (cancer features) การแปรผันทางพันธุกรรมที่ถูกค้นพบแล้วว่าเกี่ยวข้องกับโรค (known disease variants) และคุณลักษณะอื่นๆที่เกี่ยวข้องกับจีโนม (genomic features) ดังรูปที่ 23 รวมทั้งสิ้น 5,277,234 ตำแหน่งที่ไม่ซ้ำกันบนจีโนม

Data source	Number of genomic coordinate	Source and reference
Known CNVs		
Common CNVs	356,817	Common CNVs from DGV database [25]
Disease CNVs	181,261	Disease CNVs from CNVD database [26]
Known variants		
dbVar	2,716,881	Genomic structural variants in dbVAR [27]
GWASdb	137,111	Human genetic variants by GWAS [31]
GWAS Catalog	6381	Etiologic and functional variants [30]
GAD	3057	Genetic variants by association studies [32]
Gene fusion	1198/1103[a]	Experimentally validated gene fusion events from ChimerDB [29]
NGS Catalog	1071	Genetic variants from NGS-based studies in human [28]
Coding and non-coding genes		
microRNA target	52,920	Targeting gene for all human miRNAs [35]
Coding gene	30,770	Protein-coding RefSeq genes [34]
Long non-coding RNA	21,033	Long non-coding genes (UCSC browser [34])
Other non-coding RNA	1337	Non-coding genes from UCSC browser (Excluding long non-coding RNAs) [34]
Genomic features		
ENCODE regulomeDB	1,880,556	Genomic functional elements from ENCODE data [37]
Segmental duplication	40,832	Global analysis result of human segmental duplications [38]
Promoter	29,119	500 bp upstream from the transcription start sites using UCSC data [34]
CpG island	28,691	CpG island data from UCSC browser [34]
Methylation	19,754	Human disease methylation sites from DiseaseMeth database [40]
Pseudogene	11,983	Pseudogene data from UCSC browser [34]
Enhancer	1478	Enhancer data from UCSC browser [34]
Cytoband	862	Cytoband data from UCSC browser [34]
Fragile site	69	Human genomic fragile sites from Entrez gene database [39]
Cancer genomic features		
COSMIC	125,753	Somatic mutations in cancer [41]
Tumor suppressor	716	Coding and non-coding tumor suppressor genes from TSGene database [42]
Oncogene	263	Coding oncogenes integrated from UniProt and TAG databases [43]

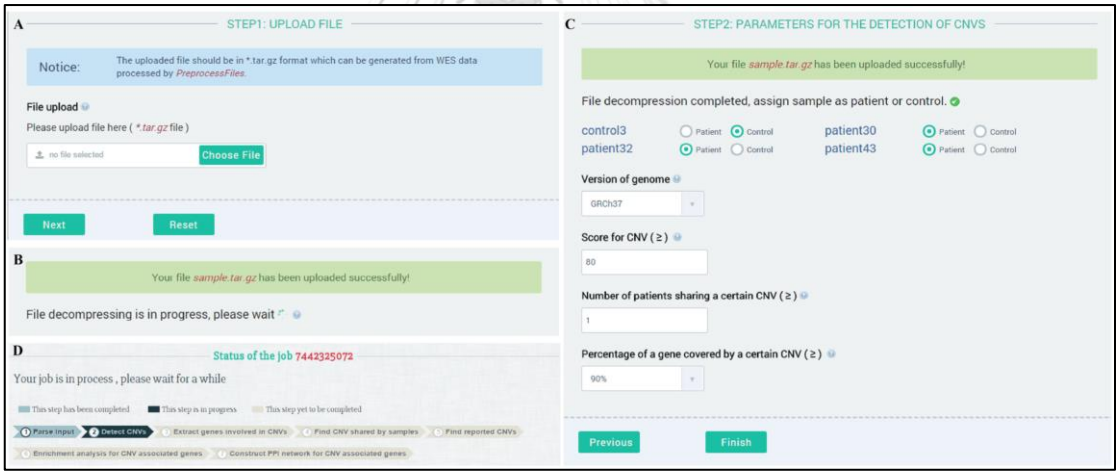
รูปที่ 23 ประเภท และแหล่งที่มาของคำอธิบายจีโนมทั้งหมดของเว็บเซิร์ฟเวอร์ CNVannotator
(ที่มา: ตารางที่ 1 ของ [76])

3.2 DeAnnCNV: a tool for online detection and annotation of copy number variations from whole-exome sequencing data

DeAnnCNV [27] เป็นเว็บเซิร์ฟเวอร์ที่ใช้หาซีเอ็นวีของกลุ่มตัวอย่าง (samples) และแสดงผลลัพธ์ในรูปแบบกราฟฟิก โดยแบ่งออกเป็น 2 โมดูล คือ (1) โมดูลตรวจจับ และแสดงผลกราฟฟิกของซีเอ็นวี และ (2) โมดูลแสดงคำอธิบายประกอบ หมายเหตุ งานวิจัยของ DeAnnCNV ได้นำเสนอผลงานวิจัยโดยใช้ข้อมูลเอ็กโซมของหนูที่ป่วย 4 ตัวในการทดสอบ (ไม่ได้ใช้ข้อมูลเอ็กโซมของมนุษย์ในการนำเสนอ)

3.2.1 โมดูลตรวจจับ และแสดงผลกราฟฟิกของซีเอ็นวี

ผู้ใช้ต้องนำข้อมูลลำดับเบสทั้งหมดบนเอ็กโซม (whole exome sequencing: WES) ของกลุ่มตัวอย่างมาบีบอัดด้วยแพ็กเกจชื่อ “ProcessFiles” ที่ DeAnnCNV เตรียมไว้ให้อยู่ในรูปแบบไฟล์เดียวนามสกุล “.tar.gz” แล้วจึงอัพโหลดไฟล์เข้าระบบ หลังจากนั้นผู้ใช้ต้องกำหนดค่าพารามิเตอร์ (parameters) ที่จะใช้ตรวจจับซีเอ็นวี (รูปที่ 24) แล้วระบบจึงจะตรวจจับซีเอ็นวี แล้วแสดงผลในโมดูลแสดงคำอธิบายประกอบซึ่งจะพูดถึงในหัวข้อถัดไป



รูปที่ 24 ตัวอย่างหน้าการนำเข้าข้อมูล และพารามิเตอร์ที่ใช้ในการตรวจจับซีเอ็นวี
(ที่มา: รูปที่ 1 ของ [27])

3.2.2 โมดูลแสดงคำอธิบายประกอบ

เมื่อ DeAnnCNV ตรวจจับซีเอ็นวีจากกลุ่มตัวอย่างเสร็จแล้ว จะแสดงผลลัพธ์ซีเอ็นวีที่ตรวจจับได้พร้อมกับคำอธิบายประกอบซีเอ็นวีเหล่านั้น ดังรูปที่ 25 โดยการให้คำอธิบายประกอบ (annotation) ของ DeAnnCNV มีขั้นตอนดังดังต่อไปนี้

1) พิจารณาว่าบริเวณนั้นถูกรายงานไว้ในฐานข้อมูลซีเอ็นวีหรือไม่ โดยใช้ข้อมูลจากฐานข้อมูลดีบีวาร์ (dbVar)

2) พิจารณาหาข้อมูลรายละเอียดของยีนที่เกี่ยวข้องกับซีเอ็นวีตัวนั้น

3) พิจารณาหาการแปรผันทางพันธุกรรมของยีนเหล่านั้น และพิจารณาว่ามีผลรายงานว่าเกี่ยวข้องกับโรคในมนุษย์หรือไม่ โดยใช้ข้อมูลจากฐานข้อมูลคลินวาร์ (ClinVar)

4) พิจารณาลักษณะที่แสดงออกในหนูที่ขาดยีนตัวนั้นว่าเป็นอย่างไร โดยใช้ข้อมูลจาก Mouse Genome Informatics: MGI

5) พิจารณาการแสดงออกของเอ็มอาร์เอ็นเอบนยีนเหล่านี้ในเนื้อเยื่อมนุษย์ หรือ ในเซลล์ไลน์ (cell lines)

6) วิเคราะห์การทำงานของยีนเหล่านั้นในมิติของยีนออนโทโลยี พาธ์เวย์ และโปรตีนในโดเมนที่เกี่ยวข้อง

7) สร้างเครือข่ายแสดงปฏิสัมพันธ์ระหว่างโปรตีน (protein-protein interactions: PPIs) สำหรับยีนที่เกี่ยวข้องกับซีเอ็นวีเพื่อพิจารณาว่ายีนตัวนั้นเกี่ยวข้องกับความผิดปกติของมนุษย์หรือไม่

A

CNV Associated Results

Chromosome	CNV Start	CNV End	Copy number	Gain/Loss	Score	Share Number	Sample ID	dbVar
				[All] x				
7	150972200	151082309	1	loss	52.0	1	patient16	
11	2398780	2428530	3	gain	65.6	1	patient16	
19	43702149	43763287	4	gain	54.7	1	patient15	essv45534 essv61803 essv6
7	143955789	144074283	1	loss	52.7	2	patient16	essv63495 essv69702 essv3
7	143880597	144074283	1	loss	31.2	2	patient15	essv63495 essv69702 essv3
19	43702149	43763287	4	gain	54.7	1	patient15	essv45534 essv61803 essv6
19	40376631	40400823	1	loss	54.3	1	patient15	essv76175 essv66632 essv3

รูปที่ 25 ตัวอย่างผลลัพธ์จากการใช้ DeAnnCNV ตรวจจับซีเอ็นวีของหนูที่ป่วย 4 ตัว

(ที่มา: รูปที่ 2 ของ [27])

3.3 CNView: a visualization and annotation tool for copy number variation from whole-genome sequencing

CNView [25] เป็นเครื่องมือที่ช่วยแสดงผลตำแหน่งของซีเอ็นวีบนจีโนมในระดับประชากร (population-scale WGS) ในรูปแบบกราฟฟิก โดยการระบุว่ามีซีเอ็นอยู่บนโครโมโซมชื่ออะไร อยู่บนตำแหน่งเบสที่เท่าใดบนโครโมโซมนั้น และอยู่บนยีนชื่ออะไร การแสดงผลสามารถแสดงให้ดูได้ทั้งแบบตัวอย่างเดียว และแบบหลายตัวอย่าง

เครื่องมือตัวนี้ถูกพัฒนาด้วยภาษาอาร์ (R language) แสดงผลลัพธ์ออกมาได้เป็นรูปภาพบนไฟล์พีดีเอฟ (PDF) และมีข้อมูลนำเข้าเป็นแมทริกซ์รูปแบบฟอร์แมตเบด (BED file format) ของค่าความครอบคลุม (coverage) ที่อยู่ในบินหรือช่วงลำดับเบสที่พิจารณา ซึ่งได้มาจากการใช้เครื่องมือ bedtools หาค่าความครอบคลุมจากไฟล์แบมของตัวอย่างที่สนใจ และของตัวอย่างอื่น ๆ อย่างน้อย 20 ตัวอย่าง แล้วนำผลลัพธ์มารวมกันในแมทริกซ์รูปแบบฟอร์แมตเบด

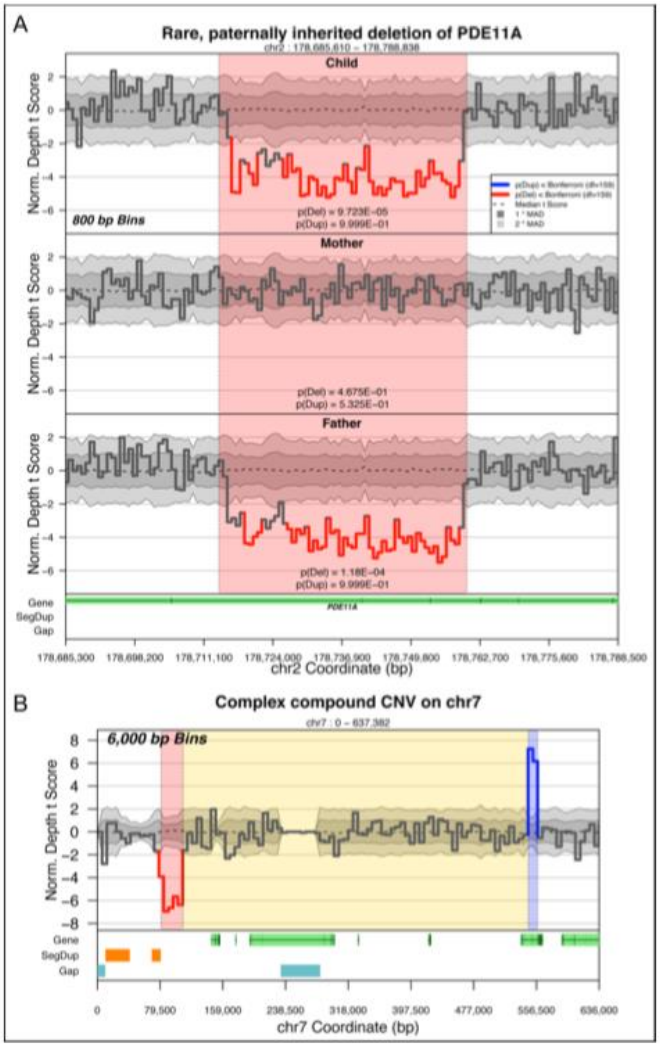
CNView มีขั้นตอนการทำงาน ดังนี้

1) การกรองค่าแมทริกซ์ (matrix filtering)
2) การบีบอัดแมทริกซ์ (matrix compression) เพื่อลดค่ารบกวนบริเวณข้างเคียง (local noise) และปริมาณการคำนวณในขั้นตอนถัดไป
3) การปรับค่าข้อมูลภายในตัวอย่างให้เป็นมาตรฐานเดียวกัน (intra-sample normalization)
4) การปรับค่าข้อมูลระหว่างตัวอย่างให้เป็นมาตรฐานเดียวกัน (inter-sample normalization)
5) การแสดงภาพความครอบคลุม (coverage visualization) คือ แสดงภาพกราฟฟิกตำแหน่งที่เกิดซีเอ็นวี
6) การให้คำอธิบายจีโนม (genome annotation) เช่น การระบุชื่อยีนบนตำแหน่งที่เกิดซีเอ็นวี

หมายเหตุ CNView สามารถประเมินผลลัพธ์เฉพาะส่วนของการแสดงผลกราฟฟิก (ไม่รวมการจัดการข้อมูลนำเข้าก่อนหน้า) ณ ตำแหน่งใด ๆ บนโครโมโซมที่สนใจพร้อมกัน 300 ตัวอย่างได้ภายในเวลา 1 นาที บนเครื่องคอมพิวเตอร์แบบตั้งโต๊ะที่มีตัวประมวลผล (processor) 2.3 กิกะเฮิรตซ์ (GHz) หน่วยประมวลผล 2 คอร์ (dual-core processor) และแรม (RAM) 8 กิกะไบต์ (GB)

ตัวอย่างผลลัพธ์จากเครื่องมือ CNView แสดงดังรูปที่ 26 ซึ่งแสดงซีเอ็นวีที่ประกอบร่วมกัน (compound CNVs) โดยรูปที่ 26 (A) แสดงโครโมโซมที่ 2 ช่วงตำแหน่งเบสที่ 178,685,610 – 178,788,838 พบว่ามีซีเอ็นวีแบบ duplication ในหน้าต่างสีแดงบนยีน PDE11A ในลูก (บน) และ

พ่อ (ล่าง) แต่ไม่แสดงซีเอ็นวีในแม่ (กลาง) และรูปที่ 26 (B) แสดงซีเอ็นวีประกอบร่วมกันแบบซับซ้อน (complex compound CNVs) บนโครโมโซมที่ 7 โดยบอกว่าในโครโมโซมที่ 7 พบตำแหน่งที่เป็นซีเอ็นวี 2 ตำแหน่ง ตำแหน่งแรก (สีแดง) เป็นซีเอ็นวีประเภท deletion ตำแหน่งที่สอง (สีน้ำเงิน) เป็นซีเอ็นวีประเภท duplication



รูปที่ 26 ตัวอย่างผลลัพธ์จากเครื่องมือ CNView

(ที่มา: รูปที่ 1 ของ [25])

3.4 iCopyDAV: Integrated platform for copy number variations—Detection, annotation and visualization

iCopyDAV [26] เป็นซอฟต์แวร์แพลตฟอร์มที่สามารถตรวจจับซีเอ็นวี ให้คำอธิบายประกอบและแสดงผลลัพธ์ในรูปแบบกราฟฟิกได้ พัฒนาโดยใช้ภาษาซีพลัสพลัส (C++ language) ภาษาอาร์

(R language) และภาษาแบช (bash language) ผู้ใช้สามารถติดตั้งโปรแกรมได้จากทั้งแบบสร้างโปรเจคขึ้นด้วยตัวเอง และจากโปรเจคอิมเมจที่ได้ทำไว้แล้วในดอกเกอร์ (docker) ศักยภาพของเครื่องมือนี้ถูกประเมินบนข้อมูล sequence depth (หมายถึง ข้อมูลเส้นรีดที่ได้กล่าวในหัวข้อก่อนหน้านี้ที่ชื่อ “2.1.3 เทคโนโลยีการหาลำดับเบสแบบเอ็นจีเอส”) ของข้อมูลที่จำลองขึ้นมา และข้อมูลจริง โดยลักษณะเด่นของ iCopyDAV มีดังนี้

1) ผู้ใช้สามารถปรับแต่งขั้นตอนการทำงานได้ เช่น ผู้ใช้สามารถเลือกขนาดบินเพื่อใช้ในการคัดกรองค่าอคติจากจีซีคอนเทนต์ และค่าอคติจากการแมปปิ้งรีดไปยังจีโนมอ้างอิงออก เพื่อให้เหมาะกับลักษณะของซีเอ็นวีที่ตรวจจับ และ sequencing coverage ของข้อมูล

2) ขั้นตอนการแบ่งส่วนในการหาซีเอ็นวี (segmentation) สามารถคำนวณแบบขนานได้ ทำให้ตรวจจับซีเอ็นวีได้เร็วสำหรับการตรวจจับซีเอ็นวีบนข้อมูลจีโนมที่มีระดับความลึกมาก (high sequence coverage) ทำให้เครื่องมือนี้สามารถใช้งานผ่านเครื่องคอมพิวเตอร์ตั้งโต๊ะธรรมดาได้

3) คุณสมบัติที่เด่นที่สุดคือ โมดูลคำอธิบายประกอบจีโนม (genomic annotation) ที่ช่วยให้ผู้ใช้ระบุชื่อ และจัดลำดับความสำคัญของซีเอ็นวีให้มีความสำคัญสูง กลาง และต่ำได้ โดยพิจารณาข้อมูลบนตำแหน่งที่เกิดซีเอ็นวี และบริเวณข้างเคียง โมดูลนี้แบ่งข้อมูลออกเป็น 4 ลักษณะ คือ องค์ประกอบในการทำงาน (functional elements) ลักษณะจีโนมและความเกี่ยวข้องกันกับโรค (clinical associations) ซีเอ็นวีที่ถูกรายงานไว้ในฐานข้อมูลการแปรผันที่เกี่ยวกับจีโนมโดยอ้างอิงจากฐานข้อมูลดีจีวี และการแปรผันเชิงโครงสร้างอื่น ๆ (other structural variations)

ขั้นตอนการทำงานโดยภาพรวมของ iCopyDAV แสดงไว้ดังรูปที่ 27 ซึ่งมีข้อมูลนำเข้าเป็นสีเขียว ขั้นตอนการทำงานเป็นสีเหลือง และข้อมูลขาออกเป็นสีฟ้า โดยสามารถอธิบายโมดูลการทำงาน (สีเหลือง) ได้ดังนี้

1) โมดูล Bin size estimation เป็นโมดูลสำหรับคำนวณหาขนาดของบินที่เหมาะสมที่สุดในการตรวจจับซีเอ็นวีด้วยวิธีการใช้ความลึกของรีด (read depth) โดยมีข้อมูลนำเข้าสองอย่าง คือ

    a. Aligned sequence reads (ไฟล์แบม) คือ ผลจากการนำข้อมูลรีดที่ได้จากการหาลำดับเบสทั้งหมดบนจีโนม หรือ บนเอ็กโซมของมนุษย์ หรือของสายพันธุ์อื่น ๆ ด้วยเทคโนโลยีเอ็นจีเอสมาเทียบกับจีโนมอ้างอิง

b. Configuration file คือ ไฟล์ข้อมูลการตั้งค่า

2) โมดูล GC correction เป็นโมดูลสำหรับปรับค่าจีซีคอนเทนต์ ให้เป็นมาตรฐานเดียวกัน โดยใน iCopyDAV จะใช้ 2 วิธีการ คือ อัลกอริทึม Local Polynomial Regression fitting (Loess) [77] และ Median approach [78] โมดูลนี้เป็นหนึ่งในตัวเลือกที่ผู้ใช้งานสามารถเลือกไม่ใช้งานขั้นตอนนี้ได้ โมดูลนี้ใช้ข้อมูลนำเข้าคือ

   a. ไฟล์เก็บค่าจีซีคอนเทนต์ในแต่ละบิน คือ ไฟล์ที่ระบุคะแนนจีซีคอนเทนต์ในแต่ละบินที่ได้จากบราวเซอร์จีโนมของมหาวิทยาลัยแห่งแคลิฟอเนียร์ซานตาครูส (UCSC genome browser) [79]

   b. ไฟล์ที่ระบุตำแหน่งเริ่มต้นและสิ้นสุดของแต่ละบินที่ไม่ทับซ้อนกัน

3) โมดูล Filter mappability score เป็นโมดูลที่ตัดเอารีดเส้นที่มีค่าความสามารถในการแมปปิ้งต่ำออกไป โดยใช้ข้อมูลนำเข้าคือ ค่าคะแนนแมปปิ้ง (Mappability score) ที่ได้มาจากบราวเซอร์จีโนมของมหาวิทยาลัยแห่งแคลิฟอเนียร์ซานตาครูส (UCSC genome browser) [79]

4) โมดูล Segmentation เป็นโมดูลสำหรับระบุว่าบริเวณข้างเคียงบริเวณใดที่มีจำนวนรีดเท่ากับตำแหน่งที่สนใจ iCopyDAV ใช้วิธีการ 2 วิธี คือ Total Variation Minimization (TVM) [80] และ Circular Binary Segmentation (CBS) approach โดยทั้งสองวิธีสามารถทำงานแบบขนาน คือ ทำงานแบบมัลติเธรด (multithread) และสามารถปรับแต่งจำนวนเธรด (thread) ที่ต้องการใช้ได้ตามจำนวนซีพียูคอร์ (CPU cores) ที่ว่างอยู่ นอกจากนี้ผู้ใช้ยังสามารถเลือกได้ว่าจะใช้เพียงวิธีการเดียว หรือใช้ทั้งสองวิธีการตามที่ได้กล่าวไว้ข้างต้น

5) โมดูล Call Variants เป็นโมดูลสำหรับใช้ในการตรวจจับซีเอ็นวีว่าเกิดขึ้นที่ตำแหน่งใดบนจีโนม และเป็นซีเอ็นวีประเภทอะไร

6) โมดูล Annotation เป็นโมดูลสำหรับให้คำอธิบายองค์ประกอบโครงสร้าง และการทำงานบนตำแหน่งของซีเอ็นวีที่ทำนายได้ และบริเวณข้างเคียง โดยสรุปคำอธิบายจีโนมไว้เป็น 4 ประเภท คือ

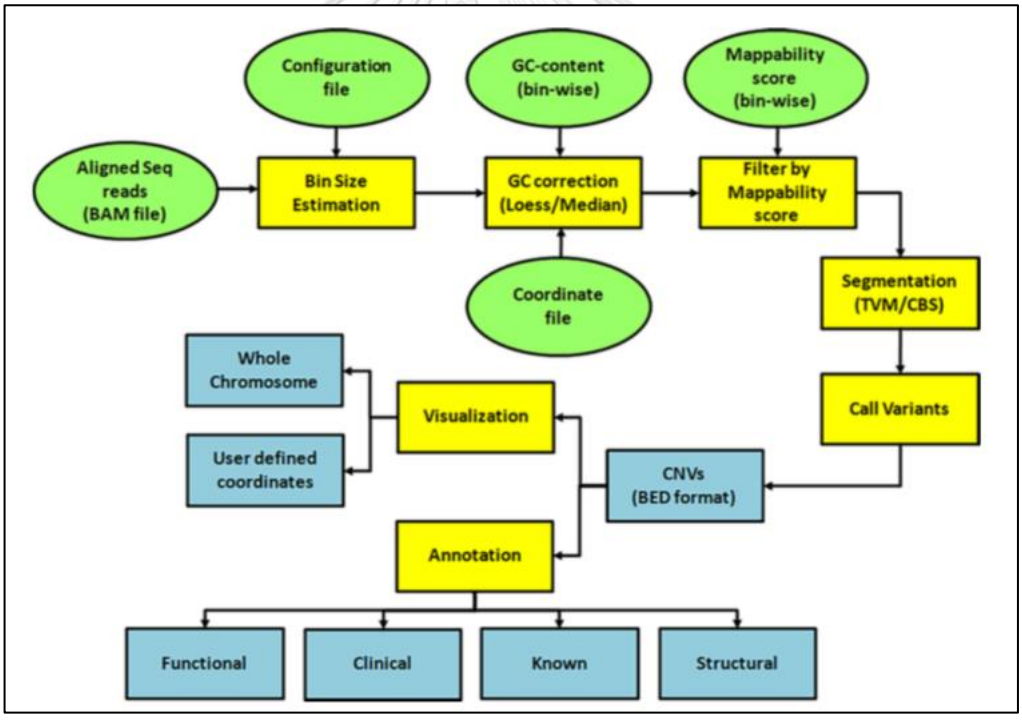
   a. คำอธิบายในเชิงฟังก์ชันการทำงานของยีนว่ายีนตัวนั้นมีหน้าที่อะไรในร่างกาย (functional annotation) เช่น ยีนตัวนั้นเกี่ยวข้องกับเมตาบอลิซึม หรือเกี่ยวข้องกับการสร้างภูมิคุ้มกัน เป็นต้น

b. คำอธิบายว่ายีนตัวนั้นเกี่ยวข้องกับโรคอะไร (clinical annotation)

c. คำอธิบายว่ายีนตัวนั้นมีการรายงานไว้ในฐานข้อมูลแล้วหรือไม่ (known annotation)

d. คำอธิบายว่ายีนตัวนั้นเกี่ยวข้องกับการแปรผันเชิงโครงสร้างประเภทไหน (structural annotation)

7) โมดูล Visualization เป็นโมดูลสำหรับแสดงผลลัพธ์ของซีเอ็นวีด้วยการวาดภาพการกระจายของซีเอ็นวีที่เกิดขึ้น โดยผู้ใช้สามารถเลือกตำแหน่งการกระจายของซีเอ็นวีบนโครโมโซมใด ๆ หรือเลือกตำแหน่งใด ๆ ที่สนใจได้ และสามารถดาวน์โหลดไฟล์ผลลัพธ์เป็นรูปภาพนามสกุล '.png' ได้



รูปที่ 27 แผนภูมิการทำงานของ iCopyDAV

(ที่มา: รูป 1 ของ [26])

3.5 เปรียบเทียบงานวิจัยที่เกี่ยวข้อง

วิทยานิพนธ์นี้ได้เสนอเครื่องมือใหม่ที่มีชื่อว่า “อินซีเอ็นวี (inCNV)” โดยทำการเปรียบเทียบเครื่องมือใหม่ กับเครื่องมือจากทฤษฎีที่เกี่ยวข้องตามตารางที่ 2 – 3 ซึ่งมีรายละเอียดดังนี้

ตารางที่ 2 เปรียบเทียบข้อมูลนำเข้า และไฟล์ที่ส่งออกของเครื่องมือตรวจจับซีเอ็นวีแต่ละตัว และตารางที่ 3 เปรียบเทียบฟังก์ชันการทำงานของเครื่องมือตรวจจับซีเอ็นวีแต่ละตัว

ตารางที่ 2 เปรียบเทียบข้อมูลนำเข้า และไฟล์ที่ส่งออกของเครื่องมือตรวจจับซีเอ็นวีแต่ละตัว

เครื่องมือ	ข้อมูลนำเข้า	รูปแบบไฟล์ข้อมูลนำเข้า	รูปแบบไฟล์ส่งออก
CNVannotator [76]	ข้อมูลซีเอ็นวี	ไฟล์ข้อความในรูปแบบตารางตามแบบฟอร์มที่ CNVannotator กำหนดไว้	✓ (ข้อความธรรมดา)
DeAnnCNV [27]	ข้อมูลเส้นรีด	ชุดของไฟล์แบมที่ถูกบีบอัดด้วยแพ็กเกจที่ DeAnnCNV กำหนด	✗ (ดูข้อมูลได้จากหน้าเว็บเท่านั้น)
CNView [25]	ค่าความครอบคลุมของแต่ละบิน (binned coverage values) ที่ระบุชื่อตัวอย่างและตำแหน่งจีโนม	ไฟล์ข้อความในแมทริกซ์รูปแบบฟอร์แมตเบด	✓ (ส่งออกรูปภาพนามสกุล ‘.pdf’)
iCopyDAV [26]	ข้อมูลเส้นรีด	ไฟล์แบม	✓ (ไฟล์รูปภาพนามสกุล ‘.png’)
inCNV	ข้อมูลซีเอ็นวีของตัวอย่างใด ๆ โดยระบุชื่อตัวอย่าง ตำแหน่งซีเอ็นวีบนโครโมโซม ประเภทของซีเอ็นวี และชื่อเครื่องมือตรวจจับซีเอ็นวี	ไฟล์ข้อความนามสกุล “.txt” ในรูปแบบคั่นด้วยแท็บ (tab-delimited format)	✓ (ไฟล์นามสกุล ‘.txt’)

ตารางที่ 3 เปรียบเทียบฟังก์ชันการทำงานของเครื่องมือตรวจจับซีเอ็นวีแต่ละตัว

เครื่องมือ	ประเภท	ฟังก์ชันการทำงาน			
		ตรวจจับซีเอ็นวี	คำอธิบายจีโนม	จัดลำดับความสำคัญของซีเอ็นจากคำอธิบายจีโนม	แสดงผลลัพธ์ในรูปแบบกราฟฟิก
CNVannotator [76]	เว็บเซิร์ฟเวอร์	✗ (รับข้อมูลซีเอ็นวีจากเครื่องมือเดียว)	✓	✗	✗
DeAnnCNV [27]	เว็บเซิร์ฟเวอร์	✓ (ใช้เพียงอัลกอริทึมจากเครื่องมือตนเองตรวจจับซีเอ็นวี)	✓	✗	✓
CNView [25]	แอปพลิเคชัน (คอมมานไลน์)	✓ (ใช้เพียงอัลกอริทึมจากเครื่องมือตนเองตรวจจับซีเอ็นวี)	✓	✗	✓
iCopyDAV [26]	แอปพลิเคชัน (คอมมานไลน์)	✓ (ใช้เพียงอัลกอริทึมจากเครื่องมือตนเองตรวจจับซีเอ็นวี)	✓	✓ (ไม่สามารถปรับแต่งได้)	✓
inCNV	เว็บแอปพลิเคชัน	✓ (ใช้อัลกอริทึมจากหลายเครื่องมือตรวจจับซีเอ็นวี)	✓	✓ (สามารถปรับแต่งได้)	✓

บทที่ 4
วิธีการดำเนินงานวิจัย

4.1 แนวคิดการรวมผลลัพธ์จากเครื่องตรวจจับซีเอ็นวีหลายตัว

4.1.1 ผลลัพธ์จากเครื่องตรวจจับซีเอ็นวี

จากการศึกษาผลลัพธ์ของเครื่องตรวจจับซีเอ็นวีหลายตัวพบว่าเครื่องมือตรวจจับซีเอ็นวีมักให้ผลลัพธ์เป็นไฟล์ข้อความที่มีแท็บเป็นตัวคั่น (tab-delimited file) และบางเครื่องมือแสดงผลลัพธ์ด้วยไฟล์วีซีเอฟ ส่งผลให้ผลลัพธ์ซีเอ็นวีเหล่านี้มีลักษณะการแสดงผลที่แตกต่างกัน และยากต่อการอ่านค่า และทำความเข้าใจ ดังรูปที่ 28 – 31

sample_name	chr	cnv	st_bp	ed_bp	length_kb	st_exon	ed_exon
G2223.remdup.uniqMap.TS.bam	chr1	dup	12908237	13036800	128.564	1366	1376
G2223.remdup.uniqMap.TS.bam	chr1	dup	25669451	25689044	19.594	2744	2749
G2223.remdup.uniqMap.TS.bam	chr1	del	110232892	110233186	0.295	8822	8823
G2223.remdup.uniqMap.TS.bam	chr1	dup	245912644	246093239	180.596	17886	17892
G2227-PJ.remdup.uniqMap.TS.bam	chr1	dup	16972863	17090975	118.113	1682	1714
G2227-PJ.remdup.uniqMap.TS.bam	chr1	dup	19655065	19666111	11.047	2070	2071
G2227-PJ.remdup.uniqMap.TS.bam	chr1	del	110232892	110233186	0.295	8822	8823
G2227-PJ.remdup.uniqMap.TS.bam	chr1	dup	160769542	160772481	2.94	12298	12299
G2227-PJ.remdup.uniqMap.TS.bam	chr1	del	40773863	40777784	3.922	4539	4548
G2227-PJ.remdup.uniqMap.TS.bam	chr1	dup	17322470	17322991	0.522	1757	1759
G2227-PJ.remdup.uniqMap.TS.bam	chr1	dup	248801587	248814185	12.599	18016	18017
G2227-PJ.remdup.uniqMap.TS.bam	chr1	dup	16817504	16918551	101.048	1666	1668
G2227-PJ.remdup.uniqMap.TS.bam	chr1	del	17264135	17275421	11.287	1724	1732
G2228-M.remdup.uniqMap.TS.bam	chr1	dup	21807422	21811392	3.971	2262	2264
G2228-M.remdup.uniqMap.TS.bam	chr1	del	220439520	220445843	6.324	16290	16291
G2228-M.remdup.uniqMap.TS.bam	chr1	del	110232892	110233186	0.295	8822	8823
G2229-F.remdup.uniqMap.TS.bam	chr1	del	171810620	172062015	251.396	13154	13166

รูปที่ 28 ตัวอย่างไฟล์ผลลัพธ์ที่ได้จากเครื่องมือตรวจจับซีเอ็นวี CODEX

sampleID	chromosome	start	stop	state
G2998-F.remdup.uniqMap.TS.bam.rpkm	chr1	148015624	148021628	del
G2223.remdup.uniqMap.TS.bam.rpkm	chr1	12855586	12939926	dup
G2223.remdup.uniqMap.TS.bam.rpkm	chr1	25678116	25683344	dup
G2223.remdup.uniqMap.TS.bam.rpkm	chr1	245848706	246490639	dup
G3586.remdup.uniqMap.TS.bam.rpkm	chr1	148008572	148015797	dup
G2309.remdup.uniqMap.TS.bam.rpkm	chr1	661139	853100	dup
G2309.remdup.uniqMap.TS.bam.rpkm	chr1	25669451	25689044	dup
G2309.remdup.uniqMap.TS.bam.rpkm	chr1	145323653	145413426	dup
G2309.remdup.uniqMap.TS.bam.rpkm	chr1	161479609	161559609	dup
G2361-F.remdup.uniqMap.TS.bam.rpkm	chr1	109490232	109742859	dup
G3154.remdup.uniqMap.TS.bam.rpkm	chr1	161559351	161599825	del
G3100.remdup.uniqMap.TS.bam.rpkm	chr1	161487764	161520413	del
G2516.remdup.uniqMap.TS.bam.rpkm	chr1	22310186	22336350	del
G2516.remdup.uniqMap.TS.bam.rpkm	chr1	104120111	104163306	dup
G2516.remdup.uniqMap.TS.bam.rpkm	chr1	155183616	155204891	dup
G2227-PJ.remdup.uniqMap.TS.bam.rpkm	chr1	16957384	16976914	dup
G2227-PJ.remdup.uniqMap.TS.bam.rpkm	chr1	17081401	17087368	dup

รูปที่ 29 ตัวอย่างไฟล์ผลลัพธ์ที่ได้จากเครื่องมือตรวจจับซีเอ็นวี CoNIFER

Targeted.Region.ID	Exon.Nu	Gene.Sym	Chr	OriStCoordinate	OriEndCoordinate	Mean.o	Adjust	SD.of.Log	Media	number	P.Value	Adjustec	gain.loss
1	0	unknown	chr1	30210	30330	1.023	0.525	0.132	1.04	120	0.267593	0.9999	gain
2	0	unknown	chr1	30552	30672	0.609	0.052	0.141	0.63	120	0.912766	0.9999	gain
3	0	unknown	chr1	69068	70028	0.282	-0.29	0.274	0.27	960	0.534888	0.9999	loss
4	0	unknown	chr1	367647	368607	-0.105	-0.31	0.335	-0.04	960	0.506711	0.9999	loss
5	0	unknown	chr1	565985	566465	0.889	-1.15	0.386	1.04	480	0.015319	0.9999	loss
6	0	unknown	chr1	621084	622044	-0.086	-0.34	0.355	-0.11	960	0.46869	0.9999	loss
7	0	unknown	chr1	861236	861476	-0.11	0.596	0.472	-0	239	0.370231	0.9999	gain
8	0	unknown	chr1	865564	865744	0.292	1.515	0.511	0.35	141	NA	NA	gain
9	0	unknown	chr1	866323	866563	1.04	2.416	0.517	1.13	140	NA	NA	gain
10	0	unknown	chr1	871093	871333	-0.373	0.585	0.667	-0.09	191	0.470104	0.9999	gain
11	0	unknown	chr1	874343	874583	-0.233	0.773	0.751	-0.42	240	0.355972	0.9999	gain
12	0	unknown	chr1	874626	874866	-1.482	0.073	1.826	-0.7	88	NA	NA	gain
13	0	unknown	chr1	876484	876724	0.105	1.553	0.208	0.06	54	NA	NA	gain
16	0	unknown	chr1	877887	878487	-0.553	0.925	0.998	-0.54	153	NA	NA	gain
17	0	unknown	chr1	878574	878814	-0.285	0.985	0.963	-0.07	237	NA	NA	gain
18	0	unknown	chr1	879012	879589	-1.009	0.077	1.051	-0.99	450	NA	NA	gain
19	0	unknown	chr1	880006	880246	-1.165	0.403	0.876	-1.49	55	NA	NA	gain
20	0	unknown	chr1	880844	881084	0.379	0.953	0.49	0.3	240	0.105936	0.9999	gain
21	0	unknown	chr1	881488	881728	-0.77	-0	0.519	-0.77	240	0.997581	0.9999	loss
22	0	unknown	chr1	881732	881972	-0.716	0.331	0.727	-0.74	233	NA	NA	gain

รูปที่ 30 ตัวอย่างไฟล์ผลลัพธ์ที่ได้จากเครื่องมือตรวจจับซีเอ็นวี CONTRA

##fileformat=VCFv4.1
##ALT=<ID=DIP,Description="Diploid copy number">
##ALT=<ID=CNV,Description="Copy Number Polymorphism">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplication">
##INFO=<ID=AC,Number=2,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=2,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=END,Number=1,Type=Integer,Description="End coordinate of this variant">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=TPOS,Number=1,Type=Integer,Description="Start coordinate of target used to genotype this variant">
##INFO=<ID=TEND,Number=1,Type=Integer,Description="End coordinate of target used to genotype this variant">
##INFO=<ID=NUMT,Number=1,Type=Integer,Description="Number of targets used to genotype this variant">
##INFO=<ID=GQT,Number=1,Type=Float,Description="CNV-specific genotyping quality threshold, calculated as the minimal Q_EXACT of discovered
##INFO=<ID=PREVTARGSTART,Number=1,Type=Integer,Description="Start coordinate of target preceding the first target used to genotype this varia
##INFO=<ID=PREVTARGEND,Number=1,Type=Integer,Description="End coordinate of target preceding the first target used to genotype this variant"
##INFO=<ID=POSTTARGSTART,Number=1,Type=Integer,Description="Start coordinate of target following the last target used to genotype this varia
##INFO=<ID=POSTTARGEND,Number=1,Type=Integer,Description="End coordinate of target following the last target used to genotype this variant">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=NDQ,Number=1,Type=Float,Description="Phred-scaled quality of =N=on =D=iploidy">
##FORMAT=<ID=DQ,Number=1,Type=Float,Description="Phred-scaled quality of =D=iploidy">
##FORMAT=<ID=EQ,Number=2,Type=Float,Description="Phred-scaled qualities of =E=xact CNV event of allele types, in order given in ALT field">
##FORMAT=<ID=SQ,Number=2,Type=Float,Description="Phred-scaled qualities of =S=ome CNV event of allele types, in order given in ALT field">
##FORMAT=<ID=NQ,Number=2,Type=Float,Description="Phred-scaled qualities of =N=o CNV event of allele types, in order given in ALT field">
##FORMAT=<ID=LQ,Number=2,Type=Float,Description="Phred-scaled qualities of =L=eft breakpoint of CNV event of allele types, in order given in Al
##FORMAT=<ID=RQ,Number=2,Type=Float,Description="Phred-scaled qualities of =R=ight breakpoint of CNV event of allele types, in order given in
##FORMAT=<ID=PL,Number=3,Type=Float,Description="Normalized, Phred-scaled relative likelihoods for DIP,DEL,DUP genotypes, capped at 255">
##FORMAT=<ID=RD,Number=1,Type=Float,Description="Mean Read Depth over region">
##FORMAT=<ID=ORD,Number=1,Type=Float,Description="Mean Original (unnormalized) Read Depth over region">
##FORMAT=<ID=DSCVR,Number=1,Type=Character,Description="Was this CNV locus discovered in this sample? (Y or N)">

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
22	17071768	22:17071768-17073440	<DIP>	<DEL>,<DUP:	.	.	AC=0,1;AF=0.00,0.03;AN=30;END=17073440;
22	18898402	22:18898402-18913235	<DIP>	<DEL>,<DUP:	.	.	AC=1,0;AF=0.03,0.00;AN=30;END=18913235;

รูปที่ 31 ตัวอย่างไฟล์ผลลัพธ์ที่ได้จากเครื่องมือตรวจจับซีเอ็นวี XHMM

4.1.2 ความสัมพันธ์ของผลลัพธ์จากเครื่องตรวจจับซีเอ็นวีแต่ละตัว

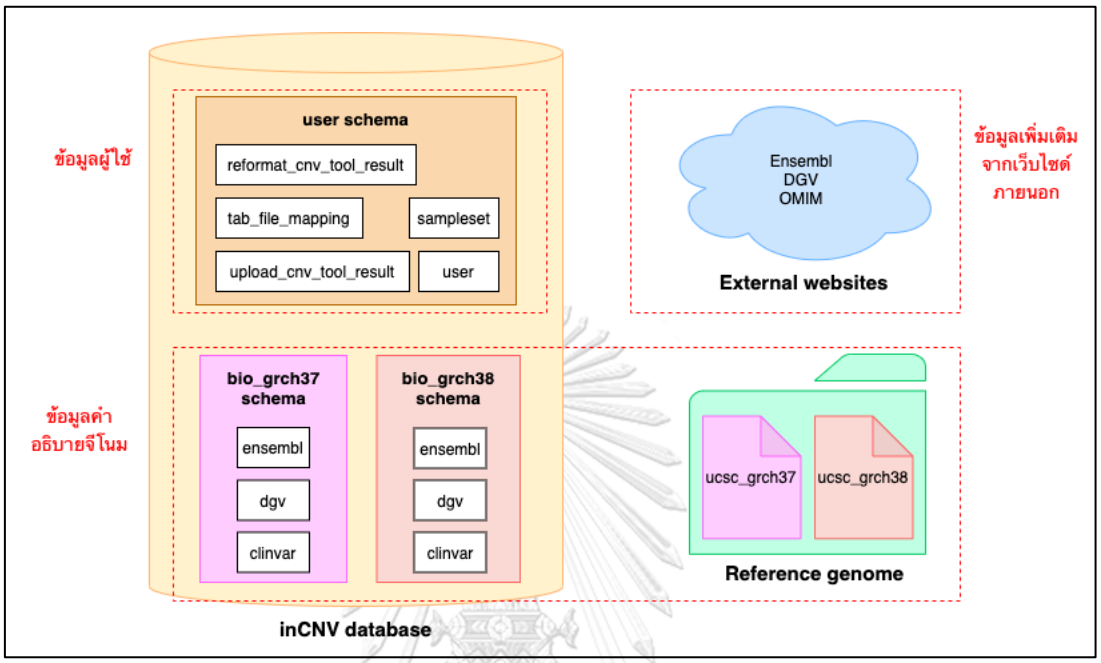
จากรูปที่ 28-31 สามารถสรุปความสัมพันธ์ในการแสดงไฟล์ผลลัพธ์ของซีเอ็นวีจากเครื่องมือแต่ละเครื่องมือได้ตามตารางที่ 4 ข้อมูลในตารางที่ 4 แสดงให้เห็นว่าผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวีแต่ละเครื่องมือมีความคล้ายคลึงกัน กล่าวคือ มีข้อมูลชื่อตัวอย่าง ข้อมูลโครโมโซม ข้อมูลตำแหน่งเบสของซีเอ็นวี และข้อมูลประเภทของซีเอ็นวี แต่ถูกนำเสนอด้วยรูปแบบที่แตกต่างกัน เช่น ชื่อฟิลด์ที่มีความหมายเหมือนกันถูกนำเสนอด้วยคำที่แตกต่างกัน จากความสัมพันธ์นี้ทำให้ผู้วิจัยสามารถสร้างระบบวิเคราะห์ซีเอ็นวีจากเครื่องมือตรวจจับซีเอ็นวีใด ๆ หรือรวมผลลัพธ์ซีเอ็นวีจากเครื่องตรวจจับซีเอ็นวีหลายตัวเข้าด้วยกันได้ โดยการสร้างโมดูลการเตรียมข้อมูลนำเข้าเพื่อให้ซอฟต์แวร์อินซีเอ็นวีสามารถเข้าใจไฟล์ผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจจับเครื่องใด ๆ ซึ่งจะถูกกล่าวถึงในหัวข้อชื่อ “5.1 โมดูลการเตรียมข้อมูลนำเข้า”

ตารางที่ 4 ตัวอย่างความสัมพันธ์ของฟิลด์ผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวีแต่ละตัว

		เครื่องมือตรวจจับซีเอ็นวี			
		CODEX	CoNIFER	CONTRA	XHMM
ชื่อฟิลด์	ชื่อตัวอย่าง	sample_name	sampleID	Target.Region.ID	-
	โครโมโซม	chr	chromosome	chr	CHROM
	ตำแหน่งเริ่มต้น	st_bp	start	OriStrCoordinate	POS
	ตำแหน่งสิ้นสุด	ed_bp	stop	OriEndCoordinate	END
	ประเภทของซีเอ็นวี	cnv	state	gain.loss	ALT
ข้อมูลผลลัพธ์ตามชื่อฟิลด์	โครโมโซม	chr22	chr22	chr22	22
	ประเภทของซีเอ็นวี	dup	dup	gain	<DUP>
		del	del	loss	<DEL>
		-	-	-	<DUP>,<DEL>
ประเภทไฟล์		txt	txt	txt	VCF

4.2 การออกแบบฐานข้อมูล

ซอฟต์แวร์อินซีเอ็นวีแบ่งข้อมูลตามฟังก์ชันการทำงานเป็น 3 ส่วนหลัก คือ ข้อมูลผู้ใช้ ข้อมูลคำอธิบายจีโนม และข้อมูลเพิ่มเติมจากเว็บไซต์ภายนอก ดังรูปที่ 32



รูปที่ 32 แผนภาพแสดงการแบ่งข้อมูลที่ใช้ในซอฟต์แวร์อินซีเอ็นวีตามฟังก์ชันการทำงาน

4.2.1 ข้อมูลผู้ใช้

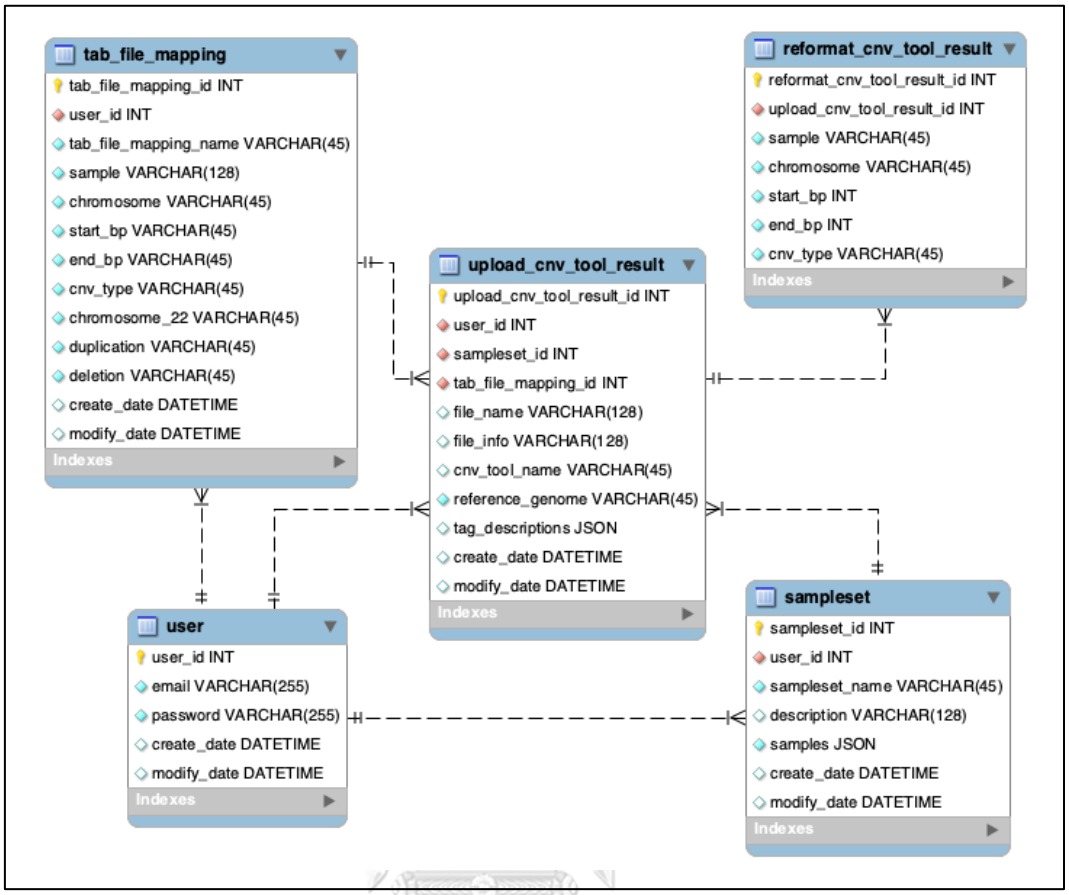
ระบบจะทำการเก็บข้อมูลผู้ใช้ลงในฐานข้อมูลอินซีเอ็นวี (inCNV database) ภายในสคีมาผู้ใช้ (user schema) ซึ่งประกอบด้วยตารางดังนี้

- ตารางผู้ใช้งานระบบ (user) เก็บข้อมูลทั่วไปของผู้ใช้ เช่น อีเมล และรหัสผ่าน
- ตารางเทมเพลตของเครื่องมือตรวจจับซีเอ็นวี (tab_file_mapping) เก็บข้อมูลเทมเพลตของเครื่องมือตรวจจับซีเอ็นวีที่ผู้ใช้สร้างขึ้น สามารถดูข้อมูลเพิ่มเติมได้ที่หัวข้อชื่อ “5.1.2 เทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวี”
- ตารางเทมเพลตของกลุ่มตัวอย่าง (sampleset) เก็บข้อมูลเทมเพลตกลุ่มตัวอย่างที่ผู้ใช้สร้างขึ้น สามารถดูข้อมูลเพิ่มเติมได้ที่หัวข้อชื่อ “5.1.3 เทมเพลตกลุ่มตัวอย่าง”
- ตารางไฟล์ผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวี (upload_cnv_tool_result) เก็บข้อมูลทั่วไปของไฟล์ผลลัพธ์ซีเอ็นวีที่ผู้ใช้อัพโหลดเข้าระบบ เช่น อัพโหลดโดยใคร

ไฟล์มีชื่อว่าอะไร ผลลัพธ์นี้ได้มาจากเครื่องมือตรวจจับซีเอ็นวีเครื่องมือใด และต้องใช้ “เทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวี” รวมถึง “เทมเพลตกลุ่มตัวอย่าง” ใดในการอ่านไฟล์นี้

- ตารางข้อมูลผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวีที่ถูกปรับเปลี่ยนโครงสร้างแล้ว (reformat_cnv_tool_result) เก็บข้อมูลผลลัพธ์ซีเอ็นวีที่ได้จากการปรับเปลี่ยนโครงสร้างข้อมูลของไฟล์ที่ถูกอัพโหลดโดยการแมปด้วย “เทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวี” และ “เทมเพลตกลุ่มตัวอย่าง” และข้อมูลของตารางนี้จะมีความสัมพันธ์กับ “ตารางไฟล์ผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวี” ในลักษณะ cascade หากมีการลบข้อมูลทั่วไปของไฟล์ใน “ตารางไฟล์ผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวี” จะส่งผลให้ข้อมูลผลลัพธ์ซีเอ็นวีของไฟล์นั้นซึ่งถูกบรรจุใน “ตารางข้อมูลไฟล์ผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวีที่ถูกปรับเปลี่ยนโครงสร้างแล้ว” ถูกลบด้วยโดยอัตโนมัติ

รูปที่ 33 แสดงแผนภาพอีอาร์ (ER diagram) ซึ่งบอกรายละเอียดแอตทริบิวต์ (attribute) และความสัมพันธ์ (relationship) ของตารางต่าง ๆ ในฐานข้อมูลผู้ใช้



รูปที่ 33 แผนภาพอีอาร์ของสคีมาผู้ใช้ (user schema)

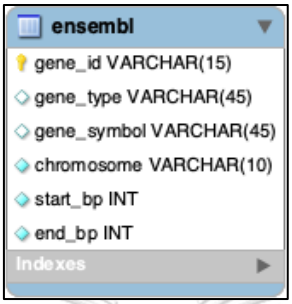
4.2.2 ข้อมูลคำอธิบายจีโนม (Genome annotation)

เนื่องจากเครื่องมือตรวจจับซีเอ็นวีส่วนใหญ่ไม่มีฟังก์ชันแสดงคำอธิบายจีโนมให้กับซีเอ็นวีที่ตรวจจับได้ ทำให้ขาดความสะดวกในการใช้งาน ผู้ใช้ต้องทำการค้นหาความหมาย และความสำคัญของซีเอ็นวีตำแหน่งใด ๆ ผ่านทางฐานข้อมูลสาธารณะด้วยตนเองทีละตำแหน่ง ดังนั้น ผู้วิจัยจึงได้ศึกษา และรวมรวมข้อมูลจากฐานข้อมูลสาธารณะที่มีความน่าเชื่อถือ และมีความหลากหลายมาใช้งานร่วมกันกับการหาตำแหน่งซีเอ็นวีเพื่อให้ซอฟต์แวร์อินซีเอ็นวีสามารถให้คำอธิบายจีโนมที่มีความหมาย และความหลากหลายแก่ผู้ใช้งานได้

อย่างไรก็ตามตำแหน่งของคำอธิบายจีโนมมีความสัมพันธ์กับจีโนมอ้างอิง ดังนั้น ในงานวิจัยนี้ จึงทำการแบ่งข้อมูลคำอธิบายจีโนมเป็น 2 สคีมาตามเวอร์ชันหลักของจีโนมอ้างอิง คือ “bio_grch37 schema” สำหรับจีโนมอ้างอิง GRCh37 และ “bio_grch38 schema” สำหรับจีโนมอ้างอิง GRCh38 ซึ่งฐานข้อมูลทั้งสองจะประกอบไปด้วยคำอธิบายจีโนมดังนี้

4.2.2.1 คำอธิบายชื่อยีน

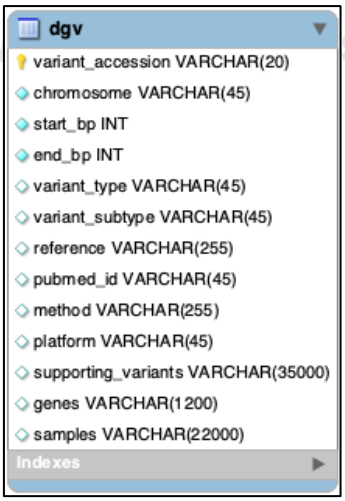
ซอฟต์แวร์อินซีเอ็นวีเลือกใช้ข้อมูลคำอธิบายชื่อยีนจากฐานข้อมูลสาธารณะอองซอมเบลอ (Ensembl) เพื่อระบุชื่อยีนที่ตรงกับตำแหน่งซีเอ็นวีที่ตรวจจับได้ โดยข้อมูลที่ซอฟต์แวร์อินซีเอ็นวีนำมาใช้ประกอบคำอธิบายจีโนมแสดงในรูปที่ 33



รูปที่ 34 ตาราง ensemble เก็บข้อมูลจากฐานข้อมูลอองซอมเบลอโดยแสดงแอตทริบิวต์ที่ซอฟต์แวร์อินซีเอ็นวีนำมาใช้ประกอบคำอธิบายจีโนม

4.2.2.2 คำอธิบายการแปรผันของคนปกติ

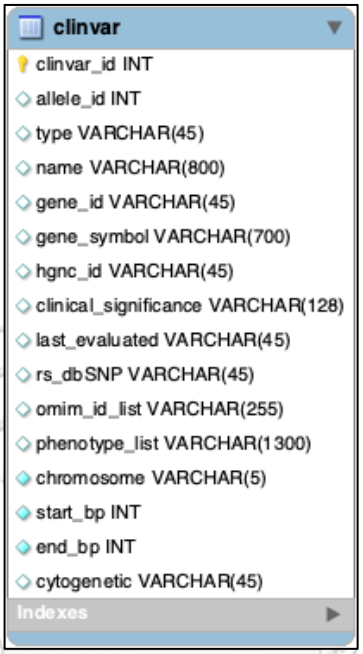
ซอฟต์แวร์อินซีเอ็นวีเลือกใช้ข้อมูลคำอธิบายการแปรผันของคนปกติจากฐานข้อมูลสาธารณะดีจีวี (DGV) เพื่อวิเคราะห์ตำแหน่งซีเอ็นวีที่ตรวจจับได้ว่าเคยมีการรายงานมาก่อนในฐานข้อมูลอ้างอิงดีจีวีหรือไม่ และเพื่อใช้ในการคัดกรองซีเอ็นวีตำแหน่งใหม่ที่ยังไม่เคยถูกรายงาน (novel CNVs) โดยข้อมูลที่ซอฟต์แวร์อินซีเอ็นวีนำมาใช้ประกอบคำอธิบายจีโนมมีดังรูปที่ 35



รูปที่ 35 ตาราง dgv เก็บข้อมูลจากฐานข้อมูลดีจีวีโดยแสดงแอตทริบิวต์ที่ซอฟต์แวร์อินซีเอ็นวีนำมาใช้ประกอบคำอธิบายจีโนม

4.2.2.3 คำอธิบายการแปรผันที่เกี่ยวกับโรค

ซอฟต์แวร์อินซีเอ็นวีเลือกใช้ข้อมูลคำอธิบายการแปรผันที่เกี่ยวกับโรคจากฐานข้อมูลสาธารณะคลินวาร์ (ClinVar) เพื่อวิเคราะห์ว่าซีเอ็นวีที่ตรวจจับได้เกี่ยวข้องกับโรคอะไรหรือไม่ โดยข้อมูลที่ซอฟต์แวร์อินซีเอ็นวีนำมาใช้ประกอบคำอธิบายจีโนมมีดังรูปที่ 36



รูปที่ 36 ตาราง clinvar จากฐานข้อมูลคลินวาร์แสดงแอตทริบิวต์ที่ซอฟต์แวร์อินซีเอ็นวีนำมาใช้ประกอบคำอธิบายจีโนม

4.2.2.4 ลำดับเบสจีโนมอ้างอิงของมนุษย์ (Human reference genome sequences)

ซอฟต์แวร์อินซีเอ็นวีใช้ข้อมูลลำดับเบสจีโนมอ้างอิงของมนุษย์จากบราวเซอร์จีโนมของมหาวิทยาลัยแห่งแคลิฟอเนียร์ซานตาครูส (UCSC genome browser) เพื่อใช้สกัดลำดับเบสบนจีโนมที่อยู่ก่อนหน้าและอยู่ต่อท้าย (flanking sequences) จากบริเวณที่ถูกสงสัยว่าเกิดซีเอ็นวี เพื่อให้นักวิจัยสามารถนำลำดับเบสเหล่านี้ไปใช้เป็นสายดีเอ็นเอตั้งต้นในการออกแบบไพรเมอร์ (primer) และถอดรหัสลำดับเบสในบริเวณดังกล่าวในห้องปฏิบัติการ (wet lab) เพื่อใช้ในการตรวจสอบและยืนยันซีเอ็นวีที่ตรวจพบ ทั้งนี้คำอธิบายลำดับเบสของจีโนมอ้างอิงไม่ได้ถูกจัดเก็บในฐานข้อมูลของอินซีเอ็นวีแต่อยู่ในรูปแบบไฟล์

4.2.3 ข้อมูลเพิ่มเติมจากเว็บไซด์ภายนอก

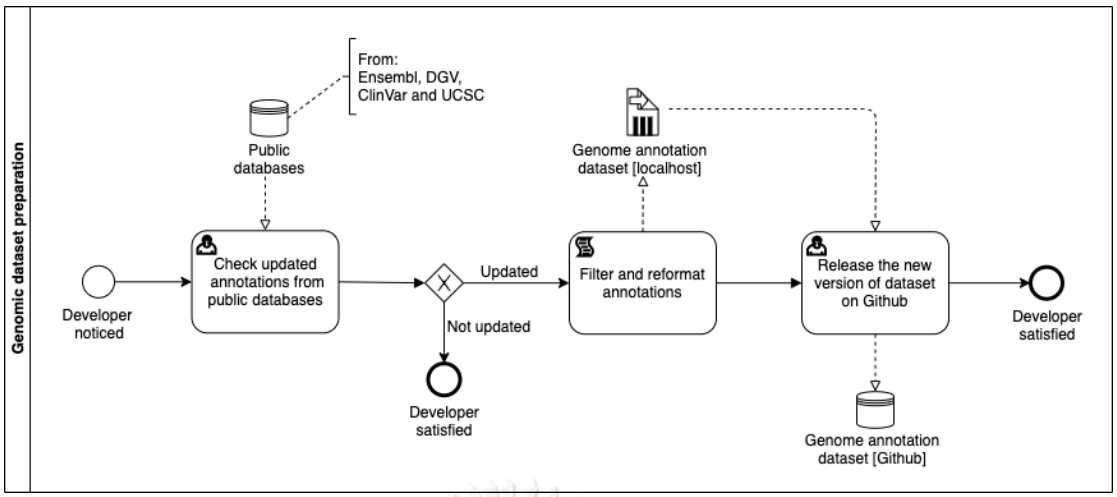
ซอฟต์แวร์อินซีเอ็นวีอนุญาตให้ผู้ใช้สามารถเข้าถึงข้อมูลนอกเหนือจากคำอธิบายจีโนมที่รวบรวมไว้ในระบบ โดยการเตรียมลิงค์เชื่อมต่อซอฟต์แวร์อินซีเอ็นวีกับฐานข้อมูลสาธารณะทางอินเตอร์เน็ตที่สอดคล้องกับตำแหน่งซีเอ็นวีที่ทำนายไว้ ซึ่งจะอธิบายต่อไปในหัวข้อการวิเคราะห์ข้อมูลซีเอ็นวีแบบบูรณาการ (Integrated CNV analyses)

4.3 การจัดการคำอธิบายจีโนม

จากการศึกษาคำอธิบายจีโนม พบว่าข้อมูลคำอธิบายจีโนมที่จะนำเข้ามาในระบบมีหลายประเภท มาจากหลายฐานข้อมูล มีลักษณะไฟล์หลายรูปแบบ และข้อมูลมักมีการปรับปรุงแก้ไขอยู่เสมอ ดังนั้นสำหรับงานวิจัยนี้ซอฟต์แวร์อินซีเอ็นจึงได้ถูกออกแบบให้สามารถปรับปรุงแก้ไขข้อมูลเหล่านี้ได้เพื่อรักษาผลลัพธ์จากการวิเคราะห์ให้มีความทันสมัย และน่าเชื่อถือต่อไปในอนาคต ซึ่งมีขั้นตอนการทำงานดังนี้

4.3.1 การเตรียมคำอธิบายจีโนม

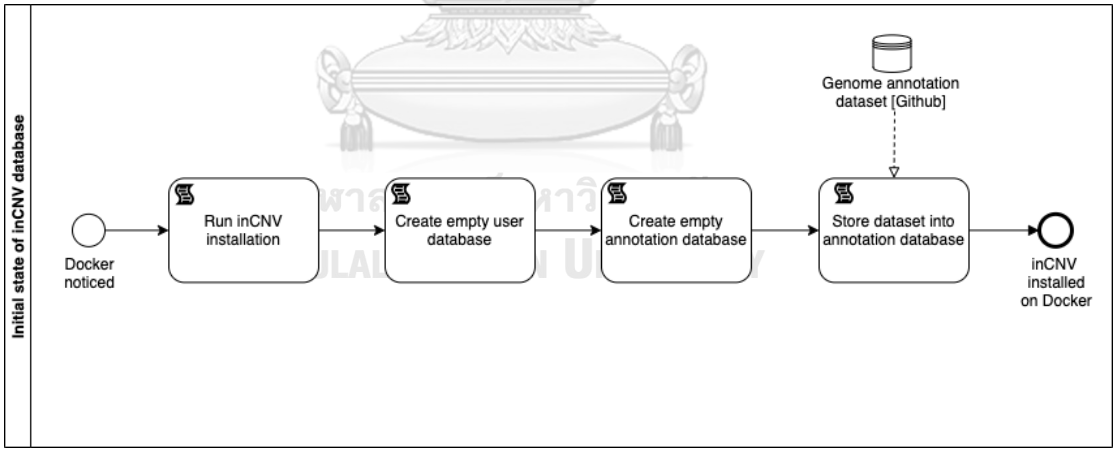
เริ่มต้นผู้วิจัยระบบรวบรวมข้อมูลจากฐานข้อมูลอองซอมเบลอ (Ensembl) ฐานข้อมูลดีจีวี (DGV) ฐานข้อมูลคลินวาร์ (ClinVar) และ ฐานข้อมูลจีโนมจากมหาวิทยาลัยแห่งแคลิฟอเนียร์ซานตาครูส มาแปลงข้อมูลให้อยู่ในรูปแบบที่ต้องการ และเรียกข้อมูลเหล่านั้นว่า “กลุ่มข้อมูลของคำอธิบายจีโนม (Genome annotation dataset)” แล้วอัพโหลดข้อมูลเหล่านั้นไว้บน GitHub repository หากในอนาคตผู้วิจัยทำการตรวจสอบข้อมูลจากฐานข้อมูลดังกล่าวแล้วพบว่ามีการปรับปรุงก็จะทำตามขั้นตอนเดิม คือ รวบรวมข้อมูลที่ต้องการแล้วอัพโหลดขึ้นบน GitHub repository ต่อไป รายละเอียดได้อธิบายไว้ในแผนภาพบีพีเอ็มเอ็น (business process model and notation: BPMN) ดังรูปที่ 37



รูปที่ 37 แผนภาพบีพีเอ็มเอ็นของการเตรียมคำอธิบายจีโนม

4.3.2 การติดตั้งคำอธิบายจีโนม

เมื่อผู้ใช้ติดตั้งซอฟต์แวร์อินซีเอ็นวี ตัวซอฟต์แวร์อินซีเอ็นวีจะเรียกดาวน์โหลดชุดข้อมูลคำอธิบายจีโนมจาก GitHub ผ่านทาง GitHub API แล้วนำข้อมูลเหล่านั้นใส่ลงในฐานข้อมูลคำอธิบายจีโนมของซอฟต์แวร์อินซีเอ็นวี ดังรูปที่ 38

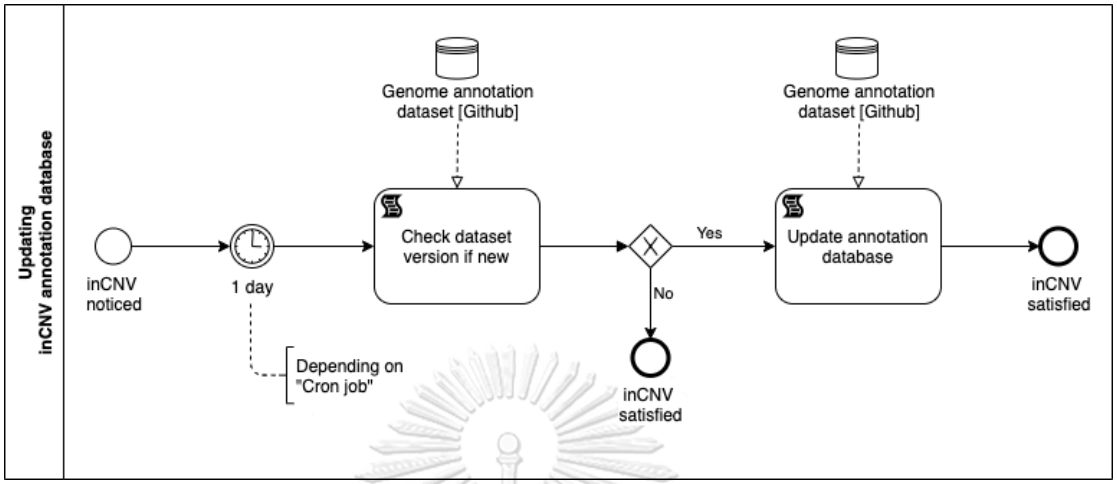


รูปที่ 38 แผนภาพบีพีเอ็มเอ็นแสดงสถานะเริ่มต้นของฐานข้อมูลซอฟต์แวร์อินซีเอ็นวี

4.3.3 การอัพเดทคำอธิบายจีโนม

ซอฟต์แวร์อินซีเอ็นวีจะมีตรวจสอบคำอธิบายจีโนมบน GitHub ว่ามีเวอร์ชันใหม่ หรือไม่ตามตารางเวลาที่กำหนดไว้ใน “Cron job” ซึ่งซอฟต์แวร์อินซีเอ็นวีจะกำหนดค่าเริ่มต้นเป็นทุกเที่ยงคืนตามเวลาท้องที่บนเครื่องคอมพิวเตอร์ หรือเครื่องเซิร์ฟเวอร์ที่ซอฟต์แวร์อินซีเอ็นวีถูกติดตั้ง (ผู้ใช้

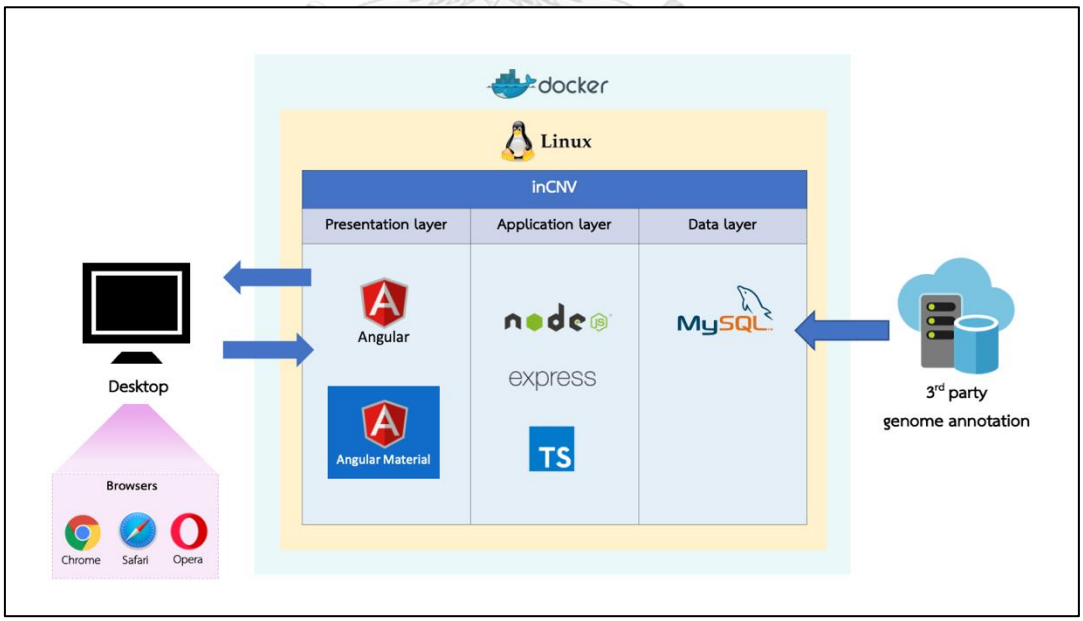
สามารถปรับเปลี่ยนตารางเวลาได้ในภายหลัง) หากพบว่าชุดข้อมูลพันธุกรรมมีการปรับปรุงเป็นเวอร์ชันใหม่ซอฟต์แวร์อินซีเอ็นวีก็จะทำการดาวน์โหลดชุดข้อมูลเหล่านั้นติดตั้งในระบบ ดังรูปที่ 39



รูปที่ 39 แผนภาพบีพีเอ็มเอ็นแสดงการอัพเดทฐานข้อมูลคำอธิบายจีโนม

4.4 เทคโนโลยีที่ใช้

ซอฟต์แวร์อินซีเอ็นวีเป็นเว็บแอปพลิเคชันถูกออกแบบด้วยสถาปัตยกรรม 3 เลเยอร์ (3-layer architecture) และติดตั้งผ่านทางด็อกเกอร์ (docker) (รูปที่ 40)



รูปที่ 40 แผนภาพแสดงเทคโนโลยีที่ใช้ของระบบ

4.4.1 สถาปัตยกรรม 3 เลเยอร์ (3-layer architecture)

ซอฟต์แวร์อินซีเอ็นวีออกแบบตามสถาปัยกรรม 3 เลเยอร์ โดยแบ่งเป็นชั้นแสดงผล (presentation layer) สถาปัตยกรรมชั้นโปรแกรมประยุกต์ (application layer) และชั้นข้อมูล (data layer) ดังนี้

4.4.1.1 ชั้นแสดงผล (Presentation layer)

ในชั้นแสดงผลซอฟต์แวร์อินซีเอ็นวีพัฒนาเว็บไซด์ด้วยแองกูลาร์เฟรมเวิร์ค (angular framework) เวอร์ชัน 9.0.0 เพื่อสร้างเว็บแอปพลิเคชันแบบหน้าเดียว (single page applications: SPAs) และในส่วนของตัวปฏิสัมพันธ์กับผู้ใช้ (graphical user interface: GUI) ซอฟต์แวร์อินซีเอ็นวีใช้ไลบรารีของแองกูลาร์แมตทีเรียล (angular material library) เวอร์ชัน 9.0.0 และดีทรีดอทเจเอส (d3.js library) เวอร์ชัน 5.14.2

ซอฟต์แวร์อินซีเอ็นวีเวอร์ชันปัจจุบันสนับสนุนบราวเซอร์ของเครื่องคอมพิวเตอร์ตั้งโต๊ะ (desktop / notebook) ด้วยกัน 3 ตัว คือ บราวเซอร์โครม (Chrome browser) บราวเซอร์โอเปรา (Opera browser) และบราวเซอร์ซาฟารี (Safari browser)

4.4.1.2 ชั้นโปรแกรมประยุกต์ (Application layer)

ซอฟต์แวร์อินซีเอ็นวีพัฒนาสถาปัตยกรรมชั้นโปรแกรมประยุกต์ด้วยจาวาสคริปรันไทม์ (javascript runtime) ที่ชื่อว่าโหนดเจเอส (NodeJS) โดยใช้เอ็กเพรสเฟรมเวิร์ค (express framework) พัฒนาร่วมกับภาษาไทป์สคริป (typescript) รวมถึงทำงานร่วมกับแพ็กเกจ indexedfasta-js เวอร์ชัน 1.0.12 จาก JBrowse [81] เพื่ออ่านไฟล์ฟอร์แมตฟาสตา (FASTA file format)

4.4.1.3 ชั้นข้อมูล (Data layer)

ซอฟต์แวร์อินซีเอ็นวีใช้มายเอสคิวแอล (MySql) เป็นระบบจัดการฐานข้อมูล (database management system: DBMS)

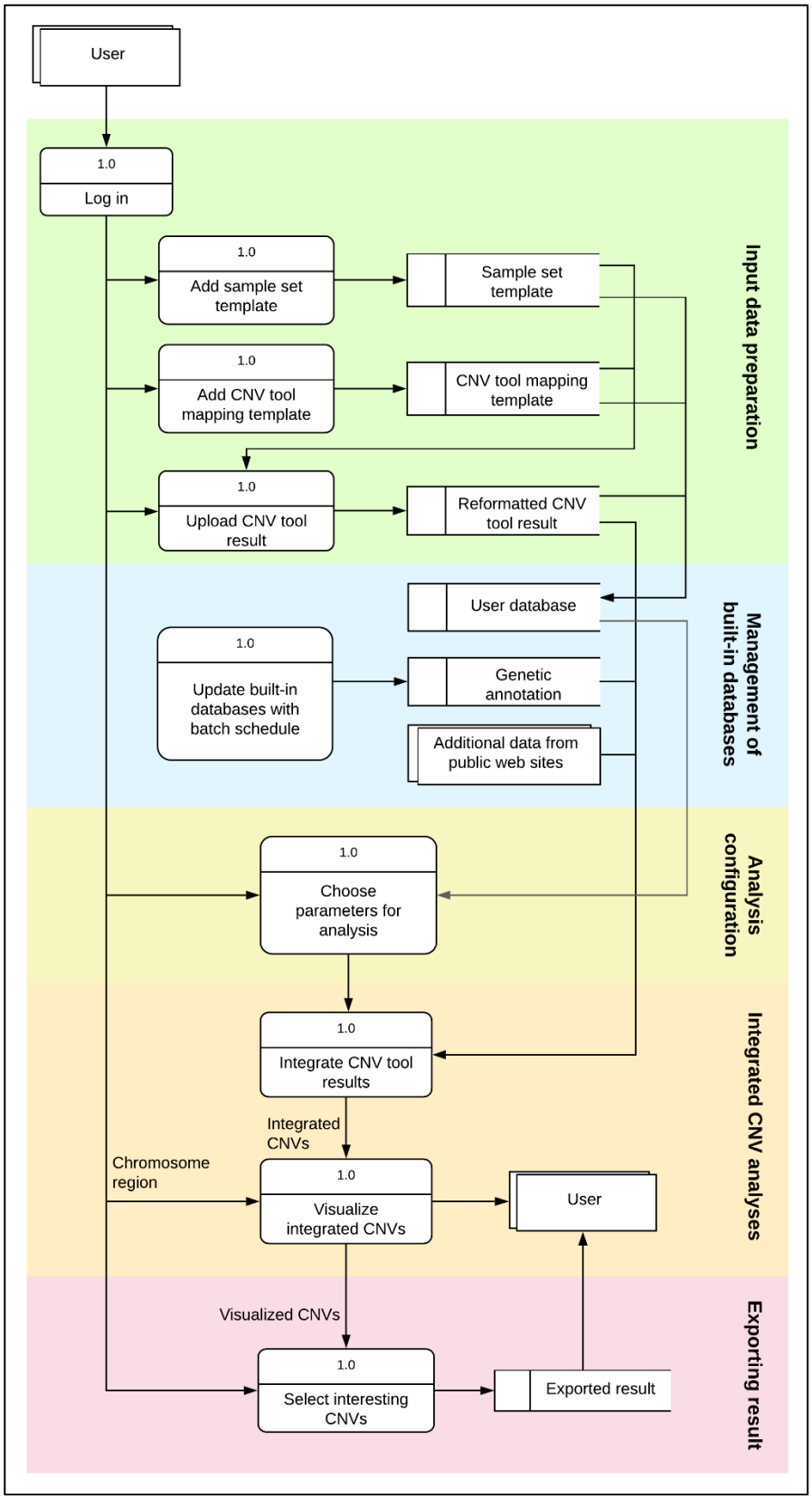
4.4.2 ด็อกเกอร์ (Docker)

ผู้วิจัยบรรจุซอฟต์แวร์อินซีเอ็นวีไว้ในด็อกเกอร์อิมเมจ (docker image) เพื่อให้ผู้ใช้สามารถติดตั้งได้โดยง่าย ผู้ใช้สามารถติดตั้งโดยการโคลนด็อกเกอร์อิมเมจของซอฟต์แวร์อินซีเอ็นวีได้ที่ https://github.com/saowwapark/inCNV แล้วติดตั้งผ่านทางด็อกเกอร์เอ็นจิน (docker engine) บนเครื่องผู้ใช้ (desktop / notebook) หรือบนเซิร์ฟเวอร์ (server) บนแพล็ตฟอร์มใด ๆ ก็ได้ (cross-platform software)

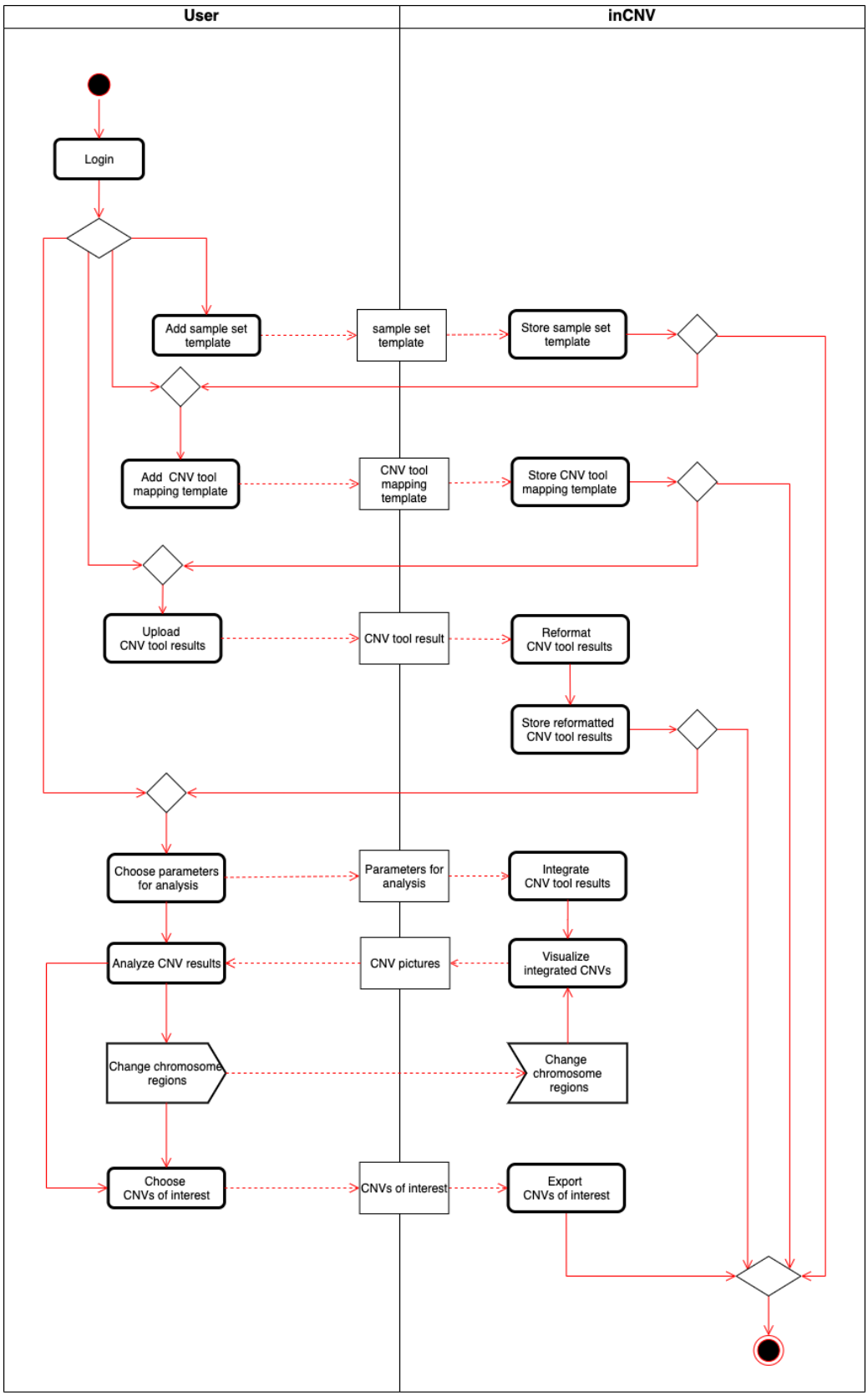
4.5 ภาพรวมการทำงานของซอฟต์แวร์อินซีเอ็นวี

จากวิธีการดำเนินการวิจัยทั้งหลายที่ได้กล่าวถึงในหัวข้อก่อนหน้านี้ ผู้วิจัยจึงได้นำเสนอซอฟต์แวร์อินซีเอ็นวีที่ใช้ในการบูรณาการข้อมูลในการวิเคราะห์ผลซีเอ็นวี โดยมีภาพรวมการทำงานดังแสดงในรูปที่ 41 – 42

จากรูปที่ 41 ซอฟต์แวร์อินซีเอ็นวีถูกแบ่งออกเป็น 5 โมดูลย่อยได้แก่ (1) โมดูลการเตรียมข้อมูลนำเข้า (Input data preparation module) (2) โมดูลการจัดการฐานข้อมูลในระบบ (Management of built-in databases module) (3) โมดูลโครงแบบการวิเคราะห์ (Analysis configuration module) (4) โมดูลการวิเคราะห์ซีเอ็นวีแบบบูรณาการ (Integrated CNV analyses module) และ (5) โมดูลการส่งออกผลลัพธ์ (Exporting result module) ซึ่งจะอธิบายรายละเอียดต่อไปในหัวข้อ “บทที่ 5 ผลการวิจัย และการพัฒนาระบบ”



รูปที่ 41 แผนภาพกระแสข้อมูล (data flow diagram) ระหว่างโมดูลหลักของซอฟต์แวร์อินซีเอ็นวี



รูปที่ 42 แผนภาพกิจกรรม (activity diagram) แสดงภาพรวมการทำงานของซอฟต์แวร์อินซีเอ็นวี

บทที่ 5
ผลการวิจัย และการพัฒนาระบบ

5.1 โมดูลการเตรียมข้อมูลนำเข้า (Input data preparation module)

ซอฟต์แวร์อินซีเอ็นวีมีรูปแบบการนำเข้าไฟล์ผลลัพธ์ที่มีความยืดหยุ่น ผู้ใช้สามารถนำเข้าไฟล์ผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวีเครื่องมือใด ๆ ก็ได้ที่มีเทมเพลต (template) ตามที่ผู้ใช้ได้เคยกำหนดไว้ โดยการนำเข้าข้อมูลของซอฟต์แวร์อินซีเอ็นวีแบ่งออกได้เป็น 4 ส่วนหลัก ๆ ดังนี้

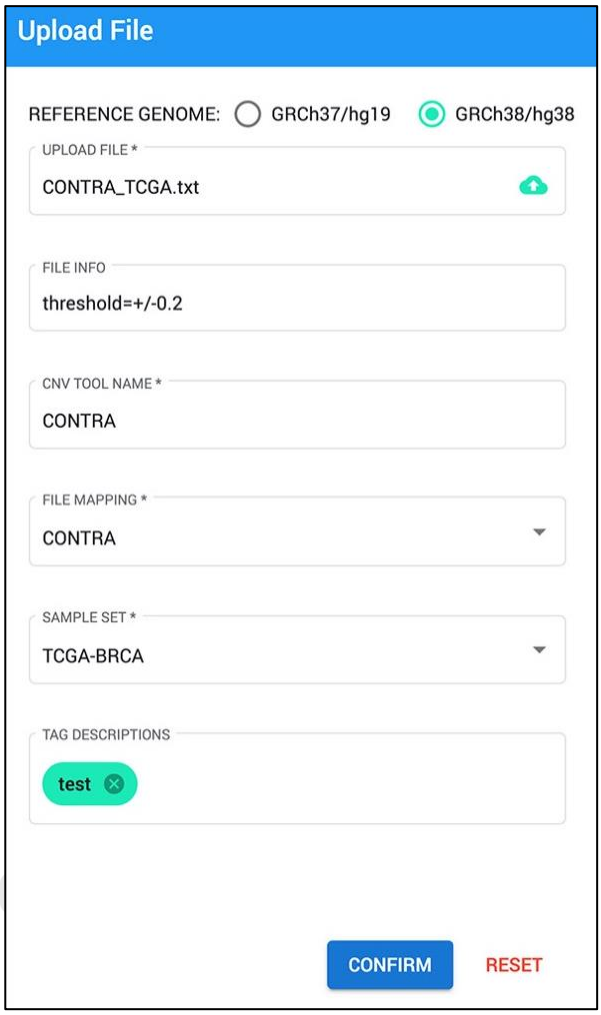
5.1.1 การอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี (Uploading CNV tool results)

องค์ประกอบนี้อนุญาตให้ผู้ใช้อัพโหลดไฟล์ผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวีเพื่อนำไปวิเคราะห์หาซีเอ็นวีที่น่าเชื่อถือต่อไป ซึ่งมีเงื่อนไขการทำงานดังต่อไปนี้ (รูปที่ 43)

1. หัวข้อ “reference genome” สำหรับให้ผู้ใช้ระบุเวอร์ชันของจีโนมอ้างอิงของมนุษย์ที่เครื่องมือตรวจจับซีเอ็นวีใช้ในการตรวจจับซีเอ็นวี
2. หัวข้อ “upload file” สำหรับให้ผู้ใช้อัพโหลดไฟล์ผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวีเข้าซอฟต์แวร์อินซีเอ็นวี โดยไฟล์ที่นำเข้านี้จะต้องเป็นไฟล์ข้อความธรรมดา (plain text) นามสกุล “.txt” ในรูปแบบคั่นด้วยแท็บ (tab-delimited format)
3. หัวข้อ “CNV tool name” สำหรับ ให้ผู้ใช้ระบุชื่อเครื่องมือตรวจจับซีเอ็นวีที่ใช้ในการให้ได้มาซึ่งไฟล์ผลลัพธ์ที่ถูกอัพโหลดเข้าระบบ
4. หัวข้อ “file mapping” และ “sample set” สำหรับให้ผู้ใช้เลือกเทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวี และเลือกเทมเพลตกลุ่มตัวอย่าง (อธิบายในหัวข้อถัดไป) ให้สอดคล้องกับข้อมูลในไฟล์ผลลัพธ์ซีเอ็นวีที่จะอัพโหลด เพื่อให้ซอฟต์แวร์อินซีเอ็นวีสามารถเข้าใจไฟล์ผลลัพธ์ซีเอ็นวี รวมถึงปรับเปลี่ยนรูปแบบไฟล์ผลลัพธ์ซีเอ็นวี (reformat) ทุกไฟล์ให้อยู่ในรูปแบบเดียวกัน และจัดเก็บไฟล์ผลลัพธ์เหล่านั้นไว้ในระบบ
5. หัวข้อ “file info” สำหรับให้ผู้ใช้ระบุข้อมูลเพิ่มเติมของไฟล์เพื่อช่วยแยกแยะผลลัพธ์ซีเอ็นวีที่ได้มาจากเครื่องมือตรวจจับซีเอ็นวีตัวเดียวกัน แต่มีข้อมูลเพิ่มเติมหรือค่าพารามิเตอร์ที่ต่างกัน เช่น กรณีผู้ใช้ต้องการวิเคราะห์ซีเอ็นวีจากไฟล์ผลลัพธ์ซีเอ็นวีหลายไฟล์ที่มาจากเครื่องมือตรวจจับซีเอ็นวีชื่อ CONTRA ซึ่งรันโปรแกรมโดยใช้ชุดข้อมูลเดียวกัน แต่ใช้ค่า threshold ในการแยกประเภทของซีเอ็นวีต่างกัน

ผู้ใช้สามารถระบุค่า threshold เป็นพารามิเตอร์ที่ใช้สำหรับแต่ละไฟล์ได้ที่คอมโพเน้นต์นี้ ซึ่งจะอธิบายเพิ่มเติมในหัวข้อ “5.4.4.2.2 แผนภูมิข้อมูลนำเข้าซีเอ็นวี”

6. หัวข้อ “tag descriptions” สำหรับให้ผู้ใช้ระบุข้อมูลเพิ่มเติมเพื่อช่วยในการจำไฟล์ผลลัพธ์ที่จะทำการอัพโหลด เช่น ชื่อโรค กลุ่มอายุ หรือแหล่งที่มาของไฟล์ เป็นต้น



รูปที่ 43 เทมเพลตเพื่อใช้ในการอัพโหลดไฟล์ผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวี

5.1.2 เทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวี (CNV tool mapping templates)

องค์ประกอบนี้อนุญาตให้ผู้ใช้สามารถออกแบบเทมเพลตของไฟล์นำเข้า และเพิ่มเทมเพลตนั้นลงในซอฟต์แวร์อินซีเอ็นวี โดยผู้ใช้จะต้องระบุค่าให้กับเทมเพลต (รูปที่ 44-45) ดังนี้

1. หัวข้อ “file mapping name” สำหรับให้ผู้ใช้ระบุชื่อเทมเพลตที่ผู้ใช้ต้องการเพิ่มลงในระบบ (สามารถใช้ชื่อเครื่องมือตรวจจับซีเอ็นวีได้)

2. หัวข้อ Header Column Mapping สำหรับให้ผู้ใช้ระบุชื่อคอลัมน์ของไฟล์ผลลัพธ์ซีเอ็นวีที่จะอัพโหลดเข้าระบบ ซึ่งประกอบด้วย

    a. หัวข้อ “sample name” สำหรับให้ผู้ใช้ระบุชื่อคอลัมน์ของไฟล์ที่จะอัพโหลดที่มีความหมายว่าชื่อตัวอย่าง

    b. หัวข้อ “chromosome” สำหรับให้ผู้ใช้ระบุชื่อคอลัมน์ของไฟล์ที่จะอัพโหลดที่ความหมายว่าเป็นชื่อโครโมโซม

    c. หัวข้อ “start position” สำหรับให้ผู้ใช้ระบุชื่อคอลัมน์ของไฟล์ที่จะอัพโหลดที่มีความหมายว่าเป็นตำแหน่งเบสเริ่มต้นของซีเอ็นวี

    d. หัวข้อ “end position” สำหรับให้ผู้ใช้ระบุชื่อคอลัมน์ของไฟล์ที่จะอัพโหลดที่มีความหมายว่าเป็นตำแหน่งเบสสิ้นสุดของซีเอ็นวี

    e. หัวข้อ “cnv type” สำหรับให้ผู้ใช้ระบุชื่อคอลัมน์ของไฟล์ที่จะอัพโหลดที่มีความหมายว่าประเภทของซีเอ็นวี

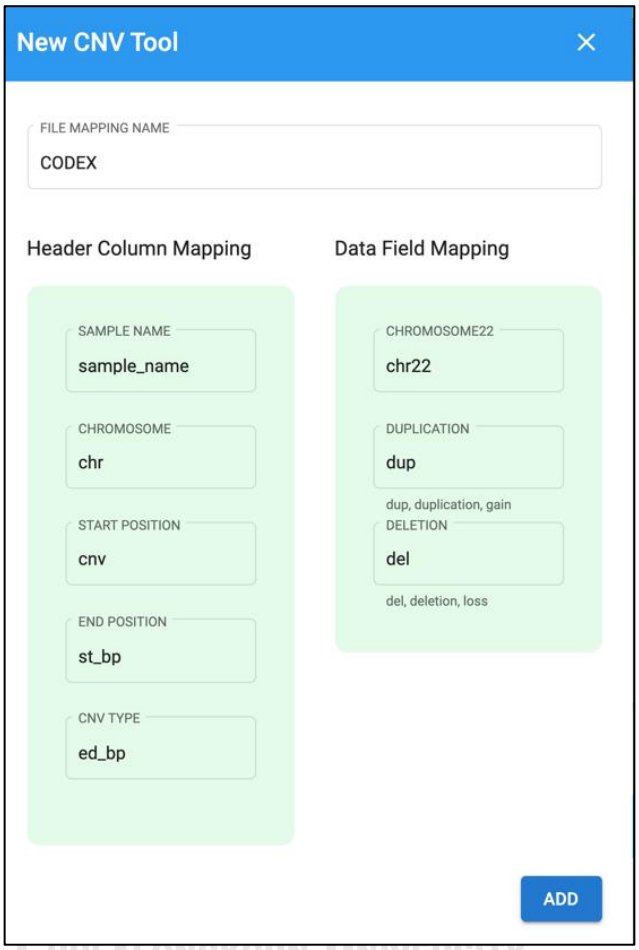
3. ส่วน Data Field Mapping หมายถึง การกำหนดรูปแบบของข้อมูลในไฟล์นำเข้าซึ่งประกอบด้วย

    a. หัวข้อ “chromosome22” สำหรับให้ผู้ใช้ระบุคำแทนโครโมโซม 22 ของไฟล์ที่จะนำเข้าระบบ เช่น จากรูปที่ 45 ไฟล์ผลลัพธ์ซีเอ็นวีระบุชื่อโครโมโซม 22 ว่า “chr22” ดังนั้นที่กล่องข้อความนี้ ผู้ใช้จะต้องกรอก “chr22” ดังรูปที่ 44 โดยระบบจะตีความว่าชื่อโครโมโซมทั้งหมดในไฟล์จะต้องขึ้นต้นด้วย “chr”

    b. หัวข้อ “duplication” สำหรับให้ผู้ใช้ระบุคำแทนซีเอ็นวีประเภท duplication ของไฟล์ที่จะนำเข้าระบบ สำหรับหัวข้อนี้ผู้พัฒนาได้มีฟังก์ชันการตรวจสอบข้อมูลโดยกำหนดให้ผู้ใช้ใส่ข้อความได้เฉพาะคำว่า “dup”, “duplication” หรือ “gain” (ได้ทั้งตัวอักษรภาษาอังกฤษตัวใหญ่ และตัวเล็ก)

    c. หัวข้อ “deletion” สำหรับให้ผู้ใช้ระบุคำแทนซีเอ็นวีประเภท deletion ของไฟล์ที่จะนำเข้าระบบ สำหรับหัวข้อนี้ผู้พัฒนาได้มีฟังก์ชันการตรวจสอบข้อมูลโดยกำหนดให้ผู้ใช้ใส่ข้อความได้เฉพาะคำว่า คือ “del”,

“deletion” หรือ “loss” (ได้ทั้งตัวอักษรภาษาอังกฤษตัวใหญ่ และตัวเล็ก)



รูปที่ 44 เทมเพลตเพื่อระบุการแมปฟิลด์ผลลัพธ์จากเครื่องมือตรวจจับซีเอ็นวีใด ๆ กับฟิลด์พื้นฐานของข้อมูลซีเอ็นวีที่กำหนดโดยซอฟต์แวร์อินซีเอ็นวี

header

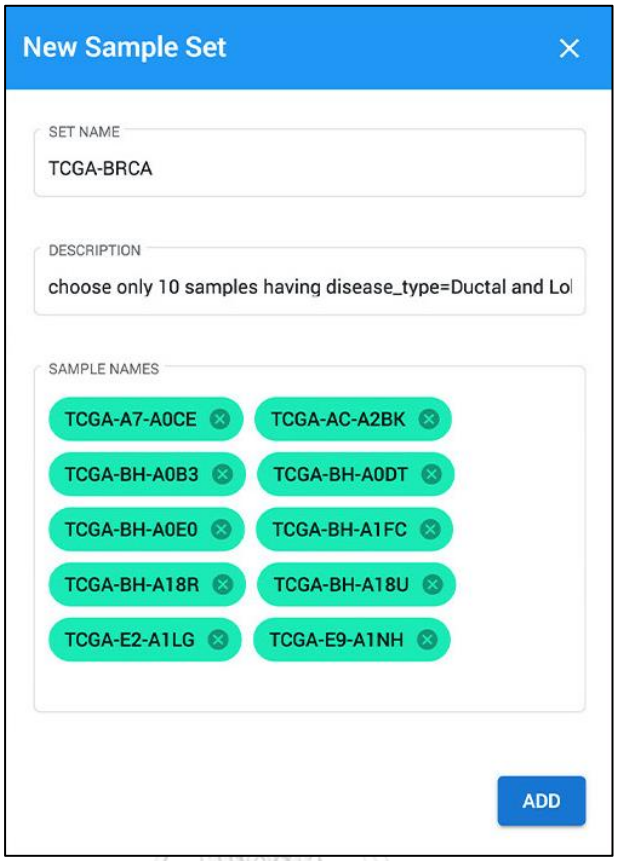
data

sample_name	chr	cnv	st_bp	ed_bp	length_kb	st_exon	ed_exon
NA06994	chr22	del	22782035	23237895	455.861	100941	100967
NA06994	chr22	del	106829563	107078845	249.283	50666	50690
NA06994	chr22	del	89156854	89247146	90.293	86826	86835
NA06994	chr22	del	106232232	106322346	90.115	50629	50635
NA06994	chr22	del	22599158	22764644	165.487	100932	100940
NA06994	chr22	dup	102222899	102269268	46.37	21505	21533
NA06994	chr22	del	55370902	55419327	48.426	26720	26722
NA06994	chr22	dup	55325110	55329066	3.957	81904	81906
NA06994	chr22	del	20344418	20404773	60.356	45251	45253
NA06994	chr22	dup	152573139	152586610	13.472	9827	9828
NA06994	chr22	del	106066799	106175028	108.23	50621	50626
NA06994	chr22	dup	248756119	248790433	34.315	16817	16818
NA06994	chr22	dup	55237480	55286943	49.464	81894	81900
NA06994	chr22	del	89476572	89477749	1.178	7078	7079
NA06994	chr22	dup	48084162	17265323	-30818.838	100402	100403

รูปที่ 45 ตัวอย่างไฟล์ผลลัพธ์ที่ได้จากการรันโปรแกรม CODEX บนโครโมโซม 22

5.1.3 เทมเพลตกลุ่มตัวอย่าง (Sample set templates)

องค์ประกอบนี้อนุญาตผู้ใช้ให้สามารถออกแบบเทมเพลตของกลุ่มตัวอย่างที่ผู้ใช้สนใจ และเพิ่มเทมเพลตนั้นลงในซอฟต์แวร์อินซีเอ็นวี โดยผู้ใช้จะต้องกำหนด ชื่อกลุ่ม คำอธิบายชื่อกลุ่ม และชื่อตัวอย่าง ซึ่งผู้ใช้สามารถสร้างได้เอง (รูปที่ 46)



รูปที่ 46 เทมเพลตเพื่อใช้สร้างกลุ่มตัวอย่าง

5.1.4 ระบบล็อกอิน (Login system)

ซอฟต์แวร์อินซีเอ็นวีมีระบบล็อกอินเพื่อช่วยอำนวยความสะดวกในการจัดการข้อมูลนำเข้าของผู้ใช้ ช่วยให้ผู้ใช้แต่ละคนจะสามารถมองเห็น และจัดการ (1) เทมเพลตการแมปของเครื่องมือตรวจจับซีเอ็นวีที่ตนเองเคยออกแบบไว้ (2) เทมเพลตกลุ่มตัวอย่างที่ตนเองเคยออกแบบไว้ และ (3) ข้อมูลจากไฟล์ผลลัพธ์ซีเอ็นวีที่ตนเองเคยนำเข้าระบบ ทำให้ผู้ใช้แต่ละคนสามารถใช้งานซอฟต์แวร์อินซีเอ็นวีได้โดยไม่รบกวนการทำงานของผู้ใช้คนอื่น สามารถอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีของตนลงในซอฟต์แวร์อินซีเอ็นวีได้สะดวกโดยไม่ต้องออกแบบเทมเพลตใหม่ หรืออัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีไฟล์เดิมซ้ำ รวมทั้งสามารถปรับเปลี่ยน และลบไฟล์ของตนได้

5.2 โมดูลการจัดการฐานข้อมูลในระบบ (Management of built-in databases module)

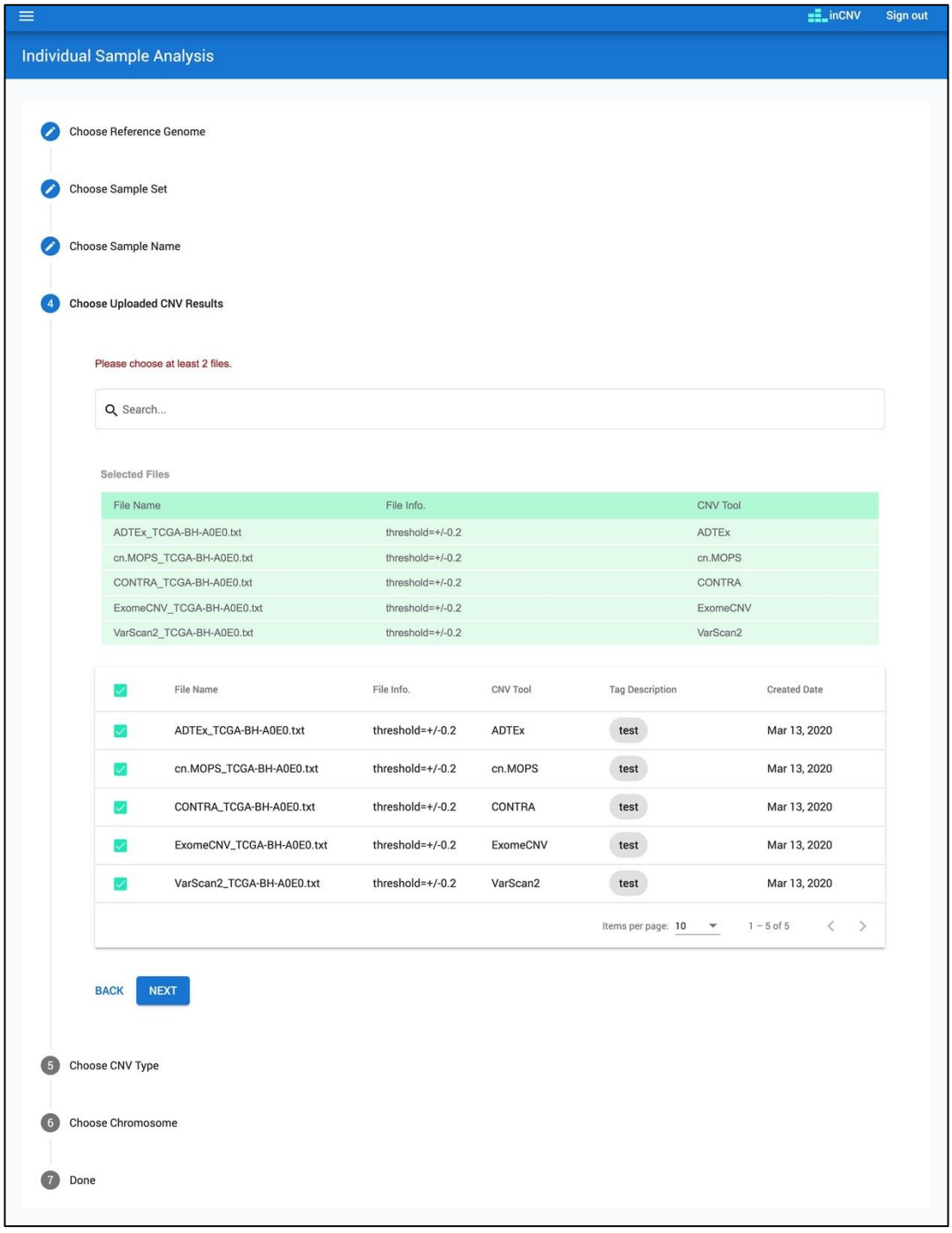
ซอฟต์แวร์อินซีเอ็นวีได้พัฒนาโมดูลนี้ตามที่ได้ออกแบบไว้หัวข้อ “4.2 การออกแบบฐานข้อมูล” และ หัวข้อ “4.3 การจัดการคำอธิบายจีโนม”

5.3 โมดูลโครงแบบการวิเคราะห์ (analysis configuration module)

โมดูลนี้เป็นกระบวนการเลือกผลลัพธ์ซีเอ็นวีสำหรับใช้ในโมดูลถัดไป (โมดูลการวิเคราะห์ซีเอ็นวีแบบบูรณาการ) โดยโครงแบบการวิเคราะห์จะแบ่งได้เป็น 2 ส่วนคือ

5.3.1 โครงแบบสำหรับการวิเคราะห์ตัวอย่างเดียว (individual-sampled analysis)

โครงแบบนี้มุ่งเน้นไปที่การรวมผลลัพธ์ซีเอ็นวีของ 1 ตัวอย่าง (1 คน) จากหลาย ๆ เครื่องมือตรวจจับซีเอ็นวี ซอฟต์แวร์อินซีเอ็นวีสามารถคัดกรองผลลัพธ์ซีเอ็นวีที่ต้องการโดยให้ผู้ใช้เลือกจีโนมอ้างอิง กลุ่มตัวอย่าง และชื่อตัวอย่างที่ต้องการนำไปวิเคราะห์ แล้วเลือกไฟล์ผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจจับซีเอ็นวีหลาย ๆ เครื่องมือ เลือกโครโมโซม และเลือกประเภทของซีเอ็นวีที่ต้องการวิเคราะห์ ดังรูปที่ 47

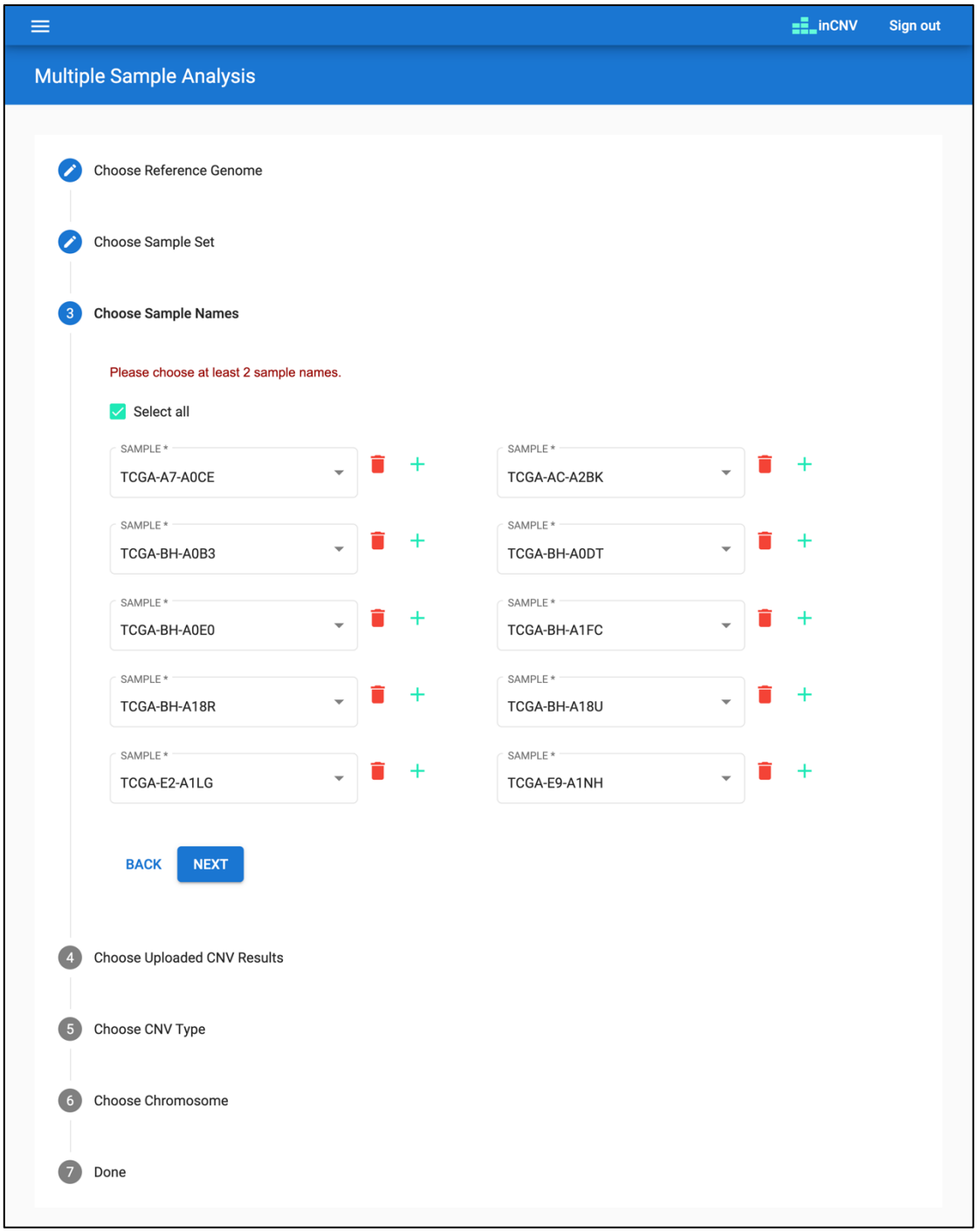


รูปที่ 47 โครงแบบการวิเคราะห์ตัวอย่างเดียว

5.3.2 โครงแบบสำหรับการวิเคราะห์หลายตัวอย่าง (multiple-sampled analysis)

โครงแบบนี้มุ่งเน้นไปที่การรวมผลลัพธ์ซีเอ็นวีของหลาย ๆ ตัวอย่าง (หลายคน) จากเครื่องมือตรวจจับซีเอ็นวีเครื่องมือเดียว ซอฟต์แวร์อินซีเอ็นวีคัดกรองผลลัพธ์ซีเอ็นวีที่ต้องการโดยให้

ผู้ใช้เลือกจีโนมอ้างอิง และกลุ่มตัวอย่าง หลังจากนั้นระบุชื่อตัวอย่างทั้งหมดที่ต้องการนำไปวิเคราะห์ เลือกไฟล์ผลลัพธ์ซีเอ็นวีที่มีข้อมูลซีเอ็นวีของตัวอย่างที่ต้องการเหล่านั้น เลือกโครโมโซม และเลือกประเภทของซีเอ็นวีที่ต้องการวิเคราะห์ ดังรูปที่ 48



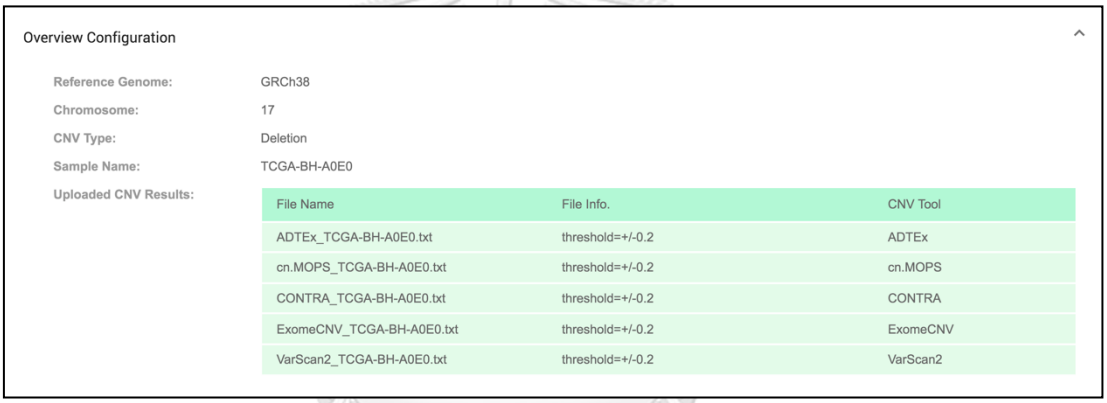
รูปที่ 48 โครงแบบการวิเคราะห์หลายตัวอย่าง

5.4 โมดูลการวิเคราะห์ซีเอ็นวีแบบบูรณาการ (Integrated CNV analysis module)

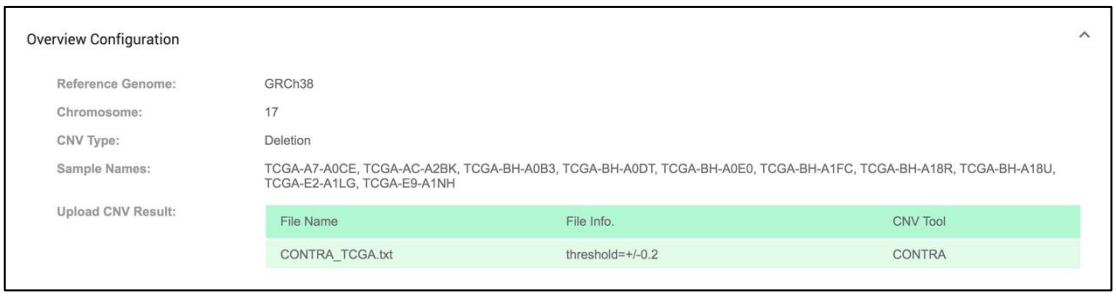
ซอฟต์แวร์อินซีเอ็นวีมีรูปแบบการวิเคราะห์ 2 รูปแบบหลักๆ คือ การวิเคราะห์แบบตัวอย่างเดียว (individual-sampled analysis) และ การวิเคราะห์แบบหลายตัวอย่าง (multiple-sampled analysis) โดยแต่ละรูปแบบมีฟีเจอร์ที่คล้ายคลึงกัน ต่างกันที่กรณีการใช้งาน (รายละเอียดเกี่ยวกับการใช้งานจะอธิบายต่อไปในหัวข้อผลการวิจัย และการอภิปรายผล) สำหรับโมดูลนี้ในทั้งสองรูปแบบการวิเคราะห์จะแบ่งเป็น 4 ส่วนการทำงาน ดังนี้

5.4.1 โครงแบบภาพรวม (Overview configuration)

โครงแบบภาพรวมแสดงข้อมูลสรุปของโมดูลโครงแบบการวิเคราะห์ (analysis configuration module) ที่ผู้ใช้ได้เลือกในขั้นตอนก่อนหน้านี้ เช่น จีโนมอ้างอิง, กลุ่มตัวอย่าง, ชื่อตัวอย่าง, ชื่อไฟล์ผลลัพธ์ซีเอ็นวี, โครโมโซม และประเภทของซีเอ็นวี ดังรูปที่ 49 – 50



รูปที่ 49 โครงแบบภาพรวมสรุปข้อมูลจากโครงแบบการวิเคราะห์แบบตัวอย่างเดียว

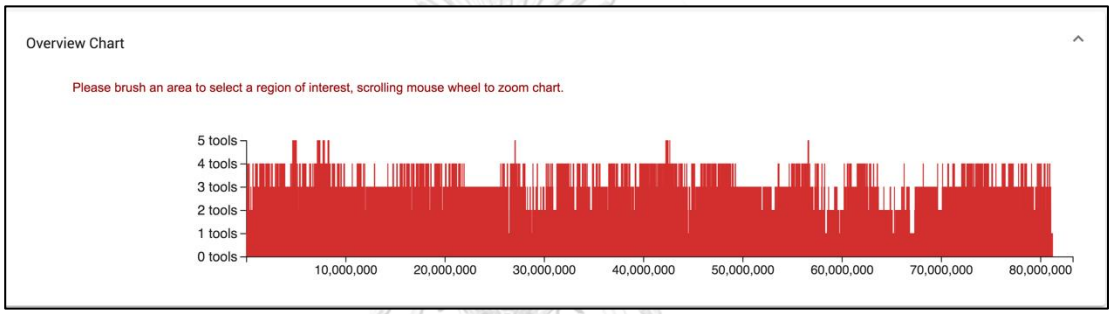


รูปที่ 50 โครงแบบภาพรวมสรุปข้อมูลจากโครงแบบการวิเคราะห์แบบหลายตัวอย่าง

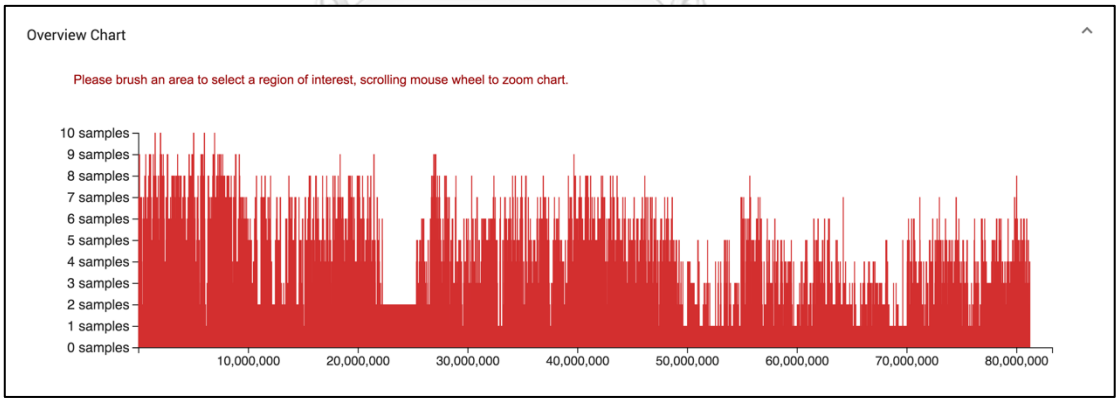
5.4.2 แผนภูมิภาพรวม (Overview chart)

แผนภูมิภาพรวมแสดงให้เห็นถึงการกระจายของซีเอ็นวีตามโครโมโซม แผนภูมินี้แสดงผลลัพธ์การรวมของซีเอ็นวีชนิด duplication หรือ deletion อย่างใดอย่างหนึ่ง ในมุมมองแบบ

กว้าง แกน X แสดงตำแหน่งเบสบนโครโมโซม และแกน Y แสดงจำนวนซีเอ็นวีที่ซ้อนทับกันบนโครโมโซม ณ ตำแหน่งเดียวกัน ทั้งนี้สำหรับการวิเคราะห์แบบตัวอย่างเดียว (individual-sampled analysis) แกน Y จะแทนจำนวนผลลัพธ์ซีเอ็นวีที่ซ้อนทับกันจากหลายเครื่องมือตรวจจับซีเอ็นวี (ของหนึ่งตัวอย่าง) ดังรูปที่ 51 ในขณะที่การวิเคราะห์แบบหลายตัวอย่าง (multiple-sampled analysis) แกน Y จะแทนจำนวนผลลัพธ์ซีเอ็นวีที่ซ้อนทับของหลายตัวอย่าง (จากหนึ่งเครื่องมือตรวจจับซีเอ็นวี) ดังรูปที่ 52 นอกจากนี้ ผู้ใช้สามารถเลื่อนเม้าส์เพื่อย่อ หรือขยายแผนภูมิตามแนวโครโมโซมได้ เพื่อให้ผู้ใช้สามารถเห็นภาพการกระจายตัวของซีเอ็นวีในภาพรวมได้สะดวกขึ้น ท้ายสุด ผู้ใช้สามารถไฮไลท์เลือกขอบเขตตำแหน่งของซีเอ็นวีบนแผนภาพรวม เพื่อดูรายละเอียดการรวมกันของซีเอ็นวีเฉพาะบริเวณที่สนใจซึ่งแสดงด้วยภาพกราฟฟิก และจะอธิบายต่อไปในหัวข้อ “แผนภูมิหลัก (main chart)”



รูปที่ 51 แผนภูมิภาพรวมการวิเคราะห์แบบตัวอย่างเดียว



รูปที่ 52 แผนภูมิภาพรวมการวิเคราะห์แบบหลายตัวอย่าง

5.4.3 ผลลัพธ์การรวมซีเอ็นวีทั้งหมด (All merged CNVs)

ส่วนนี้แสดงซีเอ็นวีทั้งหมดที่ถูกรวมเข้าด้วยกันแล้วของการวิเคราะห์แบบตัวอย่างเดียว (individual-sampled analysis) หรือ แบบหลายตัวอย่าง (multiple-sampled analysis) นำเสนอในรูปแบบตารางที่สามารถคัดกรองข้อมูลที่ต้องการได้ โดยประกอบด้วย 2 คอมโพเน้นต์ย่อย ได้แก่

คอมโพเน้นต์คัดกรอง (filtering component) และคอมโพเน้นต์ตารางรายละเอียด (detailed table component)

5.4.3.1 คอมโพเน้นต์คัดกรอง (Filtering component)

สำหรับการวิเคราะห์แบบตัวอย่างเดียว (individual-sampled analysis) ผู้ใช้สามารถคัดกรองผลรวมของซีเอ็นวีจากคำอธิบายพันธุกรรมด้วย “gene symbols” จากฐานข้อมูลอองซอมเบลอ (Ensembl) “variant accession” จากฐานข้อมูลดีจีวี (DGV) รหัสโอมิม “OMIM” และคำอธิบายฟีโนไทป์ (phenotypes) จากฐานข้อมูลคลินวาร์ (ClinVar) และสามารถคัดกรองผลรวมของซีเอ็นวีผ่านทางชื่อเครื่องมือตรวจจับซีเอ็นวี ดังรูปที่ 53 สำหรับการวิเคราะห์แบบหลายตัวอย่าง (multiple-sampled analysis) ผู้ใช้สามารถคัดกรองผลรวมของซีเอ็นวีโดยใช้ชุดของแอตทริบิวต์ชุดเดียวกับการวิเคราะห์แบบตัวอย่างเดียว ยกเว้นเปลี่ยนจากการคัดกรองผ่านชื่อเครื่องมือตรวจจับซีเอ็นวีเป็นการคัดกรองผ่านชื่อตัวอย่าง ดังรูปที่ 54

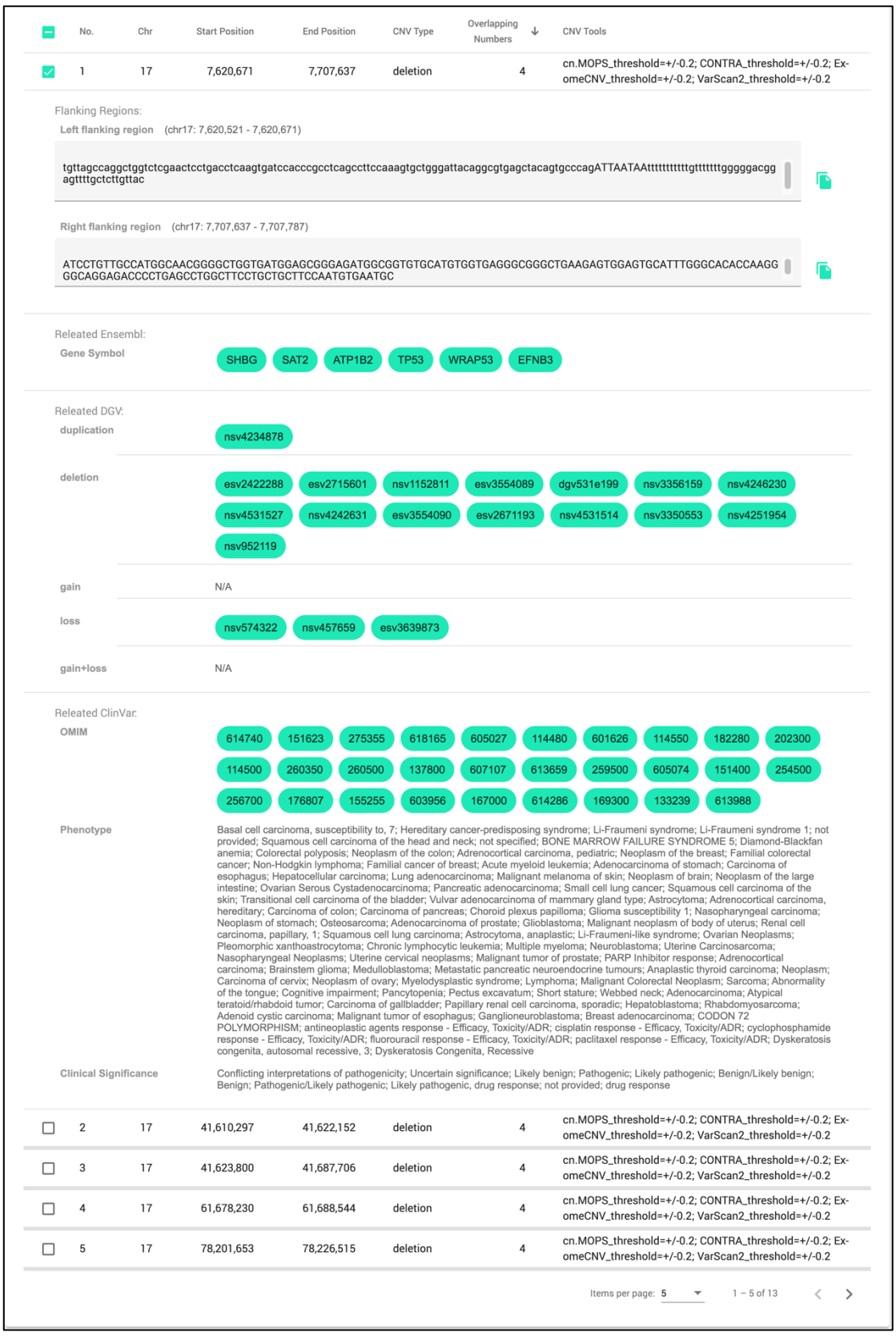
รูปที่ 53 คอมโพเน้นต์คัดกรองของการวิเคราะห์แบบตัวอย่างเดียว

รูปที่ 54 คอมโพเน้นต์คัดกรองของการวิเคราะห์แบบหลายตัวอย่าง

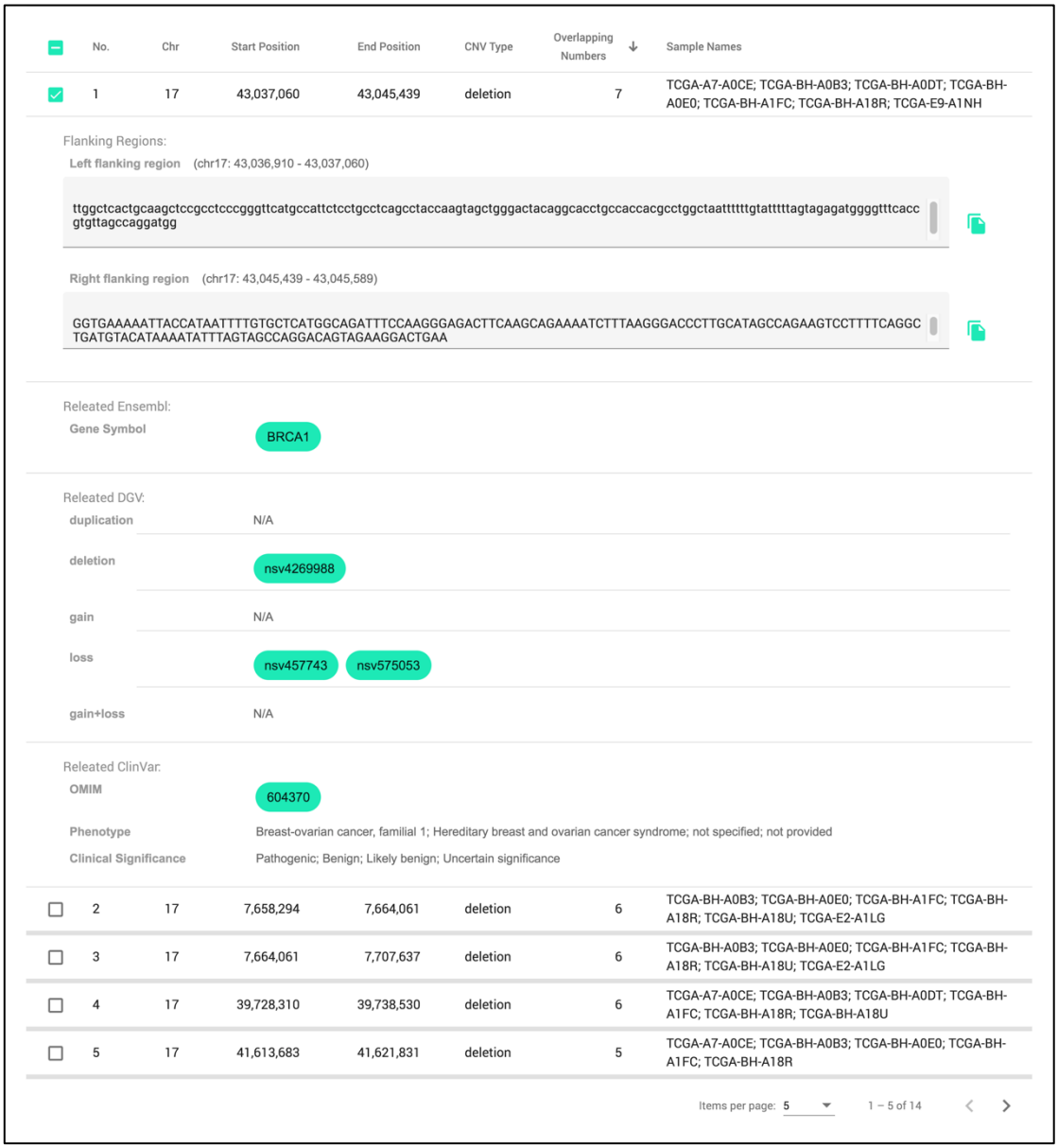
5.4.3.2 คอมโพเน้นต์ตารางรายละเอียด (Detailed table component)

คอมโนเน้นต์นี้แสดงข้อมูลผลลัพธ์ซีเอ็นวีทั้งหมดที่ถูกรวมเข้าด้วยกันในรูปแบบตาราง ผู้ใช้สามารถคลิกที่แถวใด ๆ ในตาราง (แต่ละแถว แสดงตำแหน่งซีเอ็นวีที่ตรวจจับได้ 1 ตำแหน่ง) เพื่อดูข้อมูลรายละเอียดของซีเอ็นวีที่ตรวจจับได้ อันได้แก่ คำอธิบายจีโนมจากฐานข้อมูลดีจีวี (DGV) ฐานข้อมูลอองซอมเบลอ (Ensembl) และฐานข้อมูลคลินวาร์ (ClinVar) รวมถึงลำดับเบสขนาบข้าง (flanking sequences) ของซีเอ็นวีนั้นๆ นอกจากนี้ ผู้ใช้ยังสามารถคลิกที่ไอคอนคำอธิบายจีโนมสีเขียวใด ๆ เพื่อดูรายละเอียดของคำอธิบายจีโนมนั้นซึ่งเชื่อมต่อกับฐานข้อมูลสาธารณะภายนอกผ่านอินเตอร์เน็ต ท้ายสุดผู้ใช้สามารถคลิกที่หัวตารางเพื่อเรียงลำดับข้อมูลตามตัวอักษรได้

หัวตารางจะประกอบไปด้วยคอลัมน์ดังนี้ ตำแหน่งเริ่มต้น (start position), ตำแหน่งสิ้นสุด (end position) และจำนวนซีเอ็นวีที่ซ้อนทับกัน (overlapping numbers) อย่างน้อย 1 เบส หมายเหตุ จำนวนซีเอ็นวีที่ซ้อนทับกัน โดยในการวิเคราะห์แบบตัวอย่างเดียว (individual-sampled analysis) หมายถึงจำนวนซีเอ็นวีที่ซ้อนทับกันจากหลายเครื่องมือตรวจจับซีเอ็นวีของตัวอย่าง 1 ตัวอย่าง ดังรูปที่ 55 โดยซีเอ็นวีลำดับแรกในตารางถูกตรวจพบใน 4 เครื่องมือ ในขณะที่จำนวนซีเอ็นวีที่ซ้อนทับกันในการวิเคราะห์แบบหลายตัวอย่าง (multiple-sampled analysis) หมายถึงจำนวนซีเอ็นวีที่ซ้อนทับกันจากหลายตัวอย่างที่มาจากเครื่องมือตรวจจับซีเอ็นวีตัวเดียวกัน ดังรูปที่ 56 โดยซีเอ็นวีลำดับแรกในตารางถูกตรวจพบใน 7 ตัวอย่าง เป็นต้น



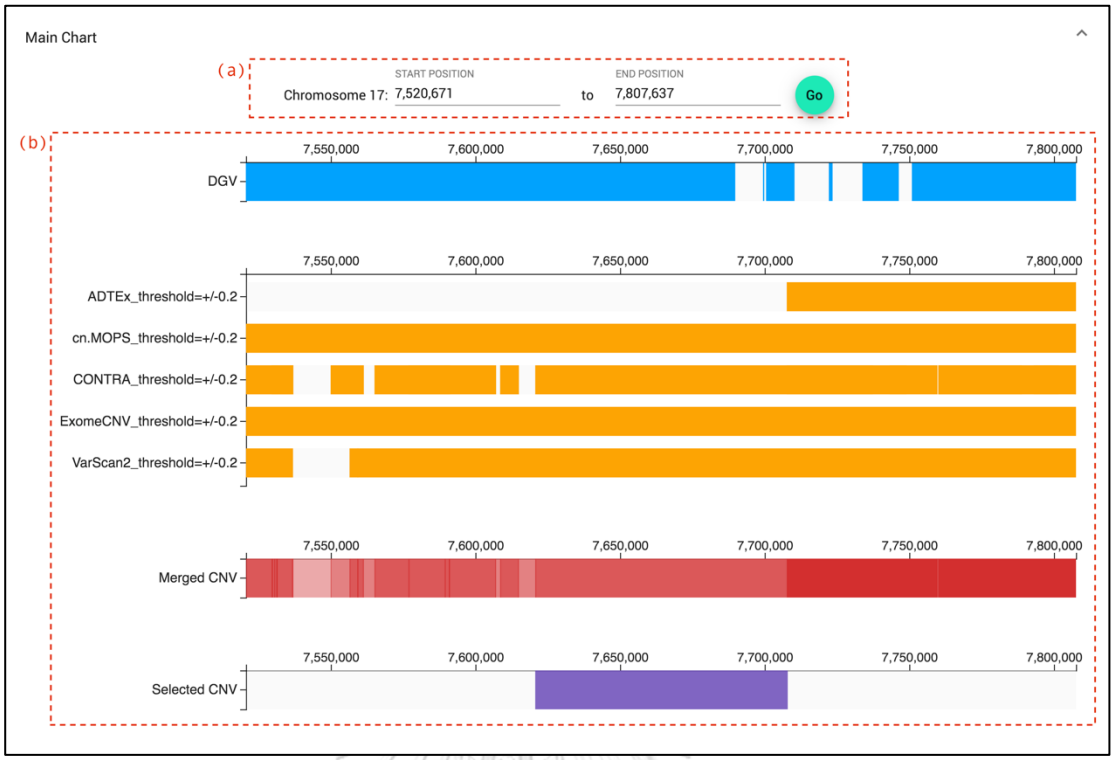
รูปที่ 55 คอมโพเน้นต์ตารางรายละเอียดการวิเคราะห์แบบตัวอย่างเดียว



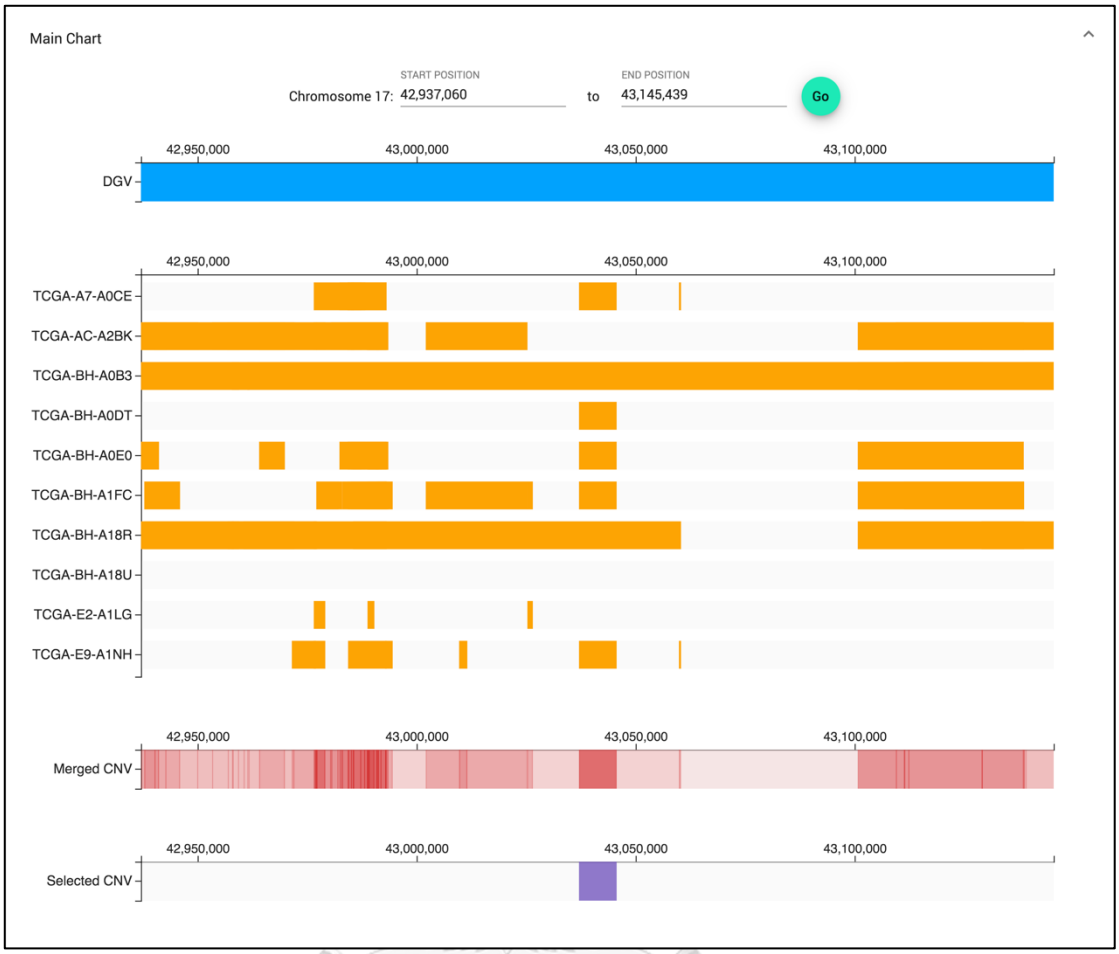
รูปที่ 56 คอมโพเน้นต์ตารางรายละเอียดการวิเคราะห์แบบหลายตัวอย่าง

5.4.4 แผนภูมิหลัก (Main chart)

แผนภูมิหลักมีไว้เพื่อแสดงข้อมูลผลลัพธ์การรวมซีเอ็นวีในรูปแบบของภาพเพื่อให้ผู้ใช้สามารถเปรียบเทียบวิเคราะห์ และจัดการผลลัพธ์เหล่านั้นได้ง่าย ดังรูปที่ 57 – 58 แผนภูมินี้ประกอบไปด้วย 2 ส่วน คือ การเลือกขอบเขตตำแหน่งเบส (region-based selection) และกลุ่มแผนภูมิรูปภาพ (Visualization charts)



รูปที่ 57 แผนภูมิหลักของการวิเคราะห์แบบตัวอย่างเดียว โดยที่ (a) แทนส่วนการเลือกขอบเขตต่ำแหน่งเบส และ (b) คือ กลุ่มแผนภูมิรูปภาพ



รูปที่ 58 แผนภูมิหลักของการวิเคราะห์แบบหลายตัวอย่าง

5.4.4.1 การเลือกขอบเขตตำแหน่งเบส (Region-based selection)

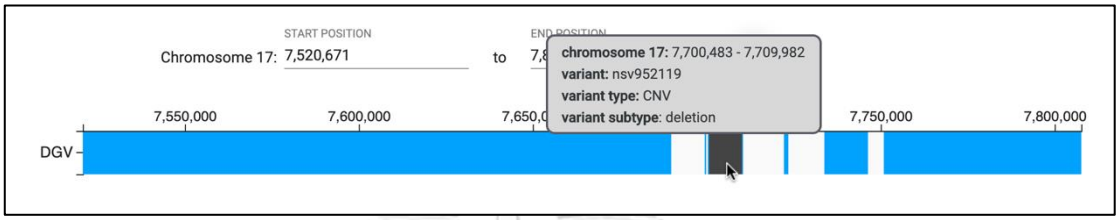
คอมโพเน้นต์นี้อนุญาตให้ผู้ใช้เลือกขอบเขตของเบสที่สนใจบนโครโมโซม โดยระบุตำแหน่งเบสเริ่มต้น และเบสสิ้นสุด ซึ่งขอบเขตตำแหน่งนี้จะเป็นตัวกำหนดพิกัดเริ่มต้น และสิ้นสุดบนแกน X ของกลุ่มแผนภูมิรูปภาพซึ่งจะอธิบายในหัวข้อถัดไป

5.4.4.2 กลุ่มแผนภูมิรูปภาพ (Visualization charts)

กลุ่มแผนภูมิรูปภาพประกอบไปด้วย แผนภูมิดีจีวี (DGV chart) แผนภูมิข้อมูลนำเข้าซีเอ็นวี แผนภูมิผลรวมซีเอ็นวี และแผนภูมิซีเอ็นที่ถูกเลือก ซึ่งมีแกน X แสดงตำแหน่งเบสบนโครโมโซม ขอบเขตของตำแหน่งบนแกน X ขึ้นอยู่กับ “การเลือกขอบเขตตำแหน่งเบส (Based-region selection)” (หัวข้อก่อนหน้า) “ขอบเขตซีเอ็นวีที่ถูกไฮไลท์ในแผนภูมิภาพรวม (Overview chart)” (หัวข้อก่อนหน้า) หรือตำแหน่งซีเอ็นวีที่เลือกล่าสุดในผลลัพธ์การรวมซีเอ็นวีทั้งหมด (All merged CNVs component) (หัวข้อถัดไป)

5.4.4.2.1 แผนภูมิดีจีวี (DGV chart)

แผนภูมิดีจีวีแสดงตำแหน่งของซีเอ็นวีที่สัมพันธ์กับฐานข้อมูลดีจีวี (DGV) ซึ่งประกอบไปด้วยซีเอ็นวี 5 ประเภท ได้แก่ “duplication”, “deletion”, “gain”, “loss”, และ “gain+loss” ผู้ใช้สามารถดูข้อมูลจากฐานข้อมูลดีจีวีซึ่งประกอบด้วย “chromosome”, “variant”, “variant type” และ “variant subtype” บนแผนภูมิดีจีวีได้โดยการวางเม้าส์ไว้เหนือแผนภูมิ ดังรูปที่ 59

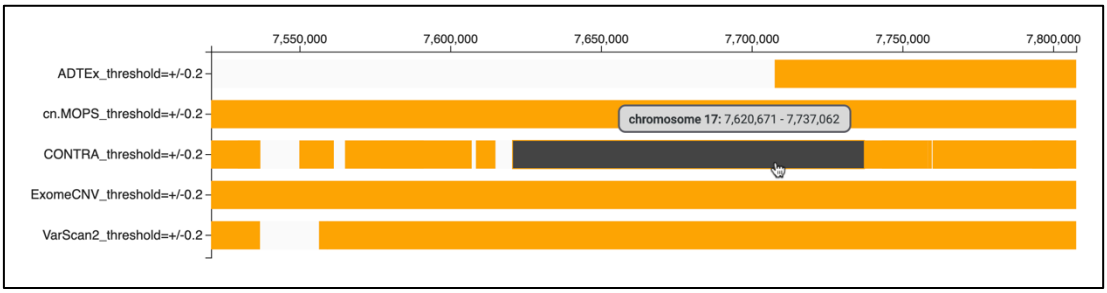


รูปที่ 59 ข้อมูลจากฐานข้อมูลดีจีวีบนแผนภูมิดีจีวี

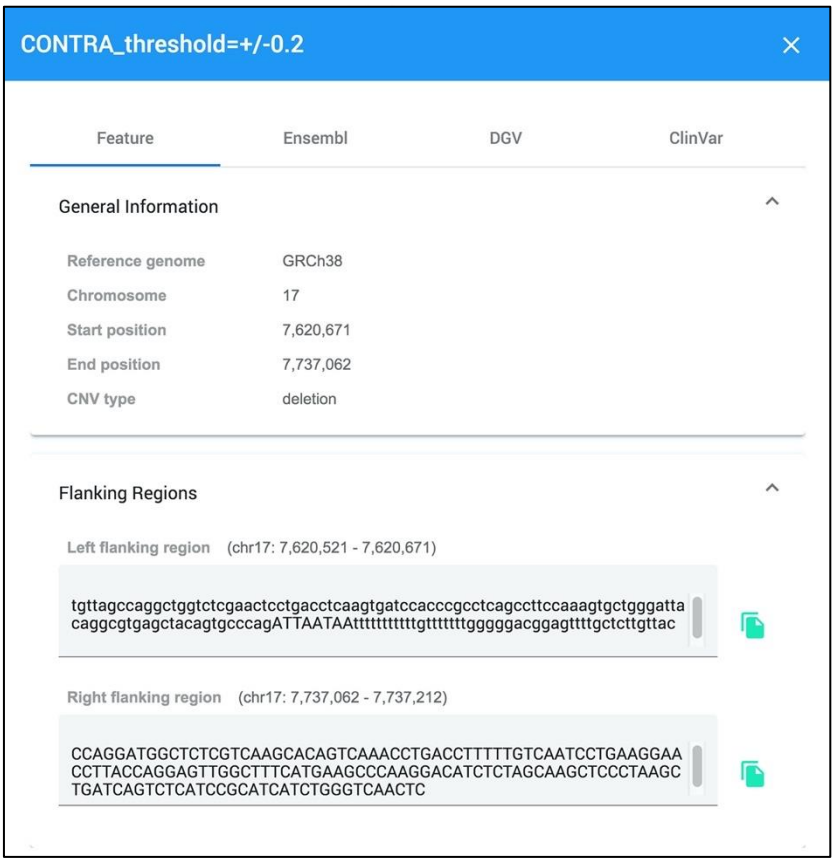
5.4.4.2.2 แผนภูมิข้อมูลนำเข้าซีเอ็นวี (Inputted-CNV charts)

แผนภูมินี้แสดงผลลัพธ์ซีเอ็นวีของแต่ละเครื่องมือตรวจจับซีเอ็นวี หรือของแต่ละตัวอย่าง สำหรับการวิเคราะห์แบบตัวอย่างเดียว (individual-sampled analysis) แกน Y จะแสดงชื่อของเครื่องมือตรวจจับซีเอ็นวีในฟอร์แมต “[ชื่อเครื่องมือตรวจจับซีเอ็นวี]_[รายละเอียดไฟล์ หรือ ค่าพารามิเตอร์]” เช่น “CONTRA_threshold=+/0.2” หมายถึงข้อมูลนี้ได้มาจากเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA” ซึ่งใช้ค่าพารามิเตอร์เพื่อให้ได้ไฟล์นี้คือ “threshold=±0.2” ในขณะที่การวิเคราะห์แบบหลายตัวอย่าง (multiple-sampled analysis) แกน Y จะแสดงชื่อตัวอย่างในฟอร์แมต “[ชื่อเครื่องมือตรวจจับซีเอ็นวี]_[รายละเอียดไฟล์ หรือ ค่าพารามิเตอร์]” เช่น “TCGA-A7-A0CE_threshold=+/0.2” หมายถึงข้อมูลนี้เป็นของตัวอย่างชื่อ “TCGA-A7-A0CE” ซึ่งใช้ค่าพารามิเตอร์เพื่อให้ได้ไฟล์นี้คือ “threshold=±0.2”

นอกจากนี้ ผู้ใช้สามารถวางเม้าส์เหนือแผนภูมิเพื่อดูตำแหน่งเบสของซีเอ็นวีว่าอยู่ตำแหน่งใดบนโครโมโซม และด้วยการคลิกบนแผนภูมิ ผู้ใช้สามารถเห็นไดอะล็อก (dialog) ซึ่งประกอบด้วยข้อมูลดังนี้ คำอธิบายจีโนมจากฐานข้อมูลอองซอมเบลอ (Ensembl) ฐานข้อมูลดีจีวี (DGV) และฐานข้อมูลคลินวาร์ (ClinVar) รวมถึงลำดับเบสขนาบข้าง (flanking sequences) ที่สัมพันธ์กับซีเอ็นวีนั้น ๆ ดังรูปที่ 60 -61



รูปที่ 60 แผนภูมิข้อมูลนำเข้าซีเอ็นวีเมื่อวางเม้าส์เหนือแผนภูมิของเครื่องมือตรวจจับชื่อ “CONTRA” จะมีทูลทิปแสดงพิกัดของผลลัพธ์ซีเอ็นวีนั้น ๆ ซึ่งมีพารามิเตอร์เป็น “threshold=+/-0.2”



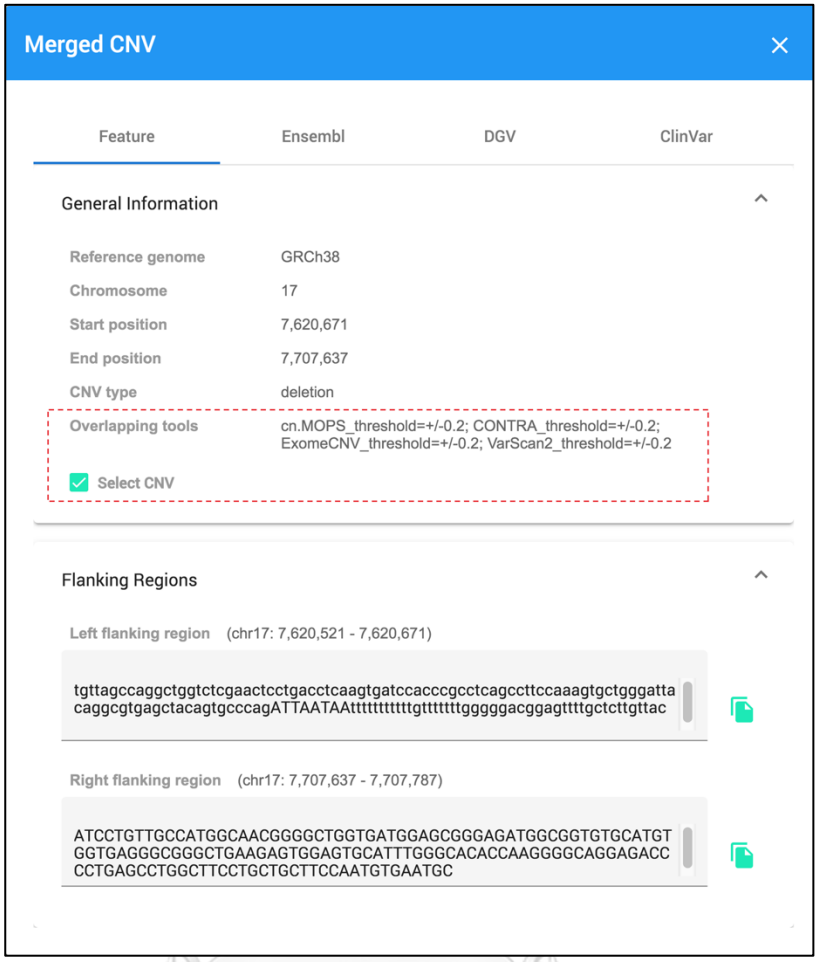
รูปที่ 61 ไดอะล็อกจากแผนภูมิข้อมูลนำเข้าซีเอ็นวีซึ่งแสดงรายละเอียดของซีเอ็นวีจากเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA” ซึ่งมีพารามิเตอร์เป็น “threshold=+/-0.2” (การวิเคราะห์แบบตัวอย่างเดียว)

5.4.4.2.3 แผนภูมิผลรวมซีเอ็นวี (Merged CNV chart)

แผนภูมิผลรวมซีเอ็นวีแสดงผลลัพธ์การรวมของซีเอ็นวีด้วยฮีทแมป (heatmap) ซึ่งสัมพันธ์กับตำแหน่งเบสบนโครโมโซม แผนภูมินี้ใช้สีเพื่อแสดงการซ้อนทับกันของซีเอ็นวีจากหลายเครื่องมือตรวจจับซีเอ็นวี หรือจากหลายตัวอย่าง โดยสีที่เข้มขึ้นหมายถึงการซ้อนทับกันของจำนวนซีเอ็นวีบนตำแหน่งเดียวกันที่เพิ่มขึ้น ดังแผนภูมิชื่อ “Merged CNV” ในรูปที่ 57 - 58 อย่างไรก็ตามสีที่เข้มขึ้นนี้อาจหมายถึง ค่าผลบวกเท็จ (false positive: FP) ซึ่งขึ้นอยู่กับประสิทธิภาพของอัลกอริทึมของเครื่องมือตรวจจับซีเอ็นวีแต่ละตัว ดังนั้นเพื่อที่จะระบุซีเอ็นวีอย่างมีประสิทธิผล ผู้ใช้จำเป็นต้องตระหนักถึงลักษณะของเครื่องมือตรวจจับซีเอ็นวีที่ใช้แต่ละตัวว่าเหมาะสมกับซีเอ็นวีแต่ละประเภทอย่างไร

นอกจากนี้ การวางเม้าส์ หรือคลิกบนแผนภูมินี้ยังให้ข้อมูลเหมือนที่ทำในแผนภูมิข้อมูลนำเข้าซีเอ็นวี (inputted-CNV charts) ยกเว้นมีข้อมูลเพิ่มขึ้นมา คือ กล่องกาเครื่องหมาย (check box) ชื่อ “select CNV” และสำหรับการวิเคราะห์แบบตัวอย่างเดียว (individual-sampled analysis) จะมีข้อมูล “overlapping tools” ดังรูปที่ 62 และสำหรับการวิเคราะห์แบบหลายตัวอย่าง (multiple-sampled analysis) จะมีข้อมูล “overlapping samples”

ท้ายสุด ผู้ใช้สามารถเลือกซีเอ็นวีนี้เพื่อส่งออกข้อมูลโดยการกาเลือก “select CNV” ในกล่องกาเครื่องหมาย หลังจากนั้นซีเอ็นวีที่ถูกเลือกจะถูกแสดงในแผนภูมิซีเอ็นวีที่ถูกเลือก (selected CNVs chart) และตารางซีเอ็นวีทั้งหมดที่ถูกเลือก (all selected CNVs table) ซึ่งจะอธิบายต่อไปในโมดูลการส่งออกผลลัพธ์



รูปที่ 62 ไดอะล็อกแสดงรายละเอียดของซีเอ็นวีจากการรวมผลลัพธ์เครื่องมือตรวจจับซีเอ็นวีชื่อได้แก่ “cn.MOPS”, “CONTRA”, “ExomeCNV” และ “VarScan2” ซึ่งมีพารามิเตอร์เป็น “threshold=+/-0.2” (การวิเคราะห์แบบตัวอย่างเดียว)

5.4.4.2.4 แผนภูมิซีเอ็นที่ถูกเลือก (Selected CNV chart)

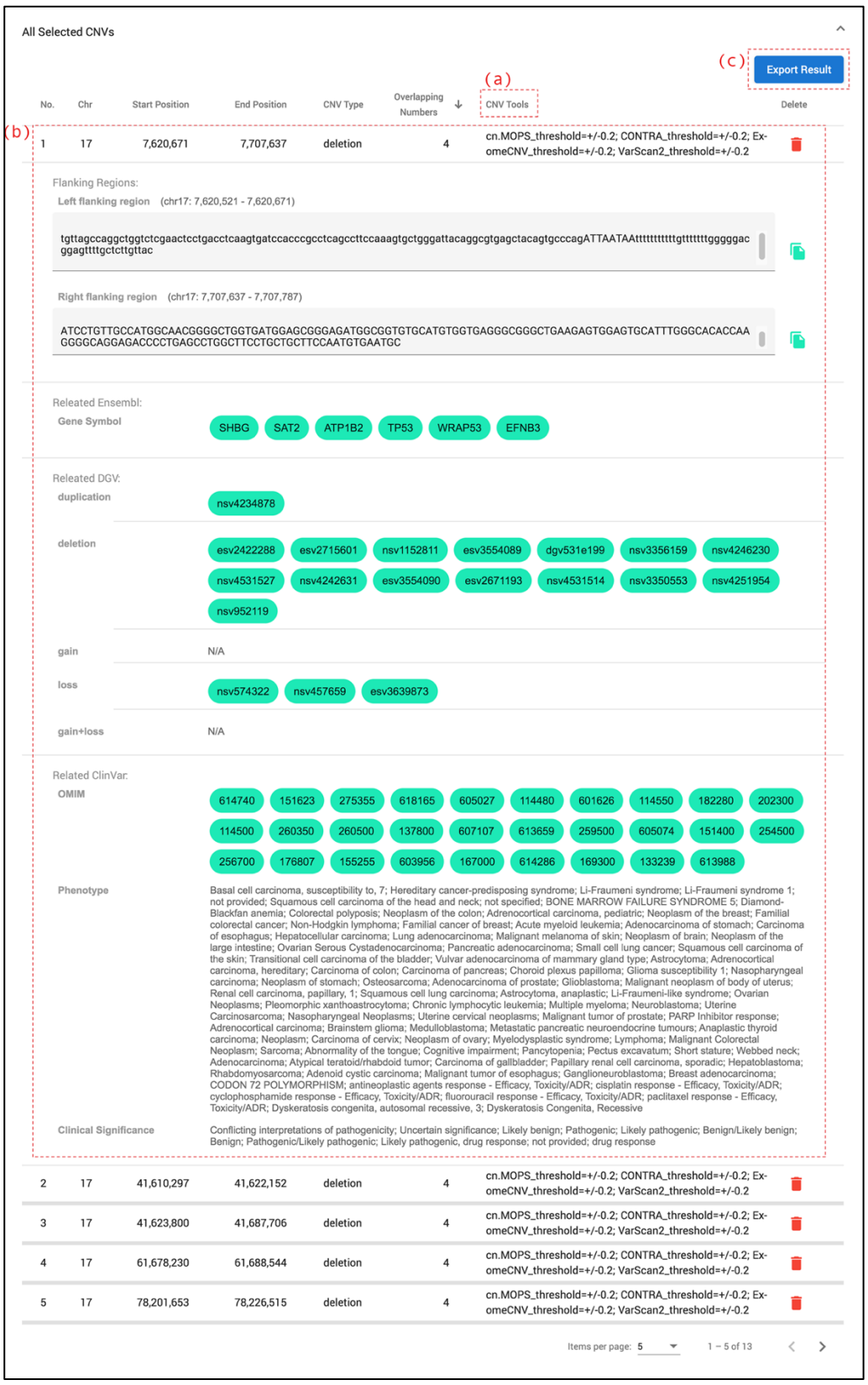
แผนภูมินี้แสดงซีเอ็นวีที่ถูกเลือก การวางเม้าส์ และคลิกบนแท่งข้อมูลบนแผนภูมิจะให้ข้อมูลเหมือนกับที่ทำในแผนภูมิผลรวมซีเอ็นวี (Merged CNV chart) ดังแผนภูมิชื่อ “Selected CNV” ในรูปที่ 56 – 57

5.5 โมดูลการส่งออกผลลัพธ์ (Exporting result module)

โมดูลนี้รวบรวมผลลัพธ์ของซีเอ็นวีที่ถูกเลือกลงในตารางชื่อ “All Selected CNVs” นำเสนอข้อมูลเหล่านั้นในหน้าเว็บ และส่งออกข้อมูลเหล่านั้นในรูปแบบไฟล์ข้อความที่มีแท็บเป็นตัวคั่น (tab-delimited file) ดังรูปที่ 63 แสดงตารางซีเอ็นวีที่ถูกเลือกทั้งหมดของการวิเคราะห์แบบตัวอย่าง

เดียวซึ่งประกอบด้วยหลายคอมโพเน้นดังนี้ (a) คอลัมน์เครื่องมือตรวจจับซีเอ็นวี (CNV tools column) ซึ่งแสดงรายชื่อเครื่องมือตรวจจับซีเอ็นวีที่ตรวจพบซีเอ็นวีในบริเวณนั้น ๆ (b) รายละเอียดของซีเอ็นวีที่ถูกขยายออกเมื่อคลิกบนแถวใด ๆ ในตาราง และ (c) ปุ่มส่งออกข้อมูลซีเอ็นวีที่ถูกเลือก

ผู้ใช้สามารถเลือก หรือยกเลิกการเลือกซีเอ็นวีจากตารางชื่อ “All Selected CNVs” ผ่านทางหลายคอมโพเน้นต์ ได้แก่ คอมโพเน้นต์ตารางรายละเอียด (detailed table component) ในผลลัพธ์การรวมซีเอ็นวีทั้งหมด (all merged CNVs) , แผนภูมิผลรวมซีเอ็นวี (merged CNV chart) และแผนภูมิซีเอ็นวีที่ถูกเลือก (selected CNV chart) ตามที่ได้อธิบายไปก่อนหน้านี้



รูปที่ 63 ตารางซีเอ็นวีที่ถูกเลือกทั้งหมดของการวิเคราะห์แบบตัวอย่างเดียว

5.6 ทดสอบการจัดการความผิดพลาดในระบบ

ผู้วิจัยได้ออกแบบการจัดการความผิดพลาดของผู้ใช้งาน 3 แบบ คือ การจัดการความผิดพลาดจากไฟล์ผลลัพธ์ซีเอ็นวีที่ถูกอัพโหลดเข้าระบบ การจัดการความผิดพลาดจากการกรอกข้อมูลผิดพลาด และความผิดพลาดจากการทำงานของจียูไอ เนื่องจากกรณีทดสอบมีจำนวนมาก ดังนั้นผู้วิจัยจึงขอเลือกเฉพาะ “ข้อมูลบางส่วนจากการจัดการความผิดพลาดจากไฟล์ผลลัพธ์ซีเอ็นวีที่ถูกอัพโหลดเข้าระบบ” และ “การจัดการความผิดพลาดจากการกรอกข้อมูลผิดพลาด” มาแสดงกรณีทดสอบ (test cases) บางส่วน ดังต่อไปนี้

5.6.1 การจัดการความผิดพลาดจากไฟล์ผลลัพธ์ซีเอ็นวีที่ถูกอัพโหลดเข้าระบบ

ระบบจะนำไฟล์ที่ถูกอัพโหลดมาอ่านทีละบรรทัด แล้วเก็บให้อยู่ในรูปแบบอะเรย์ (array) หน่วยความจำชั่วคราว แล้วจึงประมวลผลข้อมูลเหล่านั้นโดยการแมปกับ “เทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวี” และ “เทมเพลตกลุ่มตัวอย่าง” เพื่อให้ได้ผลลัพธ์ซีเอ็นวีตามรูปแบบที่ซอฟต์แวร์อินซีเอ็นวีได้ออกแบบไว้ แล้วบันทึกข้อมูลลงบนตาราง “upload_cnv_tool_result” และ “reformat_cnv_tool_result” ซึ่งได้มีอธิบายเพิ่มเติมไว้ในหัวข้อชื่อ “4.2.1 ฐานข้อมูลผู้ใช้” หากพบว่าข้อมูลจากไฟล์ที่ถูกอัพโหลดมีความผิดพลาดเกิดขึ้นก็จะทำการย้อนกลับการทำงานบนฐานข้อมูล (roll back) โดยการลบข้อมูลทั้งหมดของไฟล์ออกจาก 2 ตารางดังกล่าว

ผู้วิจัยเลือกตัวอย่างกรณีทดสอบการอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่สำคัญ และสามารถทดสอบได้บนหน้าเว็บของซอฟต์แวร์อินซีเอ็นวี แสดงรายละเอียดตามตารางที่ 5 – 10

ตารางที่ 5 กรณีทดสอบการอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ถูกต้องเข้าระบบ

ตารางที่ 6 กรณีทดสอบการอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีเข้าระบบแต่เลือกเทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวีไม่ตรงกับที่ระบุในไฟล์

ตารางที่ 7 กรณีทดสอบการอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีเข้าระบบแต่เลือกเทมเพลตกลุ่มตัวอย่างไม่ตรงกับที่ระบุในไฟล์

ตารางที่ 8 กรณีทดสอบการอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ใส่ข้อมูลโครโมโซมผิด

ตารางที่ 9 กรณีทดสอบการอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ใส่ข้อมูลตำแหน่งเบสเริ่มต้นผิด

ตารางที่ 10 กรณีทดสอบการอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ใส่ข้อมูลประเภทของซีเอ็นวีผิด

ตารางที่ 5 กรณีทดสอบอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ถูกต้องเข้าระบบ

กรณีทดสอบ 1	สถารการณ์ทดสอบ	อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ถูกต้องเข้าระบบ
	ขั้นตอนการทดสอบ	1. ไปที่หน้าอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 2. อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 3. ระบุชื่อเครื่องมือตรวจจับซีเอ็นวีที่เป็นที่มาของไฟล์ที่ถูกอัพโหลด 4. เลือกเทมเพลตเครื่องมือตรวจจับซีเอ็นวี 5. เลือกเทมเพลตกลุ่มตัวอย่าง 6. กดปุ่มอัพโหลด
	อินพุต	● ไฟล์ผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจชื่อ “CONTRA” และกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE” ที่ถูกต้อง ● ชื่อเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA” ● เทมเพลตเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA” ● เทมเพลตกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE”
	ผลลัพธ์ที่คาดหวัง	● หน้าเว็บแสดงผลลัพธ์ซีเอ็นวีของไฟล์ที่ถูกอัพโหลด ซึ่งถูกปรับเปลี่ยนตามเทมเพลตเครื่องมือตรวจจับซีเอ็นวี และเทมเพลตกลุ่มตัวอย่างที่ผู้ใช้เลือก เพื่อให้ผู้ใช้ตรวจสอบ และยืนยันการนำผลลัพธ์เข้าระบบ ● หน้า “My Files” มีข้อมูลไฟล์ที่ทำการทดสอบอยู่
	ผลลัพธ์ที่เกิดขึ้นจริง	ตามที่คาดหวัง
	ผลการทดสอบ (ผ่าน / ไม่ผ่าน)	ผ่าน

ตารางที่ 6 กรณีทดสอบอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีเข้าระบบแต่เลือกเทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวีไม่ตรงกับที่ระบุในไฟล์

กรณีทดสอบ 2	สถารการณ์ทดสอบ	อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีเข้าระบบแต่เลือกเทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวีไม่ตรงกับที่ระบุในไฟล์
	ขั้นตอนการทดสอบ	1. ไปที่หน้าอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 2. อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 3. ระบุชื่อเครื่องมือตรวจจับซีเอ็นวีที่เป็นที่มาของไฟล์ที่ถูกอัพโหลด 4. เลือกเทมเพลตเครื่องมือตรวจจับซีเอ็นวี 5. เลือกเทมเพลตกลุ่มตัวอย่าง 6. กดปุ่มอัพโหลด
	อินพุต	● ไฟล์ผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจชื่อ “CONTRA” และกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE” ● ชื่อเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA” ● เทมเพลตเครื่องมือตรวจจับซีเอ็นวีชื่อ “CODEX” ● เทมเพลตกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE”
	ผลลัพธ์ที่คาดหวัง	● มีข้อความแจ้งแสดงความผิดพลาดบนหน้าเว็บว่า “Cannot map header column named 'Sample Name’, CNV Type, ‘Start Basepair’, ‘End Basepair’. Please check file headers whether matching with the configuration of tab file mapping.” ● หน้า “My Files” ไม่มีข้อมูลไฟล์ที่ทำการทดสอบอยู่
	ผลลัพธ์ที่เกิดขึ้นจริง	ตามที่คาดหวัง
	ผลการทดสอบ (ผ่าน / ไม่ผ่าน)	ผ่าน

ตารางที่ 7 กรณีทดสอบอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีเข้าระบบแต่เลือกเทมเพลตกลุ่มตัวอย่างไม่ตรงกับที่ระบุในไฟล์

กรณีทดสอบ 3	สถารการณ์ทดสอบ	อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีเข้าระบบแต่เลือกเทมเพลตกลุ่มตัวอย่างไม่ตรงกับที่ระบุในไฟล์
	ขั้นตอนการทดสอบ	1. ไปที่หน้าอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 2. อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 3. ระบุชื่อเครื่องมือตรวจจับซีเอ็นวีที่เป็นที่มาของไฟล์ที่ถูกอัพโหลด 4. เลือกเทมเพลตเครื่องมือตรวจจับซีเอ็นวี 5. เลือกเทมเพลตกลุ่มตัวอย่าง 6. กดปุ่มอัพโหลด
	อินพุต	● ไฟล์ผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจชื่อ “CONTRA” และกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE” ● ชื่อเครื่องมือตรวจจับซีเอ็นวีชื่อ CONTRA ● เทมเพลตเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA” ● เทมเพลตกลุ่มตัวอย่างชื่อ “TCGA-AC-A2BK”
	ผลลัพธ์ที่คาดหวัง	● หน้า “My Files” มีข้อมูลไฟล์ว่าได้ถูกอัพโหลดเข้าระบบแต่ไม่มีเนื้อหาไฟล์ซึ่งระบุข้อมูลซีเอ็นวี
	ผลลัพธ์ที่เกิดขึ้นจริง	ตามที่คาดหวัง
	ผลการทดสอบ (ผ่าน / ไม่ผ่าน)	ผ่าน

ตารางที่ 8 กรณีทดสอบอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ใส่ข้อมูลโครโมโซมผิด

กรณีทดสอบ 4	สถารการณ์ทดสอบ	อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ใส่ข้อมูลโครโมโซมผิด
	ขั้นตอนการทดสอบ	1. ไปที่หน้าอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 2. อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 3. ระบุชื่อเครื่องมือตรวจจับซีเอ็นวีที่เป็นที่มาของไฟล์ที่ถูกอัพโหลด 4. เลือกเทมเพลตเครื่องมือตรวจจับซีเอ็นวี 5. เลือกเทมเพลตกลุ่มตัวอย่าง 6. กดปุ่มอัพโหลด
	อินพุต	● ไฟล์ผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจชื่อ “CONTRA” และกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE” ที่บรรทัดที่ 3 ใส่ชื่อข้อมูลโครโมโซมเป็น “test” 1 sample chr start end cnv_type 2 TCGA-A7-A0CE 1 11868 14412 del 3 TCGA-A7-A0CE test 14362 29806 del 4 TCGA-A7-A0CE 1 29553 33264 del ● ชื่อเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA” ● เทมเพลตเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA” ● เทมเพลตกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE”
	ผลลัพธ์ที่คาดหวัง	● มีข้อความแจ้งแสดงความผิดพลาดบนหน้าเว็บว่า “At line: 3. Cannot map chromosome column. Please check file content whether matching with the configuration of tab file mapping.” ● หน้า “My Files” ไม่มีข้อมูลไฟล์ที่ทำการทดสอบอยู่
	ผลลัพธ์ที่เกิดขึ้นจริง	ตามที่คาดหวัง
	ผลการทดสอบ (ผ่าน / ไม่ผ่าน)	ผ่าน

ตารางที่ 9 กรณีทดสอบอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ใส่ข้อมูลตำแหน่งเบสเริ่มต้นผิด

กรณีทดสอบ 5	สถารการณ์ทดสอบ	อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ใส่ข้อมูลตำแหน่งเบสเริ่มต้นผิด
	ขั้นตอนการทดสอบ	1. ไปที่หน้าอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 2. อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 3. ระบุชื่อเครื่องมือตรวจจับซีเอ็นวีที่เป็นที่มาของไฟล์ที่ถูกอัพโหลด 4. เลือกเทมเพลตเครื่องมือตรวจจับซีเอ็นวี 5. เลือกเทมเพลตกลุ่มตัวอย่าง 6. กดปุ่มอัพโหลด
	อินพุต	● ไฟล์ผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจชื่อ “CONTRA” และกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE” ที่บรรทัดที่ 3 ใส่ชื่อข้อมูลเบสเริ่มต้นเป็น “test” 1 sample chr start end cnv_type 2 TCGA-A7-A0CE 1 11868 14412 del 3 TCGA-A7-A0CE 1 test 29806 del 4 TCGA-A7-A0CE 1 29553 33264 del ● ชื่อเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA” ● เทมเพลตเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA” ● เทมเพลตกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE”
	ผลลัพธ์ที่คาดหวัง	● มีข้อความแจ้งแสดงความผิดพลาดบนหน้าเว็บว่า “At line: 3. ‘start basepair: ‘test’ is not a number.” ● หน้า “My Files” ไม่มีข้อมูลไฟล์ที่ทำการทดสอบอยู่
	ผลลัพธ์ที่เกิดขึ้นจริง	ตามที่คาดหวัง
	ผลการทดสอบ (ผ่าน / ไม่ผ่าน)	ผ่าน

ตารางที่ 10 กรณีทดสอบอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ใส่ข้อมูลประเภทของซีเอ็นวีผิด

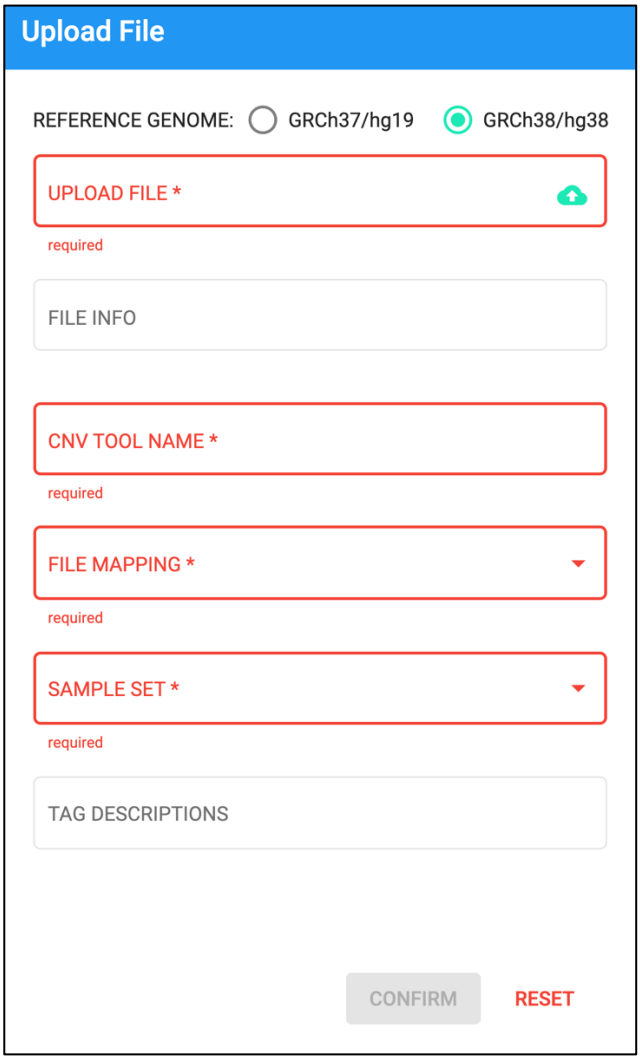
กรณีทดสอบ 6	สถารการณ์ทดสอบ	อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีที่ใส่ข้อมูลประเภทของซีเอ็นวีผิด
	ขั้นตอนการทดสอบ	1. ไปที่หน้าอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 2. อัพโหลดไฟล์ผลลัพธ์ซีเอ็นวี 3. ระบุชื่อเครื่องมือตรวจจับซีเอ็นวีที่เป็นที่มาของไฟล์ที่ถูกอัพโหลด 4. เลือกเทมเพลตเครื่องมือตรวจจับซีเอ็นวี 5. เลือกเทมเพลตกลุ่มตัวอย่าง 6. กดปุ่มอัพโหลด

	อินพุต	● ไฟล์ผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจชื่อ “CONTRA” และกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE” ที่บรรทัดที่ 3 ใส่ชื่อข้อมูลเบสเริ่มต้นเป็น “test”
		`1 sample chr start end cnv_type`
		`2 TCGA-A7-A0CE 1 11868 14412 del`
		`3 TCGA-A7-A0CE 1 14362 29806 test`
		`4 TCGA-A7-A0CE 1 29553 33264 del`
		● ชื่อเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA”
		● เทมเพลตเครื่องมือตรวจจับซีเอ็นวีชื่อ “CONTRA”
		● เทมเพลตกลุ่มตัวอย่างชื่อ “TCGA-A7-A0CE”

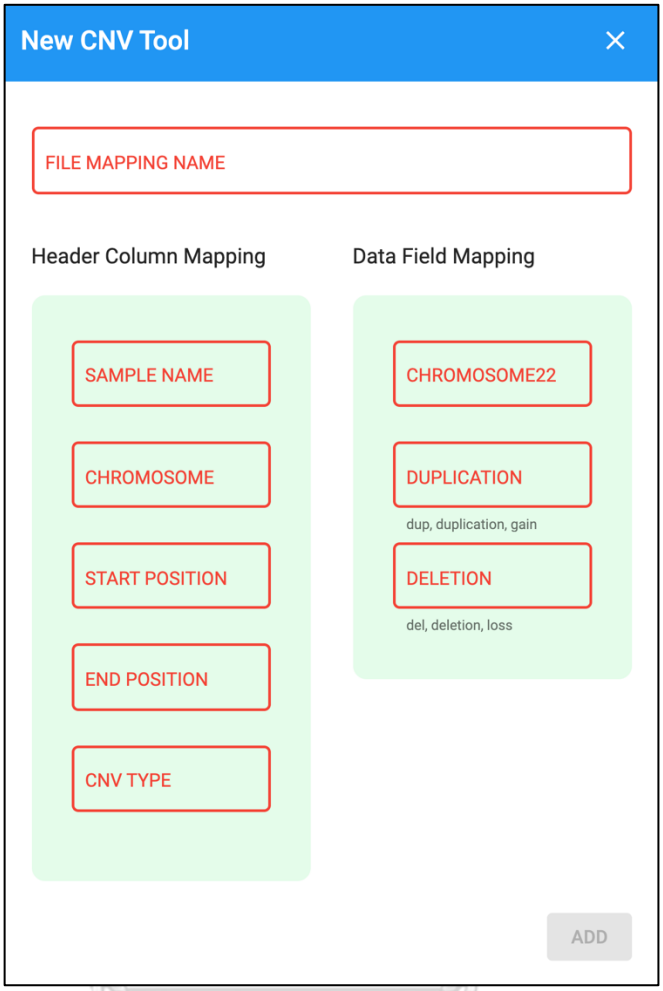
	ผลลัพธ์ที่คาดหวัง	● มีข้อความแจ้งแสดงความผิดพลาดบนหน้าเว็บว่า “At line: 3. Cannot map data named ‘test’. Please check file headers whether matching with the configuration of tab file mapping.” ● หน้า “My Files” ไม่มีข้อมูลไฟล์ที่ทำการทดสอบอยู่
	ผลลัพธ์ที่เกิดขึ้นจริง	ตามที่คาดหวัง
	ผลการทดสอบ (ผ่าน / ไม่ผ่าน)	ผ่าน

5.6.2 การจัดการความผิดพลาดจากการกรอกข้อมูลผิดพลาด

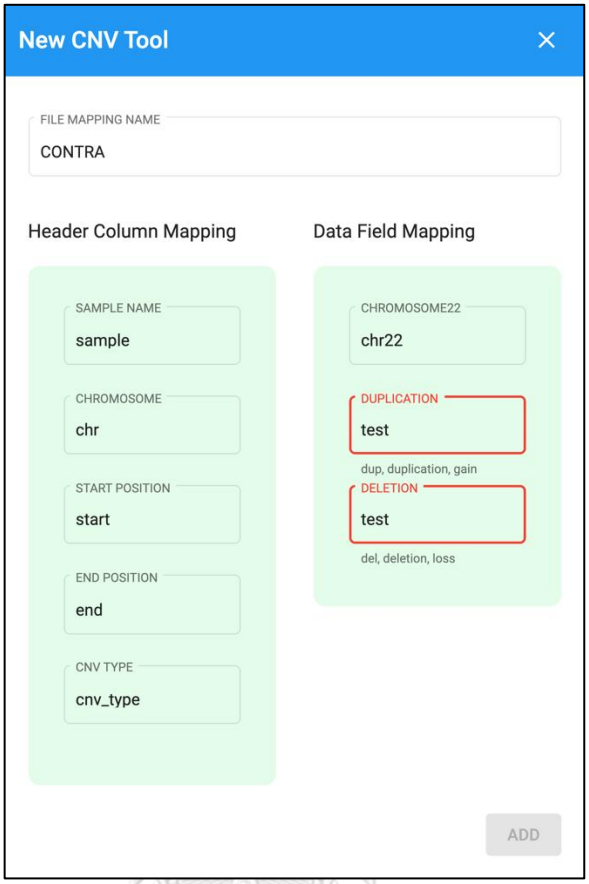
ผู้วิจัยของแสดงตัวอย่างการตรวจสอบความถูกต้องของข้อมูลก่อนนำข้อมูลเข้าระบบบางส่วนดังรูปที่ 64 – 66



รูปที่ 64 หน้าการอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีซึ่งแสดงการจัดการความผิดพลาดจากการไม่กรอกข้อมูลที่จำเป็นลงไปในระบบด้วยการไฮไลท์กล่องข้อความด้วยสีแดง



รูปที่ 65 ไดอะล็อกของเทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวีซึ่งแสดงการจัดการความผิดพลาดจากการไม่กรอกข้อมูลที่จำเป็นลงไปในระบบด้วยการไฮไลท์กล่องข้อความด้วยสีแดง



รูปที่ 66 ไดอะล็อกของเทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวีซึ่งแสดงการจัดการความผิดพลาดจากการกรอกข้อมูลผิดลงไปด้วยการไฮไลท์กล่องข้อความด้วยสีแดง

5.7 ผลการทดสอบความต้องการทรัพยากรเชิงคำนวณขั้นต่ำของระบบ (Minimum system requirements)

ซอฟต์แวร์อินซีเอ็นวีเวอร์ชันปัจจุบันสนับสนุนการทำงานแบบ interactive processing สำหรับการทดลองในงานวิจัยนี้ ใช้เครื่องเซิร์ฟเวอร์ที่มีซีพียู 4 คอร์ ของ Intel(R) Xeon(R) Gold 6140 CPU @ 2.30 GHz และในส่วนของมุมมองผู้ใช้ ผู้วิจัยได้ทำการทดลองบนเครื่องคอมพิวเตอร์โน้ตบุ๊ค MacBook Pro ซึ่งมีซีพียู 2 คอร์ของ Intel Core i5 และแสดงผลลัพธ์บนบราวเซอร์โอเปร่า (Opera web browser) ผู้วิจัยทำการทดลองทั้งการวิเคราะห์แบบตัวอย่างเดียว (individual-sampled analysis) และแบบหลายตัวอย่าง (multiple-sampled analysis) หมายเหตุ ในส่วนของการวิเคราะห์แบบตัวอย่างเดียว เนื่องจากข้อมูลที่ใช้มีจำนวนน้อยกว่าการวิเคราะห์แบบหลายตัวอย่างมาก ดังนั้นผู้วิจัยจะกำหนดความต้องการขั้นต่ำของระบบโดยอ้างอิงจากการวิเคราะห์แบบหลาย

ตัวอย่างเท่านั้น และจะทำการทดสอบการทำงานผ่านเครือข่ายอินเตอร์เน็ตเป็นจำนวน 10 ครั้ง และครั้งละ 1 request

สำหรับการทดลองในการวิเคราะห์แบบหลายตัวอย่าง โดยใช้ข้อมูลจากเครื่องมือตรวจจับซีเอ็นวีชื่อ CONTRA ของ 10 ตัวอย่าง ซึ่งมีขนาดไฟล์รวม 16.7 เมกะไบต์ พบว่าเพื่อให้เซิร์ฟเวอร์ประมวลผลทั้งหมดเซิร์ฟเวอร์ใช้ซีพียู และแรมสูงที่สุดที่ 377.02% CPU (จาก 4 คอร์) โดยใช้แรม 2.232 กิกะไบต์ และใช้เวลาในการตอบสนอง (response time) ทั้งหมดประมาณ 27 วินาที สำหรับในส่วนของมุมมองผู้ใช้เราพบว่าบราวเซอร์ใช้ซีพียู และแรมสูงที่สุดที่ 31.5% CPU (จาก 2 คอร์) ใช้แรม 188 เม็กกะไบต์ และใช้เวลาประมาณ 1 วินาที เพื่อแสดงผลหน้าเว็บ

นอกจากนี้ผู้วิจัยได้ประเมินประสิทธิภาพของซอฟต์แวร์อินซีเอ็นวีด้วยการวิเคราะห์ผลลัพธ์ของซีเอ็นวี 100 ตัวอย่างซึ่งมีขนาดไฟล์รวม 176 เมกะไบต์ โดยการจำลองข้อมูลใหม่ขึ้นมาโดยใช้ข้อมูล 10 ตัวอย่างจากการทดลองก่อนหน้านี้มาทำซ้ำเป็นข้อมูล 100 ตัวอย่าง หลังจากรวมผลลัพธ์ซีเอ็นวีเหล่านั้นผ่านซอฟต์แวร์อินซีเอ็นวี พบว่าเซิร์ฟเวอร์ใช้ซีพียู และแรมสูงที่สุดที่ 384.45% CPU (จาก 4 คอร์) โดยใช้แรม 4.495 กิกะไบต์ และใช้เวลาในการตอบสนองของเซิร์ฟเวอร์เพื่อประมวลผลลัพธ์ทั้งหมดประมาณ 458 วินาที ขณะที่บนเว็บบราวเซอร์ทางฝั่งผู้ใช้ต้องใช้ซีพียู 104.7% CPU (จาก 2 คอร์) แรม 1 กิกะไบต์ และใช้เวลา 8 วินาที เพื่อแสดงผลลัพธ์

เนื่องจากเวอร์ชันปัจจุบันของซอฟต์แวร์อินซีเอ็นวีสนับสนุนเพียงการทำงานแบบ interactive processing ทางผู้พัฒนาจึงแนะนำให้ติดตั้งซอฟต์แวร์อินซีเอ็นวีบนเครื่องเซิร์ฟเวอร์ที่มีซีพียูอย่างน้อย 4 คอร์ และมีแรมอย่างน้อย 8 กิกะไบต์ เพื่อประสบการณ์ที่ดีในการใช้งาน นอกจากนี้ผู้วิจัยได้วางแผนให้ซอฟต์แวร์อินซีเอ็นวีสามารถวิเคราะห์ผ่านทาง batch processing ได้ในอนาคต

บทที่ 6
การประยุกต์ใช้อินซีเอ็นวีกับการวิเคราะห์ข้อมูลเอ็กโซม

6.1 กลุ่มข้อมูลที่ใช้ในการวิจัย (Data sets)

ซอฟต์แวร์อินซีเอ็นวีใช้ข้อมูลผลลัพธ์ซีเอ็นวีจากงานวิจัยของ Zare et al. [14] ผลลัพธ์นี้ได้มาจากการรันเครื่องมือตรวจจับซีเอ็นวีได้แก่ ADTEx [82], cn.MOPS [83], CONTRA, ExomeCNV [84] และ VarScan2 [85] บนเอ็กโซม (whole exome sequences: WES) ของผู้ป่วยโรคมะเร็งเต้านม 10 ราย ในโครงการ BRCA project ซึ่งถูกสร้างโดย The Cancer Genome Atlas (TCGA) [86].

ตามที่ Zare et al. [14] ได้กล่าวไว้ ไฟล์ผลลัพธ์จาก ADTEx, cn.MOPS และ ExomeCNV อธิบายประเภทของซีเอ็นวีด้วยตัวเลขมาตรฐาน โดยที่ “1” หมายถึง ซีเอ็นวีประเภท deletion, “2” หมายถึง ลำดับเบสปกติซึ่งไม่เกิดซีเอ็นวีขึ้น (no CNV / normal)” และ “3” หมายถึง ซีเอ็นวีประเภท duplication และตัวเลขที่มากกว่า “3” จะหมายถึง การแปรผันเชิงโครงสร้างแบบ amplification ดังนั้นเพื่อให้สอดคล้องกับงานวิจัยของ Zare et al. [14] เราจึงคัดเลือกข้อมูลจากไฟล์ผลลัพธ์ซีเอ็นวีเฉพาะข้อมูลที่อธิบายประเภทของซีเอ็นวีด้วยตัวเลข “1” มาแทนด้วยซีเอ็นวีประเภท deletion และคัดเลือกข้อมูลที่อธิบายประเภทของซีเอ็นวีด้วยตัวเลข “3” และตัวเลขที่มากกว่านี้มาแทนซีเอ็นวีประเภท duplication

จากการศึกษาผลลัพธ์ซีเอ็นวีจาก CONTRA และ VarScan2 เราพบว่าผลลัพธ์เหล่านั้นแสดงประเภทของซีเอ็นวีด้วย log2ratio ซึ่งงานวิจัยของ Zare et al.[14] ใช้ค่า thresholds ±0.2 ในการแบ่งประเภทของซีเอ็นวี ดังนั้นเพื่อให้การทดลองของเราสอดคล้องกับงานวิจัยของ Zare et al.[14] เราจึงคัดเลือกผลลัพธ์ซีเอ็นวีเฉพาะผลลัพธ์ที่มีค่า log2ratio < -0.2 เป็นซีเอ็นวีประเภท deletion และคัดเลือกผลลัพธ์ซีเอ็นวีเฉพาะผลลัพธ์ที่มีค่า log2ratio > +0.2 เป็นซีเอ็นวีประเภท duplication และผลลัพธ์ซีเอ็นวีที่นอกเหนือจากสองค่าดังกล่าวจะถูกตัดทิ้งไม่นำไปใช้ในการทดลอง

นอกจากนี้เราได้ปรับเปลี่ยนโครงสร้างข้อมูลจากไฟล์ผลลัพธ์ของ *Zare et al. [14]* ก่อนใช้ในการทดลอง โดยทำขั้นตอนดังต่อไปนี้ (1) คัดกรองข้อมูลที่ไม่จำเป็นในการใช้งานซอฟต์แวร์อินซีเอ็นวีออก และ (2) ปรับเปลี่ยนโครงสร้างข้อมูลที่เหลือให้เหมาะสมกับเทมเพลตของเครื่องมือตรวจจับซีเอ็นวีซึ่งได้กำหนดไว้ก่อนหน้านี้ซึ่งได้อธิบายในหัวข้อเทมเพลตการแมปเครื่องมือตรวจจับซีเอ็นวี (CNV tool mapping templates) ก่อนหน้านี้

ท้ายสุด หลังจากนำไฟล์ผลลัพธ์ซีเอ็นวีจากหลายเครื่องมือตรวจจับ และหลายตัวอย่างมาผ่านกระบวนการข้างต้นแล้ว เราจึงทำการบีบอัดไฟล์เหล่านั้น และกำหนดชื่อกลุ่มข้อมูลดังกล่าวว่า

“demo data” และอัพโหลดขึ้นที่ https://github.com/saowwapark/inCNV/tree/master/demo-data เพื่อให้ผู้ใช้งานซอฟต์แวร์อินซีเอ็นวีสามารถเข้าถึงข้อมูลตัวอย่างที่ใช้ในการทดลองได้

6.2 ผลการวิจัย และการอภิปราย

ซอฟต์แวร์อินซีเอ็นวีแบ่งการวิเคราะห์ออกเป็น 3 รูปแบบ คือ การวิเคราะห์แบบตัวอย่างเดียว (individual-sampled analysis), การวิเคราะห์แบบหลายตัวอย่าง (multiple-sampled analysis) และ การวิเคราะห์แบบรวมกระบวนการ (combined-processed analysis)

6.2.1 การวิเคราะห์แบบตัวอย่างเดียว (Individual-sampled analysis)

การวิเคราะห์แบบตัวอย่างเดียวจะเน้นที่การคัดกรองแบบอินเตอร์เซคชัน (intersection) บนผลลัพธ์การรวมซีเอ็นวีจากเครื่องมือตรวจจับซีเอ็นหลายตัว โดยมีขั้นตอนการวิเคราะห์ดังต่อไปนี้

1) อัพโหลดผลลัพธ์ซีเอ็นวีของตัวอย่าง TCGA-BH-A0E0 จากเครื่องมือตรวจจับซีเอ็น 5 ตัว คือ ADTEx, cn.MOPS, CONTRA, ExomeCNV และVarScan2
2) รวมผลลัพธ์ซีเอ็นวีประเภท “deletion” บนโครโมโซม 17 จากเครื่องมือตรวจจับซีเอ็นวีดังกล่าว
3) คัดกรองผลลัพธ์ซีเอ็นวีเฉพาะซีเอ็นวีที่อยู่บนยีนมะเร็งบนโครโมโซม 17 ได้แก่ BIRC5, BRCA1, BRIP1, ERBB2, GRB7, KPNA2, KRT17, RAD51C และ TP53 โดยอ้างอิงจาก [87]
4) เรียงลำดับผลลัพธ์จากจำนวนของซีเอ็นวีที่ซ้อนทับกันในบริเวณใด ๆ บนโครโมโซม

จากการทดลอง เราพบว่าเครื่องมือตรวจจับซีเอ็นวีทั้ง 5 ตัว ให้ผลลัพธ์ซีเอ็นวีชนิด “deletion” ซ้อนทับกันตามตำแหน่งยีนมะเร็งบนโครโมโซมที่ 17 เป็นจำนวน 13 ซีเอ็นวี ดังรูปที่ 67 โดย 5 ซีเอ็นวีถูกตรวจจับด้วย 4 เครื่องมือ 7 ซีเอ็นวี ถูกตรวจจับด้วย 3 เครื่องมือ และมี 1 ซีเอ็นวีถูกตรวจจับด้วย 2 เครื่องมือ ผลเหล่านี้แสดงให้เห็นว่าซอฟต์แวร์อินซีเอ็นวีสามารถสร้างความสะดวกในการรวม, คัดกรอง และจัดลำดับความสำคัญของผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจจับซีเอ็นวีหลายเครื่องได้ หมายเหตุ (1) ผู้ใช้สามารถดูรายละเอียดของซีเอ็นวีที่ตรวจจับได้โดยการคลิกที่แถวของตาราง ตามที่ได้กล่าวไว้ในหัวข้อ “5.4.3.2 คอมโพเน้นต์ตารางรายละเอียด (Detailed table component)” (2) ซอฟต์แวร์อินซีเอ็นวีมีฟีเจอร์ของการหาบริเวณขนาบข้าง (flanking regions /

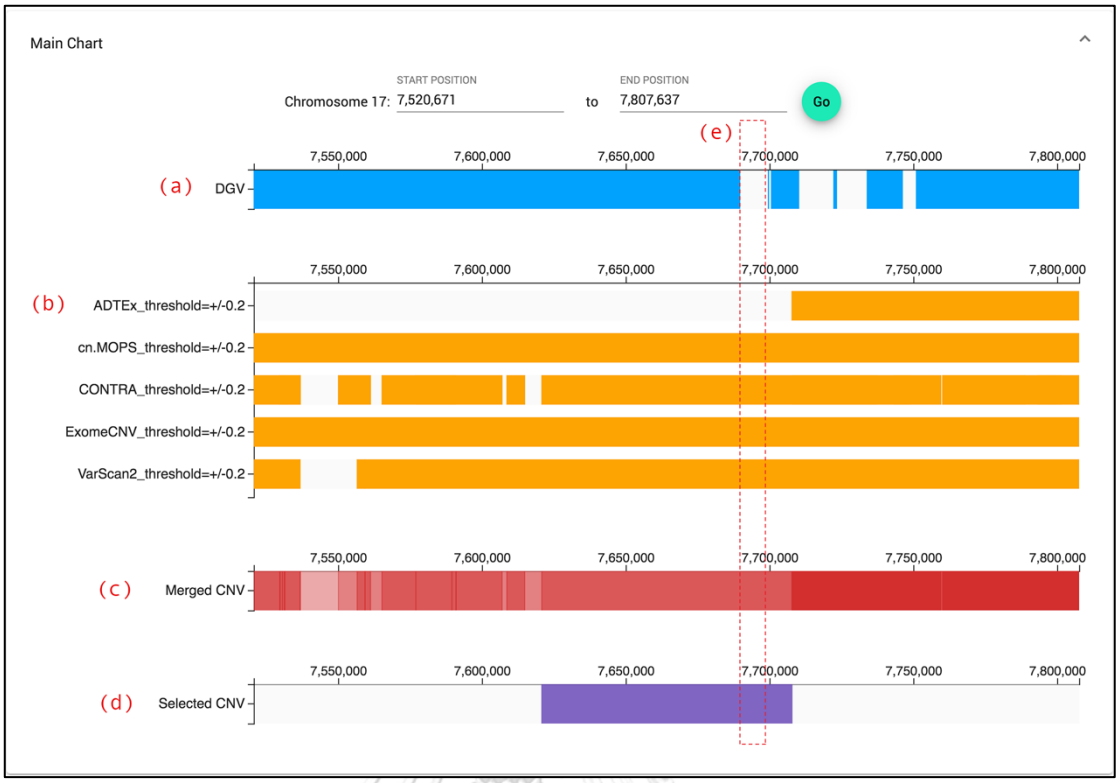
flanking sequences) ของตำแหน่งที่ถูกสงสัยว่าเกิดซีเอ็นวี ซึ่งสนับสนุนให้นักวิจัยนำไปทดสอบในห้องปฏิบัติการเพื่อตรวจสอบซีเอ็นวีที่ทำนายไว้ได้ ตามที่ได้อธิบายในหัวข้อ “5.2.2.4 ลำดับเบสบนจีโนมอ้างอิงของมนุษย์”

นอกจากนี้มีกรณีที่น่าสนใจ คือ ซอฟต์แวร์อินซีเอ็นวีพบบางซีเอ็นวีที่ถูกตรวจจับโดยหลายเครื่องมือ แต่บริเวณของซีเอ็นวีเหล่านั้นไม่มีความสัมพันธ์กับฐานข้อมูลดีจีวี ดังรูปที่ 68 ซึ่งลักษณะดังกล่าวสามารถสรุปได้ว่าซอฟต์แวร์อินซีเอ็นวีสามารถช่วยผู้ใช้จำกัดขอบเขตของการหาซีเอ็นวีใหม่ (novel CNVs) ที่ยังไม่เคยถูกรายงานมาก่อนหน้า

No.	Chr	Start Position	End Position	CNV Type	Overlapping Numbers	CNV Tools
1	17	7,620,671	7,707,637	deletion	4	CONTRA_threshold=+/-0.2; cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
2	17	41,610,297	41,622,152	deletion	4	CONTRA_threshold=+/-0.2; cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
3	17	41,623,800	41,687,706	deletion	4	CONTRA_threshold=+/-0.2; cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
4	17	61,678,230	61,688,544	deletion	4	CONTRA_threshold=+/-0.2; cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
5	17	78,201,653	78,226,515	deletion	4	CONTRA_threshold=+/-0.2; cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
6	17	39,684,560	39,722,095	deletion	3	cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
7	17	39,728,310	39,738,530	deletion	3	cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
8	17	39,743,173	39,775,688	deletion	3	cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
9	17	43,037,060	43,045,439	deletion	3	CONTRA_threshold=+/-0.2; cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2
10	17	43,138,386	43,180,330	deletion	3	cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
11	17	61,790,804	61,896,792	deletion	3	cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
12	17	67,351,273	68,164,813	deletion	3	cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2; VarScan2_threshold=+/-0.2
13	17	58,639,522	59,234,816	deletion	2	cn.MOPS_threshold=+/-0.2; ExomeCNV_threshold=+/-0.2

Items per page: 25 1 – 13 of 13

รูปที่ 67 ตารางรวมผลลัพธ์ซีเอ็นวีประเภท deletion บนโครโมโซม 17 ของตัวอย่าง TCGA-BH-A0E0 จากเครื่องมือตรวจจับซีเอ็นวี ADTEx, cn.MOPS, CONTRA, ExomeCNV และVarScan2 ที่ผ่านการคัดกรองข้อมูลเลือกเฉพาะซีเอ็นวีที่อยู่บนยีนมะเร็งเต้านม BIRC5, BRCA1, BRIP1, ERBB2, GRB7, KPNA2, KRT17, RAD51C และ TP53



รูปที่ 68 แผนภูมิหลักของการรวมผลลัพธ์ซีเอ็นวีประเภท deletion บนโครโมโซม 17 ในช่วงตำแหน่งเบสที่7,520,671 - 7,807,637 ของตัวอย่าง TCGA-BH-A0E0 จากเครื่องมือตรวจจับซีเอ็นวี ADTEx, cn.MOPS, CONTRA, ExomeCNV และVarScan2 โดยมีรายละเอียดดังนี้ (a) แผนภูมิดีจีวีแสดงตำแหน่งซีเอ็นวีที่มีรายงานในฐานข้อมูลดีจีวี (b) แผนภูมิข้อมูลนำเข้าผลลัพธ์ซีเอ็นวีซึ่งแสดงตำแหน่งซีเอ็นวีที่เครื่องมือตรวจจับซีเอ็นวีแต่ละตัวตรวจจับได้ (c) แผนภูมิผลรวมซีเอ็นวีซึ่งแสดงความหนาแน่นของซีเอ็นวีจากเครื่องมือตรวจจับซีเอ็นวีทั้ง 5 ตัว (d) แผนภูมิแสดงซีเอ็นวีที่ถูกเลือกเพื่อแสดงตำแหน่งซีเอ็นวีที่เลือกไว้ (e) บริเวณที่น่าสนใจสำหรับการหาซีเอ็นวีตำแหน่งใหม่ที่ยังไม่เคยถูกรายงาน

6.2.2 การวิเคราะห์แบบหลายตัวอย่าง (Multiple-sampled analysis)

การวิเคราะห์แบบหลายตัวอย่างจะเน้นที่การคัดกรองแบบอินเตอร์เซคชัน (intersection) ของผลลัพธ์การรวมซีเอ็นวีจากหลายตัวอย่างซึ่งได้มาจากเครื่องมือตรวจจับซีเอ็นวีตัวเดียวกัน ในการวิเคราะห์นี้ ผู้วิจัยใช้ผลลัพธ์ซีเอ็นวีที่ได้จากเครื่องมือตรวจจับซีเอ็นวีชื่อ CONTRA ของผู้ป่วยโรคมะเร็ง 10 รายได้แก่ TCGA-A7-A0CE, TCGA-AC-A2BK, TCGA-BH-A0B3, TCGA-BH-A0DT, TCGA-BH-A0E0, TCGA-BH-A1FC, TCGA-BH-A18R, TCGA-BH-A18U, TCGA-E2-A1LG และ TCGA-E9-A1NH แลใช้ซอฟต์แวร์อินซีเอ็นวีในการรวมผลลัพธ์เข้าด้วยกันแล้วค้นหาซีเอ็นวีที่เกี่ยวข้องกับยีนมะเร็งเต้านมดังที่ทดลองในการวิเคราะห์แบบตัวอย่างเดียวซึ่งได้อธิบายก่อนหน้า โดย

ในการทดลองนี้พบ 14 ซีเอ็นวีตกอยู่ในบริเวณที่เป็นตำแหน่งของยีนมะเร็งเต้านม BIRC5, BRCA1, BRIP1, ERBB2, GRB7, KPNA2, KRT17, RAD51C และ TP53 ได้แก่ 1 ซีเอ็นวีพบในผู้ป่วย 7 ราย, 3 ซีเอ็นวี พบในผู้ป่วย 6 ราย, 3 ซีเอ็นวีพบในผู้ป่วย 5 ราย และ 4 ซีเอ็นวีพบในผู้ป่วย 3 ราย สิ่งเหล่านี้อธิบายได้ว่าซอฟต์แวร์อินซีเอ็นวีช่วยระบุความสัมพันธ์ของกลุ่มคนที่ป่วยเป็นโรคเดียวกันได้ (รูปที่ 69)

นอกจากนี้ซอฟต์แวร์อินซีเอ็นวียังสามารถถูกนำมาให้ได้ในกรณีอื่น ๆ อีกด้วย เช่น

1) ซอฟต์แวร์อินซีเอ็นวีสามารถถูกใช้ในการหา *de novo* CNVs ของตัวอย่างที่อยู่ในครอบครัวเดียวกัน [5, 29] และเพื่อที่จะทำเช่นนั้นได้ ผู้ใช้ต้องดำเนินการตามขั้นตอนดังนี้ (1) นำไฟล์ผลลัพธ์ซีเอ็นวีของพ่อ แม่ และลูกซึ่งเป็นตัวอย่างที่สนใจเข้าซอฟต์แวร์อินซีเอ็นวี (2) ใช้ซอฟต์แวร์อินซีเอ็นวีรวมผลลัพธ์ซีเอ็นเหล่านั้นเข้าด้วยกัน (3) ใช้ซอฟต์แวร์อินซีเอ็นวีคัดกรอง common CNVs หรือ known CNVs ออก (4) สำรวจว่าซีเอ็นวีที่เหลือของลูกว่าแตกต่างจากซีเอ็นวีของพ่อแม่หรือไม่ หากพบว่าแตกต่างจากพ่อและแม่ เราสามารถสรุปได้ว่าซีเอ็นวีนั้นมีแนวโน้มที่จะเป็น *de novo* CNVs อย่างไรก็ตามการสรุปเช่นนี้จำเป็นต้องมีการทำการทดลองในห้องปฏิบัติการเพิ่มเติมด้วยเพื่อยืนยันความถูกต้องของสมมติฐาน

2) ซอฟต์แวร์อินซีเอ็นวีสามารถถูกใช้ในการระบุตัวอย่างที่มีแนวโน้มจะเป็นโรคที่เราสนใจได้ โดยทำตามขั้นตอนดังนี้ (1) นำไฟล์ผลลัพธ์ซีเอ็นวีของของกลุ่มตัวอย่างที่เป็นโรคที่เราสนใจกับผลลัพธ์ซีเอ็นวีของตัวอย่างที่เราต้องการทำการทดลองเข้าซอฟต์แวร์อินซีเอ็นวี (2) ใช้ซอฟต์แวร์อินซีเอ็นวีรวมผลลัพธ์ซีเอ็นวีเหล่านั้นเข้าด้วยกัน (3) ใช้ซอฟต์แวร์อินซีเอ็นวีคัดกรองผลลัพธ์การรวมซีเอ็นวีด้วยการค้นหายีนที่เกี่ยวข้องกับโรค (4) ใช้ซอฟต์แวร์อินซีเอ็นวีคัดกรองผลลัพธ์การรวมซีเอ็นวีด้วยการค้นหาตัวอย่างที่สนใจ (5) ใช้ซอฟต์แวร์อินซีเอ็นวีเรียงลำดับผลลัพธ์ด้วยจำนวนตัวอย่างที่ซ้อนทับกันจากมากไปหาน้อย (รายละเอียดอธิบายในหัวข้อ “5.4.3 ผลลัพธ์การรวมซีเอ็นวีทั้งหมด (All merged CNVs)”) ดังนั้นหลังจากการคัดกรองแล้วหากยังเหลือซีเอ็นวีจำนวนมาก และซีเอ็นวีเหล่านั้นมีจำนวนที่ซ้อนทับกันที่หนาแน่น เราสามารถสรุปได้ว่าตัวอย่างที่เราทำการทดลองมีแนวโน้มที่จะเป็นโรคที่เราสนใจ อย่างไรก็ตามการสรุปเช่นนี้จำเป็นต้องมีการทำการทดลองในห้องแลปเพิ่มเติมด้วยยืนยันความถูกต้องของสมมติฐาน

No.	Chr	Start Position	End Position	CNV Type	Overlapping Numbers	Sample Names
1	17	43,037,060	43,045,439	deletion	7	TCGA-A7-A0CE; TCGA-BH-A0B3; TCGA-BH-A0DT; TCGA-BH-A0E0; TCGA-BH-A1FC; TCGA-BH-A18R; TCGA-E9-A1NH
2	17	7,658,294	7,664,061	deletion	6	TCGA-BH-A0B3; TCGA-BH-A0E0; TCGA-BH-A1FC; TCGA-BH-A18R; TCGA-BH-A18U; TCGA-E2-A1LG
3	17	7,664,061	7,707,637	deletion	6	TCGA-BH-A0B3; TCGA-BH-A0E0; TCGA-BH-A1FC; TCGA-BH-A18R; TCGA-BH-A18U; TCGA-E2-A1LG
4	17	39,728,310	39,738,530	deletion	6	TCGA-A7-A0CE; TCGA-BH-A0B3; TCGA-BH-A0DT; TCGA-BH-A1FC; TCGA-BH-A18R; TCGA-BH-A18U
5	17	41,613,683	41,621,831	deletion	5	TCGA-A7-A0CE; TCGA-BH-A0B3; TCGA-BH-A0E0; TCGA-BH-A1FC; TCGA-BH-A18R
6	17	41,623,800	41,656,988	deletion	5	TCGA-A7-A0CE; TCGA-BH-A0B3; TCGA-BH-A0E0; TCGA-BH-A1FC; TCGA-BH-A18R
7	17	39,684,560	39,705,857	deletion	3	TCGA-BH-A0B3; TCGA-BH-A0DT; TCGA-BH-A18R
8	17	39,743,173	39,766,029	deletion	3	TCGA-BH-A0B3; TCGA-BH-A0DT; TCGA-BH-A18U
9	17	43,163,903	43,170,998	deletion	3	TCGA-AC-A2BK; TCGA-BH-A0B3; TCGA-BH-A18R
10	17	78,201,653	78,226,515	deletion	3	TCGA-BH-A0B3; TCGA-BH-A0E0; TCGA-BH-A18U
11	17	61,678,230	61,688,131	deletion	2	TCGA-BH-A0B3; TCGA-BH-A0E0
12	17	61,853,711	61,896,792	deletion	2	TCGA-BH-A0B3; TCGA-BH-A18R
13	17	68,014,251	68,047,417	deletion	2	TCGA-BH-A0B3; TCGA-E9-A1NH
14	17	58,663,984	58,754,813	deletion	1	TCGA-BH-A0B3

Items per page: 25 1 – 14 of 14

รูปที่ 69 ตารางการรวมผลลัพธ์ซีเอ็นวีประเภท deletion บนโครโมโซม 17 จากเครื่องมือตรวจจับซีเอ็นวี CONTRA ของตัวอย่าง TCGA-A7-A0CE, TCGA-AC-A2BK, TCGA-BH-A0B3, TCGA-BH-A0DT, TCGA-BH-A0E0, TCGA-BH-A1FC, TCGA-BH-A18R, TCGA-BH-A18U, TCGA-E2-A1LG และ TCGA-E9-A1NH ที่ผ่านการคัดกรองข้อมูลเลือกเฉพาะซีเอ็นวีที่อยู่บนยีนมะเร็งเต้านม BIRC5, BRCA1, BRIP1, ERBB2, GRB7, KPNA2, KRT17, RAD51C และ TP53

6.2.3 การวิเคราะห์แบบรวมกระบวนการ (Combined-processed analysis)

ผู้ใช้สามารถรวมการวิเคราะห์แบบตัวอย่างเดียว (individual-sampled analysis) และการวิเคราะห์แบบหลายตัวอย่าง (multiple-sampled analysis) เข้าไว้ด้วยกันเพื่อช่วยระบุหาซีเอ็นวีที่สนใจได้ ยกตัวอย่างเช่น ผู้ใช้สามารถประยุกต์การวิเคราะห์แบบหลายตัวอย่าง เพื่อหาตัวอย่างซึ่งเสี่ยงต่อการเกิดโรค แล้วทำการจัดลำดับความสำคัญของซีเอ็นวีด้วยอัลกอริทึมจากเครื่องมือตรวจจับซีเอ็นวีหลายตัวด้วยการวิเคราะห์แบบตัวอย่างเดียวเพื่อช่วยในการคัดกรองผลลัพธ์ก่อนนำไปทำการทดลองในห้องปฏิบัติการเพื่อสรุปผลต่อไป

บทที่ 7
สรุปผลการวิจัย

งานวิจัยนี้ได้แสดงให้เห็นว่าซอฟต์แวร์อินซีเอ็นวีสามารถรวมผลลัพธ์ซีเอ็นวีจากเครื่องมือตรวจจับจำนวนมากเพื่อระบุซีเอ็นวีให้ครอบคลุมกับลักษณะที่หลากหลายของซีเอ็นวีได้ (หมายเหตุ ความถูกต้องในการระบุตำแหน่งซีเอ็นวีขึ้นอยู่กับอัลกอริทึมที่เลือกใช้ในเครื่องมือตรวจจับซีเอ็นวี) รวมถึงสามารถรวมผลลัพธ์ซีเอ็นวีจากหลายตัวอย่างเพื่อหาความสัมพันธ์ของตัวอย่างกับโรคที่สนใจได้ นอกจากนี้ ซอฟต์แวร์อินซีเอ็นวีสามารถช่วยผู้ใช้คัดกรอง จัดลำดับความสำคัญ เลือก และให้คำอธิบายกับลำดับเบสที่มีแนวโน้มที่จะเป็นซีเอ็นวีเพื่อทุ่นแรง ทรัพยากร และเงินทุนในการทดลองหาซีเอ็นวีในห้องปฏิบัติการ (wet-lab experiments)

ซอฟต์แวร์อินซีเอ็นวีสามารถใช้งานได้ในระยะยาว และผู้ใช้สามารถมั่นใจในความถูกต้องของข้อมูลที่นำเสนอ ผู้พัฒนาทำการเตรียมกลุ่มข้อมูลคำอธิบายทางจีโนม (Genome annotation dataset) ซึ่งระบุหมายเลขเวอร์ชัน และคำอธิบายถึงแหล่งที่มาของข้อมูล อัพโหลดไว้ที่ GitHub repository เพื่อให้ซอฟต์แวร์อินซีเอ็นวีสามารถเข้าถึงข้อมูลดังกล่าวและอัพเดทเวอร์ชันฐานข้อมูลในซอฟต์แวร์อินซีเอ็นวีได้อัตโนมัติ นอกจากนี้ ผู้ใช้ยังสามารถมั่นใจในความถูกต้องของข้อมูล โดยการตรวจสอบความถูกต้องของข้อมูลที่ถูกอัพโหลดไว้บน GitHub repository

ซอฟต์แวร์อินซีเอ็นวียังเหมาะกับผู้ที่ขาดความรู้ทางด้านการเขียนโปรแกรม ผู้ใช้สามารถติดตั้งซอฟต์แวร์อินซีเอ็นวีได้โดยง่ายโดยการติดตั้งผ่านทางด็อกเกอร์ ส่งผลให้ผู้ใช้ไม่ต้องคอมไพล์ (compile) โค้ดโปรแกรม (source codes) ดาวน์โหลดไลบรารีของโปรแกรม (program libraries) เช็คความเข้ากันได้ของเวอร์ชันไลบรารีด้วยตนเอง นอกจากนี้ ผู้ใช้สามารถใช้งานซอฟต์แวร์อินซีเอ็นวีได้โดยง่ายผ่านทางการใช้ภาพเป็นตัวปฏิสัมพันธ์กับผู้ใช้ (graphical user interface: GUI) ทำให้ผู้ใช้ไม่ต้องเขียนสคริปต์ (script) เพื่อเรียกใช้งานซอฟต์แวร์อินซีเอ็นวีด้วยตนเอง

ท้ายสุดนอกจากซอฟต์แวร์อินซีเอ็นวีสามารถถูกใช้กับลำดับเบสทั้งหมดบนเอ็กโซม (whole exome sequencing: WES) ผู้ใช้สามารถใช้งานซอฟต์แวร์อินซีเอ็นวีร่วมกับลำดับเบสทั้งหมดบนจีโนม (whole genome sequencing: WGS) ได้อีกด้วย โดยเปลี่ยนข้อมูลนำเข้าจากข้อมูลลำดับเบสทั้งหมดบนเอ็กโซม เป็นข้อมูลลำดับเบสทั้งหมดบนจีโนม แล้วดำเนินขั้นตอนการใช้งานตาม “บทที่ 4 วิธีการดำเนินการวิจัย” และ “บทที่ 6 การทดลอง และผลการทดลอง” ที่ได้กล่าวไว้ข้างต้นของงานวิจัยนี้

บทที่ 8
แนวทางวิจัยในอนาคต

ผู้วิจัยได้เสนอแนวทางในการปรับปรุงและพัฒนาระบบเพิ่มเติมในหัวข้อต่อไปนี้

8.1 พัฒนาส่วนหลังของโปรแกรมด้วยภาษาจาวาร่วมกับโหนดเจเอส

งานวิจัยนี้มีการแบ่งการทำงานส่วนหลังของโปรแกรม (backend) ออกเป็น 2 ส่วน คือ ส่วนเว็บเซิร์ฟเวอร์สำหรับวิเคราะห์ผลลัพธ์ซีเอ็นวี และส่วนสคริปต์สำหรับจัดการคำอธิบายจีโนม (อันได้แก่ เตรียม สร้าง และอัพเดทคำอธิบายจีโนม) ซึ่งถูกพัฒนาร่วมกันในซอฟต์แวร์ตัวเดียวด้วยโหนดเจเอส อย่างไรก็ตามการผนวกงานสองส่วนนี้เข้าด้วยกันในซอฟต์แวร์ตัวเดียวทำให้ขาดความยืดหยุ่นในการทำงานลักษณะ multi-thread processing ดังนั้นผู้วิจัยจึงเสนอว่า ควรแยกพัฒนาโปรแกรมสองส่วนออกจากกันอย่างสิ้นเชิง โดยพัฒนาส่วนของเว็บเซิร์ฟเวอร์สำหรับการวิเคราะห์ผลลัพธ์ซีเอ็นวีด้วยภาษาจาวา (Java) ร่วมกับสปริงเฟรมเวิร์ค (Spring framework) เพื่อให้เซิร์ฟเวอร์สามารถทำงานลักษณะ multi-thread processing ได้สะดวก และพัฒนาส่วนสคริปต์สำหรับจัดการคำอธิบายจีโนมด้วยการแยกเซอร์วิส (service) ออกมา

8.2 สร้างระบบ batch processing สำหรับรวมผลลัพธ์ซีเอ็นวี

งานวิจัยนี้ทำงานในลักษณะ interactive ซึ่งมีข้อเสียคือ หากข้อมูลผลลัพธ์ของซีเอ็นวีมีขนาดใหญ่ เซิร์ฟเวอร์ก็จะใช้เวลานานในการให้ได้มาซึ่งผลลัพธ์ส่งผลให้ผู้ใช้ต้องรอผลลัพธ์ทางหน้าเว็บบราวเซอร์เป็นเวลานาน และอาจเกิดภาวะสูญเสียการเชื่อมต่อกับเครื่องเซิร์ฟเวอร์ได้ (connection loss) นอกจากนี้หากผู้ใช้งานมีจำนวนมากก็จะทำให้เกิดคอขวด และผู้ใช้ก็จะต้องรอผลลัพธ์นานมากขึ้น และเซิร์ฟเวอร์อาจทำงานหนักเกินไปเกิดการใช้ซีพียู และหน่วยความจำเต็มได้ ดังนั้นผู้วิจัยจึงเห็นว่าควรเปลี่ยนระบบจากการทำงานในลักษณะ interactive processing ให้เป็น batch processing เพื่อให้ระบบสามารถรองรับการประมวลผลข้อมูลขนาดใหญ่ และจำนวนผู้ใช้งานจำนวนมากได้ โดยเสนอให้ระบบสามารถแจ้งสถานะของงานที่ผู้ใช้ขอให้ระบบทำได้ทางอีเมล

8.3 ส่งข้อมูล และประมวลผลลัพธ์ด้วยระบบ stream processing

งานวิจัยนี้ส่งข้อมูลระหว่างส่วนหน้า (frontend) และส่วนหลังของโปรแกรม (backend) ด้วยข้อมูลทั้งหมดภายในครั้งเดียว ไม่ได้แบ่งข้อมูลเป็นส่วนย่อยแล้วทยอยส่งจนครบ รวมถึงประมวลผลข้อมูลบนหน่วยความจำชั่วคราว หรือแรม (ram) ทั้งหมดทีเดียว ไม่ได้แบ่งผลลัพธ์ในการประมวลผลเป็นส่วนย่อยแล้วเก็บลงในหน่วยความจำถาวร (storage) ทำให้หากข้อมูลมีขนาดใหญ่เกินกว่าที่จะที่ระบบรองรับได้ก็จะเกิดปัญหาขึ้น เช่น แรมบนเครื่องเซิร์ฟเวอร์ไม่สามารถประมวลผลข้อมูลทั้งหมดได้

ภายในครั้งเดียวทำให้เซิร์ฟเวอร์ตายได้ แรมบนหน้าเว็บบราวเซอร์ไม่สามารถรองผลลัพธ์การวิเคราะห์ซีเอ็นวีทั้งหมดได้ทำให้เว็บบราวเซอร์ไม่ตอบสนองได้ หรือปัญหาเรื่องความถูกต้องของข้อมูลที่วิเคราะห์รวมถึงการแสดงผล เป็นต้น ดังนั้นผู้วิจัยจึงเสนอให้เพิ่มการทำงานด้วยระบบ stream processing ในบางส่วนของโปรแกรม เช่น การอัพโหลดไฟล์ผลลัพธ์ซีเอ็นวีเข้าระบบ การรวมผลลัพธ์ซีเอ็นวีที่ได้จากเครื่องมือตรวจจับซีเอ็นวี และการส่งผลลัพธ์การประมวลผลไปยังเว็บบราวเซอร์

เนื่องจากซอฟต์แวร์อินซีเอ็นวีเน้นการนำเสนอในรูปแบบการสร้างภาพกราฟฟิกที่สามารถมีปฏิสัมพันธ์กับผู้ใช้ ดังนั้นสำหรับการส่งผลลัพธ์การประมวลผลไปยังเว็บบราวเซอร์ ผู้วิจัยเสนอให้เซิร์ฟเวอร์ส่งผลลัพธ์เพียงบางส่วนเท่าที่ใช้ในการแสดงผลบนหน้าจอ ณ เวลาหนึ่งเท่านั้น โดยให้เก็บผลลัพธ์การประมวลผลทั้งหมดในรูปแบบของไฟล์ พร้อมทั้งสร้างไฟล์ที่เก็บอินเด็กของผลลัพธ์นั้นไว้ (index file) เพื่อใช้ในการเข้าถึงตำแหน่งของข้อมูลผลลัพธ์ได้อย่างรวดเร็ว เมื่อผู้ใช้งานสั่งเว็บบราวเซอร์ให้ร้องขอดูข้อมูลผลลัพธ์ที่ประมวลผลได้ เบราเซอร์ก็จะสั่งงานเซิร์ฟเวอร์ให้อ่านไฟล์ผลลัพธ์ด้วยไฟล์อินเด็ก แล้วข้อมูลผลลัพธ์เท่าที่จำเป็นจะถูกส่งไปยังเว็บบราวเซอร์เพื่อให้แสดงผลต่อไป

8.4 เก็บข้อมูลคำอธิบายจีโนมทั้งหมดในรูปแบบไฟล์

งานวิจัยนี้เก็บข้อมูลคำอธิบายจีโนม เช่น คำอธิบายชื่อยีน คำอธิบายการแปรผันของคนปกติ และคำอธิบายการแปรผันที่เกี่ยวกับโรค เก็บข้อมูลลงในฐานข้อมูล และเนื่องจากประมวลผลลัพธ์การตรวจจับซีเอ็นวีจำเป็นต้องยิง http request ไปยังฐานข้อมูลเหล่านี้จำนวนมากทำให้เกิดผลเสีย คือ เกิดทราฟฟิค (traffic) หนาแน่นในการอ่านข้อมูลจากฐานข้อมูล ทำให้ในระหว่างที่เซิร์ฟเวอร์กำลังประมวลผลของข้อมูลผู้ใช้คนแรกยังไม่เสร็จ ผู้ใช้คนที่สองจะไม่สามารถเข้าถึงฐานข้อมูลนี้ได้ นอกจากนี้การอ่านข้อมูลจากฐานข้อมูลยังกินทรัพยากรของเครื่องคอมพิวเตอร์มากกว่าการอ่านจากไฟล์ ดังนั้นผู้วิจัยจึงเสนอแนวทางวิจัยในอนาคตให้ผู้สนใจบันทึกข้อมูลคำอธิบายจีโนมเหล่านี้ลงไฟล์ พร้อมทั้งสร้างไฟล์อินเด็ก และอัลกอริทึมในการเข้าถึงคำอธิบายจีโนมเหล่านี้เพื่อแก้ปัญหาที่กล่าวมาข้างต้น

8.5 สร้าง unit test บนโปรแกรมส่วนหน้า

งานวิจัยนี้พัฒนาโปรแกรมส่วนหน้าด้วยแองกูลาร์เฟรมเวิร์คซึ่งมีอาศัยหลักการทำงานของ dependency injection ทำให้สามารถ reuse โปรแกรมให้เหมาะกับการทำ unit test ได้ ผู้วิจัยจึงเสนอแนวทางวิจัยในอนาคตให้ผู้สนใจสร้าง unit test เพื่อใช้ประโยชน์จากเฟรมเวิร์คนี้ได้อย่างเต็มประสิทธิภาพ ในการสร้างซอฟต์แวร์ที่มีคุณภาพมากขึ้น

8.6 สร้าง GUI สำหรับการอัพเดทคำอธิบายจีโนม

งานวิจัยนี้ทำการอัพเดทคำอธิบายจีโนมโดยการใช้ batch processing ตามตารางเวลาที่กำหนดไว้ใน “Cron job” ว่าจะให้ตรวจสอบ และอัพเดทคำอธิบายจีโนมได้เวลาไหนตามเวลาบนเครื่องเซิร์ฟเวอร์ที่ซอฟต์แวร์อินซีเอ็นวีถูกติดตั้ง ซึ่งวิธีดังกล่าวอาจมีกรณีการอัพเดทฐานข้อมูลในขณะที่ผู้ใช้ใช้งานซอฟต์แวร์อินซีเอ็นวีวิเคราะห์หาซีเอ็นวีอยู่ได้ ทำให้การวิเคราะห์หาซีเอ็นวีอาจเกิดการหยุดชะงัก หรือได้ผลลัพธ์ที่ผิดพลาดได้ ดังนั้นเพื่อความสะดวกในการใช้งาน ผู้วิจัยเสนอแนวทางวิจัยในอนาคตให้ผู้สนใจเพิ่มหน้าเว็บไซด์บนซอฟต์แวร์อินซีเอ็นวีเพื่อให้สามารถจัดการอัพเดทคำอธิบายจีโนมได้ผ่านทาง GUI โดยมีขั้นตอนดังนี้

1. มีหน้าเว็บไซด์สำหรับบอกที่มา และเวอร์ชันของคำอธิบายจีโนมที่ใช้อยู่ในปัจจุบันแก่ผู้ใช้ทุกคน
2. กำหนดสิทธิ์ผู้ใช้งานระบบ ให้มีสิทธิ์ของผู้ดูแลระบบ ซึ่งสามารถอัพเดทคำอธิบายจีโนมได้
3. สร้าง GUI สำหรับผู้ดูแลระบบให้สามารถอัพเดทคำอธิบายจีโนมได้ผ่านทางหน้าเว็บของซอฟต์แวร์อินซีเอ็นวี โดยให้ผู้ใช้สามารถเลือกได้ว่าจะทำการอัพเดทฐานคำอธิบายจีโนมประเภทใดบ้าง

บรรณานุกรม

1. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.
2. Beckmann, J.S., X. Estivill, and S.E. Antonarakis, *Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability.* Nat Rev Genet, 2007. **8**(8): p. 639-46.
3. Hastings, P.J., et al., *Mechanisms of change in gene copy number.* Nat Rev Genet, 2009. **10**(8): p. 551-64.
4. Buchanan, J.A. and S.W. Scherer, *Contemplating effects of genomic structural variation.* Genet Med, 2008. **10**(9): p. 639-47.
5. Bacchelli, E., et al., *An integrated analysis of rare CNV and exome variation in Autism Spectrum Disorder using the Infinium PsychArray.* Sci Rep, 2020. **10**(1): p. 3198.
6. Szatkiewicz, J.P., et al., *Detecting large copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample.* Mol Psychiatry, 2013. **18**(11): p. 1178-84.
7. Dajani, R., et al., *CNV Analysis Associates AKNAD1 with Type-2 Diabetes in Jordan Subpopulations.* Sci Rep, 2015. **5**: p. 13391.
8. Glessner, J.T., et al., *Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data.* Circ Res, 2014. **115**(10): p. 884-896.
9. Park, G., et al., *Multiphasic analysis of whole exome sequencing data identifies a novel mutation of ACTG1 in a nonsyndromic hearing loss family.* BMC Genomics, 2013. **14**: p. 191.
10. Shearer, A.E., et al., *Copy number variants are a common cause of non-syndromic hearing loss.* Genome Med, 2014. **6**(5): p. 37.
11. Zampaglione, E., et al., *Copy-number variation contributes 9% of pathogenicity in the inherited retinal degenerations.* Genet Med, 2020.
12. Majewski, J., et al., *What can exome sequencing do for you?* Journal of Medical

Genetics, 2011. **48**(9): p. 580.

13. *National Human Genome Research Institute. The Cost of Sequencing a Human Genome. https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/ Accessed: 2020- 03-20.*
14. Zare, F., et al., *An evaluation of copy number variation detection tools for cancer using whole exome sequencing data.* BMC Bioinformatics, 2017. **18**(1): p. 286.
15. Hong, C.S., et al., *Assessing the reproducibility of exome copy number variations predictions.* Genome Med, 2016. **8**(1): p. 82.
16. Yao, R., et al., *Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data.* Mol Cytogenet, 2017. **10**: p. 30.
17. Li, J., et al., *CONTRA: copy number analysis for targeted resequencing.* Bioinformatics, 2012. **28**(10): p. 1307-13.
18. Abyzov, A., et al., *CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.* Genome Res, 2011. **21**(6): p. 974-84.
19. Zhang, L., et al., *Comprehensively benchmarking applications for detecting copy number variation.* PLoS Comput Biol, 2019. **15**(5): p. e1007069.
20. Jiang, Y., et al., *CODEX: a normalization and copy number variation detection method for whole exome sequencing.* Nucleic Acids Res, 2015. **43**(6): p. e39.
21. Magi, A., et al., *EXCAVATOR: detecting copy number variants from whole-exome sequencing data.* Genome Biology, 2013. **14**(10).
22. Fromer, M. and S.M. Purcell, *Using XHMM software to detect copy number variation in whole-exome sequencing data.* Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.], 2014. **81**: p. 7.23.1-7.23.21.
23. Krumm, N., et al., *Copy number variation detection and genotyping from exome sequence data.* Genome Res, 2012. **22**(8): p. 1525-32.
24. Skidmore, Z.L., et al., *GenVisR: Genomic Visualizations in R.* Bioinformatics, 2016. **32**(19): p. 3012-4.
25. Collins, R.L., et al., *CNView: a visualization and annotation tool for copy*

number variation from whole-genome sequencing. bioRxiv, 2016.

26. Dharanipragada, P., S. Vogeti, and N. Parekh, *iCopyDAV: Integrated platform for copy number variations-Detection, annotation and visualization.* PLoS One, 2018. **13**(4): p. e0195334.
27. Zhang, Y., et al., *DeAnnCNV: a tool for online detection and annotation of copy number variations from whole-exome sequencing data.* Nucleic Acids Res, 2015. **43**(W1): p. W289-94.
28. Garvin, T., et al., *Interactive analysis and assessment of single-cell copy-number variations.* Nat Methods, 2015. **12**(11): p. 1058-60.
29. Leppa, V.M., et al., *Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families.* Am J Hum Genet, 2016. **99**(3): p. 540-554.
30. *Dougherty E. R, et al., Genomic signal processing and statistics.* EURASIP Book Series on Signal Processing and Communications, 2005.
31. Fickett, J.W. and C.S. Tung, *Assessment of protein coding measures.* Nucleic Acids Res, 1992. **20**(24): p. 6441-50.
32. Fickett, J.W., *The gene identification problem: an overview for developers.* Comput Chem, 1996. **20**(1): p. 103-18.
33. Vaidyanathan, P.P. and B.J. Yoon, *The role of signal-processing concepts in genomics and proteomics.* Journal of the Franklin Institute-Engineering and Applied Mathematics, 2004. **341**(1-2): p. 111-135.
34. Peter N. Robinson , R.M.P., Marten Jager, *Computational Exome and Genome Analysis*. Chapman & Hall/CRC Mathermatical and Computational Biology Series. 2018: CRC Press.
35. Saberkari, H., et al., *A Novel Fast Algorithm for Exon Prediction in Eukaryotic Genes Using Linear Predictive Coding Model and Goertzel Algorithm Based on the Z-Curve.* International Journal of Computer Applications, 2013. **67**(17).
36. Sanger, F., et al., *Nucleotide sequence of bacteriophage φX174 DNA.* Nature, 1977. **265**: p. 687.
37. McPherson, J.D., et al., *A physical map of the human genome.* Nature, 2001. **409**(6822): p. 934-41.

38. Sachidanandam, R., et al., *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.* Nature, 2001. **409**(6822): p. 928-33.

39. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

40. *National Human Genome Research Institute. The Cost of Sequencing a Human Genome. https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/ Accessed: 2018-06-11.*

41. Bartlett, A., et al., *Mapping genome-wide transcription-factor binding sites using DAP-seq.* Nature Protocols, 2017. **12**(8): p. 1659-1672.

42. Levy, S., et al., *The Diploid Genome Sequence of an Individual Human.* PLoS Biology, 2007. **5**(10): p. e254.

43. Wheeler, D.A., et al., *The complete genome of an individual by massively parallel DNA sequencing.* Nature, 2008. **452**: p. 872.

44. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping.* Nature reviews. Genetics, 2011. **12**(5): p. 363-376.

45. Pinto, D., et al., *Convergence of genes and cellular pathways dysregulated in autism spectrum disorders.* Am J Hum Genet, 2014. **94**(5): p. 677-94.

46. Pinto, D., et al., *Functional impact of global rare copy number variation in autism spectrum disorders.* Nature, 2010. **466**(7304): p. 368-72.

47. Malhotra, D. and J. Sebat, *CNVs: harbingers of a rare variant revolution in psychiatric genetics.* Cell, 2012. **148**(6): p. 1223-41.

48. Wellcome Trust Case Control, C., et al., *Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls.* Nature, 2010. **464**(7289): p. 713-20.

49. Cantsilieris, S. and S.J. White, *Correlating multiallelic copy number polymorphisms with disease susceptibility.* Hum Mutat, 2013. **34**(1): p. 1-13.

50. Jacquemont, S., et al., *Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus.* Nature, 2011. **478**(7367): p. 97-U111.

51. Lee, C. and S.W. Scherer, *The clinical context of copy number variation in the human genome.* Expert Rev Mol Med, 2010. **12**: p. e8.

52. Firth, H.V., et al., *DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources.* American Journal of Human Genetics, 2009. **84**(4): p. 524-533.

53. Riggs, E.R., et al., *Towards an evidence-based process for the clinical interpretation of copy number variation.* Clinical Genetics, 2012. **81**(5): p. 403-412.

54. de Vries, B.B.A., et al., *Diagnostic genome profiling in mental retardation.* American Journal of Human Genetics, 2005. **77**(4): p. 606-616.

55. Eichler, E.E., *Copy Number Variation and Human Disease.* Nature Education, 2008. **1**(3): p. 1.

56. Hollox, E.J., *Psoriasis is associated with increased beta-defensin genomic copy number.* Nat Genet, 2008. **40**: p. 23-5.

57. Fellermann, K., et al., *A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon.* Am J Hum Genet, 2006. **79**(3): p. 439-48.

58. Aitman, T.J., et al., *Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans.* Nature, 2006. **439**(7078): p. 851-5.

59. Sharp, A.J., et al., *Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome.* Nat Genet, 2006. **38**(9): p. 1038-42.

60. Walsh, T., et al., *Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia.* Science, 2008. **320**(5875): p. 539-43.

61. Sebat, J., et al., *Strong association of de novo copy number mutations with autism.* Science, 2007. **316**(5823): p. 445-9.

62. Pirooznia, M., et al., *Validation and assessment of variant calling pipelines for next-generation sequencing.* Human Genomics, 2014. **8**(1): p. 14.

63. Itsara, A., et al., *Population analysis of large copy number variants and hotspots of human genetic disease.* Am J Hum Genet, 2009. **84**(2): p. 148-61.

64. *Chromosomal duplication*. [cited 2020-09-04; Available from: https://ghr.nlm.nih.gov/art/large/duplication.jpeg

65. *Chromosomal deletion*. [cited 2020-09-04; Available from: https://ghr.nlm.nih.gov/art/large/chromosomaldeletion.jpeg.

66. Magi, A., et al., *Read count approach for DNA copy number variants detection.* Bioinformatics, 2012. **28**(4): p. 470-8.

67. Szatkiewicz, J.P., et al., *Improving detection of copy-number variation by simultaneous bias correction and read-depth segmentation.* Nucleic Acids Res, 2013. **41**(3): p. 1519-32.

68. Tattini, L., R. D'Aurizio, and A. Magi, *Detection of Genomic Structural Variants from Next-Generation Sequencing Data.* Frontiers in Bioengineering and Biotechnology, 2015. **3**(92).

69. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.

70. Cunningham, F., et al., *Ensembl 2015.* Nucleic Acids Research, 2015. **43**(D1): p. D662-D669.

71. MacDonald, J.R., et al., *The Database of Genomic Variants: a curated collection of structural variation in the human genome.* Nucleic Acids Research, 2014. **42**(D1): p. D986-D992.

72. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Research, 2014. **42**(D1): p. D980-D985.

73. *Diagram of interactions within the MVC pattern*. [cited 2020 Sep 8; Available from: https://en.wikipedia.org/wiki/Model–view–controller#/media/File:MVC-Process.svg.

74. *DAO Pattern Players*. [cited 2020 Sep 9; Available from: https://webdev.jhuep.com/~jcs/ejava-javaee/coursedocs/content/html/jpa-dao-pattern.html.

75. *Class diagram representing dao of student object*. [cited 2020 Sep 9; Available from: https://www.tutorialspoint.com/design_pattern/data_access_object_pattern.htm .

76. Zhao, M. and Z. Zhao, *CNVannotator: a comprehensive annotation server for*

copy number variation in the human genome. PLoS One, 2013. **8**(11): p. e80170.

77. Cleveland, W.S. and C. Loader. *Smoothing by Local Regression: Principles and Methods*. in *Statistical Theory and Computational Aspects of Smoothing*. 1996. Heidelberg: Physica-Verlag HD.
78. Yoon, S., et al., *Sensitive and accurate detection of copy number variants using read depth of coverage.* Genome Research, 2009. **19**(9): p. 1586-1592.
79. Kent, W.J., et al., *The Human Genome Browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.
80. Duan, J., et al., *CNV-TV: A robust method to discover copy number variation from short sequencing reads.* BMC Bioinformatics, 2013. **14**(1): p. 150.
81. Buels, R., et al., *JBrowse: a dynamic web platform for genome visualization and analysis.* Genome Biol, 2016. **17**: p. 66.
82. Amarasinghe, K.C., et al., *Inferring copy number and genotype in tumour exome data.* BMC Genomics, 2014. **15**: p. 732.
83. Klambauer, G., et al., *cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate.* Nucleic Acids Res, 2012. **40**(9): p. e69.
84. Sathirapongsasuti, J.F., et al., *Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV.* Bioinformatics, 2011. **27**(19): p. 2648-2654.
85. Reble, E., et al., *VarScan2 analysis of de novo variants in monozygotic twins discordant for schizophrenia.* Psychiatr Genet, 2017. **27**(2): p. 62-70.
86. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
87. *breast cancer related genes - GeneCards Search Results [Internet]. [cited 2020 March 13]. Available from: https://www.genecards.org/Search/Keyword?queryString=breast%20cancer.*

ประวัติผู้เขียน

ชื่อ-สกุล	เสาวภาค จันทร์วิกูล
วัน เดือน ปี เกิด	23 มีนาคม 2533
สถานที่เกิด	กรุงเทพ
วุฒิการศึกษา	วิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง