เครื่องมือทางชีวสารสนเทศเพื่อตรวจหาการแปรผันเชิงโครงสร้างทางพันธุกรรม

นายศักยภาพ ผิวเหลือง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมซอฟต์แวร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

A bioinformatics tool for structural variant detection

Mr. Sakkayaphab Piwluang

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	เครื่องมือทางชีวสารสนเทศเพื่อตรวจหาการแปรผันเชิงโครงสร้างทางพันธุกรรม
โดย	นายศักยภาพ ผิวเหลือง
สาขาวิชา	วิศวกรรมซอฟต์แวร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร.ดวงดาว วิชาดากุล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศ. ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รศ. ดร.เกริก ภิรมย์โสภา)

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

..... กรรมการ
(อาจารย์ ดร.เอกพล ช่วงสุวนิช)

..... กรรมการภายนอกมหาวิทยาลัย
(อาจารย์ ดร.ศศิธร โชติวุฒิมนตรี)

ศักยภาพ ผิวเหลือง : เครื่องมือทางชีวสารสนเทศเพื่อตรวจหาการแปรผันเชิงโครงสร้างทางพันธุกรรม. (A bioinformatics tool for structural variant detection) อ.ที่ปรึกษาหลัก : อ. ดร.ดวงดาว วิชาดากุล

การแปรผันเชิงโครงสร้างทางพันธุกรรมคือการเปลี่ยนแปลงลำดับเบสของจีโนมที่ครอบคลุมบริเวณกว้าง การแปรผันเชิงโครงสร้างเหล่านี้มีโอกาสที่จะเกี่ยวข้องกับการเกิดโรค ดังนั้นการตรวจหาการแปรผันเชิงโครงสร้างจึงเป็นปัจจัยหนึ่งที่สำคัญในการหาสาเหตุของโรค อย่างไรก็ตามเครื่องมือสำหรับตรวจหาการแปรผันเชิงโครงสร้างที่มีอยู่มีประสิทธิภาพแตกต่างกันไปในการตรวจหาการแปรผันเชิงโครงสร้างแต่ละประเภท รวมทั้งไม่สามารถตรวจหาการแปรผันได้ครอบคลุมในตำแหน่งส่วนใหญ่ที่ได้ยืนยันจากการทดลองในห้องปฏิบัติการ วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการทางคอมพิวเตอร์เพื่อตรวจหาการแปรผันเชิงโครงสร้างที่เน้นการเพิ่มประสิทธิภาพความครอบคลุม โดยพยายามรักษาความแม่นยำของการแปรผันเชิงโครงสร้างที่ตรวจพบผ่านการวิเคราะห์คู่รีด การแตกรีด และการนับรีด เพื่อรวบรวมหลักฐานที่แสดงความเป็นไปได้ในการเกิดการแปรผันเชิงโครงสร้างแต่ละประเภทในแต่ละบริเวณของจีโนม และนำเสนอวิธีการคัดกรองเบรกเอ็นด์ที่แสดงถึงตำแหน่งเริ่มต้นและตำแหน่งสิ้นสุดของการเกิดการแปรผันประเภทต่างๆ วิธีการกรองจะทำการแบ่งตัวอย่างเป็นบล็อกๆ ข้อมูลของบล็อกประกอบไปด้วย จำนวนรีดที่แมพได้ในบล็อกนั้นและจำนวนการแปรผันเชิงโครงสร้างของแต่ละประเภทภายในบล็อก การคัดกรองเบรกเอ็นด์จะอาศัยข้อมูลทั้งจากบล็อกที่เบรกเอ็นด์อยู่และบล็อกที่อยู่ติดกัน ผลลัพธ์จากการเปรียบเทียบประสิทธิภาพของวิธีการที่นำเสนอกับ SvABA DELLY GROM LUMPY และ Wham พบว่าวิธีการที่นำเสนอได้ผลลัพธ์ดีกว่าเครื่องมืออื่นๆ ในส่วนของความแม่นยำในการตรวจหาลำดับเบสที่เกิดความซ้ำเป็นชุดติดๆกันและลำดับเบสที่เกิดการกลับด้าน และความครบถ้วนในส่วนของการตรวจหาลำดับเบสที่ถูกเพิ่มเข้ามาสำหรับชุดข้อมูลจริง NA12878 และ HG00514 ที่ใช้ในการทดสอบ

สาขาวิชา วิศวกรรมซอฟต์แวร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2562 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

5970324021 : MAJOR SOFTWARE ENGINEERING
KEYWORD: structural variation, genomics

Sakkayaphab Piwluang : A bioinformatics tool for structural variant detection. Advisor: Duangdao Wichadakul, Ph.D.

Genomic structural variations (SVs) represent large genomic alterations and have been reported to be associated with diseases. The detection of structural variations is an important approach for investigating the cause of diseases. While several tools for detecting structural variations are available, they achieved varied performance for each type of the variation. Moreover, many experimentally verified variations were still uncaught by these tools. This thesis proposes a computational method for SV detection aiming to increase the coverage while maintaining the precision. The method incorporated the read-pair, split-read, and read count analyses to compile the evidence for each SV type. To filter the potential breakends, the starting and ending positions of a SV, the genome was divided into blocks containing the read coverage and the number of detected SVs of each type. Our method then considers the data of a block suggested with a breakend position and the data of its adjacent blocks for breakend filtering. Based on two real datasets NA12878 and HG00514, our method outperformed SvABA, DELLY, GROM, LUMPY, and Wham in term of precision for detecting tandem duplication and inversion and got the highest recall for detecting insertion while maintaining the comparable precision.

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

กิตติกรรมประกาศ

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

ขอขอบพระคุณกรรมการสอบวิทยานิพนธ์ รศ. ดร. เกริก ภิรมย์โสภา อ.ดร. เอกพล ช่วงสุวนิช และ อ.ดร. ศศิธร โชติวุฒิมนตรี ที่กรุณาสละเวลาให้คำแนะนำ ตรวจสอบและแก้ไขวิทยานิพนธ์ ซึ่งเป็นประโยชน์อย่างยิ่งต่อการพัฒนางานวิจัยนี้

ขอขอบพระคุณ คุณเฉลิมพล ศรีจอมทอง ที่สนับสนุนแนวทางการวิเคราะห์ข้อมูลและความรู้ทางด้านจีโนมิกส์และพันธุศาสตร์

ขอบคุณทุนสนับสนุนจาก Chulalongkorn Academic Advancement into Its 2nd Century (CUAASC), ศูนย์ชีววิทยาเชิงระบบ คณะแพทย์ศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย (Chulalongkorn University Systems Biology Center)

สุดท้ายนี้ ข้าพเจ้าหวังเป็นอย่างยิ่งว่า เนื้อหาในวิทยานิพนธ์ฉบับนี้จะเป็นประโยชน์แก่ผู้อื่นบ้าง ไม่มากก็น้อย

ศักยภาพ ผิวเหลือง

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1. ที่มาและความสำคัญของปัญหา

ในปัจจุบันมีงานวิจัยมากมายที่พบว่าโรคหลากหลายชนิด มีความสัมพันธ์กับความผิดปกติของยีนและโครงสร้างทางพันธุกรรม [1] ไม่ว่าจะเป็นการเกิดโรคหัวใจและหลอดเลือด [2-4] ตับอ่อนอักเสบเฉียบพลัน [5] โรคลมบ้าหมู [6] โรคมะเร็ง [7-9] โรคออทิสซึม [10, 11] โรคจิตเภท [12, 13] โรคพาร์กินสัน [14] และ โรคอัลไซเมอร์ [15] เป็นต้น เพื่อที่จะเข้าใจได้มากขึ้นถึงสาเหตุของการเกิดและแนวทางการรักษาที่เป็นไปได้จึงต้องทำการศึกษาการแปรผันเชิงโครงสร้างทางพันธุกรรมให้ละเอียดมากขึ้น

ในปัจจุบันมีเทคโนโลยีอย่าง Next-generation sequencing (NGS) ที่ทำให้การหาลำดับเบสของสิ่งมีชีวิตรวดเร็วขึ้น และมีต้นทุนค่าใช้จ่ายที่ถูกลงอย่างมาก [16] ด้วยวิธีการตัดสารพันธุกรรม (DNA) เป็นชิ้นสั้นๆ โดยมีความยาวเบสที่ 25 – 1000 เบส หลายร้อยล้านชิ้น แล้วจึงนำมาประกอบเป็นลำดับเบสร่วมกัน และด้วยลำดับเบสที่มีขนาดสั้นนั้นย่อมส่งผลต่อความผิดพลาดได้ง่าย จึงต้องมีวิธีวิเคราะห์ข้อมูลที่ดีเพื่อให้เกิดความแม่นยำ ด้วยเหตุนี้ทำให้การค้นหาการแปรผันเชิงโครงสร้างทางพันธุกรรมให้ครอบคลุม และถูกต้องจึงเป็นไปได้ยาก รวมทั้งต้องการทรัพยากรการคำนวณจำนวนมากในการประมวลผลข้อมูล

ปัจจุบันมีงานวิจัยเรื่องเครื่องมือที่ใช้สำหรับค้นหาการแปรผันเชิงโครงสร้างในแต่ละโครโมโซมอยู่จำนวนหนึ่ง ซึ่งจะพบว่าแต่ละเครื่องมือมีทั้งข้อดี และข้อเสียแตกต่างกันไป เช่น ความไว (sensitivity) ที่เครื่องมือจะตรวจพบ ความแม่นยำถูกต้อง ความเร็วที่ใช้ในการประมวลผล ร่วมถึงความครอบคลุมประเภทของการแปรผัน ที่สามารถตรวจพบ เป็นต้น

ด้วยเหตุนี้ การปรับปรุงเครื่องมือที่ใช้สำหรับค้นหาการแปรผันทางพันธุกรรมยังมีความน่าสนใจอยู่มาก ไม่ว่าจะเป็นในด้านของการปรับปรุงประสิทธิภาพการค้นหา เพื่อที่จะต่อยอดช่วยให้หาความสัมพันธ์ของโรคต่างๆ ได้มากขึ้น หรือปรับปรุงด้านของความเร็วและการใช้ทรัพยากรในการประมวลผลเพื่อให้สามารถนำไปประยุกต์ใช้ในเชิงคลีนิกได้อย่างมีประสิทธิภาพ รวมทั้งเป็นการประหยัดค่าใช้จ่ายได้อีกด้วย

1.2. วัตถุประสงค์

เพื่อสร้างเครื่องมือที่ตรวจหาการแปรผันเชิงโครงสร้างทางพันธุกรรม โดยครอบคลุมประเภทของการแปรผันเชิงโครงสร้างต่อไปนี้ deletion, insertion, tandem duplication, inversion และ chromosomal translocation

1.3. ขอบเขตงานวิจัย

1) กำหนดข้อมูลเข้าเป็นไฟล์แบม และไฟล์ฟาสต้า (จีโนมอ้างอิง)

2) การตรวจหาการแปรผันเชิงโครงสร้างในแต่ละโครโมโซมใช้วิธีการเชิงอัลกอริทึม

3) การทดสอบเครื่องมือจะทดสอบกับชุดข้อมูลต่อไปนี้

3.1) ข้อมูลจำลอง โดยทำการสร้างตัวอย่างจำเพาะด้วยโปรแกรม SURVIVOR [17] และ WGSim [18]

3.2) ข้อมูลจริงชุดที่ 1 NA12878 โดยมีจำนวนข้อมูล 10 ตัวอย่างจำเพาะคือ ตัวอย่าง SRR1910366 นำมาจากฐานข้อมูล DDBJ (DNA Data Bank of Japan) [19] และตัวอย่าง ERR174336, ERR174337, ERR174338, ERR174339, ERR174340, ERR091571, ERR091572, ERR091573 และ ERR091574 นำมาจากฐานข้อมูล SRA (Sequence Read Archive) ของ NCBI [20] ซึ่งทั้ง 10 ตัวอย่างจำเพาะเป็นตัวแทนของจีโนมมนุษย์ NA12878 ที่มีการถอดรหัสโดยใช้ความลึกของจำนวนรีดที่แตกต่างกัน โดยตรวจสอบผลการตรวจหาการแปรผันเชิงโครงสร้าง โดยใช้ข้อมูลการแปรผันเชิงโครงสร้างของจีโนม NA12878 จาก DGV [21]

3.3) ข้อมูลจริงชุดที่ 2 HG00514 โดยมีจำนวนข้อมูล 2 ตัวอย่างจำเพาะคือ ERR894729 และ ERR903030 นำมาจากฐานข้อมูล SRA [20] โดยตรวจสอบผลตรวจหาการแปรผันเชิงโครงสร้าง โดยใช้ข้อมูลการแปรผันเชิงโครงสร้างของจีโนม HG00514 จาก [22]

4) ผลลัพธ์ที่ได้จะอยู่ในรูปแบบของวีซีเอฟเวอร์ชัน 4.2

5) เครื่องมือทำงานได้หลายแพลตฟอร์ม

1.4. ขั้นตอนและวิธีการดำเนินการวิจัย

1) ศึกษาองค์ความรู้ และทฤษฎีที่เกี่ยวข้องกับงานวิจัย

2) วิเคราะห์ข้อดีข้อเสียของแต่ละเครื่องมือที่เคยทำมาก่อน

3) ทำการออกแบบ และเครื่องมือพัฒนาอัลกอริทึมและเครื่องมือ

4) ทดสอบเปรียบเทียบประสิทธิภาพกับเครื่องมือที่มีมาก่อน

5) สรุปผลการวิจัย

6) เรียบเรียง และจัดทำบทความวิชาการ

7) เรียบเรียง และจัดทำวิทยานิพนธ์

1.5. ประโยชน์ที่คาดว่าจะได้รับ

ได้เครื่องมือใหม่ที่ช่วยในการวิเคราะห์หาการแปรผันเชิงโครงสร้างทางพันธุกรรมที่มีประสิทธิภาพมากขึ้นในเชิงของความครอบคลุมโดยยังรักษาประสิทธิภาพของความแม่นยำ

1.6. ผลงานตีพิมพ์จากงานวิจัย

ส่วนหนึ่งของวิทยานิพนธ์ได้รับการตีพิมพ์และนำเสนอในงานประชุมวิชาการระดับนานาชาติ Sakkayaphab and Duangdao Wichadakul, "iPRIns:A Tool with the Improved Precision and Recall for Insertion Detection in the Human Genome" ในรายงานการประชุมวิชาการนานาชาติ 2020 8th International Conference on Bioinformatics and Computational Biology (ICBCB 2020), Taiyuan, China May 16-18, 2020

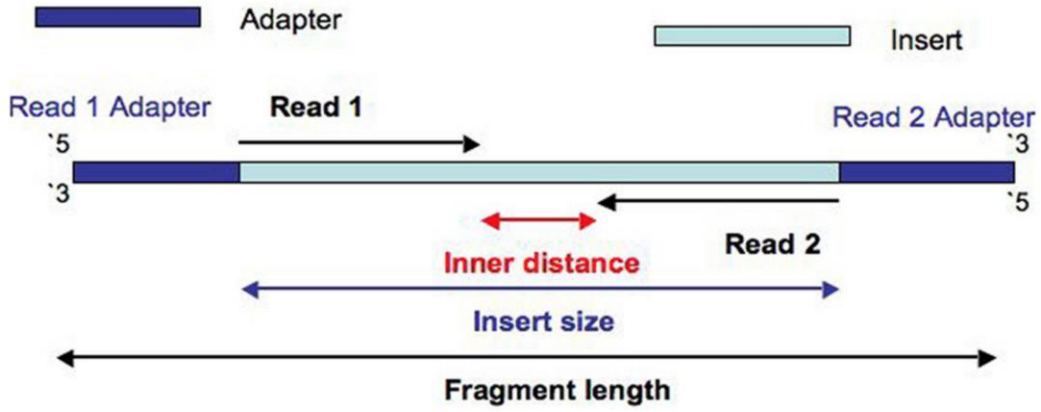
S

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การถอดรหัสพันธุกรรม (DNA Sequencing)

การถอดรหัสพันธุกรรมในปัจจุบันมีชื่อเรียกว่า เทคโนโลยีเอ็นจีเอส (Next generation sequencing) ซึ่งเป็นกระบวนการสร้างลำดับเบสที่รวดเร็ว และค่าใช้จ่ายที่ถูกลงกว่าการหาลำดับเบสแบบแซงเกอร์ (Sanger sequencing) [23] เทคโนโลยีเอ็นจีเอสมีหลายแพลตฟอร์มโดยแตกต่างกันในเรื่องของ ค่าใช้จ่าย ขั้นตอนการหาลำดับเบส ระยะเวลาในการทำงาน ความแม่นยำ ขนาดเส้นลำดับเบส และรูปแบบของสายลำดับเบส เป็นต้น โดยที่รูปแบบของสายลำดับเบสจะมีทั้งที่เป็นแบบเทคโนโลยีถอดรหัสพันธุกรรมเป็นสายเดี่ยว (single-end sequencing) ที่เป็นการอ่านแบบทิศทางเดียว คือจาก 5’ ไป 3’ และเทคโนโลยีถอดรหัสพันธุกรรมเป็นสายคู่ (paired-end sequencing) ที่อ่านทั้ง 2 ทิศทาง อ่านไปข้างหน้า (forward strand) และอ่านย้อนกลับ (reverse strand) (รูปที่ 1) สำหรับ paired-end sequencing บนแพลตฟอร์ม Illumina ผลลัพธ์ที่ได้จะเป็นสายรหัสพันธุกรรมดีเอ็นเอหรือรีดจำนวนมาก ถูกเก็บในไฟล์รูปแบบ FASTQ โดยที่แต่ละรีดจะมีอยู่ 4 บรรทัด คือ ป้ายชื่อ ลำดับเบส เครื่องหมาย + และคุณภาพเบส (รูปที่ 2) เมื่อเป็น paired-end sequencing จะมี 2 ไฟล์ คือไฟล์ที่อ่านจากสายดีเอ็นเอที่อ่านไปข้างหน้า และสายดีเอ็นเอที่อ่านย้อนกลับโดยรีดที่เป็นคู่กันจะแยกกันอยู่คนละไฟล์ โดยมีป้ายชื่อ (label) ที่เหมือนกัน



รูปที่ 1 องค์ประกอบของ Fragment

(ที่มา : รูปที่ 1 ของ [24])

```
1 @SRR062635.1 HWI-EAS110_103327062:4:1:1071:15970/1
2 CAGGAAAGACAATTCCAAAATCAGTTAGAGTCCTGTTGGCGCGTGTAATACATCTCCACTTTGAAAATGAAGACAGGGGGTTACGAGTGTTATTAATGAG
3 +
4 EAE:A?EE6EE:E:BED5E?BC---?CAA5ADBBBA,<--+A31<=CC:5@##################################################
  @SRR062635.2 HWI-EAS110_103327062:4:1:1072:21126/1
  TGGGAATGTAAATTAGTCCAGCCACTCTGGAGAACCGTATGGAGGTTCCTCCAAAAATTACAAATAGAACTACCATATGATCCAGCAATCCCATGCTATG
  +
  DD=?BBDD:DD=AA?>=-?=AAB5?=C5C::C==A4<<<;5:::=ACAAA:4:??3,&91693376;)7;-<A?<)8;2+8;;+:A:-5:-?:<:??=)?
  @SRR062635.3 HWI-EAS110_103327062:4:1:1075:18579/1
  AGATTTCCCTGAGAAAGTCATATTTAAGCTGCCATTTGAAGACCAAGGAATCATGACTAGAGACAAGAAGAGAGAACATAGAGTGATTATGGAGAATCTT
  +
  <5;3;;1@=@BD=D5D;DDDD::DADB:=DAD5::9==4..4=>?DA=-->.@=?A-CC?-B>AA);/65*-13>CCC5C####################
  @SRR062635.964 HWI-EAS110_103327062:4:1:1543:1046/2
  AGTATCAGTCCAGTCCTCAGTGACGGGACCCTAACTGACCTGCCCTTCTTTGGCTTAGATTGCTTAAATGGTTCTGGATGTGATGATGGTGCACCTTGCC
  +
  CC:CC@.>@.C..==.6;;6C-5AA.6,>->5AA?@:55-A555A=--CC?AA*?<:55>?C#####################################
  @SRR062635.1459 HWI-EAS110_103327062:4:1:1622:910/2
  TATATTAGAGTAGAGTCTAAAGATTAGAATGATCCACAGGTTAATATGGGCCATTATAAAGAGATTAGTGATATTAACAATNTAGTATCAACATGGAGAT
  +
  ?C-A>6@@:@CC:@CDADD:D:DD=D?D55D?DD:ADADDC5-@?:@?CC5@?B?-B:?6>2A>>?5??@###########!##################
```

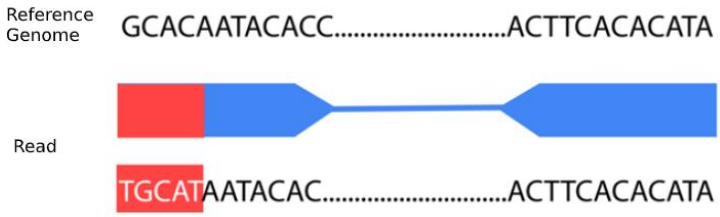
รูปที่ 2 ตัวอย่างไฟล์ FASTQ

2.1.2 การเทียบรีดสายสั้นกับจีโนมอ้างอิง

การถอดรหัสพันธุกรรมจะได้มาเพียงแค่รีดสั้นๆ จำนวนมาก ที่จะบอกถึงลำดับเบส และคุณภาพของเบส แต่ไม่ทราบถึงตำแหน่งที่มา และโครโมโซมของรีด ดังนั้นการที่จะทราบได้ คือการใช้จีโนมอ้างอิง (รูปที่ 3) โดยนำรีดไปเทียบหาตำแหน่งที่ใกล้เคียงมากที่สุด ซึ่งบางกรณีอาจจะมีบางเบสที่ไม่ตรงกัน จนไปถึงขั้นที่มีเบสจำนวนหนึ่งของรีดที่ไม่ตรงกับลำดับเบสในจีโนมอ้างอิงเลย แต่ลำดับเบสที่เหลือทั้งหมดของรีดตรงกับบริเวณนั้นของจีโนมอ้างอิงมากที่สุด ลำดับเบสของรีดในส่วนที่ไม่ตรงนี้จะถูกเรียกว่าลำดับเบสซอฟต์คลิป (soft-clipped) (รูปที่ 4) แต่ถ้ารีดนั้นไม่ตรงเลยรีดนั้นก็จะกลายเป็นรีดที่ไม่ถูกแมพ (unmapped read) ซึ่งปัจจุบันมีเครื่องมือมากมายในการทำหน้าที่ในการเทียบรีดสายสั้นกับจีโนมอ้างอิง [25] เช่น BWA [26, 27] Bowtie [28, 29] SOAP [30, 31] และ Novoalign เป็นต้น ซึ่ง 3 เครื่องมือแรกใช้ FM-index ที่เป็นการนำเอาข้อความจาก Burrow wheeler transform (BWT) มาค้นหาลำดับเบส ยกเว้น Novoalign ใช้ตารางแฮช (Hash table) ซึ่งแต่ละเครื่องมือก็มีคุณสมบัติที่แตกต่างกันออกไป เช่น การรองรับจำนวนลำดับเบสที่ไม่ตรงกัน การรองรับช่องว่าง แนวคิดในการเลือกจุดที่เหมาะสมที่สุด และเวลาที่ใช้ในการคำนวณ เป็นต้น ซึ่งแต่ละเครื่องมือเมื่อทำการเทียบรีดสายสั้นกับจีโนมอ้างอิงเสร็จผลลัพธ์ที่ได้ออกมาจะอยู่ในรูปแบบของ SAM (Sequence Alignment Map) [32]

```
>chr4_ctg9_hap1
GAATTCTTCACATTTCCTGGCTTTTAAAAGTTCTCCTTCCACAAATCTTC
TATTACTATATATCCGTGTAAACCCCAGTCATATAATCTCTCCCCACTGA
CCTTGGCAAACACGATGGCTTGTTGTGGATAACAGGTGGAGGCAGCTAAT
CCCATGAAACCAGGTGGACACACTAGCTTCTTTATTTTTGAACCTCTAGC
CAAAAAGAGTCCATTAAGGCCAGCAAAACCAATAACACCAAGTCTTGGAA
AAAATCCAGGAGGTGCATTTTAGAGATACTCATAGCTGTCTAATCCCCAC
TGAACCAAACTTTGCATCTTAGGCTTAGTTTGGGAGTACGTTTCCTGACA
CAAATTTCTATGTGGCTCACAATAGCGTCGGAACTGTGAGATGCCTACTT
CAAGCTGGCTCCTTGACTCTTCCACACACTTCGATTGACCCTCGGGAACT
GAGTACAGTGAAAGCTCATCAACCTTCATGGGATTTTGAGGAGGTGAGTC
CTTTTTTGGTGCTGCATAGACTTTGAAGGTGAACAAGCGTAGGCTGGTTG
GCCTCACGGACCTCCGAATTACCTTGAACATGTCGCTGTCCGTGGTGGCT
CTGGCAGGATTTACTTTCTTTTATTCCAGTTTTATTCCAGTATTTCTTTT
TTTAAAATTTCTTTTATTCCAGTATTTCTTTTTTAAACCTATTCTTTCTG
AAGATCCCAAACTTTTACATATTAGAGAATACAATGGTGGTTTACCTTAT
GAGAGGAAGTGTTTTCTGCCTTATGCCCAGACTAATAGTACACCTCAGTA
AAGCTCATTTTGGGAACATGTGTCTTGCTATATCACAGCTGCTTGATAGA
GAAGCATAGAAAAATGTACAGTGCATGTATAATAAAGACCAAATAATTTC
```

รูปที่ 3 ตัวอย่างจีโนมอ้างอิง



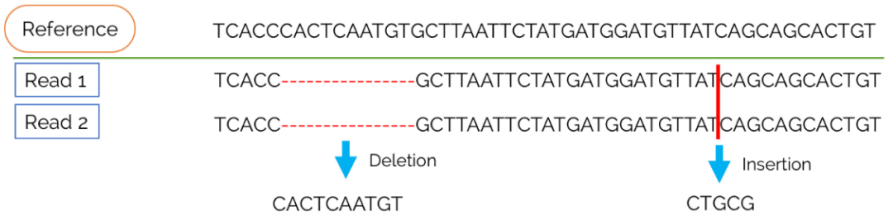
รูปที่ 4 ตัวอย่างลำดับเบสซอฟต์คลิป (บริเวณสีแดงในรูป)

2.1.3 การแปรผันทางพันธุกรรม (Genetic variation)

การแปรผันทางพันธุกรรม [33, 34] มีรูปแบบที่หลากหลาย ได้แก่ single-nucleotide variant (SNV) ที่เกิดการแทนที่ตำแหน่งของเบส 1 เบส อินเดลที่เกิดการแปรผันในแบบ deletion และ insertion ตั้งแต่ระดับ 1-49 เบส [35] และสุดท้ายการแปรผันเชิงโครงสร้างในแต่ละโครโมโซม (Structural variation: SV) ที่เกิดการแปรผันตั้งแต่ระดับ 50 เบสเป็นต้นไป [36]

2.1.3.1 อินเดล (Indel)

อินเดล คือส่วนที่เกิด insertion และ deletion ที่มีขนาดเล็ก (รูปที่ 5) โดยประกอบด้วยจำนวนเบสน้อยกว่า 50 เบส ซึ่งเป็นการแปรผันทางพันธุกรรมประเภทหนึ่ง

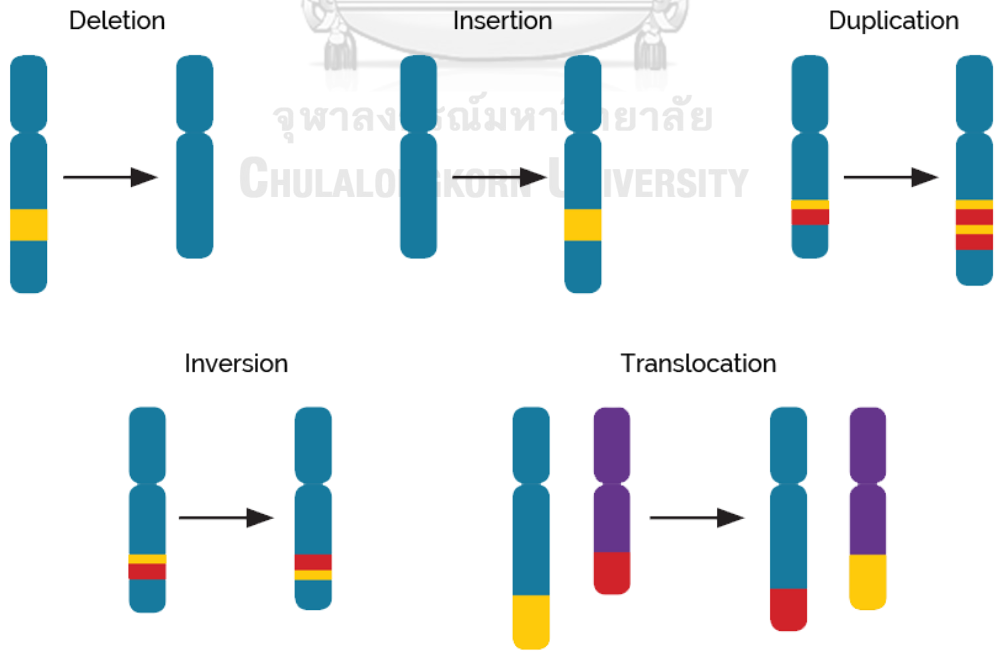


รูปที่ 5 รูปแบบของอินเดล

2.1.3.2 การแปรผันเชิงโครงสร้าง (Structural variation)

การค้นหาการแปรผันเชิงโครงสร้างสามารถเกิดขึ้นในแต่ละโครโมโซมหรือระหว่างโครโมโซม โดยที่การแปรผันเชิงโครงสร้างในแต่ละโครโมโซมจะมีขนาดตั้งแต่ระดับ 50 เบสขึ้นไป และมี 5 รูปแบบหลักๆ (รูปที่ 6) ได้แก่

1) Deletion คือ ส่วนของดีเอ็นเอหายไปจากโครโมโซมเมื่อเทียบกับจีโนมอ้างอิง
2) Insertion คือ ส่วนของดีเอ็นเอมีลำดับเบสแทรกเพิ่มเติมเมื่อเทียบกับจีโนมอ้างอิง
3) Tandem duplication คือ ส่วนของดีเอ็นเอเกิดการซ้ำเมื่อเทียบกับจีโนมอ้างอิง โดยส่วนที่ซ้ำติดกัน
4) Inversion คือ ส่วนของดีเอ็นเอเกิดการกลับด้านของลำดับเบสเมื่อเทียบกับจีโนมอ้างอิง โดยไม่มีการเพิ่มหรือลดจำนวนเบส
5) Chromosomal translocation คือ ส่วนของดีเอ็นเอมีการย้ายจากโครโมโซมหนึ่งไปอีกโครโมโซมหนึ่งเมื่อเทียบกับจีโนมอ้างอิง



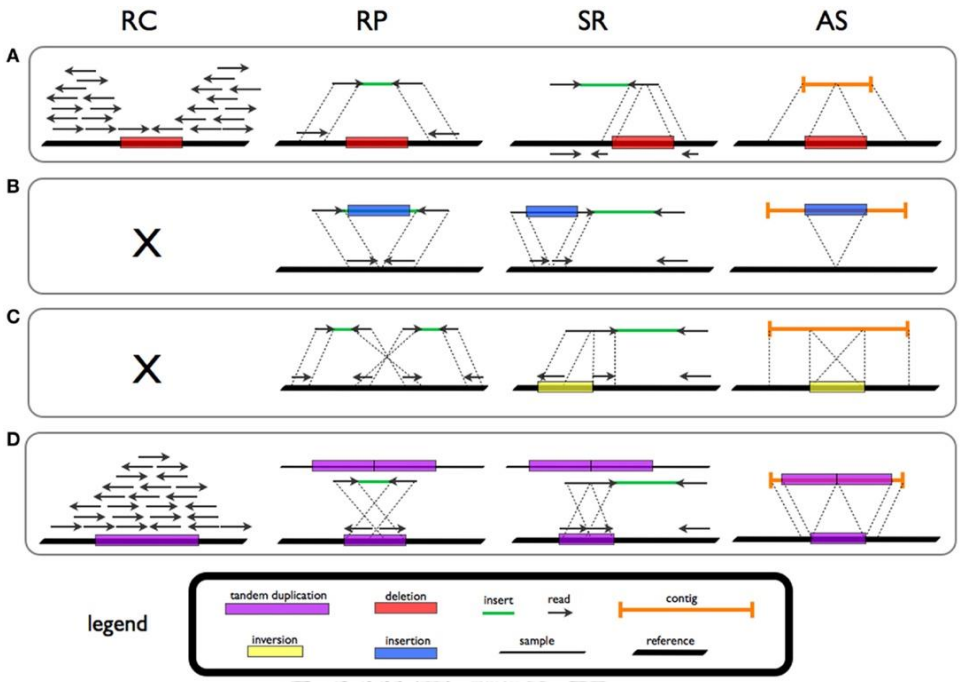
รูปที่ 6 รูปแบบการแปรผันเชิงโครงสร้างในแต่ละโครโมโซม โดยโครโมโซมด้านซ้ายคือจีโนมอ้างอิง

(ที่มา : รูปที่ 2 ของ [37])

2.1.4 การตรวจหาการแปรผันเชิงโครงสร้างในแต่ละโครโมโซม

การตรวจหาการแปรผันเชิงโครงสร้างในแต่ละโครโมโซม จะแบ่งออกได้ 2 แบบ [38] คือ แบบทดลอง (experimental) กับ แบบเชิงคำนวณ (computational) โดยที่แบบทดลอง จะใช้วิธีการทางไฮบริไดเซชัน (hybridization-based) วิธีการทางพีซีอาร์ (PCR-based) และวิธีการวิเคราะห์โมเลกุลเดี่ยว (single-molecule analysis) ซึ่งแต่ละวิธีมีข้อจำกัด และค่าใช้จ่ายที่แตกต่างกันไป สำหรับการตรวจหาการแปรผันเชิงโครงสร้างเชิงคำนวณจะใช้การวิเคราะห์ลำดับเบสดีเอ็นเอที่เกิดจากการถอดรหัสพันธุกรรม (sequencing-based) มี 4 วิธีหลักๆ [36] (รูปที่ 7), คือ read count (RC), read-pair (RP), split-read (SR), และ *de novo* assembly (AS) ซึ่งทั้ง 4 วิธีมีลักษณะการใช้งาน ที่แตกต่างกันดังนี้

1) Read count (RC) หรือนับรีด เป็นวิธีการหนึ่งในการหาแปรผันเชิงโครงสร้างในแต่ละโครโมโซม โดยอาศัยการอ่านจากความลึก (depth) ซึ่งหมายถึงนับจำนวนรีดที่มีลำดับเบสตรงกับจีโนมในบริเวณเดียวกัน โดยที่การนับรีดนี้ นอกจากจะใช้ในการคัดกรองรีดในบางบริเวณที่มีจำนวนรีดน้อยเกินไปแล้ว ยังมักถูกใช้กับการหาการแปรผันของจำนวนชุดดีเอ็นเอ (copy number variation)
2) Read-pair (RP) หรือคู่รีด เป็นวิธีการหนึ่งในการหาการแปรผันเชิงโครงสร้างของโครโมโซม โดยพิจารณาจากความยาวคู่รีด และทิศทางระหว่างรีดกับรีดที่เข้าคู่ (mate) โดยที่จะมองหาคู่รีดที่มีความผิดปกติของความยาวคู่รีด ยกตัวอย่างเช่น รีดที่มีรีดที่เข้าคู่ห่างกันหรือใกล้กันกว่าที่ควรจะเป็น โดยสามารถตัดสินว่ามีความผิดปกติของระยะห่างความยาวคู่รีด โดยเทียบกับค่าเฉลี่ย ของความยาวคู่รีด (ตั้งแต่จุดเริ่มต้นรีด 1 จนจุดสิ้นสุดรีด 2) หรือ insert size จากข้อมูลการออกแบบการทดลองก่อนนำดีเอ็นเอเข้าเครื่องถอดรหัสพันธุกรรม ดังนั้นการพิจารณาคู่รีดจึงเหมาะกับการใช้ในการประเมินว่า บริเวณดังกล่าวอาจจะเกิดการแปรผันได้ แต่ไม่สามารถรู้จุดเริ่มต้น และสิ้นสุดของการเกิดอย่างแน่ชัดได้
3) Split-read (SR) หรือแตกรีด เป็นวิธีการค้นหาการการแปรผันของโครโมโซม โดยการแยกลำดับเบสของรีดเป็นส่วนย่อยๆ แล้วเทียบตำแหน่งที่เข้ากันได้มากที่สุดของแต่ละส่วนย่อยๆ นั้น ดังนั้นจึงเหมาะกับการใช้ในการระบุจุดเริ่มต้น และสิ้นสุดของการแปรผันทางพันธุกรรม หรือเรียกว่าเบรกเอ็นด์ แต่แลกกับการประมวลผลที่เพิ่มขึ้น
4) *de novo* assembly หรือต่อรีด เป็นวิธีการที่จะพยายามสร้างเส้นจีโนมขึ้นมาใหม่โดยอาศัยการนำรีดมาต่อๆ กันให้ยาว ด้วยลักษณะนี้ทำให้ต้องการขนาดของรีดที่ยาวประมาณ 75 – 1000 เบส ซึ่งถ้ามีรีดที่ยาวจะก็ช่วยให้เชื่อมต่อกันระหว่างรีดได้ง่ายขึ้น



รูปที่ 7 การตรวจหาการแปรผันเชิงโครงสร้างในแต่ละโครโมโซม

(ที่มา : รูปที่ 1 ของ [36])

2.1.5 Binary Sequence Alignment/Mapping (BAM file format)

ไฟล์แบม (BAM) [32] เป็นไฟล์ข้อมูลเข้าสำหรับงานวิจัยนี้ โดยไฟล์แบมที่มีลักษณะคล้ายกับไฟล์แซม (SAM) (Sequence Alignment Map) [39] แต่ถูกทำการบีบอัดด้วย BGZF (Blocked GNU Zip Format) ทำให้ไฟล์มีขนาดเล็กลง และมีมาตรฐานรองรับการทำดัชนี โดยที่ไฟล์แบมประกอบด้วย 2 ส่วนหลักคือส่วนของเฮดเดอร์ และส่วนของ alignment (รูปที่ 8) ซึ่งลักษณะของเฮดเดอร์แต่ละบรรทัดจะขึ้นต้นด้วยตัวอักษร “@” ในขณะที่ในส่วนของ alignment จะไม่มีตัวอักษร “@”นำหน้า ส่วนของ alignment ประกอบไปด้วย 12 ฟิลด์ ดังตารางที่ 1

ตารางที่ 1 ฟิลด์หลักในไฟล์แบม

คอลัมน์	ฟิลด์	แบบชนิด	คำอธิบาย
1	QNAME	String	ชื่อของลำดับเบส
2	FLAG	Int	ระบุลักษณะของลำดับเบสรวมถึงความสัมพันธ์กับรีดที่เข้าคู่
3	RNAME	String	ชื่อของโครโมโซม เช่น chr1, chr2
4	POS	Int	ระบุตำแหน่งเริ่มต้นที่รีดแมพกับจีโนมอ้างอิง
5	MAPQ	Int	ระบุคุณภาพของแมพปิง (mapping)
6	CIGAR	String	ระบุลักษณะของลำดับเบสเมื่อแมพปิงกับตำแหน่งนี้
7	RNEXT	String	ชื่อของโครโมโซมของรีดที่เข้าคู่ เช่น chr1, chr2

8	PNEXT	Int	ระบุตำแหน่งของรีดที่เข้าคู่
9	TLEN	Int	ระบุขนาดเทมเพลต
10	SEQ	String	ระบุลำดับเบส เช่น CACTGT
11	QUAL	String	ระบุคุณภาพของเบส
12	OPT	String	เป็นฟิลด์เพิ่มเติมที่อยู่ในรูปแบบของ TAG:TYPE:VALUE

คอลัมน์ OPT จะเป็นแบบทางเลือกที่ alignment ให้มาโดยจะอยู่ในรูปแบบของ TAG:TYPE:VALUE ยกตัวอย่างเช่น SA tag บ่งชี้รีดที่ถูกแตกออก (split-read) จะอยู่ในรูปแบบ SA:Z:chr5,18606884,-,52M75S,10,0; ที่หมายถึงการแตกรีดที่โครโมโซม 5 (chr5) ตำแหน่งที่ 18606884 สแตรนด์ (strand) เป็นแบบอ่านย้อนกลับ และมี รหัส CIGAR เป็น 52M75S คุณภาพแมพปิงของรีดเท่ากับ 10 และ NM บอกค่าจำนวนความต่างระหว่างลำดับเบสและจีโนมอ้างอิงเท่ากับ 0 ทั้งนี้รหัส CIGAR 52M75S หมายถึง 52 เบสแรกของรีดตรง (match) กับลำดับเบสของจีโนมและอีก 75 เบสถัดมาเป็นเบสซอฟต์คลิป

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

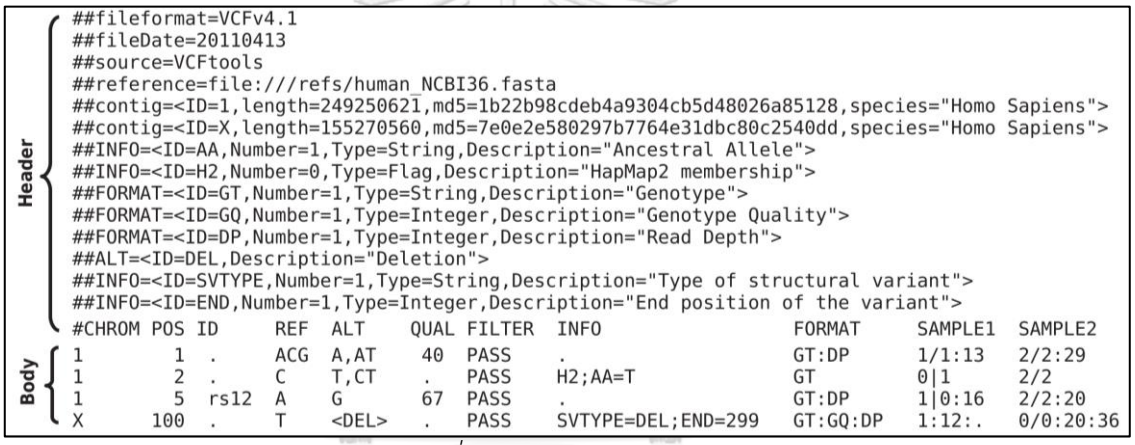
รูปที่ 8 ตัวอย่างรูปแบบไฟล์แซม (SAM)
(ที่มา : รูปที่ 2 ของ [39])

2.1.6 วีซีเอฟ (Variant call format: VCF)

วีซีเอฟ [40] เป็นไฟล์ที่เก็บผลลัพธ์ที่ได้จากโปรแกรมการวิเคราะห์การแปรผันทางพันธุกรรม โดยจะบอกถึงบริเวณที่มีความผิดปกติไม่ว่าจะเป็นตำแหน่ง ช่วงการแปรผัน หรือรูปแบบประเภทของการแปรผัน ตัวไฟล์ (รูปที่ 9) ประกอบด้วย 2 ส่วนได้แก่ ส่วนที่บอกรายละเอียดของข้อมูลภายในไฟล์ ซึ่งจะใส่คำอธิบายต่างๆ ทั้งเวอร์ชันของวีซีเอฟหรืออักษรย่อต่างๆที่จะขึ้นต้นด้วยเครื่องหมาย # โดยข้อมูลส่วนแรกนี้จะอยู่ต้นไฟล์ ส่วนที่สองเป็นรายการของผลที่เกิดตำแหน่งการแปรผัน โดยจะมีฟิลด์ต่างๆดังแสดงในตารางที่ 2 ต่อไปนี้

ตารางที่ 2 ฟิลด์หลักในไฟล์วีซีเอฟ

คอลัมน์	ฟิลด์	ประเภทของข้อมูล	คำอธิบาย
1	CHROM	String	ชื่อของโครโมโซม
2	POS	Int	ตำแหน่งของเบสแรกที่เกิดการแปรผัน
3	ID	String	ระบุชื่อของการแปรผันนี้ (เช่น DELLY001 หรือ Manta121 เป็นต้น)
4	REF	String	ระบุเบสบนจีโนมอ้างอิง
5	ALT	String	ระบุเบสที่มีการแปรผันไป
6	QUAL	Int	ระบุคุณภาพของเบส
7	FILTER	String	ระบุว่าการแปรผันนี้ผ่านการกรองผลลัพธ์หรือไม่
8	INFO	String	ระบุข้อมูลเพิ่มเติม



รูปที่ 9 ตัวอย่างไฟล์วีซีเอฟ

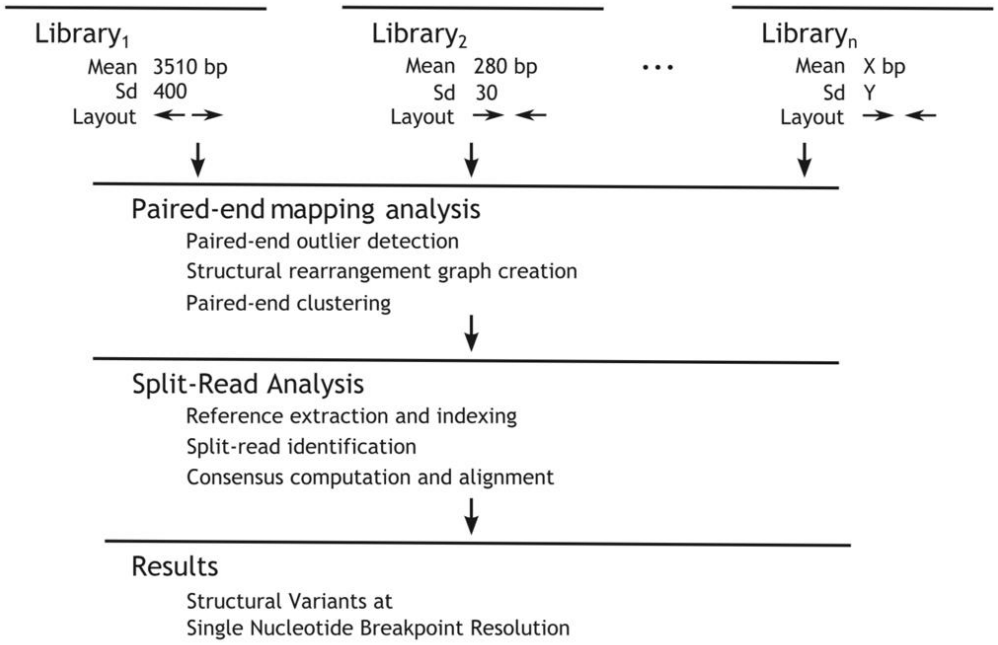
(ที่มา : รูปที่ 1 ของ [40])

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 DELLY

DELLY [41] เป็นเครื่องมือหาการแปรผันเชิงโครงสร้างในแต่ละโครโมโซม สามารถรองรับความยาวคู่รีดที่หลากหลาย เช่น paired-end, mate-pair เป็นต้น และ รองรับการหา deletion, inversion, tandem duplication และ chromosomal translocation โดย (รูปที่ 10) ขั้นตอนแรก DELLY จะทำการหาค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐาน หลังจากนั้นจะใช้คู่รีดในการวิเคราะห์ โดยที่ ถ้าเป็นกรณี deletion ลักษณะของความยาวคู่รีดต้องมีลักษณะที่ใหญ่กว่าปกติ แต่มีลักษณะของทิศทางเป็นปกติในกรณีของ inversion DELLY จะดูจากลักษณะของทิศทางที่ผิดปกติของรีดที่เข้าคู่ ถ้าเป็นกรณีของ tandem duplication จะใช้การดูตำแหน่งของรีดโดยที่รีดที่เข้าคู่จะสลับ

ลำดับตำแหน่งกัน ถ้าเป็นกรณีของ chromosomal translocation จะใช้การดูจากรีดที่เข้าคู่ที่ไปอยู่อีกโครโมโซมหนึ่ง และ insertion ดูจากลักษณะของความยาวคู่รีดต้องมีขนาดเล็กกว่าปกติ หลังจากวิเคราะห์ด้วยคู่รีดเสร็จในแต่ละกลุ่มรีดที่มีลักษณะผิดปกติจะถูกนำมาสร้างกราฟ โดยรีดที่มีลักษณะ single-anchored paired-end คือรีดหนึ่งมีลำดับเบสที่เหมือนกับบนจีโนมอ้างอิง ส่วนคู่ของรีด (mate) ไม่ถูกแมพคือไม่มีตำแหน่งบนจีโนมที่มีลำดับเบสเหมือนคู่ของรีดนี้เลย จะถูกนำมาหาด้วยวิธีการแตกรีด ซึ่งรีดที่ไม่ถูกแมพ จะต้องมี 2 ส่วนที่เทียบกันได้กับจีโนมอ้างอิง คือจุดเริ่มต้นที่เกิดการแปรผัน และจุดสิ้นสุดที่เกิดการแปรผัน ซึ่งสองจุดนี้เรียกว่าตำแหน่งเบรกเอ็นด์ (breakends) DELLY ใช้ Gotoh algorithm ในการเทียบรีดที่ไม่ถูกแมพกับจีโนมอ้างอิง ซึ่งรองรับช่องว่าง (gap) ในการเทียบลำดับเบส ขั้นตอนสุดท้าย DELLY จึงให้คะแนนกับเบรกเอ็นด์ และแสดงผลอยู่ในรูปแบบของไฟล์วีซีเอฟ



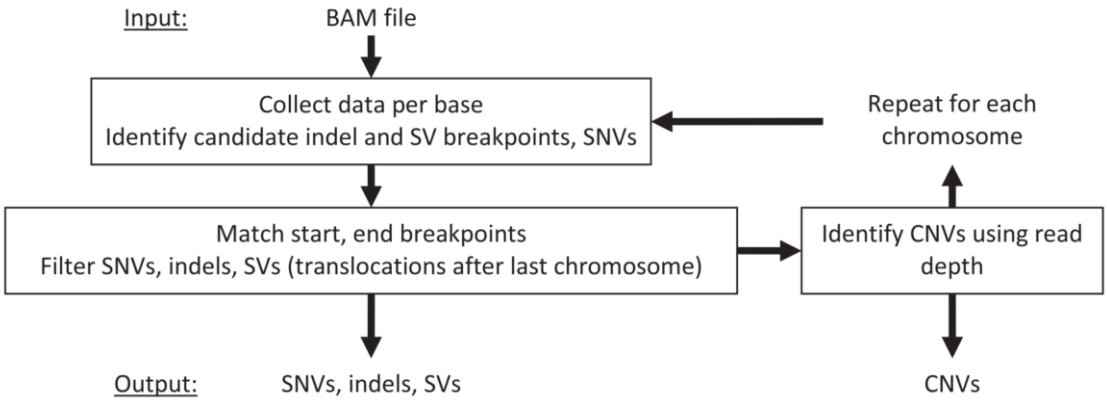
รูปที่ 10 ขั้นตอนการทำงานของ DELLY

(ที่มา : รูปที่ 1 ของ [41])

2.2.2 GROM (Genome Rearrangement OmniMapper)

GROM [42] เป็นเครื่องมือหนึ่งที่สามารถค้นหาได้ทั้ง insertion, deletion, tandem duplication, inversion, indel, SNV และ CNV โดยใช้วิธีคู่รีด แตกรีด และนับรีด และสามารถใช้ค้นหาทั้งจีโนม เฉพาะส่วนที่เป็นเอ็กโซม หรือจากลำดับเบสของอาร์เอ็นเอ ข้อมูลที่ GROM ใช้ประกอบด้วยค่าเฉลี่ยคุณภาพแมพปิงของรีด และคุณภาพของเบส กลุ่มรีดที่มีลักษณะผิดปกติ รีดที่เข้าคู่ที่ไม่ถูกแมพ รีดที่ถูกแตก และความลึกของรีด โดยขั้นตอนแรก GROM (รูปที่ 11) จะทำการหา

ค่าเฉลี่ยความยาวคู่รีดจำนวน 10 ล้านคู่รีด หลังจากนั้นจึงดูจากคู่รีดที่มีลักษณะผิดปกติ ได้แก่ ทิศทางรีด และความยาวคู่รีด หลังจากนั้นจึงเก็บรีดที่มีลำดับเบสซอฟต์คลิป มากกว่าหรือเท่ากับ 5 เบส หรือส่วนที่เป็นแตกรีด (SA tag) ต้องมากกว่าหรือเท่ากับ 20 เบส หลังจากนั้นจึงนำรีดกลุ่มนี้เทียบกับจีโนม ซึ่งในกระบวนการนี้ส่วนที่มีเบรกเอ็นด์เหมือนกันหรือทับกันอยู่จะทำการรวมกัน และในตอนประมวลผลจะแบ่งตามประเภทของการแปรผัน และขนาดของกลุ่มรีดที่ผิดปกติ ส่วนวิธีนับรีดใช้เพื่อช่วยกำหนดขอบเขตบริเวณที่คาดว่าเกิด deletion และ tandem duplication แล้วใช้ลำดับเบสซอฟต์คลิป ในบริเวณนั้นเทียบ ซึ่งคุณภาพแมพปิงของรีดที่ได้ต้องมากกว่าหรือเท่ากับ 20 ผลสรุปของ GROM จะอยู่ในไฟล์วีซีเอฟ GROM รองรับการหาการแปรผันทางพันธุกรรมโดยใช้สายเดี่ยว (single-read) ด้วย แต่ผลที่ได้จะมีข้อจำกัดตามรีดประเภทนี้ GROM ต้องการหน่วยความจำประมาณ 13 กิกะไบต์ สำหรับกรณีเธรดเดียว และ 128 กิกะไบต์ สำหรับกรณี 24 เธรด

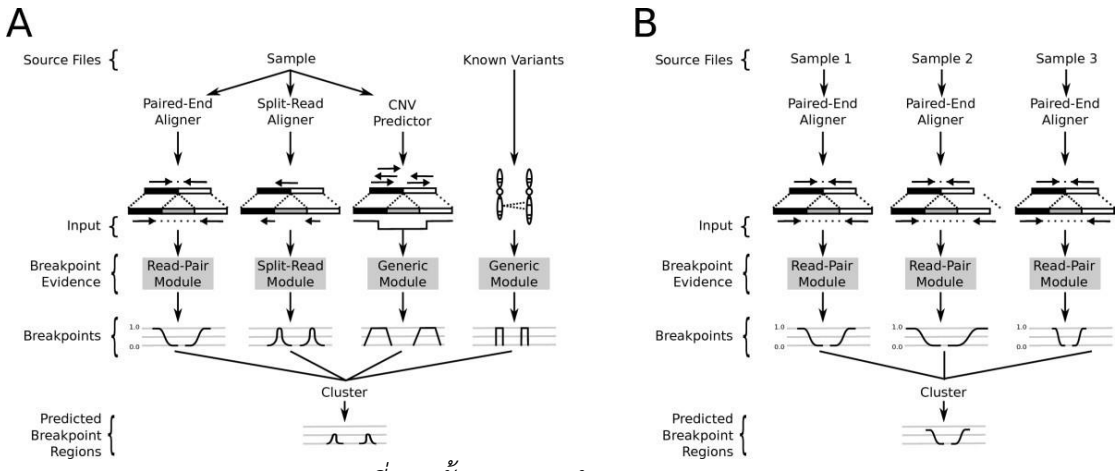


รูปที่ 11 ขั้นตอนการทำงานของ GROM

(ที่มา : รูปที่ 2 ของ [42])

2.2.3 LUMPY

LUMPY [43] เป็นเครื่องมือที่รองรับการตรวจหาการแปรผันเชิงโครงสร้างที่หลากหลาย ได้แก่ deletion, tandem duplication, inversion และ chromosomal translocation โดยใช้ทั้งวิธีการคู่รีด แตกรีด และนับรีด แล้วใช้ข้อมูลทางสถิติในการช่วยตัดสินใจเลือกผลลัพธ์ที่ได้ นอกจากนี้ LUMPY ยังรองรับการระบุตำแหน่งของเอวิเด้นซ์เพื่อให้ LUMPY หาตำแหน่งเบรกพอยต์หรือเบรกเอ็นด์ได้ด้วยการใช้ไฟล์ BEDPE (รูปที่ 12) LUMPY ได้ทำการเปรียบเทียบกับเครื่องมือ GASVPro, DELLY และ Pindel ซึ่งสำหรับข้อมูลจำลอง LUMPY มีค่าความครบถ้วนที่สูงกว่าเครื่องมืออื่นๆ ในการตรวจหาการแปรผันเชิงโครงสร้างประเภทต่างๆ ยกเว้น chromosomal translocation ส่วนข้อมูลจริง LUMPY ใช้ข้อมูล NA12878 ในการทดสอบประสิทธิภาพ โดยสามารถลดการตรวจหาที่ผิดได้ดีกว่าเครื่องมืออื่นๆ



รูปที่ 12 ขั้นตอนการทำงานของ LUMPY

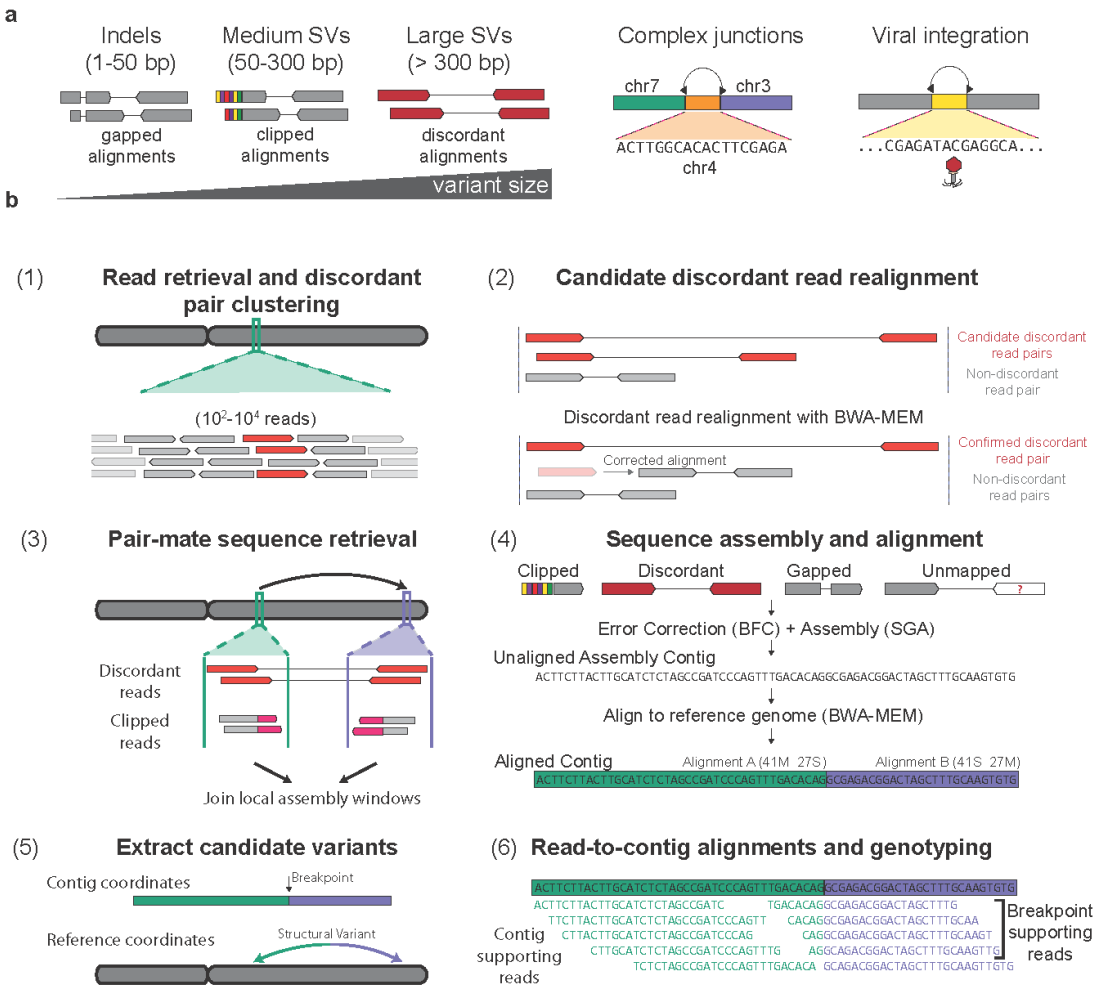
(ที่มา : รูปที่ 2 ของ[43])

2.2.4 SvABA

SvABA [44] เป็นเครื่องมือที่รองรับการตรวจหาการแปรผันเชิงโครงสร้างที่หลากหลาย ภาพรวมของวิธีการของ SvABA (รูปที่ 13) เครื่องมือเริ่มจากการหาเอวิเด้นซ์ ซอฟต์คลิปรีด คู่รีดที่มีลักษณะผิดปกติ รีดที่ไม่สามารถแมพได้ และรีดที่มีอินเดล เมื่อได้ตำแหน่งของเอวิเด้นซ์ SvABA ใช้ String Graph Assembler (SGA) เพื่อที่จะรวมส่วนของรีดที่ไม่สามารถเทียบกับจีโนมอ้างอิงได้ หลังจากนั้นจึงทำการเทียบกับจีโนมอ้างอิงด้วย BWA-MEM เพื่อหาตำแหน่งเบรกพอยต์

2.2.5 Wham

Wham [45] เป็นอีกเครื่องมือหนึ่งที่ถูกสร้างเพื่อการตรวจหาการแปรผันเชิงโครงสร้างในแต่ละโครโมโซม Wham มีจุดเด่นในการประยุกต์ใช้การเรียนรู้ด้วยเครื่อง วิธี Random forest ของ Decision trees ในการแยกแยะการแปรผันเชิงโครงสร้างในแต่ละโครโมโซม ซึ่งวิธีการแยกแยะของ Random forest จะใช้ 14 คุณลักษณะ ได้แก่ 1. Discordant 2. Mate not mapped 3. Mates mapped to same strand 4. Mates on different seqids 5. Number of split reads 6. Split read (fragment 1) on same strand as mate 7. Split read (fragment 2) on same strand as mate 8. Split read (fragment 1) and read two (fragment 2) on same strand 9. Internal insertion 10. Internal deletion 11. Mates mapped too close 12. Mates mapped to far 13. Everted pairs และ 14. Relative depth ซึ่งแต่ละคุณลักษณะจะถูกปรับค่าทำให้เป็นมาตรฐาน (normalize) ด้วยความลึกของรีด ณ ตำแหน่งนั้นๆ ชุดข้อมูลในการเรียนรู้ได้มาจากการจำลองข้อมูล โดยผลของ cross-validation เท่ากับ 0.94 อย่างไรก็ตามผู้ใช้สามารถทำการให้เครื่องมือเรียนรู้ใหม่ได้



รูปที่ 13 ขั้นตอนการทำงานของ SvABA

(ที่มา : รูปที่ 2 ของ [44])

2.2.6 เปรียบเทียบวิธีการที่นำเสนอกับเครื่องมือที่มีมาก่อน

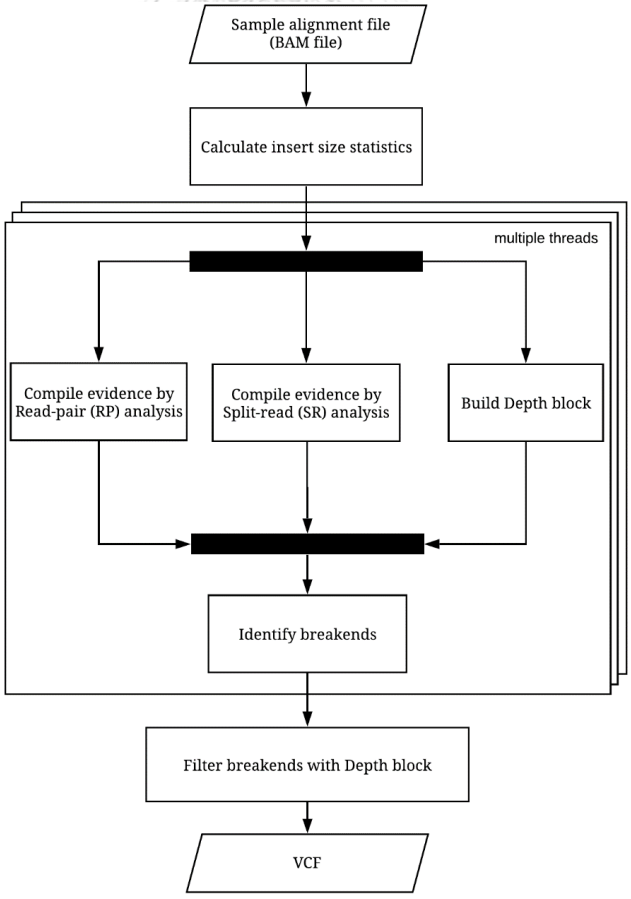
กรณีของการตรวจหาการแปรผันเชิงโครงสร้างในแต่ละโครโมโซม พบว่าทุกเครื่องมือมีพื้นฐานการตรวจจับการแปรผันเชิงโครงสร้างในภาพรวมมีวิธีการที่คล้ายกัน คือวิธีคู่รีด (read-pair) เพื่อหาเอวิเด้นซ์และขั้นตอนสุดท้ายจะใช้วิธีแตกรีด (split-read) ในการเทียบรีดกับจีโนมอ้างอิงเพื่อหาตำแหน่งเบรกเอ็นด์ อย่างไรก็ตามในแต่ละเครื่องมือยังมีรายละเอียดปลีกย่อยที่แตกต่างกัน รวมไปถึงความสามารถในการตรวจหาการแปรผันเชิงโครงสร้างบางประเภท

บทที่ 3
วิธีการดำเนินงานวิจัย

3.1 แนวคิดและวิธีการวิจัย

3.1.1 ภาพรวมขั้นตอนการทำงานของอัลกอริทึม

ภาพรวมการทำงานของอัลกอริทึมเริ่มจากรับไฟล์แบมและจีโนมอ้างอิงเป็นข้อมูลเข้า (รูปที่ 14) ทำการคำนวณหาค่าเฉลี่ยของความยาวคู่รีด และทำการแบ่งข้อมูลออกเป็นหลายๆ ชุดตามจำนวนของโครโมโซมของไฟล์แบมแล้วส่งขอบเขตที่แบ่งไว้ไปยังส่วนของการวิเคราะห์ ประกอบด้วยการรวบรวมและวิเคราะห์คู่รีด แตกรีด และการสร้าง Depth block โดยการวิเคราะห์จะทำงานแบบขนานเพื่อลดระยะเวลาในการประมวลผล หลังจากถูกส่งไปในส่วนของการวิเคราะห์ ฟังก์ชันอ่านรีดจึงทำการอ่าน แล้วทยอยส่งรีดไปยังฟังก์ชันที่ใช้ในการหาการแปรผันทางพันธุกรรมแต่ละประเภท การหาตำแหน่งเบรกเอ็นด์ และการคัดกรองตำแหน่งเบรกเอ็นด์ โดยผลลัพธ์สุดท้ายที่ได้จะส่งออกความแปรผันเชิงโครงสร้างทางพันธุกรรม insertion, deletion, tandem duplication, inversion และ chromosomal translocation ที่หาได้ในรูปแบบไฟล์วีซีเอฟ



รูปที่ 14 ขั้นตอนการทำงานของโปรแกรม

3.1.2 ขั้นตอนการคำนวณเชิงสถิติของความยาวคู่รีด

การคำนวณหาค่าเฉลี่ยของความยาวคู่รีด โดยการอ่านรีดเป็นจำนวน 100,000 รีด โดยลักษณะของรีดที่เลือกมาทำการคำนวณหาค่าเฉลี่ยต้องเป็นรีดที่ไม่มีอินเดล ไม่มีลำดับเบสที่เป็นซอฟต์คลิป ความยาวคู่รีดต้องไม่เกิน 100,000 เบส ต้องอยู่บนโครโมโซมเดียวกัน ผลการเทียบรีดต้องไม่เป็นลำดับรอง (secondary alignment) และไม่เป็น PCR duplication

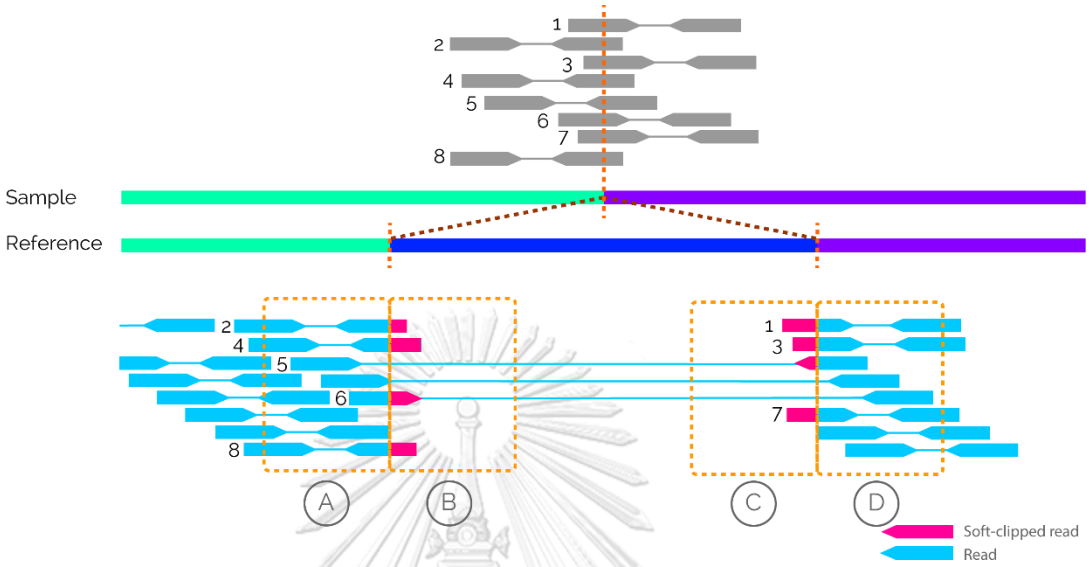
3.1.3 ขั้นตอนการรวบรวมหลักฐานโดยการวิเคราะห์คู่รีด

ภาพรวมของขั้นตอนการหาการแปรผันเชิงโครงสร้างจะเริ่มต้นด้วยวิธีการวิเคราะห์คู่รีด (read-pair) โดยขั้นตอนนี้จะมีอ็อบเจกต์ที่ชื่อว่าเอวิเด้นซ์ (evidence) ที่ได้มาจากการแปลงรีดที่มีลักษณะผิดปกติ ซึ่งภายในเอวิเด้นซ์ประกอบไปด้วยตำแหน่งที่คาดว่ามีเบรกเอ็นด์ และรีดจำนวนหนึ่งที่คาดว่าสามารถช่วยให้หาเบรกเอ็นด์ได้ สำหรับขั้นตอนวิธีการวิเคราะห์คู่รีดสำหรับแต่ละประเภทของการแปรผันเชิงโครงสร้างมีดังต่อไปนี้

1. ขั้นตอนการวิเคราะห์คู่รีดเพื่อหาการแปรผันเชิงโครงสร้างแบบ deletion

การเกิด deletion นั้นจะมีบริเวณหนึ่งถูกลบออกไป (รูปที่ 15) เมื่อทำการเทียบกับจีโนมอ้างอิงด้านทางซ้ายบริเวณ B จะพบซอฟต์คลิป (ส่วนของรีดที่ไม่ตรงกับลำดับเบสในจีโนมอ้างอิง) ที่เกิดจากการยืดออกจากกันเมื่อเทียบกับจีโนมอ้างอิง ซึ่งบริเวณนี้จะเป็นบริเวณที่เป็นจุดเริ่มต้นของเบรกเอ็นด์ และสำหรับบริเวณด้านขวาก็จะเกิดซอฟต์คลิป ในบริเวณ C โดยต้องพิจารณาเงื่อนไขต่อไปนี้จากคู่รีด คือ ถ้าเป็นทางฝั่งทางด้านซ้าย ตำแหน่งของรีดจะต้องมีตำแหน่งที่น้อยกว่าตำแหน่งของรีดที่เข้าคู่ สแตรนด์ของรีดจะต้องเป็น + (forward) และสแตรนด์ของรีดที่เข้าคู่เป็น – (reverse) เช่นตัวอย่างรีดหมายเลข 4 และความยาวคู่รีดต้องมีค่ามากกว่าค่าเฉลี่ย หลังจากนั้นจึงทำการนำลำดับเบสซอฟต์คลิป ที่ตกบริเวณ B มาทำเทียบกับจีโนมอ้างอิงที่อยู่บนบริเวณ D ถ้าลำดับเบสซอฟต์คลิป ของรีดมีจำนวนเบสเข้ากันได้มากกว่าหรือเท่ากับจำนวนของลำดับเบสซอฟต์คลิป ก็จะถูกเก็บไว้เป็นตำแหน่งเป็นไปได้ว่าเป็นจุดสิ้นสุดของเบรกเอ็นด์ ในขณะที่ตำแหน่งรีดตัวมันเองมีจุดเริ่มต้นของเบรกเอ็นด์ ส่วนถ้าเป็นฝั่งบริเวณด้านขวาเช่นรีดที่เข้าคู่หมายเลข 5 เงื่อนไขในการเกิดเหตุการณ์คือตำแหน่งรีดจะต้องมากกว่าตำแหน่งรีดที่เข้าคู่ สแตรนด์ของรีดจะต้องเป็น - และสแตรนด์ของรีดที่เข้าคู่ต้องเป็น + และความยาวคู่รีดต้องมีค่ามากกว่าค่าเฉลี่ย หลังจากเมื่อนำลำดับเบสซอฟต์คลิป ที่ตกบนบริเวณ C มาทำการเทียบกับจีโนมอ้างอิงที่อยู่บนบริเวณ A ถ้าลำดับเบสซอฟต์คลิป ของรีดมีจำนวนเบสเข้ากันได้มากกว่าหรือเท่ากับจำนวนของลำดับเบสซอฟต์คลิป ก็จะถูกเก็บไว้เป็นตำแหน่งที่เป็นไปได้ว่าเป็นจุดเริ่มต้นของเบรกเอ็นด์ และตำแหน่งรีดตัวมันเองมีจุดสิ้นสุดของเบรกเอ็นด์

ดังนั้นการหา deletion คือ การนำลำดับเบสซอฟต์คลิป บริเวณ B เทียบกับ D และเบสซอฟต์คลิปบริเวณ C เทียบกับ A โดยที่เบรกเอ็นด์เริ่มต้นจะอยู่บนบริเวณ A และเบรกเอ็นด์สิ้นสุดจะอยู่บริเวณ D

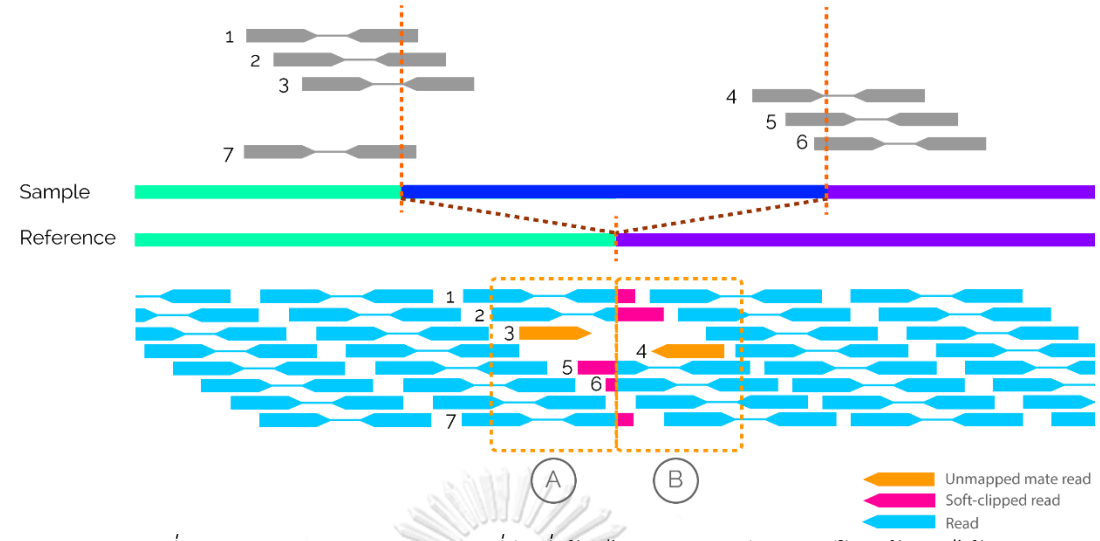


รูปที่ 15 รูปแบบของ deletion

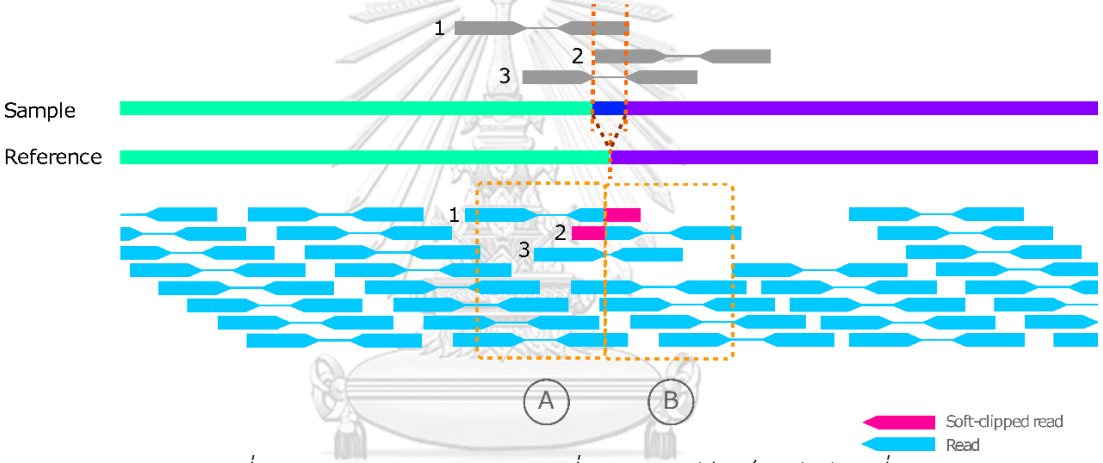
2. ขั้นตอนการวิเคราะห์คู่รีดเพื่อหาการแปรผันเชิงโครงสร้างแบบ insertion

การเกิด insertion (รูปที่ 16 และ 17) ที่มีเบสอื่นจำนวนหนึ่งเข้ามาแทรก แต่เมื่อทำการเทียบกับจีโนมอ้างอิงในบริเวณเดียวกันจะไม่มีอยู่ ทำให้มีรีดที่ไม่ถูกแมพเกิดขึ้น และเกิดลำดับเบสซอฟต์คลิปบนบริเวณ A และ B โดยที่ส่วนที่เป็นลำดับเบสซอฟต์คลิป บริเวณ B เกิดจากส่วนที่เข้ามาแทรกในจุดเริ่มต้น และลำดับเบสซอฟต์คลิปบริเวณ A เกิดจากส่วนที่เข้ามาแทรกในส่วนปลาย ดังนั้นอัลกอริทึมจึงพิจารณาเงื่อนไขต่อไปนี้จากคู่รีด (1) ถ้าเป็นกรณีที่รีดที่เข้าคู่ไม่สามารถเทียบกับจีโนมอ้างอิงได้ (รีดที่ 3 รูปที่ 16) คือ สแตรนด์ของรีดจะเป็น + และ รีดที่เข้าคู่ไม่สามารถเทียบกับจีโนมอ้างอิง (2) ลำดับเบสซอฟต์คลิป จะต้องมีทั้งบริเวณ A และบริเวณ B สำหรับกรณีที่ความยาวคู่รีดเล็กกว่าค่าเฉลี่ย (รูปที่ 17) สแตรนด์ของรีดจะเป็น + และ รีดที่เข้าคู่เป็น – และลำดับเบสซอฟต์คลิปจะต้องมีทั้งบริเวณ A และบริเวณ B

ดังนั้นการหา insertion ทั้งในแบบที่รีดที่เข้าคู่ไม่สามารถเทียบกับจีโนมอ้างอิงได้และที่ความยาวคู่รีดเล็กกว่าค่าเฉลี่ย ลักษณะเบรกเอ็นด์จะอยู่บริเวณ ระหว่าง A และ B ซึ่งกรณีของ insertion ตำแหน่งเบรกเอ็นด์เป็นตำแหน่งเดียวกัน ข้อจำกัดของ insertion คือ ถ้าขนาดของ insertion กว้างมากๆ จะไม่สามารถทราบขนาดของ insertion ได้



รูปที่ 16 รูปแบบของ insertion ที่รีดที่เข้าคู่ไม่สามารถเทียบกับจีโนมอ้างอิงได้



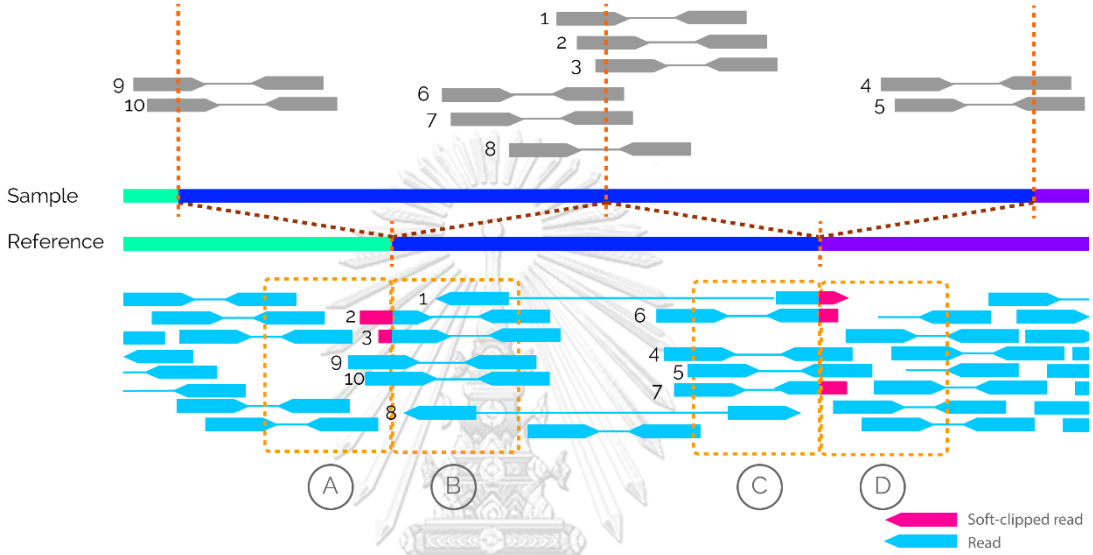
รูปที่ 17 รูปแบบของ insertion ที่ความยาวคู่รีดเล็กกว่าค่าเฉลี่ย

3. ขั้นตอนการวิเคราะห์คู่รีดเพื่อหาการแปรผันเชิงโครงสร้างแบบ tandem duplication

การเกิด tandem duplication จะมีบริเวณที่ซ้ำกันเกิดขึ้น (รูปที่ 18) แต่เมื่อทำการเทียบกับจีโนมอ้างอิงมีเพียงแค่บริเวณเดียวเท่านั้น ทำให้เวลาทำการเทียบกับจีโนมอ้างอิงส่วนที่ซ้ำกันจะต้องตกในจุดเดียวกัน แต่เมื่อรีดกับรีดที่เข้าคู่อยู่คนละส่วนที่ซ้ำกัน (รูปที่ 18 รีดหมายเลข 1, 8) ทำให้รีดทั้งคู่เกิดสแตรนด์ที่ผิดปกติ และลำดับของตำแหน่งก็จะสลับกัน ดังนั้นต้องพิจารณาเงื่อนไขต่อไปนี้จากคู่รีด คือ ถ้าเป็นฝั่งทางซ้ายคือ รีดมีสแตรนด์เป็น – มีตำแหน่งของรีดน้อยกว่าตำแหน่งของรีดที่เข้าคู่ ที่มีสแตรนด์เป็น + (รูปที่ 18 รีดหมายเลข 1,8) หลังจากนั้นเมื่อนำลำดับเบสซอฟต์คลิป ที่ตกบนบริเวณ A มาทำเทียบกับจีโนมอ้างอิงที่อยู่บนบริเวณ C ถ้าลำดับเบสซอฟต์คลิปของรีดมีจำนวนเบสเข้ากันได้มากกว่าหรือเท่ากับจำนวนของลำดับเบสซอฟต์คลิป ก็เป็นไปได้ว่าตำแหน่งบริเวณนั้นคือจุดสิ้นสุดของเบรกเอ็นด์และตัวรีดเองมีจุดเริ่มต้นของเบรกเอ็นด์ ในกรณีของฝั่งขวา เมื่อพิจารณารีดหมายเลข 1 เมื่อนำลำดับเบสซอฟต์คลิป ที่ตกบนบริเวณ D มาเทียบกับจีโนมอ้างอิงที่อยู่บนบริเวณ B

ถ้าลำดับเบสซอฟต์คลิป ของรีด มีจำนวนเบสเข้ากันได้มากกว่าหรือเท่ากับจำนวนของลำดับเบสซอฟต์คลิป ซึ่งจะเป็นไปได้ว่าตำแหน่งบริเวณนั้นคือจุดเริ่มต้นของเบรกเอ็นด์ และตำแหน่งรีดตัวมันเองมีจุดสิ้นสุดของเบรกเอ็นด์

ดังนั้นการหา tandem duplication คือการนำลำดับเบสซอฟต์คลิป บริเวณ A เทียบกับ C และนำลำดับเบสซอฟต์คลิป บริเวณ D เทียบกับ B โดยลักษณะเบรกเอ็นด์จุดเริ่มต้นจะอยู่บนบริเวณ B และจุดสิ้นสุดจะอยู่บริเวณ C



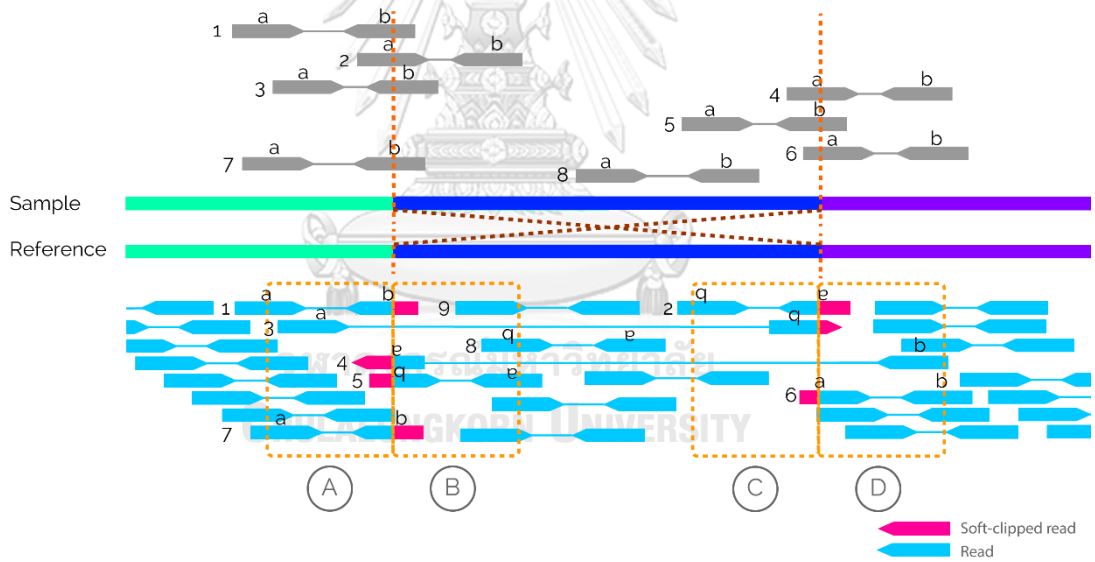
รูปที่ 18 รูปแบบของ tandem duplication

4. ขั้นตอนการวิเคราะห์คู่รีดเพื่อหาการแปรผันเชิงโครงสร้างแบบ inversion

การเกิด inversion โดยจะมีบริเวณหนึ่งที่กลับด้านกัน ซึ่งเมื่อต้องทำการเทียบกับจีโนมอ้างอิง จึงต้องทำ reverse complement (เป็นการกลับด้านกันของลำดับเบส หลังจากนั้นทำการเปลี่ยนอักษรเบส จาก A เป็น T หรือ T เป็น A และ G เป็น C หรือ C เป็น G ยกตัวอย่างเช่น ACTG เปลี่ยนเป็น CAGT เป็นต้น) ถึงจะเทียบกันได้ โดยเงื่อนไขในการตรวจสอบการแปรผันนี้คือ (รูปที่ 19) สำหรับบริเวณด้านซ้ายตำแหน่งของรีดจะต้องน้อยกว่าตำแหน่งของรีดที่เข้าคู่ สแตรนด์ของรีดจะต้องเป็น - และสแตรนด์ของรีดที่เข้าคู่จะต้องเป็น – เช่น รีด 4 หลังจากนั้นถ้าเรานำซอฟต์คลิปบริเวณ A มาทำ reverse complement จะต้องเทียบเข้ากันได้กับส่วนของจีโนมอ้างอิงบนตำแหน่ง D ได้ หรือนำซอฟต์คลิป บริเวณ B มาทำ reverse complement ต้องเทียบเข้ากันได้กับส่วนของจีโนมอ้างอิงบนตำแหน่ง C ได้ ถ้าลำดับเบสซอฟต์คลิป ของรีดมีจำนวนเบสเข้ากันได้มากกว่าหรือเท่ากับจำนวนของลำดับเบสซอฟต์คลิป ก็จะเป็นไปได้ว่าตำแหน่งบริเวณนั้นคือจุดสิ้นสุดของเบรกเอ็นด์ และตำแหน่งรีดตัวมันเองมีจุดเริ่มต้นของเบรกเอ็นด์ ส่วนถ้าเป็นทางด้านขวา ตำแหน่งของรีดจะต้องมากกว่าตำแหน่งของรีดที่เข้าคู่ เช่น รีด b รีดที่เข้าคู่ของรีด 3 สแตรนด์ของรีดจะต้องเป็น +

และสแตรนด์ของรีดที่เข้าคู่จะต้องเป็น + หลังจากนั้นถ้าเรานำซอฟต์คลิปบริเวณ C มาทำ reverse complement จะต้องเทียบเข้ากันได้กับส่วนของจีโนมอ้างอิง บนตำแหน่ง B ได้ หรือนำซอฟต์คลิปบริเวณ D มาทำ reverse complement จะต้องเทียบเข้ากันได้กับส่วนของจีโนมอ้างอิงบนตำแหน่ง A ได้ ถ้าลำดับเบสซอฟต์คลิป ของรีดมีจำนวนเบสเข้ากันได้มากกว่าหรือเท่ากับจำนวนของลำดับเบสซอฟต์คลิป ซึ่งจะเป็นไปได้ว่าตำแหน่งบริเวณนั้นคือจุดเริ่มต้นของเบรกเอ็นด์ และตำแหน่งรีดตัวมันเองมีจุดสิ้นสุดของเบรกเอ็นด์

ดังนั้นการหา inversion คือ ก่อนทำการเทียบต้องทำการ reverse complement ทุกครั้ง โดยการนำลำดับเบสซอฟต์คลิป บริเวณ A ทำ reverse complement และทำการเทียบกับ D ลำดับเบสซอฟต์คลิป บริเวณ B ทำ reserve complement แล้วเทียบกับ C ลำดับเบสซอฟต์คลิป บริเวณ C ทำ reverse complement แล้วเทียบกับ B และลำดับเบสซอฟต์คลิปบริเวณ D ทำ reverse complement แล้วเทียบกับ A โดยลักษณะเบรกเอ็นด์จุดเริ่มต้นจะอยู่ระหว่างบริเวณ A และ B และจุดสิ้นสุดจะอยู่ระหว่างบริเวณ C และ D



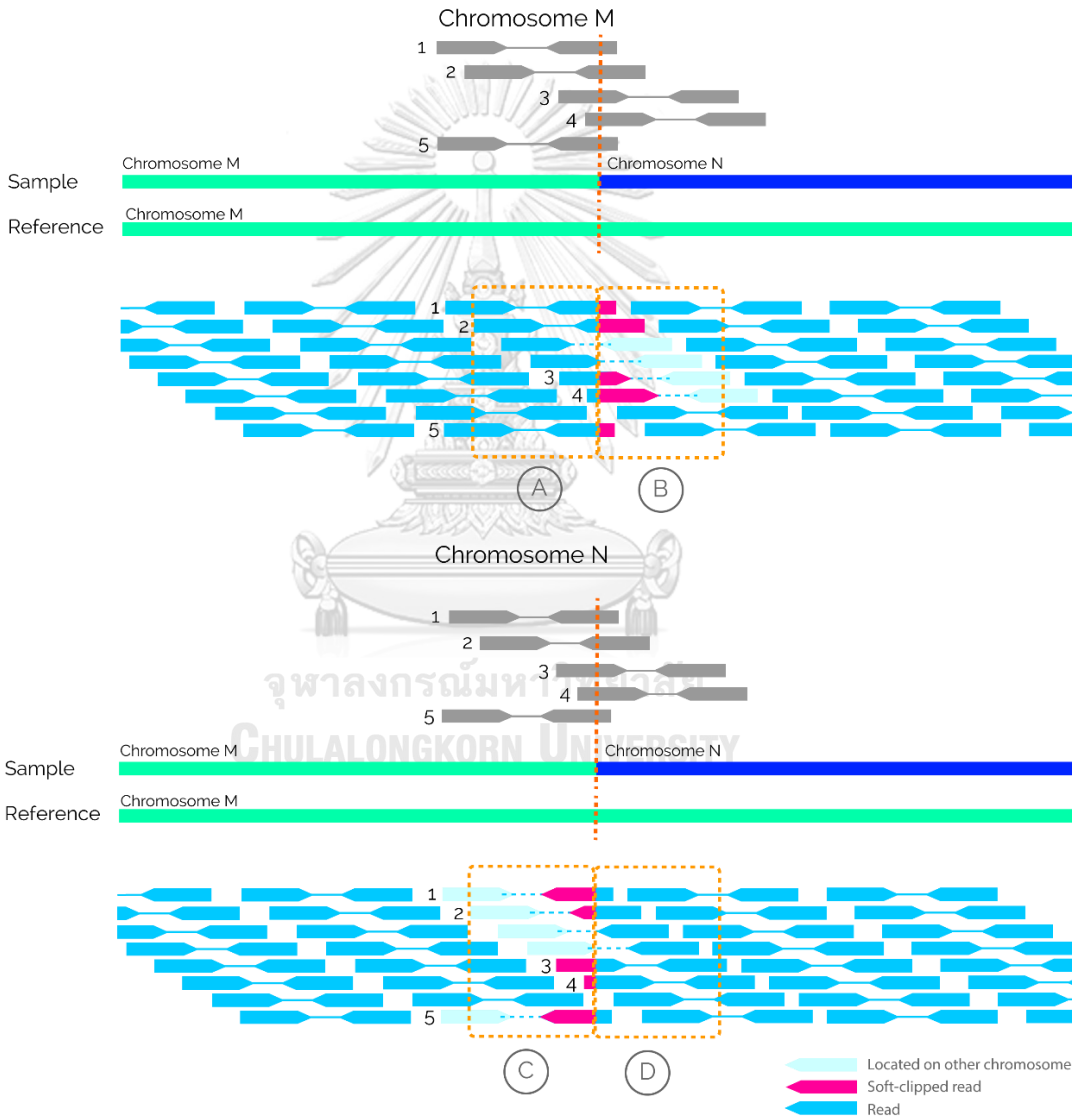
รูปที่ 19 รูปแบบของ inversion

5. ขั้นตอนการวิเคราะห์คู่รีดเพื่อหาการแปรผันเชิงโครงสร้างแบบ chromosomal translocation

การเกิด chromosomal translocation จะมีบริเวณหนึ่งของโครโมโซมย้ายสลับไปอีกโครโมโซมหนึ่ง ทำให้เมื่อย้ายไปแล้วจะต้องมีรีดบ้างคู่ที่จะแยกกันอยู่ (รูปที่ 20 เช่น รีดหมายเลข 3, 4) ดังนั้นต้องพิจารณาเงื่อนไขต่อไปนี้จากคู่รีด 1) ถ้าเป็นโครโมโซม M ซึ่งรีดจะต้องมีรีดที่เข้าคู่อยู่บนโครโมโซม N สแตรนด์ของรีดเป็น + และ สแตรนด์ของรีดที่เข้าคู่เป็น – (รูปที่ 20 โครโมโซม M รีดหมายเลข 3, 4) โดยลำดับเบสซอฟต์คลิปบริเวณ B ของโครโมโซม M จะต้องสามารถเทียบเข้ากันได้

กับบริเวณ D ของโครโมโซม N และ 2) ถ้าเป็นโครโมโซม N ซึ่งรีดจะต้องมีรีดที่เข้าคู่อยู่โครโมโซม M สแตรนด์ของรีดเป็น - และสแตรนด์ของรีดที่เข้าคู่ เป็น + (รูปที่ 20 โครโมโซม N รีดหมายเลข 1, 2) ลำดับเบสซอฟต์คลิปบริเวณ C ของโครโมโซม N จะต้องสามารถเทียบเข้ากันได้กับบริเวณ A ของโครโมโซม M

ดังนั้นการหา chromosomal translocation คือการนำลำดับเบสซอฟต์คลิป บริเวณ B เทียบกับ D และนำลำดับเบสซอฟต์คลิป บริเวณ C เทียบกับ A ลักษณะเบรกเอ็นด์จุดเริ่มต้นจะอยู่บริเวณ A และจุดสิ้นสุดจะอยู่บริเวณ D



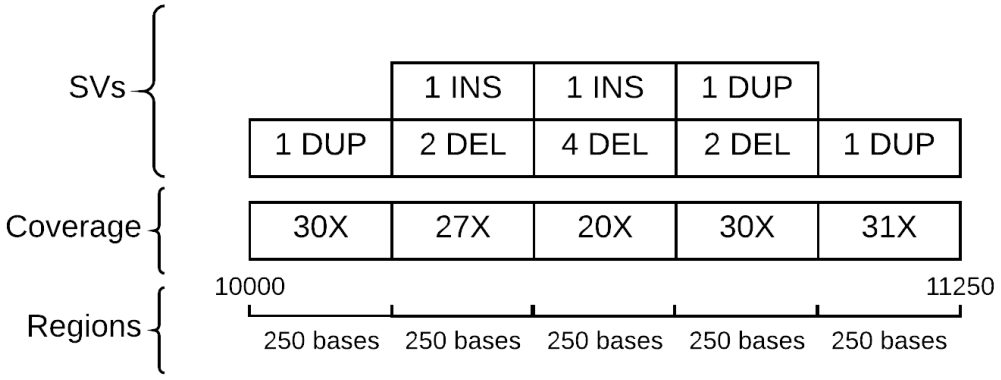
รูปที่ 20 รูปแบบของ chromosomal translocation

3.1.4 ขั้นตอนการรวบรวมหลักฐานโดยวิธีการแตกรีด

สำหรับอีกวิธีที่ใช้ในตรวจหาการแปรผันเชิงโครงสร้าง คือวิธีการแตกรีด (split-read) ที่เครื่องมือพิจารณาข้อมูลจากรีด 2 ส่วน คือ 1) รหัส CIGAR ที่ใช้อธิบายลักษณะของรีดเทียบกับจีโนมอ้างอิงยกตัวอย่างเช่น 100M50D100M ที่หมายถึงรีดนั้นมี 100 เบสแรกตรงกับจีโนมอ้างอิง (match) 50 เบสถัดมาเป็น deletion และ100 เบสท้ายตรงกับจีโนมอ้างอิง (match) ซึ่ง CIGAR สามารถใช้ในการตรวจหาการแปรผันเชิงโครงสร้าง ประเภท deletion และ insertion ได้ 2) แท็ก SA ที่ใช้อธิบายรีดที่มีบางส่วนสามารถเทียบได้กับอีกตำแหน่งหนึ่งบนจีโนมอ้างอิง ยกตัวอย่างเช่น SA:Z:scf7180000067989, 85273,-,54S47M,60,1 ความหมายของแต่ละตัวแปรคือ SA:Z:rname, pos, strand, CIGAR, mapQ, NM โดย Z คือบอกประเภทของแท็กเป็นแบบสตริง rname คือ ชื่อของโครโมโซมอีกตำแหน่งหนึ่ง pos คือตำแหน่งบนโครโมโซมอีกตำแหน่งหนึ่ง strand คือสแตรนด์ของอีกตำแหน่งหนึ่ง CIGAR คือ CIGAR ของอีกตำแหน่งหนึ่ง mapQ คือคุณภาพการแมพของอีกตำแหน่งหนึ่ง และ NM คือจำนวนความต่างระหว่างลำดับเบสและจีโนมอ้างอิงของอีกตำแหน่งหนึ่ง

3.1.5 การสร้าง Depth block

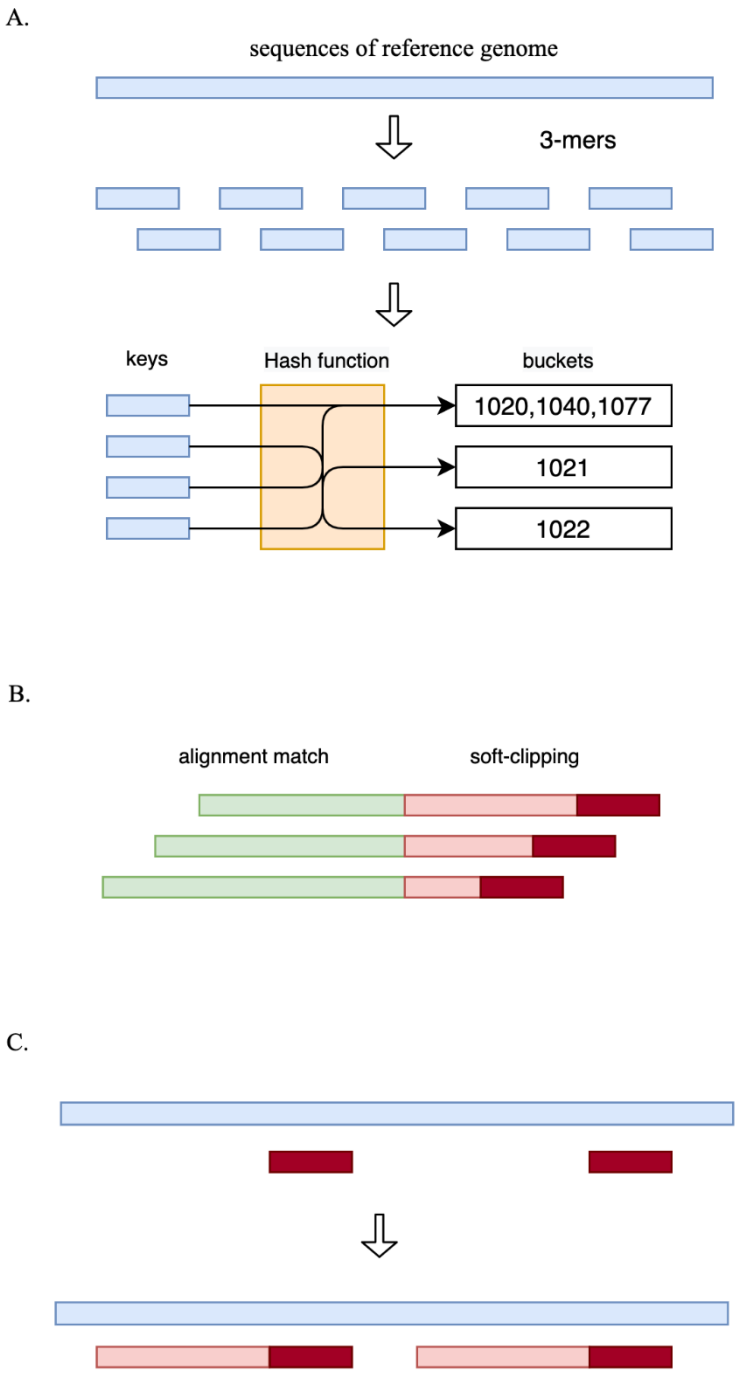
Depth block แสดงถึงภาพรวมของแต่ละโครโมโซมตามแต่ละบริเวณ (region) โดยในแต่ละบล็อกประกอบด้วยข้อมูลความลึกเฉลี่ยของรีด จำนวนการแปรผันเชิงโครงสร้างที่เป็นไปได้ที่ได้จากขั้นตอนก่อนหน้าและจำนวนของซอฟต์คลิปรีด โดยแต่ละโครโมโซมจะถูกแบ่งออกเป็นบล็อก บล็อกละ 250 เบส (รูปที่ 21) เนื่องจาก 250 เบส เป็นขนาดเฉลี่ยของความยาวรีดของเครื่องถอดรหัสจีโนมทั่วไปสำหรับการอ่านรหัสพันธุกรรมแบบสายสั้นคู่ (paired-end short reads) ซึ่งหากขนาดบล็อกใหญ่เกินไปอาจจะลดประสิทธิภาพในการกรองได้หรือหากขนาดบล็อกสั้นเกินไปก็ทำให้ไม่สามารถกรองได้ ข้อมูลในแต่ละบล็อก ประกอบด้วย (1) ค่าความลึกเฉลี่ยของรีดของบล็อกนั้น (2) จำนวนการแปรผันเชิงโครงสร้างแยกตามประเภท (3) จำนวนของซอฟต์คลิปรีดของแต่ละบล็อกที่มีความยาวมากกว่า 8 เบสโดยดูจากรหัส CIGAR เช่น 50S200M ซึ่งหมายถึง 50 เบสแรกของรีดเป็นซอฟต์คลิปและอีก 200 เบสถัดมาตรงกับจีโนมอ้างอิง หรือ 200M50S ซึ่งหมายถึง 200 เบสแรกของรีดตรงกับจีโนมอ้างอิงและอีก 50 เบสถัดมาเป็นซอฟต์คลิป อัลกอริทึมที่ใช้ในการตรวจสอบจำนวนการแปรผันเชิงโครงสร้างในแต่ละบล็อกเป็นอัลกอริทึมเดียวกับที่ใช้ในขั้นตอนการรวบรวมเอวิเด้นซ์



รูปที่ 21 รูปแบบของ Depth block

3.1.6 การระบุตำแหน่งของเบรกเอ็นด์

หลังการวิเคราะห์รีด รวบรวมหลักฐาน และสร้างบล็อกข้างต้น วิธีการที่นำเสนอจะทำการแมพรีดที่ผิดปกติที่รวบรวมได้เข้ากับจีโนมอ้างอิงใหม่อีกครั้ง เพื่อระบุตำแหน่งของเบรกเอ็นด์โดยการแมพรีดเข้ากับจีโนมอ้างอิงนี้ใช้ขั้นตอนวิธี Rabin–Karp [46] ที่แบ่งจีโนมอ้างอิงเป็นสตริง และใช้ฟังก์ชันแฮชแต่ละสตริงย่อย (รูปที่ 22 A) ซึ่งคีย์คือสตริงย่อยและ bucket คือตำแหน่งเริ่มต้นทั้งหมดบนจีโนมอ้างอิงที่พบแต่ละสตริงย่อย วิธีการที่นำเสนอใช้สตริงย่อยขนาด 3-mer (3 เบส) หลังจากนั้นจึงใช้สตริงย่อยของซอฟต์คลิป (รูปที่ 22 B) เข้าฟังก์ชันแฮชเพื่อหาตำแหน่งจีโนมอ้างอิงใน bucket ส่วนลำดับเบสที่เหลือจะใช้วิธีคล้ายวิธีการแบบซื่อตรง (Naive string matching) (รูปที่ 22 ข้อ C) โดยเพิ่มเติมให้รองรับเบสที่ไม่ตรงกัน (mismatch) โดยถ้าเบสไม่ตรงกันติดต่อกันเกิน 1 เบส จะทำการหยุดเทียบสตริงของรีดนั้น นอกจากนี้ได้ทำการกำหนดไว้ว่าในแต่ละตำแหน่งที่แมพได้อย่างน้อยต้องมีมากกว่า 8 เบสขึ้นไปจึงจะเก็บข้อมูลของรีด ณ ตำแหน่งดังกล่าว วิธีการแมพรีดใหม่นี้ใช้ได้กับการระบุตำแหน่งของเบรกเอ็นด์ของการแปรผันเชิงโครงสร้างทางพันธุกรรมประเภท deletion, inversion, tandem duplication และ chromosomal translocation



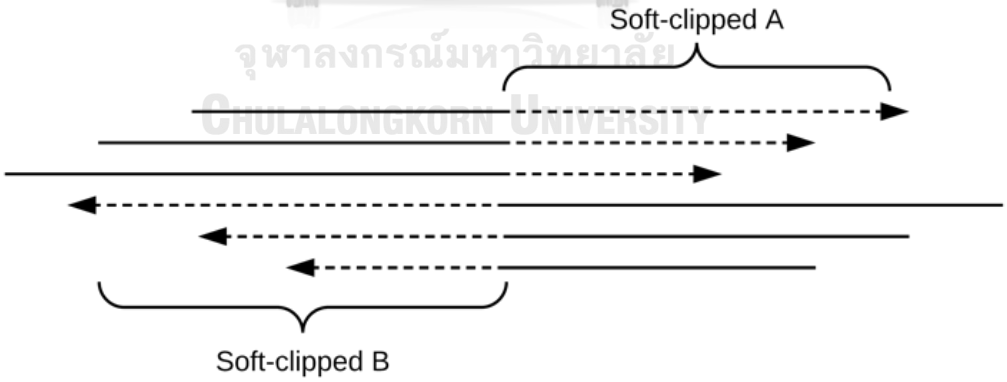
รูปที่ 22 รูปแบบการระบุตำแหน่งของเบรกเอ็นด์ (A) แตกจีโนมอ้างอิงในบริเวณที่สนใจเพื่อเข้าฟังก์ชันแฮชโดยใน bucket ประกอบไปด้วยตำแหน่งของสตริงย่อยบนจีโนม (B) นำลำดับเบสซอฟต์คลิปบริเวณหัวหรือท้ายมาเข้าเข้าฟังก์ชันแฮชเพื่อหาตำแหน่งใน bucket (C) แมพรีดใหม่ที่เหลือด้วยวิธีซื่อตรง

เนื่องจากในแต่ละรีดสามารถแมพได้หลายตำแหน่งบนจีโนมอ้างอิง เพื่อที่จะระบุได้ว่าตำแหน่งไหนคือเบรกเอ็นด์ ในวิทยานิพนธ์นี้ได้เสนอสมการ (1) ต่อไปนี้เพื่อใช้ในการให้คะแนนแต่ละเบรกเอ็นด์

$$Score_{alignment} = N_{match} + (2L_{max}) \quad (1)$$

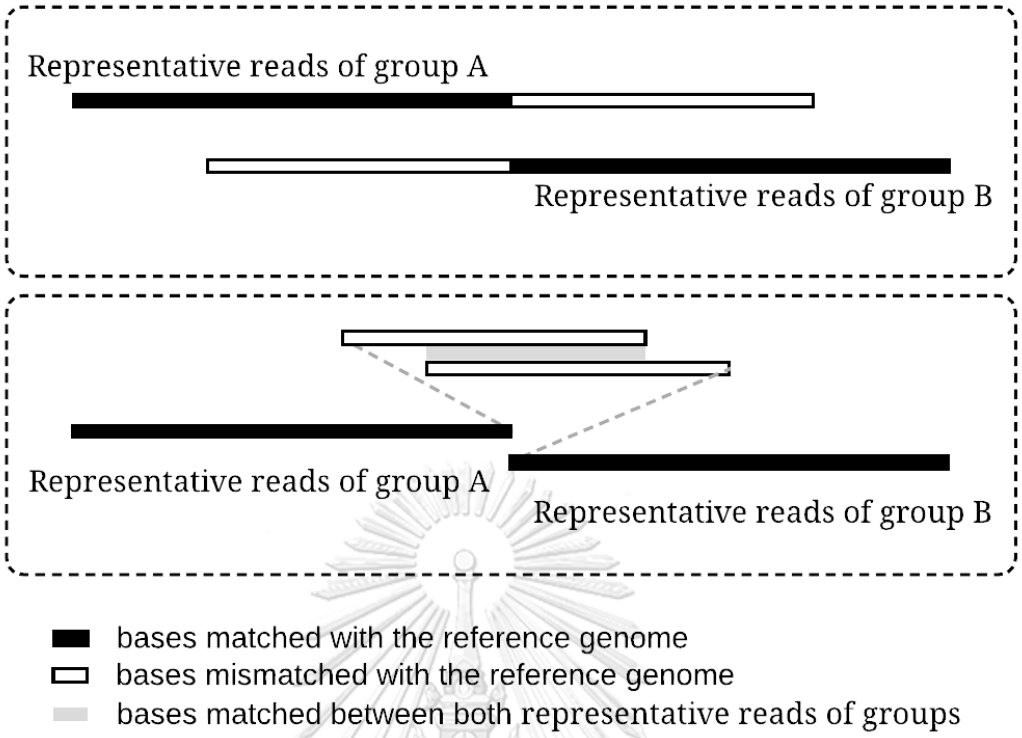
โดยที่ $Score_{alignment}$ คือคะแนนของตำแหน่งเบรกเอ็นด์ N_{match} คือจำนวนรีดที่แมพตรงตำแหน่งนี้และ L_{max} คือลำดับเบสที่ยาวที่สุดที่แมพกับตำแหน่งนี้

สำหรับวิธีการหาของ insertion ต่างจากการวิธีการหา deletion, inversion, tandem duplication และ chromosomal translocation เนื่องจาก insertion ไม่สามารถแมพเข้ากับจีโนมอ้างอิงได้ จึงจำเป็นต้องอาศัยการหาความสัมพันธ์ของลำดับเบสในแต่ละรีดด้วย Smith-Waterman [47] โดยในขั้นตอนแรกจะเป็นการรวมลำดับเบสซอฟต์คลิปของรีดที่มีตำแหน่งของลำดับเบสซอฟต์คลิปตรงกัน ให้อยู่ในกลุ่มเดียวกันก่อน เนื่องจากการใช้ Smith-Waterman ใช้ทรัพยากรในการประมวลผลสูงทำให้ต้องรวมกลุ่มของลำดับเบสซอฟต์คลิปให้อยู่ในกลุ่มเดียวกันก่อน โดยมีเงื่อนไขว่า edit distance ระหว่างซอฟต์คลิป ต้องน้อยกว่า 3 จึงจะถือว่าเป็นลำดับเบสซอฟต์คลิปเดียวกัน วิธีการรวมลำดับเบสซอฟต์คลิป ยกตัวอย่างเช่น กลุ่มของลำดับเบสซอฟต์คลิป A (รูปที่ 23) จะทำการรวมลำดับเบสซอฟต์คลิปจากเส้นสั้นไปยังเส้นยาว โดยใช้ edit distance อย่างไรก็ตามเนื่องจาก edit distance ต้องมีขนาดความยาวลำดับเบสที่เท่ากันจึงต้องมีการตัดส่วนที่เกินออกก่อนเทียบ



รูปที่ 23 รูปแบบการรวมลำดับเบสซอฟต์คลิปด้วย edit distance

หลังจากรวมกลุ่มของลำดับเบสซอฟต์คลิปเสร็จจึงทำการนำลำดับเบสซอฟต์คลิปของทั้ง 2 กลุ่มมาเทียบกันโดยใช้ขั้นตอนวิธี Smith-Waterman (รูปที่ 24) ที่ได้กำหนดคะแนนให้ +1 สำหรับเบสตรงกัน -1 สำหรับเบสไม่ตรงกัน -5 สำหรับเกิดช่องว่าง โดยได้กำหนดให้อย่างน้อยต้องมีคะแนนมากกว่า 15 เบสจึงจะถือว่าตำแหน่งดังกล่าวเป็นเบรกเอ็นด์



รูปที่ 24 การเทียบกลุ่มของรีดด้วยขั้นตอนวิธี Smith-Waterman

3.1.7 การกรองตำแหน่งเบรกเอ็นด์ด้วย Depth block

วิธีการคัดกรองการแปรผันเชิงโครงสร้างและตำแหน่งเบรกเอ็นด์ของการแปรผันเชิงโครงสร้างนั้นๆ ด้วย Depth block จะนับจำนวนการแปรผันเชิงโครงสร้างแต่ละประเภทใน 1 ขอบเขต โดย 1 ขอบเขตครอบคลุมพื้นที่ 3 บล็อก เป็นเงื่อนไขในการคัดกรอง จากตัวอย่างในรูปที่ 25 ถ้าต้องการตรวจสอบว่าการแปรผันเชิงโครงสร้าง deletion ที่มีตำแหน่งเริ่มต้นภายในบล็อก B2 และสิ้นสุดที่ตำแหน่งภายในบล็อก B4 และมีรีดสนับสนุน 5 รีด จะถูกคัดเลือกไว้เป็นคำตอบหรือไม่ จะต้องมีการคัดกรองเบรกเอ็นด์ที่ตำแหน่งเริ่มต้น โดยการนับจำนวนรวมการแปรผันเชิงโครงสร้างแต่ละประเภทของบล็อก B1 B2 และ B3 ซึ่งประกอบด้วย 6 deletion 1 duplication และ 2 insertion และเทียบกับจำนวนรีดสนับสนุนของการแปรผันนี้ (deletion) ด้วยเงื่อนไข (2) ของการเปรียบเทียบต่อไปนี้ต้องเป็นจริง

$$N_{SVi\_supp} > N_{SVj\_count} \quad (2)$$

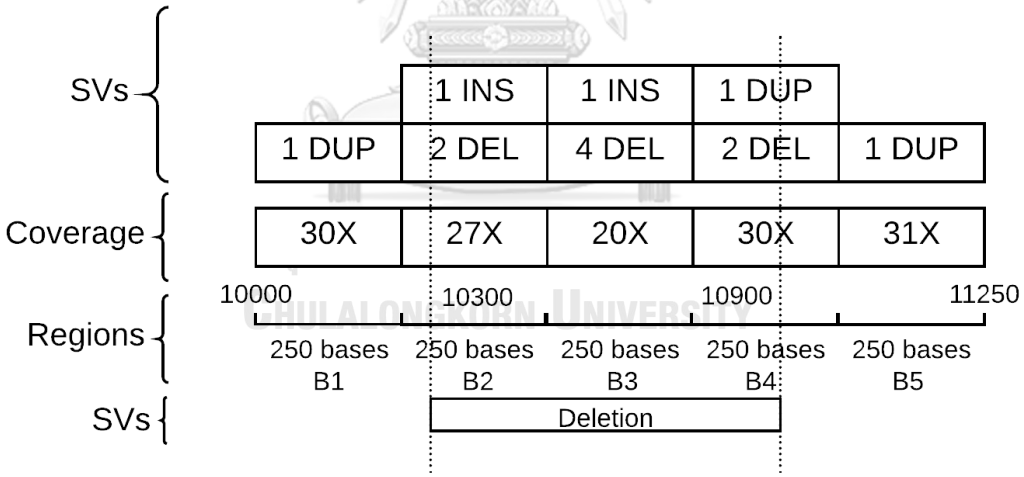
$$N_{SVi\_supp} > (N_{SVi\_count} - N_{SVi\_supp}) \quad (3)$$

โดยที่ N_{SVi_supp} คือจำนวนรีดสนับสนุนของการแปรผันเชิงโครงสร้างประเภท i หรือ SV_i ที่กำลังพิจารณา เช่น 5 รีดของ deletion ในตัวอย่างข้างต้น และ N_{SVj_count} คือ จำนวนการแปรผัน

ประเภท j หรือ SV_j ที่นับได้ใน 1 ขอบเขต เช่นนับจำนวน duplication ได้เท่ากับ 1 เป็นต้น โดย SV_i ต้องไม่เท่ากับ SV_j

ดังนั้นจากเงื่อนไขข้างต้นสรุปว่าตำแหน่งเบรกเอ็นด์ตั้งต้นของการแปรผันเชิงโครงสร้างประเภท deletion ที่พิจารณาผ่านการคัดกรอง ส่วนวิธีการกรองตำแหน่งสิ้นสุดของเบรกเอ็นด์ก็จะมีการตรวจสอบเงื่อนในลักษณะเดียวกันคือรวมการแปรผันเชิงโครงสร้างในแต่ละประเภทของบล็อก B3 B4 และ B5 ประกอบไปด้วย 6 deletion 2 duplication และ 1 insertion และพิจารณาเงื่อนไขตามสมการ (2)-(3) ข้างต้น ซึ่งในตัวอย่างนี้ เบรกเอ็นด์ที่ตำแหน่งสิ้นสุดก็ผ่านการคัดกรองเช่นกัน ดังนั้นการแปรผันเชิงโครงสร้างประเภท deletion ที่พิจารณานี้ผ่านการคัดกรองและจะถูกนำไปเป็นสมาชิกของคำตอบต่อไป

นอกจากนี้ Depth block ยังถูกใช้ในการคัดกรองในกรณีที่ขอบเขตมีจำนวนเฉลี่ยความลึก (coverage หรือ depth) หรือจำนวนรีดที่ตกในบล็อกเดียวกันหรือขอบเขตเดียวกันสูงผิดปกติเมื่อเทียบกับบริเวณอื่นๆ โดยใช้ตรวจสอบจากข้อมูลจำนวนเฉลี่ยความลึกของแต่ละบล็อกที่เบรกเอ็นด์ตกอยู่ ถ้าบล็อกมีจำนวนเฉลี่ยความลึกมากกว่า 3 เท่าของความลึกเฉลี่ยของตัวอย่าง ก็จะถูกกรองออก ยกเว้นว่าเบรกเอ็นด์นั้นเป็นการแปรผันเชิงโครงสร้างประเภท tandem duplication



รูปที่ 25 รูปแบบการกรองตำแหน่งเบรกเอ็นด์ด้วย Depth block

บทที่ 4
การทดลองและผลการทดลอง

วิธีการที่นำเสนอนี้ได้ถูกนำไปทดสอบประสิทธิภาพกับทั้งข้อมูลจำลองและข้อมูลจริง ดังรายละเอียดดังต่อไปนี้

4.1 สภาพแวดล้อมและเครื่องมือที่ใช้ในการทดลอง

เครื่องคอมพิวเตอร์ที่ใช้ทดสอบมีซีพียู Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz จำนวน 20 คอร์ และหน่วยความจำ 93 กิกะไบต์ บนระบบปฏิบัติการ CentOS Linux release 7.7.1908

4.2 เปรียบเทียบประสิทธิภาพกับชุดข้อมูลจำลอง

ในการเตรียมข้อมูลจำลอง ผู้วิจัยใช้โปรแกรม SURVIVOR [17] ในการสร้างจีโนมอ้างอิงขึ้นมาใหม่และมีการเพิ่มการแปรผันเชิงโครงสร้างประเภทต่างๆ ประกอบด้วย 948 deletion 1052 insertion 1000 inversion 1000 tandem duplication แ ล ะ 2000 chromosomal translocation เข้าไปในจีโนม และใช้ WGsim [18] ในการสร้างคู่รีดขนาดสั้น (paired-end reads) ที่มีความยาวรีดขนาด 101 เบส จากจีโนมที่จำลองขึ้นมาข้างต้น โดยสร้างความลึกรีด 3 แบบ คือ 20x 50x และ 100x และใช้โปรแกรม BWA-MEM ในการเทียบรีดกับจีโนมอ้างอิง hg19 [48] สำหรับการตั้งค่าของแต่ละเครื่องมือที่นำมาเปรียบเทียบประสิทธิภาพในการหาการแปรผันเชิงโครงสร้าง จะเป็นค่าเดิมของเครื่องมือนั้นๆ ยกเว้นกรณีของการตั้งค่าจำนวนเธรดที่จะกำหนดที่ 20 เธรด ยกเว้น DELLY ที่ไม่รองรับการทำงานแบบหลายเธรด

การวัดประสิทธิภาพจะใช้ค่าความแม่นยำ (precision) ค่าความครบถ้วน (recall) และค่ามัชฌิมฮาร์มอนิกของค่าความแม่นยำกับค่าความครบถ้วน (F1 score) สามารถหาได้ดังสมการ (4) (5) และ (6) ตามลำดับ

$$precision = \frac{TP}{(TP + FP)} \tag{4}$$

$$recall = \frac{TP}{(TP + FN)} \tag{5}$$

$$F1 = 2 \times \frac{Precision \times Recall}{(Precision + Recall)} \tag{6}$$

โดยที่ TP คือผลบวกจริง FP คือผลบวกลวง FN คือผลลบลวง โดยการพิจารณาผลบวกจริงนั้น ถ้าเป็นกรณีของ deletion, tandem duplication และ inversion ใช้ reciprocal overlap ที่ 90% โดย reciprocal overlap พิจารณาจากส่วนทับซ้อนระหว่าง บริเวณของจีโนมตั้งแต่จุดตั้งต้น

เบรกเอ็นด์ไปจนถึงจุดสิ้นสุดของเบรกเอ็นด์ของทั้งผลลัพธ์หนึ่งๆ จากเครื่องมือและจากตัววัดผล โดยจะต้องมีส่วนทับซ้อนกันอย่างน้อย 90% โดยต้องมีการเทียบกันทั้งไปและกลับ คือนำบริเวณของผลลัพธ์เป็นตัวตั้งและนำบริเวณของตัววัดผลมาเทียบต้องทับซ้อนกันอย่างน้อย 90% และเมื่อนำบริเวณของตัววัดผลเป็นตัวตั้งและนำบริเวณของผลลัพธ์มาเทียบก็ต้องมีบริเวณทับซ้อนกันอย่างน้อย 90% เช่นกัน จึงจะนับว่าผลลัพธ์นั้นๆ ถูกต้อง สำหรับกรณีของ insertion และ chromosomal translocation ถ้าตำแหน่งเบรกเอ็นด์ของผลลัพธ์ตกอยู่ในช่วง ±300 เบส ของตำแหน่งที่ระบุไว้ในตัววัดผล จึงจะนับว่าผลลัพธ์นั้นๆ ถูกต้อง

ผลการเปรียบเทียบที่ได้ (ตารางที่ 3) สำหรับ deletion เครื่องมือ SvABA สามารถทำผลลัพธ์ได้ดีในกรณีของความแม่นยำส่วน DELLY สามารถทำได้ดีในกรณีของความครบถ้วนและ F1 score สำหรับ insertion วิธีการที่นำเสนอ สามารถทำได้ดีทั้งความแม่นยำ ความครบถ้วน และ F1 score ในกรณี 20x และ 50x ส่วน Wham สามารถทำความครบถ้วนได้มากกว่าในกรณี 100x สำหรับ tandem duplication DELLY สามารถทำได้ดีกว่าเครื่องมืออื่นๆ ทั้งความแม่นยำ ความครบถ้วน และ F1 score ในกรณี 20x ในขณะที่ LUMPY ทำความแม่นยำได้สูงกว่า ในกรณี 50x และ 100x สำหรับ inversion โปรแกรม LUMPY และ SvABA ทำได้ดีในส่วนของความแม่นยำ ในขณะที่ DELLY ทำได้ดีทั้งส่วนของความครอบคลุมและ F1 score สำหรับ chromosomal translocation วิธีการที่นำเสนอทำได้ดีในกรณีของความแม่นยำและ LUMPY ทำได้ดีในกรณีของความครบถ้วนและ F1 score

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพของการตรวจหาการแปรผันเชิงโครงสร้างของชุดข้อมูลจำลอง (ค่าความแม่นยำ/ค่าความครบถ้วน/ F1 score)

อัลกอริทึม	20x			50x			100x		
	ค่าความแม่นยำ	ค่าความครบถ้วน	F1	ค่าความแม่นยำ	ค่าความครบถ้วน	F1	ค่าความแม่นยำ	ค่าความครบถ้วน	F1
Deletion									
วิธีการที่นำเสนอ	0.994	0.687	0.812	0.989	0.694	0.816	0.994	0.7	0.822
SvABA	**1.000**	0.557	0.715	**0.998**	0.53	0.692	**1**	0.493	0.660
DELLY	0.988	**0.715**	**0.830**	0.974	**0.716**	**0.826**	0.966	**0.720**	**0.825**
GROM	0.781	0.631	0.698	0.708	0.631	0.667	0.708	0.634	0.669
LUMPY	0.995	0.68	0.808	0.995	0.687	0.813	0.995	0.693	0.817
Wham	0.976	0.632	0.767	0.950	0.655	0.775	0.940	0.66	0.776

Insertion									
วิธีการที่นำเสนอ	**0.998**	**0.510**	**0.675**	**0.981**	**0.478**	**0.643**	**0.987**	0.447	**0.615**
SvABA	0	0	0	0	0	0	0	0	0
DELLY	0	0	0	0	0	0	0	0	0
GROM	0.690	0.449	0.544	0.350	0.471	0.401	0.120	0.402	0.185
LUMPY	0	0	0	0	0	0	0	0	0
Wham	0.452	0.099	0.162	0.398	0.436	0.417	0.313	**0.573**	0.405
Tandem duplication									
วิธีการที่นำเสนอ	0.995	0.645	0.783	0.991	0.677	0.805	0.985	0.676	0.802
SvABA	0	0	0	0	0	0	0	0	0
DELLY	**0.999**	**0.708**	**0.829**	0.993	**0.713**	**0.830**	0.986	**0.716**	**0.830**
GROM	0.763	0.566	0.650	0.750	0.569	0.647	0.731	0.567	0.639
LUMPY	0.998	0.639	0.779	**0.998**	0.656	0.792	**0.998**	0.66	0.795
Wham	0.990	0.593	0.742	0.989	0.625	0.766	0.989	0.631	0.770
Inversion									
วิธีการที่นำเสนอ	0.999	0.789	0.882	0.993	0.796	0.883	0.991	0.797	0.884
SvABA	**1**	0.341	0.509	**1**	0.342	0.510	**1**	0.341	0.509
DELLY	0.998	**0.817**	**0.898**	0.995	**0.819**	**0.899**	0.99	**0.823**	**0.899**
GROM	0.946	0.732	0.825	0.945	0.737	0.828	0.947	0.735	0.828
LUMPY	**1**	0.777	0.875	**1**	0.781	0.877	**1**	0.782	0.878
Wham	0.964	0.755	0.847	0.973	0.787	0.870	0.963	0.773	0.857
Chromosomal translocation									
วิธีการที่นำเสนอ	**0.899**	0.626	0.738	**0.94**	0.587	0.723	**0.954**	0.554	0.701
SvABA	0.325	0.093	0.145	0.327	0.088	0.139	0.27	0.062	0.101

DELLY	0.784	0.624	0.695	0.666	0.571	0.615	0.576	0.517	0.545
GROM	0.866	0.584	0.698	0.865	0.596	0.705	0.865	0.596	0.705
LUMPY	0.88	**0.640**	**0.741**	0.879	**0.653**	**0.749**	0.869	**0.657**	**0.748**
Wham	0	0	0	0	0	0	0	0	0

4.3 เปรียบเทียบประสิทธิภาพกับชุดข้อมูลจริง

ชุดข้อมูลจริงที่ใช้ในการเปรียบเทียบประสิทธิภาพประกอบด้วย ชุดข้อมูล NA12878 และชุดข้อมูล HG00514 โดยดาวน์โหลดจาก DDBJ [19] และ SRA [20] ชุดข้อมูล NA12878 ประกอบด้วย 10 ตัวอย่างจำเพาะ ERR174336, ERR174337, ERR174338, ERR174339, ERR174340, SRR1910366, ERR091571, ERR091572, ERR091573 และ ERR091574 โดยที่แต่ละตัวอย่างเป็นจีโนมของมนุษย์รหัส NA12878 เดียวกัน ส่วนชุดข้อมูล HG00514 ประกอบไปด้วย 2 ตัวอย่างจำเพาะ ERR894729 และ ERR903030 โดยข้อมูลทั้งหมดอยู่ในรูปแบบไฟล์ FASTQ โดยในการทดสอบผู้วิจัยใช้ BWA-MEM ในการเทียบรีดของตัวอย่างเหล่านี้กับจีโนมอ้างอิง hs37d5 [49] สำหรับชุดข้อมูลของ NA12878 ผู้วิจัยใช้ข้อมูลจาก DGV [21] (วันที่เผยแพร่ 2020-02-25 เป็นข้อมูล GRCh 37) ในการวัดผลลัพธ์และทำการดึงเฉพาะรายการที่มี NA12878 และชุดข้อมูลของ HG00514 ใช้ข้อมูลจาก [22] ในการวัดผลลัพธ์

ข้อมูล DGV ของจีโนม NA12878 ประกอบไปด้วย 8282 deletion 2409 tandem duplication และ 12640 insertion ส่วนข้อมูล [22] ของจีโนม HG00514 ประกอบไปด้วย 37215 deletion 42111 insertion 462 inversion และ 2115 tandem duplication

4.3.1 ผลการเปรียบเทียบประสิทธิภาพของการตรวจหาการแปรผันเชิงโครงสร้างของชุดข้อมูล NA12878

ตารางที่ 4 แสดงผลการเปรียบเทียบประสิทธิภาพของการตรวจหาการแปรผันเชิงโครงสร้างประเภท deletion, insertion และ tandem duplication ของชุดข้อมูล NA12878 โดยการแปรผันเชิงโครงสร้างประเภท deletion โปรแกรม SvABA สามารถทำความแม่นยำได้ดีกว่าในขณะที่มีความครบถ้วนน้อยกว่าเครื่องมืออื่น ส่วน DELLY สามารถทำได้ดีกว่าเครื่องมืออื่นในส่วนของความครบถ้วนแต่มีความแม่นยำน้อยกว่าเครื่องมืออื่นมาก สำหรับผลของวิธีการที่นำเสนอมีความแม่นยำน้อยกว่า SvABA เล็กน้อยแต่มีค่าความครบถ้วนสูงกว่ามาก ในขณะที่มีความครบถ้วนน้อยกว่า DELLY เล็กน้อยแต่มีค่าความแม่นยำสูงกว่ามาก ทำให้วิธีการที่นำเสนอมีค่าคะแนน F1 สูงกว่าทั้งสองโปรแกรมและสูงสุดในกลุ่มโปรแกรมที่นำมาเทียบทั้งหมด

สำหรับการแปรผันเชิงโครงสร้างประเภท insertion วิธีการที่นำเสนอสามารถทำความครอบคลุมและ F1 score ได้มากกว่าเครื่องมืออื่นๆ ในขณะที่ GROM สามารถทำได้ดีในส่วนของความแม่นยำแต่มีความครบถ้วนน้อยกว่าเครื่องมืออื่นๆ มาก ส่วน SvABA, DELLY และ LUMPY ไม่มีการส่งออกผลลัพธ์สำหรับการแปรผันเชิงโครงสร้างประเภทนี้

สำหรับการแปรผันเชิงโครงสร้างประเภท tandem duplication วิธีการที่นำเสนอสามารถทำความแม่นยำได้ดีในตัวอย่างที่มีความลึกรีดต่ำและความครบถ้วนอยู่ใกล้เคียงกับเครื่องมือส่วนใหญ่ ส่วน GROM สามารถหาความครอบคลุมและ F1 score ได้ดีกว่าเครื่องมืออื่นๆ แต่มีความแม่นยำต่ำกว่าวิธีการที่นำเสนอ

ตารางที่ 4 ผลการเปรียบเทียบประสิทธิภาพของการตรวจหาการแปรผันเชิงโครงสร้างของชุดข้อมูล NA12878 (ค่าความแม่นยำ/ค่าความครบถ้วน/ F1 score)

อัลกอริทึม	ERR174336	ERR174337	ERR174338	ERR174339	ERR174340	SRR1910366	ERR091571	ERR091572	ERR091573	ERR091574
Deletion										
วิธีการที่นำเสนอ	0.478/ 0.161/ **0.241**	0.473/ 0.160/ **0.240**	0.479/ 0.160/ **0.240**	0.472/ 0.154/ **0.232**	0.459/ 0.123/ **0.194**	0.496/ 0.235/ **0.319**	0.471/ 0.164/ **0.243**	0.476/ 0.166/ **0.246**	0.473/ 0.163/ **0.242**	0.479/ 0.170/ **0.251**
SvABA	0.495/ 0.132/ 0.208	0.504/ 0.135/ 0.213	0.510/ 0.134/ 0.213	0.499/ 0.128/ 0.204	0.470/ 0.102/ 0.167	0.594/ 0.100/ 0.171	0.507/ 0.140/ 0.219	0.503/ 0.136/ 0.215	0.512/ 0.140/ 0.220	0.508/ 0.140/ 0.219
DELLY	0.192/ 0.165/ 0.177	0.188/ 0.165/ 0.176	0.189/ 0.166/ 0.177	0.183/ 0.161/ 0.171	0.203/ 0.148/ 0.171	0.078/ 0.241/ 0.118	0.185/ 0.166/ 0.175	0.183/ 0.166/ 0.174	0.186/ 0.168/ 0.176	0.184/ 0.171/ 0.177
GROM	0.399/ 0.138/ 0.205	0.388/ 0.135/ 0.201	0.392/ 0.136/ 0.202	0.390/ 0.133/ 0.199	0.382/ 0.124/ 0.188	0.390/ 0.186/ 0.252	0.399/ 0.138/ 0.205	0.392/ 0.136/ 0.202	0.382/ 0.134/ 0.198	0.393/ 0.140/ 0.206
LUMPY	0.438/ 0.113/ 0.180	0.433/ 0.113/ 0.179	0.440/ 0.115/ 0.182	0.430/ 0.109/ 0.174	0.459/ 0.089/ 0.149	0.272/ 0.207/ 0.235	0.432/ 0.163/ 0.237	0.431/ 0.164/ 0.237	0.426/ 0.158/ 0.231	0.430/ 0.167/ 0.241
Wham	0.431/ 0.095/ 0.156	0.432/ 0.097/ 0.159	0.437/ 0.096/ 0.157	0.416/ 0.089/ 0.146	0.410/ 0.068/ 0.117	0.310/ 0.171/ 0.220	0.411/ 0.094/ 0.152	0.424/ 0.096/ 0.157	0.424/ 0.098/ 0.160	0.415/ 0.101/ 0.162
Insertion										
วิธีการที่นำเสนอ	0.584/ 0.046/ **0.085**	0.572/ 0.046/ **0.085**	0.549/ 0.045/ **0.083**	0.597/ 0.045/ **0.083**	0.615/ 0.035/ **0.066**	0.524/ 0.136/ **0.215**	0.547/ 0.050/ **0.091**	0.551/ 0.051/ **0.094**	0.562/ 0.053/ **0.096**	0.526/ 0.052/ **0.095**
SvABA	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0
DELLY	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0
GROM	0.868/ 0.005/	0.933/ 0.007/	0.907/ 0.005/	0.947/ 0.006/	0.816/ 0.002/	0.737/ 0.047/	0.879/ 0.007/	0.904/ 0.007/	0.896/ 0.007/	0.918/ 0.008/

	0.010	0.013	0.011	0.011	0.005	0.088	0.014	0.015	0.014	0.016
LUMPY	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0
Wham	0.197/ 0.002/ 0.004	0.259/ 0.003/ 0.007	0.163/ 0.002/ 0.004	0.256/ 0.003/ 0.006	0.147/ 0.001/ 0.002	0.076/ 0.017/ 0.028	0.281/ 0.004/ 0.007	0.262/ 0.004/ 0.007	0.302/ 0.004/ 0.008	0.291/ 0.004/ 0.008
<u>Tandem duplication</u>										
วิธีการที่นำเสนอ	0.213/ 0.022/ 0.039	0.208/ 0.023/ 0.042	0.21/ 0.02/ 0.037	0.255/ 0.023/ 0.043	0.251/ 0.020/ 0.038	0.143/ 0.032/ 0.052	0.220/ 0.023/ 0.042	0.194/ 0.022/ 0.040	0.206/ 0.021/ 0.038	0.196/ 0.022/ 0.040
SvABA	0/ 0/ 0	0/ 0/ 0!	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0	0/ 0/ 0
DELLY	0.013/ 0.026/ 0.018	0.016/ 0.030/ 0.021	0.013/ 0.025/ 0.017	0.014/ 0.026/ 0.018	0.015/ 0.023/ 0.018	0.003/ 0.028/ 0.005	0.013/ 0.027/ 0.017	0.015/ 0.031/ 0.020	0.014/ 0.028/ 0.019	0.014/ 0.029/ 0.018
GROM	0.130/ 0.042/ **0.063**	0.135/ 0.044/ **0.067**	0.142/ 0.045/ **0.069**	0.128/ 0.040/ **0.061**	0.133/ 0.036/ **0.056**	0.151/ 0.039/ **0.063**	0.136/ 0.044/ **0.067**	0.141/ 0.047/ **0.070**	0.134/ 0.045/ **0.067**	0.126/ 0.043/ **0.064**
LUMPY	0.071/ 0.022/ 0.033	0.065/ 0.021/ 0.031	0.064/ 0.021/ 0.031	0.068/ 0.021/ 0.032	0.067/ 0.016/ 0.026	0.014/ 0.017/ 0.016	0.067/ 0.023/ 0.034	0.066/ 0.022/ 0.033	0.072/ 0.024/ 0.036	0.063/ 0.022/ 0.033
Wham	0.110/ 0.021/ 0.035	0.110/ 0.021/ 0.035	0.113/ 0.022/ 0.036	0.119/ 0.022/ 0.038	0.129/ 0.022/ 0.038	0.106/ 0.014/ 0.024	0.101/ 0.020/ 0.033	0.121/ 0.023/ 0.039	0.124/ 0.024/ 0.041	0.103/ 0.022/ 0.036

4.3.2 ผลการเปรียบเทียบประสิทธิภาพของการตรวจหาการแปรผันเชิงโครงสร้างของชุดข้อมูล HG00514

ตารางที่ 5 แสดงผลการเปรียบเทียบประสิทธิภาพของการตรวจหาการแปรผันเชิงโครงสร้างประเภท deletion, insertion, tandem duplication และ inversion ของชุดข้อมูล HG00514 โดยการแปรผันเชิงโครงสร้างประเภท deletion วิธีการที่นำเสนอสามารถทำความครอบคลุมและ F1 score ได้ดีกว่าเครื่องมืออื่น ในขณะที่ SvABA สามารถทำได้ในส่วนของความแม่นยำเมื่อเทียบกับเครื่องมืออื่น อย่างไรก็ตาม เมื่อเทียบวิธีการที่นำเสนอกับ SvABA พบว่าวิธีการที่นำเสนอได้ความแม่นยำน้อยกว่า แต่ได้ความครอบคลุมและ F1 score มากกว่า

สำหรับการแปรผันเชิงโครงสร้างประเภท insertion วิธีการที่นำเสนอสามารถทำความครบครอบคลุมและ F1 score ได้มากกว่าเครื่องมืออื่น ส่วน DELLY และ GROM สามารถทำได้ดีในส่วนของความแม่นยำแต่มีความครบถ้วนที่ต่ำมากๆ ส่วน LUMPY มีความแม่นยำใกล้เคียงกันกับเครื่องมือที่นำเสนอแต่มีความครอบคลุมน้อยกว่ามาก

สำหรับการแปรผันเชิงโครงสร้างประเภท tandem duplication วิธีการที่นำเสนอสามารถทำความแม่นยำและ F1 score ได้ดีกว่าในขณะที่มีความครอบคลุมใกล้เคียงกับเครื่องมืออื่นๆ ส่วน DELLY สามารถหาความครบถ้วนได้ดีแต่มีความแม่นยำที่ต่ำกว่าเครื่องมืออื่นๆ อยู่มาก

สำหรับ inversion วิธีการที่นำเสนอสามารถทำความแม่นยำได้ดีกว่าเครื่องมืออื่น และได้ F1 score ในตัวอย่างจำเพาะ ERR894729 ดีกว่าเครื่องมืออื่นๆ ส่วน ERR903030 ได้ F1 score น้อยกว่า GROM เล็กน้อย ส่วน DELLY สามารถหาความครอบคลุมได้ดีแต่มีความแม่นยำที่ต่ำกว่าเครื่องมืออื่นๆ มาก

ตารางที่ 5 ผลผลการเปรียบเทียบประสิทธิภาพของการตรวจหาการแปรผันเชิงโครงสร้างของชุดข้อมูล HG00514 (ค่าความแม่นยำ/ค่าความครบถ้วน/ F1 score)

อัลกอริทึม	ERR894729	ERR903030
Deletion		
วิธีการที่นำเสนอ	0.858/0.124/**0.217**	0.865/0.125/**0.219**
SvABA	0.921/0.093/0.169	0.919/0.096/0.174
DELLY	0.229/0.074/0.112	0.236/0.076/0.114
GROM	0.852/0.087/0.157	0.847/0.087/0.157
LUMPY	0.740/0.112/0.194	0.736/0.113/0.196
Wham	0.722/0.049/0.091	0.728/0.050/0.094
Insertion		
วิธีการที่นำเสนอ	0.424/0.022/**0.041**	0.449/0.024/**0.046**
SvABA	0/0/0	0/0/0
DELLY	1.000/0.001/0.001	0.571/0.001/0.001
GROM	0.559/0.002/0.004	0.604/0.002/0.005
LUMPY	0/0/0	0/0/0
Wham	0.455/0.013/0.024	0.444/0.012/0.023
Tandem duplication		
วิธีการที่นำเสนอ	0.444/0.059/**0.104**	0.441/0.060/**0.106**
SvABA	0/0/0	0/0/0

DELLY	0.016/0.073/0.026	0.015/0.071/0.025
GROM	0.170/0.054/0.082	0.159/0.052/0.079
LUMPY	0.076/0.059/0.067	0.066/0.051/0.057
Wham	0.124/0.059/0.080	0.120/0.059/0.079
<u>Inversion</u>		
วิธีการที่นำเสนอ	0.280/0.065/**0.105**	0.255/0.054/0.089
SvABA	0.154/0.065/0.091	0.124/0.052/0.073
DELLY	0.005/0.162/0.010	0.005/0.173/0.010
GROM	0.138/0.065/0.088	0.147/0.071/**0.096**
LUMPY	0.091/0.043/0.059	0.099/0.045/0.062
Wham	0.031/0.078/0.045	0.120/0.056/0.077

4.4 เปรียบเทียบทรัพยากรที่ใช้ในทดสอบข้อมูล

ในการวัดประสิทธิภาพในเชิงของการใช้และการจัดการทรัพยากรการคำนวณ ผู้วิจัยทำการทดสอบประสิทธิภาพของทุกเครื่องมือบนตัวอย่างจำเพาะ ERR174336 เนื่องจากตัวอย่างจำเพาะเกือบทุกตัวอย่างที่ขึ้นต้นด้วย ERR มีความยาวคู่รีดและขนาดความลึกของตัวอย่างใกล้เคียงกันจึงเลือกใช้ ERR174336 เป็นตัวแทนของข้อมูลจำเพาะโดยใช้เครื่องทดสอบตามที่ระบุไว้ในหัวข้อที่ 4.1 และตั้งค่าพารามิเตอร์ให้ทำงาน 20 เธรด ยกเว้น LUMPY และ DELLY ที่ไม่รองรับเธรด จากผลการวัดประสิทธิภาพในเชิงของการใช้และการจัดการทรัพยากรการคำนวณ ในตารางที่ 6 พบว่า Wham ใช้เวลาทั้งหมดน้อยที่สุด ในขณะที่ LUMPY ใช้จำนวนชั่วโมงซีพียูและหน่วยความจำน้อยที่สุด สำหรับเครื่องมือที่นำเสนอใช้เวลาทั้งหมด 0.117 ชั่วโมง และหน่วยความจำสูงสุดที่ใช้ 1.277 กิกะไบต์

ตารางที่ 6 ผลการเปรียบเทียบการใช้ทรัพยากรการคำนวณของแต่ละเครื่องมือ โดยทดสอบกับข้อมูล ERR174336

อัลกอริทึม	เวลาที่ใช้ทั้งหมด (ชั่วโมง)	ชั่วโมงซีพียู (cpu hours)	หน่วยความจำสูงสุดที่ใช้ (กิกะไบต์)
วิธีการที่นำเสนอ	0.117	0.624	1.277
SvABA	0.431	8.236	8.305
DELLY	0.589	0.583	2.769

GROM	0.154	2.016	86.204
LUMPY	0.686	**0.279**	**0.654**
Wham	**0.055**	0.790	2.259

4.5 อภิปรายผล

สำหรับชุดข้อมูลจริงที่นำมาทดสอบทั้ง 2 ชุด ในภาพรวมวิธีการที่นำเสนอสามารถตรวจหาการแปรผันเชิงโครงสร้างได้ดี โดยเฉพาะชุดข้อมูล HG00514 ที่สามารถทำ F1-score ได้ดีกว่าเครื่องมืออื่นๆ เกือบทั้งหมดยกเว้น inversion ของตัวอย่างจำเพาะ ERR903030 ส่วนชุดข้อมูล NA12878 วิธีการที่นำเสนอสามารถทำ F1-score ในกรณีของ deletion และ insertion ได้ดีกว่าเครื่องมืออื่นๆ ยกเว้น tandem duplication ที่ GROM สามารถทำได้ดีกว่า

นอกจากนี้ ผลลัพธ์ทั้งหมดแสดงให้เห็นจุดเด่นและจุดด้อยของแต่ละเครื่องมือ ตัวอย่างเช่น DELLY ผลลัพธ์ที่ได้ส่วนใหญ่จะมีความครบถ้วนที่สูงกว่าเครื่องมืออื่นๆ แต่ความแม่นยำมักจะได้ต่ำกว่าเครื่องมืออื่นๆ สำหรับโปรแกรม SvABA นั้นจะเด่นในการตรวจหา deletion สามารถให้ความแม่นยำที่สูงกว่าเครื่องมืออื่นๆ แต่ได้ความครอบคลุมค่อนข้างต่ำ ส่วน GROM จากผลการทดสอบพบว่าหา tandem duplication ได้ดีในชุดข้อมูล NA12878 แต่ในทางตรงกันข้ามเมื่อทดสอบกับชุดข้อมูล HG00514 กับทำได้ไม่ดีนัก

ภาพรวมของวิธีการที่นำเสนอสามารถแสดงผลลัพธ์ได้ดี เมื่อเปรียบเทียบกับเครื่องมืออื่นๆ ไม่ว่าจะเป็น NA12878 ที่ได้ F1 score สูงกว่าเครื่องมืออื่นๆ ในกรณีของ deletion และ insertion หรือแม้กระทั่ง HG00514 ที่ได้ F1 score สูงกว่าเครื่องมืออื่นๆ ในเกือบทุกการทดสอบ แม้จะมีบางมิติที่ได้ผลลัพธ์ที่ต่ำกว่าแต่ส่วนใหญ่แล้วผลลัพธ์ที่ได้มักจะมากกว่าหรือใกล้เคียงกัน เมื่อเทียบกับเครื่องมืออื่นๆ ส่วนบางการทดสอบนั้นพบว่าอาจจะมีบางเครื่องมือที่ได้ผลลัพธ์ดีมากในมิติหนึ่งแต่อีกมิติหนึ่งกลับตรงกันข้าม ซึ่งทำให้วิธีการที่นำเสนอไม่สามารถทำให้ดีได้ทุกมิติ อย่างไรก็ตามผู้วิจัยได้เพิ่มในส่วนของตารางข้อมูลนับจำนวนจริง (ภาคผนวก ก) เพื่อให้เห็นข้อมูลที่ละเอียดขึ้น

นอกจากนี้ผลของแผนภาพเวนน์ (ภาคผนวก ข) แสดงจำนวนการแปรผันเชิงโครงสร้างแต่ละประเภท ที่แต่ละเครื่องมือสามารถตรวจหาได้ตรงกับข้อมูลผลลัพธ์จาก DGV หรือ [22] ซึ่งได้แสดงให้เห็นถึงความจำเป็นที่จะต้องทดสอบกับหลายๆเครื่องมือ เพื่อช่วยให้การตรวจหาการแปรผันเชิงโครงสร้างให้มีความครอบคลุมมากขึ้น

สำหรับข้อจำกัดในการตรวจหาการแปรผันเชิงโครงสร้างของการอ่านรหัสพันธุกรรมแบบสายสั้นคู่ จากการสังเกตของผู้วิจัยพบว่าเกิดจากการที่เครื่องมือที่ทำหน้าที่ในการเทียบรีดสายสั้นกับจีโนมอ้างอิงมีการระบุตำแหน่งที่แท้จริงคลาดเคลื่อน หรือเกิดความคลุมเครือในการเลือกตำแหน่งในจีโนมที่

รีดหนึ่งควรไปแมพ รวมไปถึงบางกรณีที่รีดไม่สามารถเทียบกับจีโนมอ้างอิงได้ เนื่องจากมีความแตกต่างกันของลำดับเบสในจีโนมข้อมูลตัวอย่างกับจีโนมอ้างอิงมากกว่าเงื่อนไขที่ใช้กำหนดในการเทียบรีด นอกจากนี้ยังพบว่าในบางตำแหน่งเบรกเอ็นด์มีจำนวนรีดสนับสนุนที่น้อยมาก ทำให้ผลลัพธ์นั้นถูกกรองออก ดังนั้นสำหรับการปรับปรุงประสิทธิภาพในการตรวจหา ควรปรับปรุงตั้งแต่การเทียบรีดสายสั้นกับจีโนมอ้างอิง หรือจำเป็นต้องใช้การอ่านรหัสพันธุกรรมแบบสายยาวเพิ่มสำหรับตัวอย่างจำเพาะนั้นๆ เพื่อช่วยตรวจหาและลดตำแหน่งที่รหัสพันธุกรรมแบบสายสั้นคู่มีการแมพกับจีโนมอ้างอิงที่ผิดพลาด

บทที่ 5
สรุปผลการวิจัย

5.1 สรุปผลการวิจัย

งานวิจัยนี้ได้แสดงถึงวิธีการตรวจหาการแปรผันเชิงโครงสร้างในแต่ละโครโมโซม ได้แก่ deletion, insertion, tandem duplication, inversion และ chromosomal translocation ซึ่งผลการทดสอบพบว่าวิธีการที่นำเสนอสามารถตรวจหาการแปรผันเชิงโครงสร้างได้ดี อาจจะมีบางผลการทดสอบที่มีบางเครื่องมือให้ผลที่ดีกว่าในบางมิติ แต่ภาพรวมนั้นวิธีการที่นำเสนอสามารถทำงานได้อย่างค่อนข้างมีประสิทธิภาพ

5.2 แนวทางวิจัยในอนาคต

1) ปรับปรุงให้รองรับในกรณีไฟล์ตัวอย่างที่มีความลึกของรีดสูง (high read coverage หรือ read depth)

2) เพิ่มเติมในการรองรับการตรวจหาการแปรผันเชิงโครงสร้างที่ซับซ้อน (complex structural variations)

3) มุ่งเน้นในการตรวจหาการแปรผันเชิงโครงสร้างโดยนำใช้เทคโนโลยีการอ่านลำดับนิวคลีโอไทด์แบบสายยาว (Long-Read Sequencing) เข้ามาประกอบ

ภาคผนวก

ภาคผนวก ก.

ผลข้อมูลนับจำนวนจริง

ตารางที่ ก.1 เปรียบเทียบชุดข้อมูล NA12878 (ผลบวกจริง/ผลบวกลวง)

อัลกอริทึม	ERR174336	ERR174337	ERR174338	ERR174339	ERR174340	SRR1910366	ERR091571	ERR091572	ERR091573	ERR091574
Deletion										
วิธีการที่นำเสนอ	1334/1456	1329/1480	1328/1447	1274/1428	1021/1203	1948/1977	1357/1523	1375/1514	1350/1503	1406/1530
SvABA	1091/1111	1119/1103	1112/1068	1062/1067	841/947	829/567	1158/1124	1130/1117	1163/1110	1158/1121
DELLY	1363/5743	1369/5910	1375/5903	1333/5970	1226/4814	1997/23618	1377/6075	1378/6169	1392/6111	1418/6300
GROM	1140/1714	1120/1764	1124/1743	1104/1725	1031/1668	1539/2403	1142/1720	1130/1751	1110/1793	1159/1790
LUMPY	938/1202	936/1225	950/1209	903/1198	739/871	1711/4587	1349/1777	1355/1786	1311/1770	1383/1836
Wham	790/1042	807/1060	794/1025	733/1030	563/809	1413/3147	775/1109	798/1085	815/1107	836/1177
Insertion										
วิธีการที่นำเสนอ	582/415	580/434	567/465	563/380	441/276	1713/1557	631/523	647/528	665/519	659/593
SvABA	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
DELLY	0/0	0/0	0/0	0/0	0/0	273/96	0/0	0/0	0/0	0/0
GROM	66/10	84/6	68/7	71/4	31/7	595/212	87/12	94/10	86/10	101/9
LUMPY	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Wham	27/110	43/123	25/128	41/119	11/64	215/2597	45/115	45/127	52/120	50/122
Tandem duplication										
วิธีการที่นำเสนอ	55/192	56/213	49/184	56/164	49/146	76/454	56/198	53/220	51/197	54/221
SvABA	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
DELLY	62/4582	73/4593	61/4594	63/4489	55/3604	67/23229	64/4857	74/4872	68/4872	70/5089
GROM	100/671	107/684	109/659	97/662	86/560	95/536	106/672	113/687	108/700	104/719
LUMPY	52/679	50/721	50/728	51/695	39/543	41/2805	55/772	54/765	57/733	53/790
Wham	50/404	52/407	54/399	54/366	33/278	111/2349	48/429	56/405	59/415	53/463

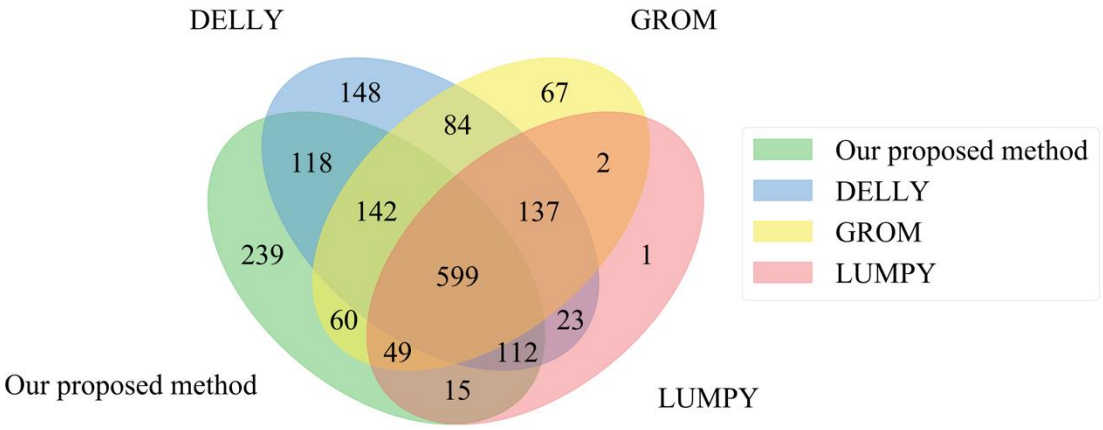
ตารางที่ ก.2 เปรียบเทียบชุดข้อมูล HG00514 (ผลบวกจริง/ผลบวกปลอม)

อัลกอริทึม	ERR894729	ERR903030
Deletion		
วิธีการที่นำเสนอ	4628/763	4668/730
SvABA	3455/295	3579/316
DELLY	2745/9240	2812/9103
GROM	3223/559	3228/584
LUMPY	4150/1455	4202/1510
Wham	1814/697	1877/700
Insertion		
วิธีการที่นำเสนอ	911/1239	1013/1245
SvABA	0/0	0/0
DELLY	3/0	4/3
GROM	90/71	102/67
LUMPY	0/0	0/0
Wham	527/632	505/633
Tandem duplication		
วิธีการที่นำเสนอ	124/155	127/161
SvABA	0/0	0/0
DELLY	155/9618	151/9611
GROM	115/562	111/586
LUMPY	125/1512	107/1518
Wham	124/878	125/921
Inversion		
วิธีการที่นำเสนอ	30/77	25/73
SvABA	30/165	24/170
DELLY	75/14836	80/14808
GROM	30/187	33/191
LUMPY	20/199	21/191
Wham	36/1114	26/1126

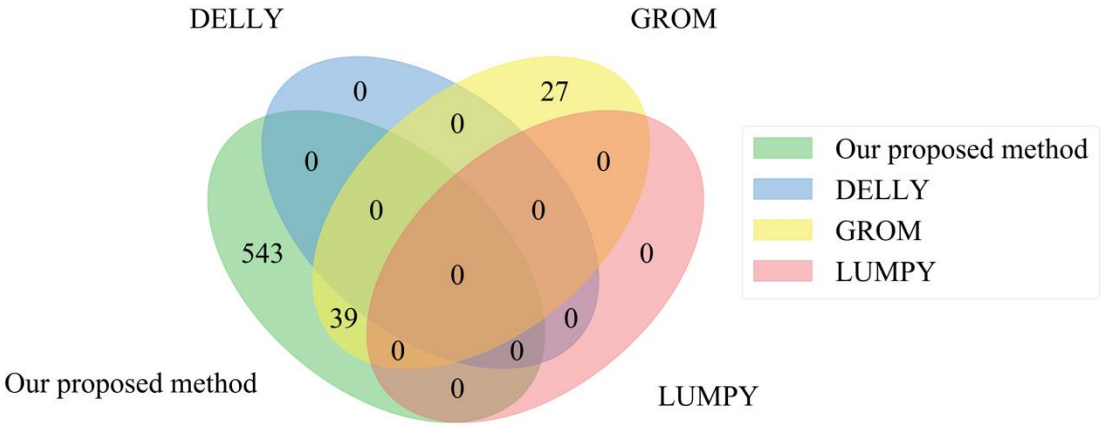
ภาคผนวก ข.

แผนภาพเวนน์

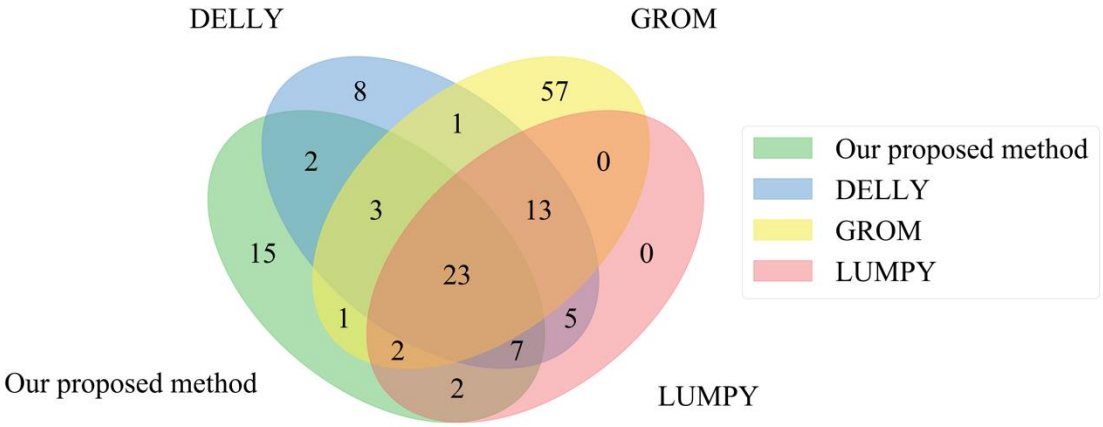
แผนภาพเวนน์ต่อไปนี้แสดงจำนวนการแปรผันเชิงโครงสร้างของแต่ละประเภท ที่แต่ละเครื่องมือสามารถตรวจหาได้ตรงกับข้อมูลผลลัพธ์จาก DGV หรือ [22]



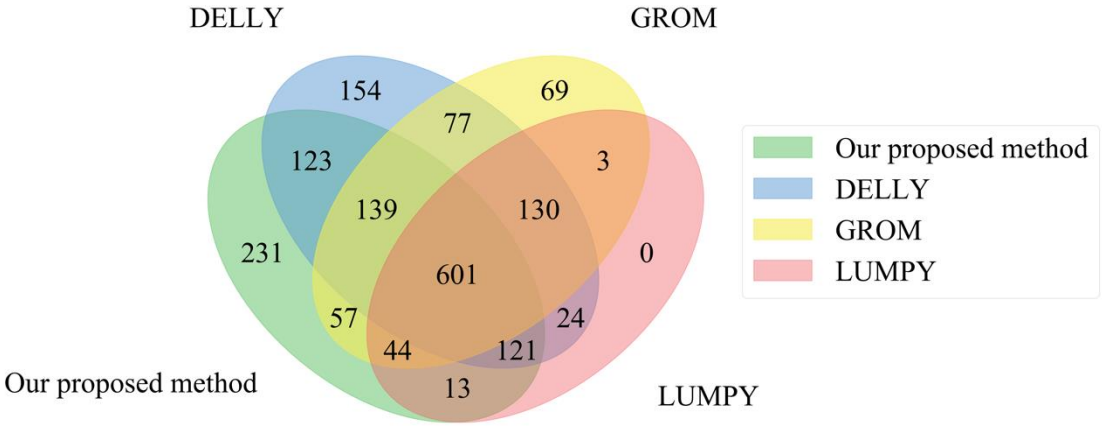
รูปที่ ข.1 แผนภาพเวนน์ของ ERR174336 ประเภท deletion



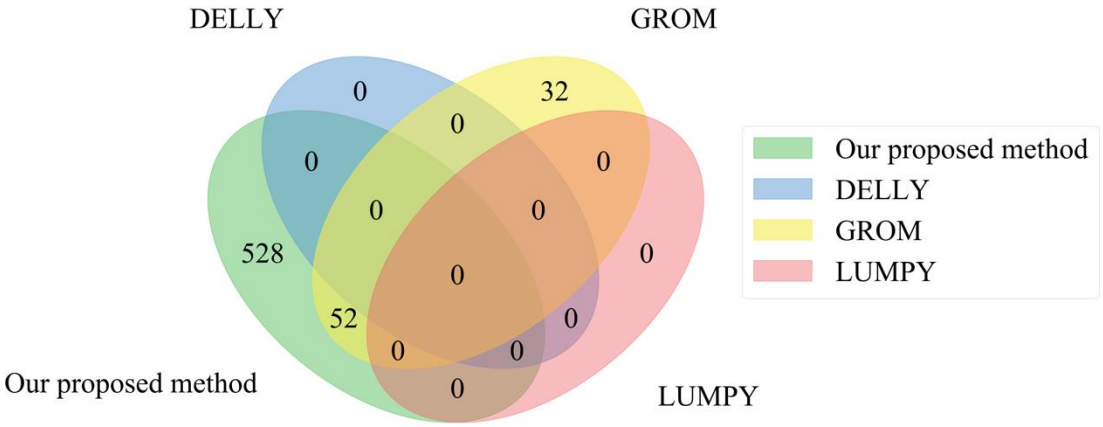
รูปที่ ข.2 แผนภาพเวนน์ของ ERR174336 ประเภท insertion



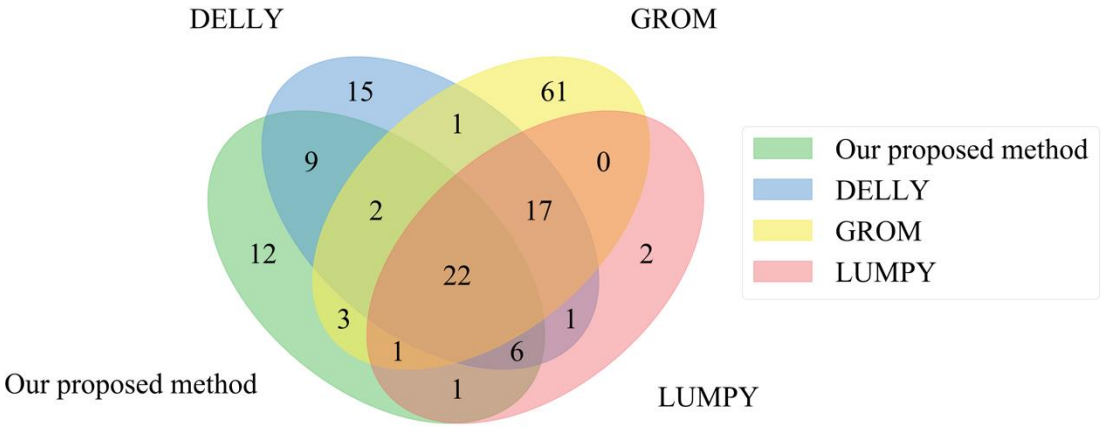
รูปที่ ข.3 แผนภาพเวนน์ของ ERR174336 ประเภท tandem duplication



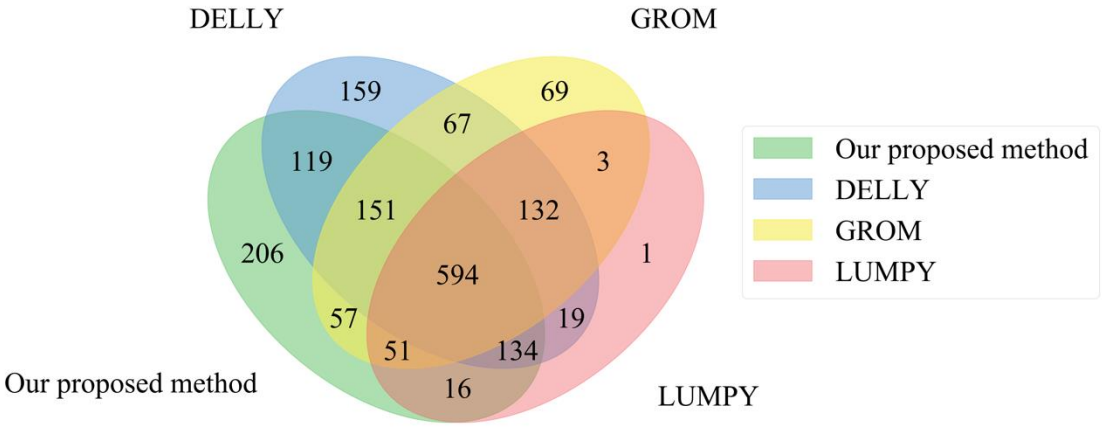
รูปที่ ข.4 แผนภาพเวนน์ของ ERR174337 ประเภท deletion



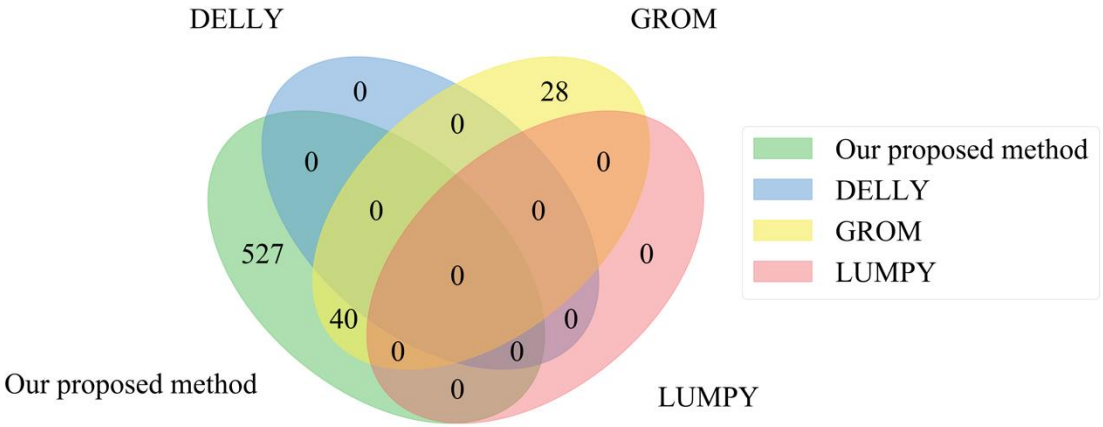
รูปที่ ข.5 แผนภาพเวนน์ของ ERR174337 ประเภท insertion



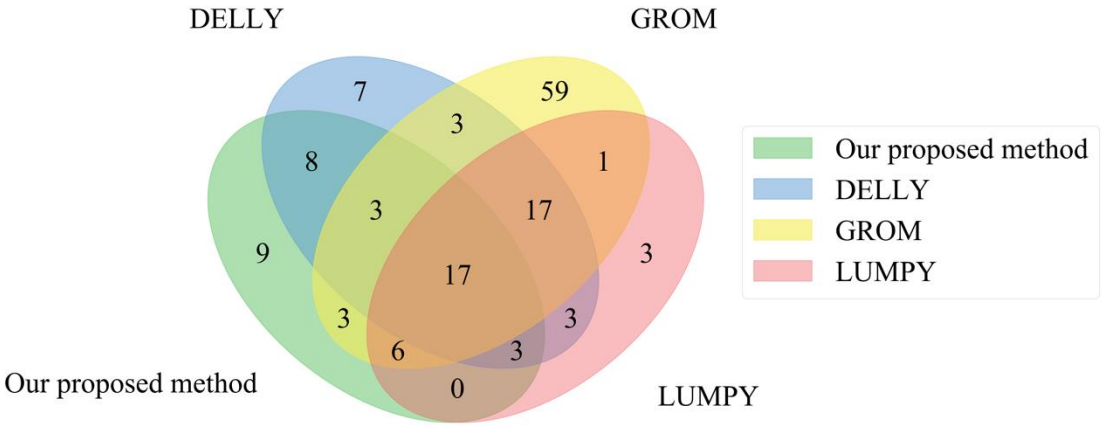
รูปที่ ข.6 แผนภาพเวนน์ของ ERR174337 ประเภท tandem duplication



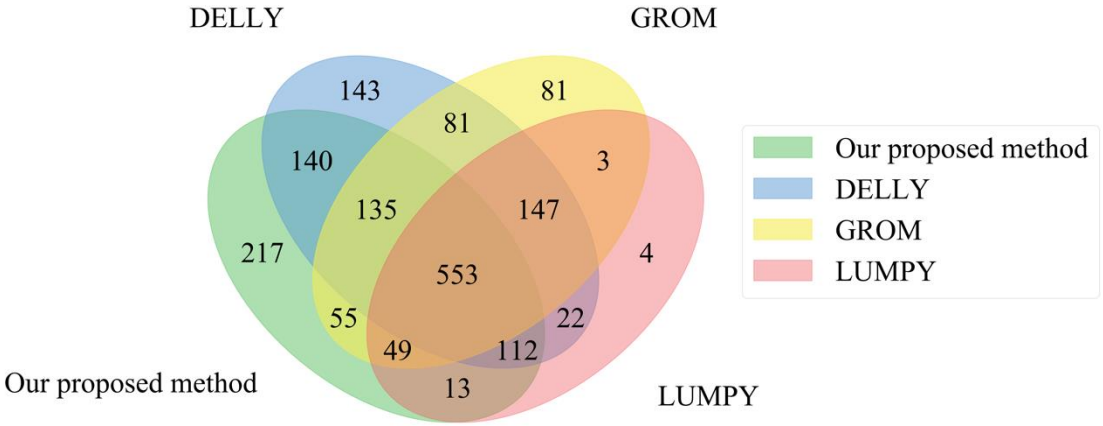
รูปที่ ข.7 แผนภาพเวนน์ของ ERR174338 ประเภท deletion



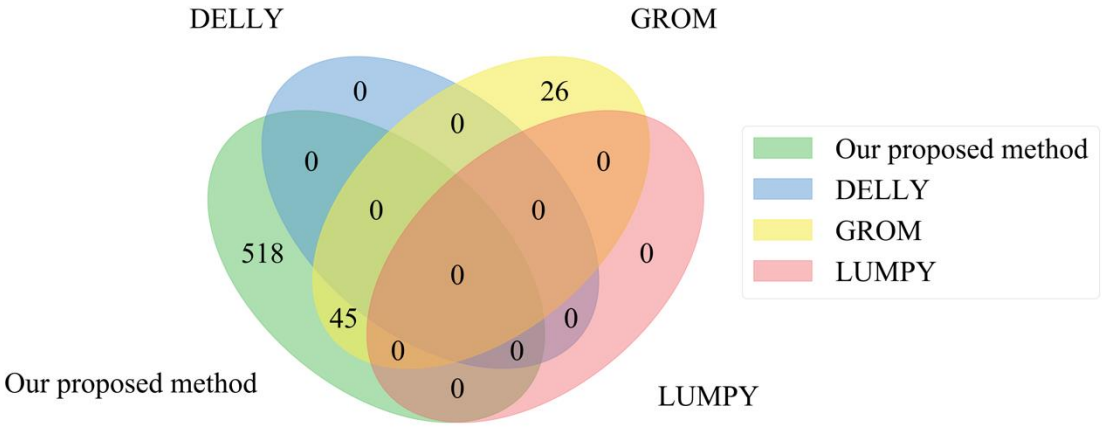
รูปที่ ข.8 แผนภาพเวนน์ของ ERR174338 ประเภท insertion



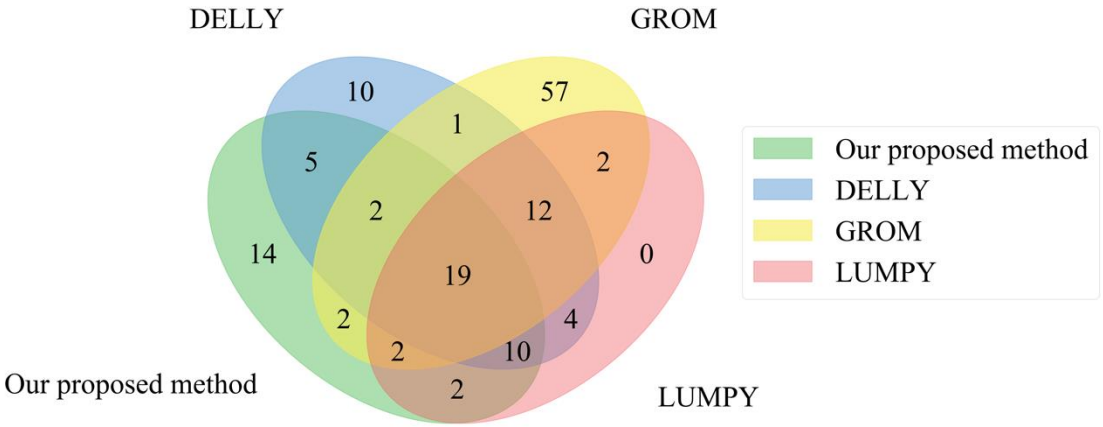
รูปที่ ข.9 แผนภาพเวนน์ของ ERR174338 ประเภท tandem duplication



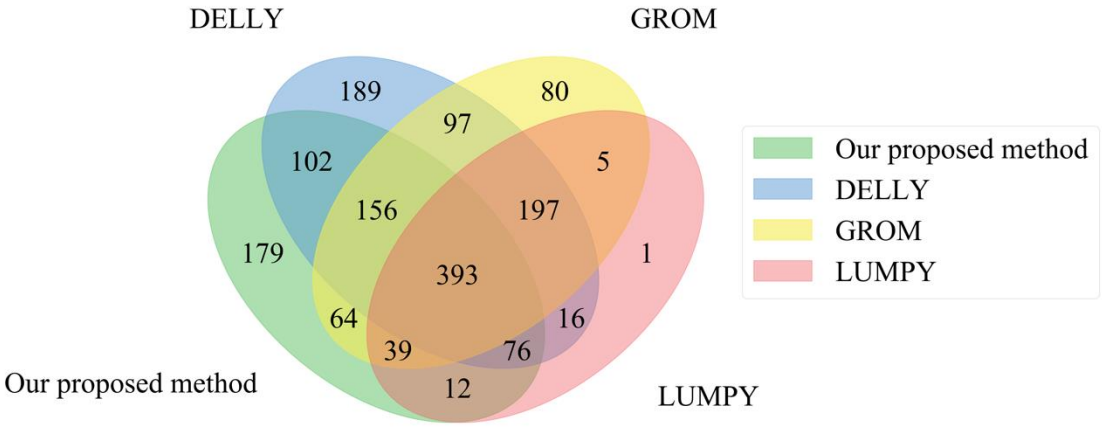
รูปที่ ข.10 แผนภาพเวนน์ของ ERR174339 ประเภท deletion



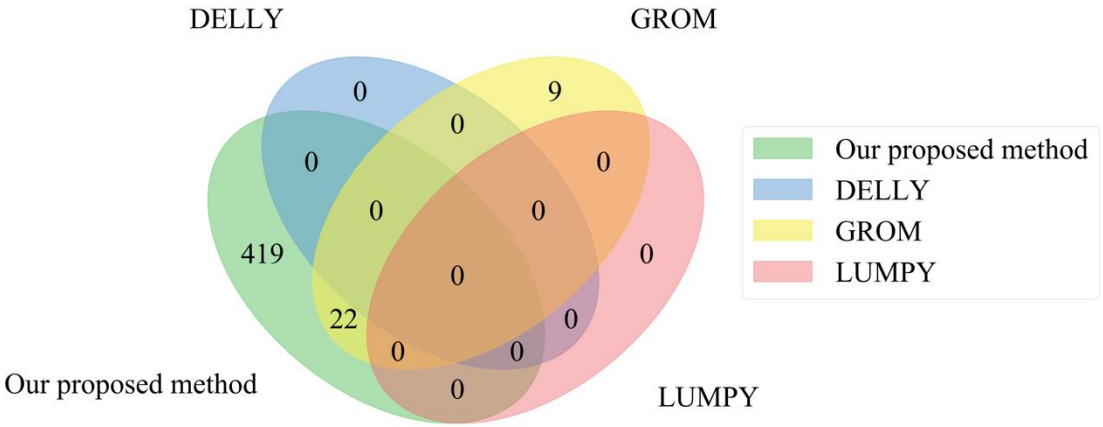
รูปที่ ข.11 แผนภาพเวนน์ของ ERR174339 ประเภท insertion



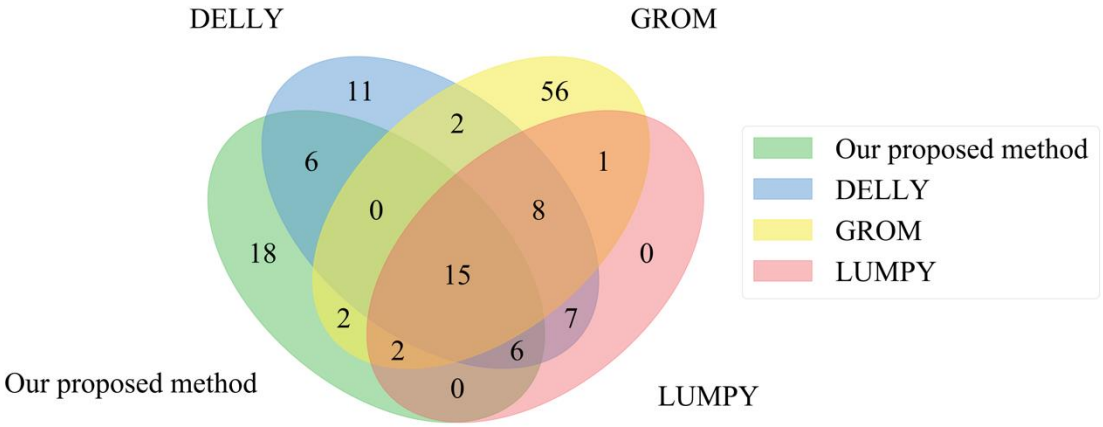
รูปที่ ข.12 แผนภาพเวนน์ของ ERR174339 ประเภท tandem duplication



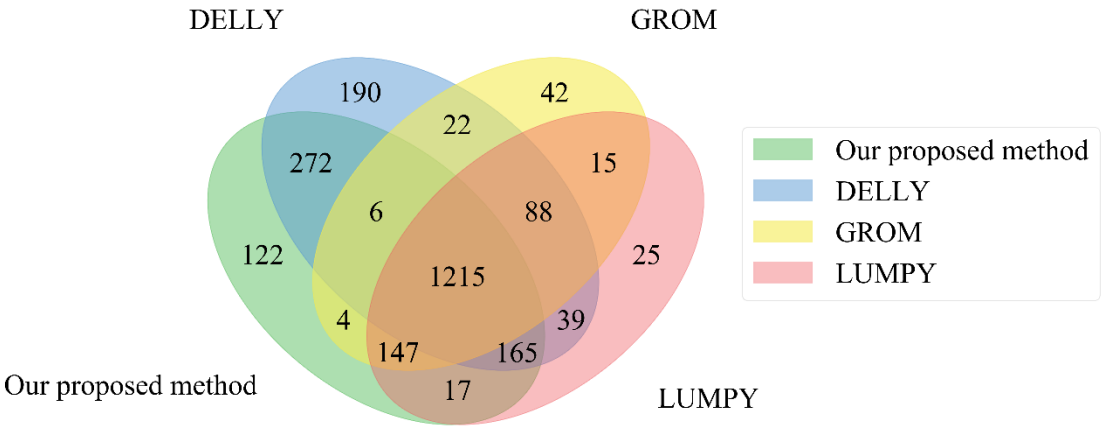
รูปที่ ข.13 แผนภาพเวนน์ของ ERR174340 ประเภท deletion



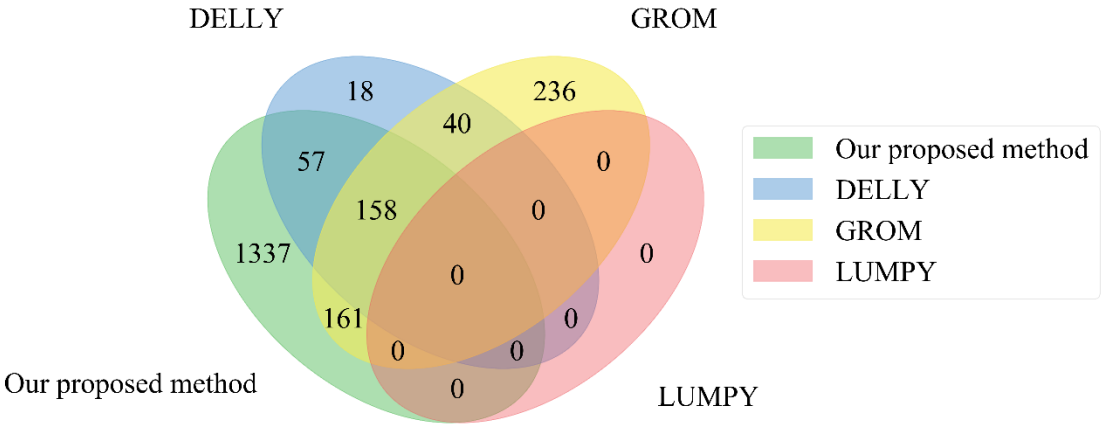
รูปที่ ข.14 แผนภาพเวนน์ของ ERR174340 ประเภท insertion



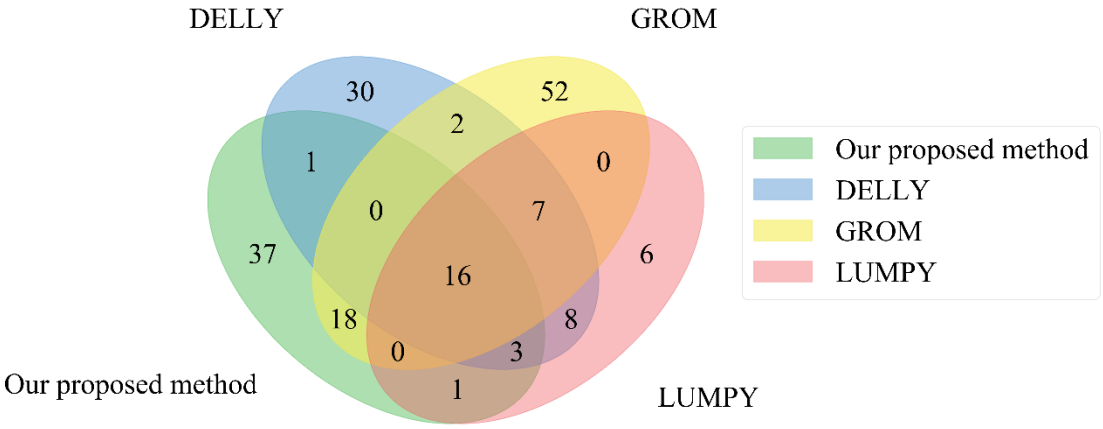
รูปที่ ข.15 แผนภาพเวนน์ของ ERR174340 ประเภท tandem duplication



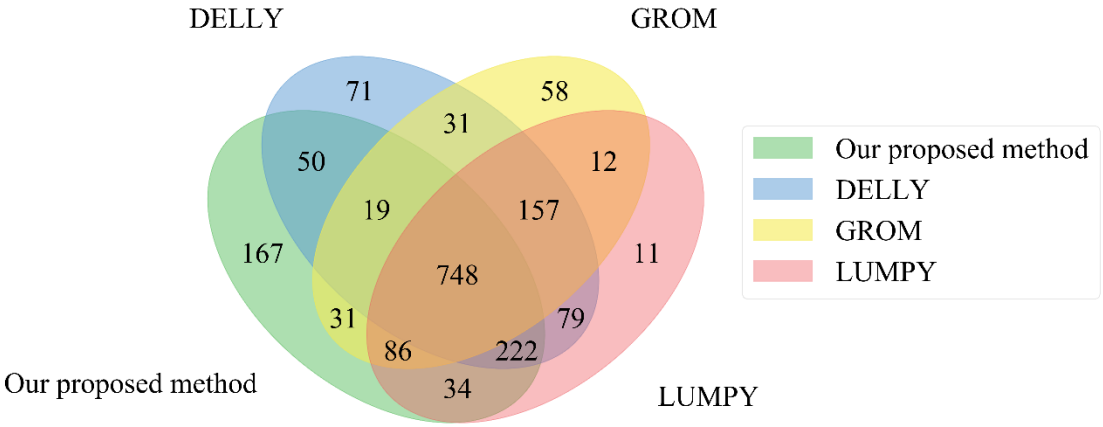
รูปที่ ข.16 แผนภาพเวนน์ของ SRR1910366 ประเภท deletion



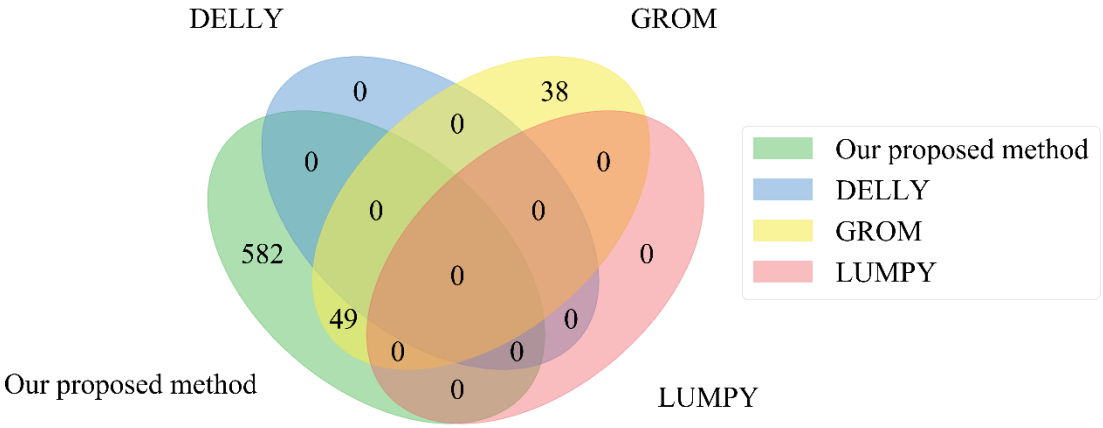
รูปที่ ข.17 แผนภาพเวนน์ของ SRR1910366 ประเภท insertion



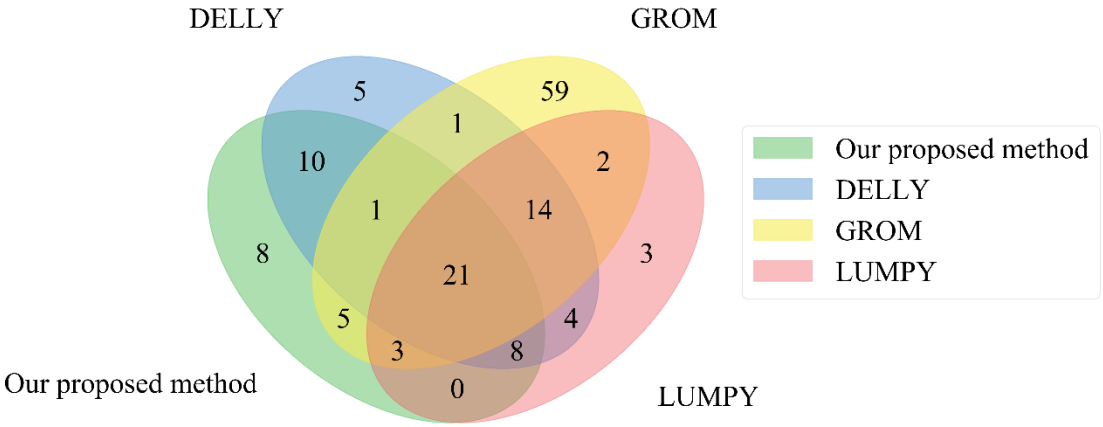
รูปที่ ข.18 แผนภาพเวนน์ของ SRR1910366 ประเภท tandem duplication



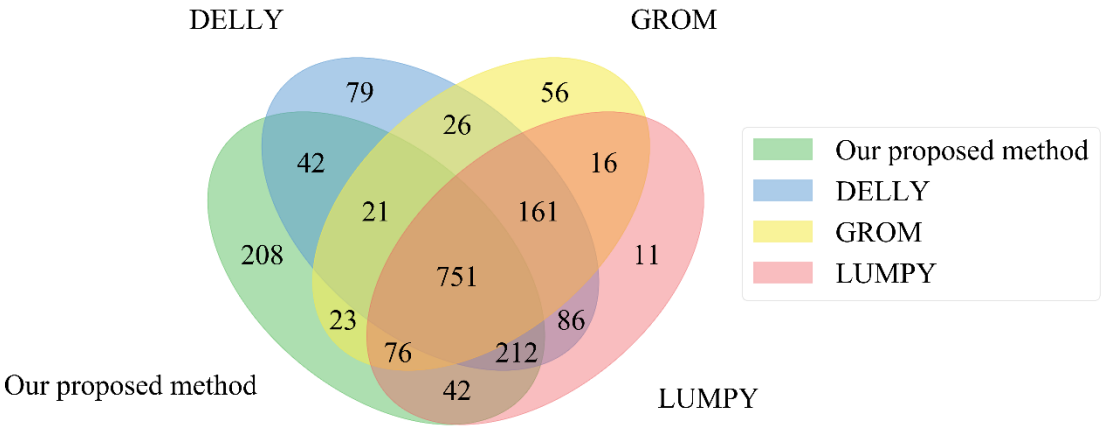
รูปที่ ข.19 แผนภาพเวนน์ของ ERR091571 ประเภท deletion



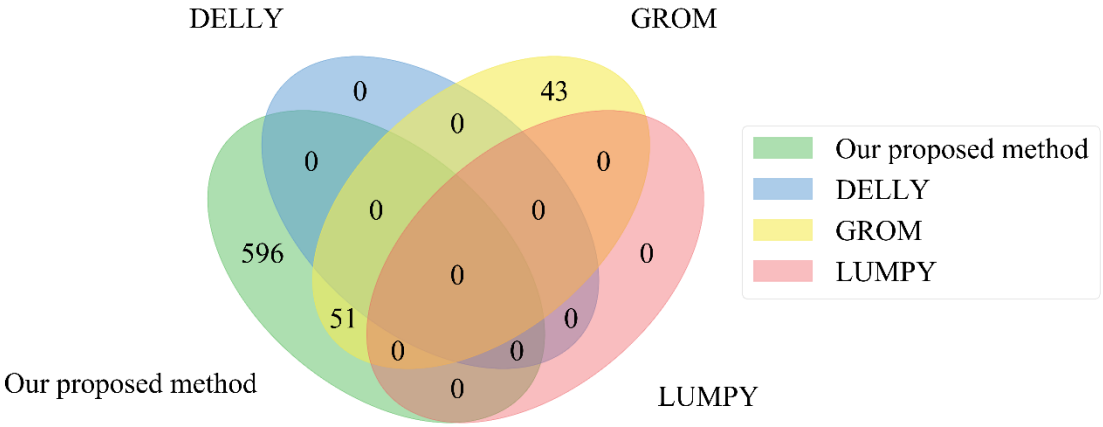
รูปที่ ข.20 แผนภาพเวนน์ของ ERR091571 ประเภท insertion



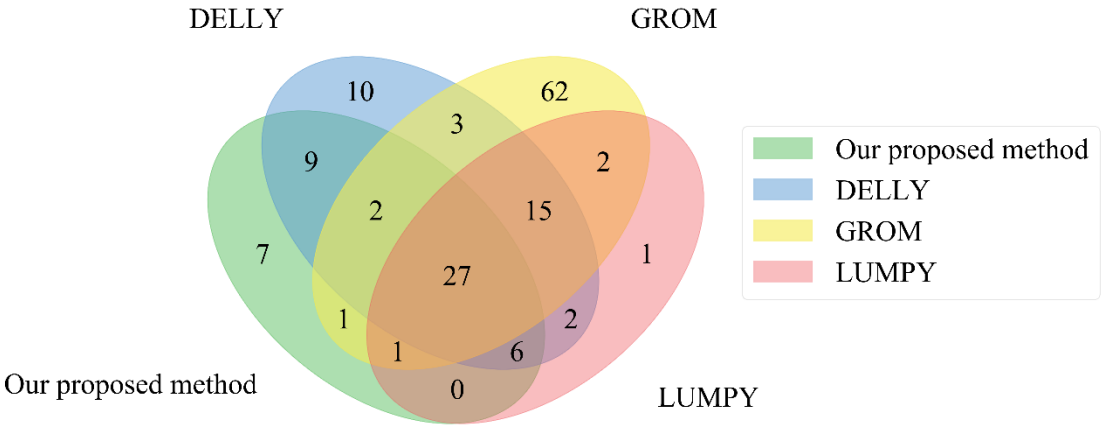
รูปที่ ข.21 แผนภาพเวนน์ของ ERR091571 ประเภท tandem duplication



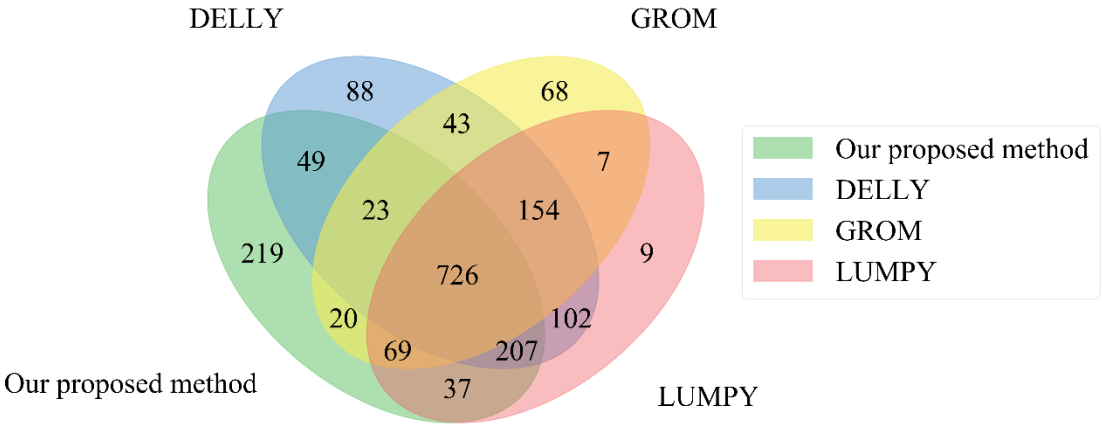
รูปที่ ข.22 แผนภาพเวนน์ของ ERR091572 ประเภท deletion



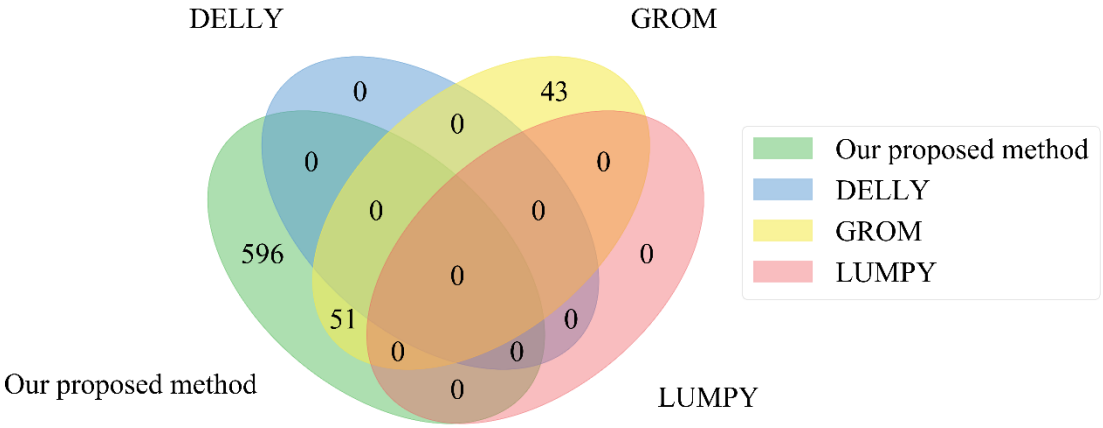
รูปที่ ข.23 แผนภาพเวนน์ของ ERR091572 ประเภท insertion



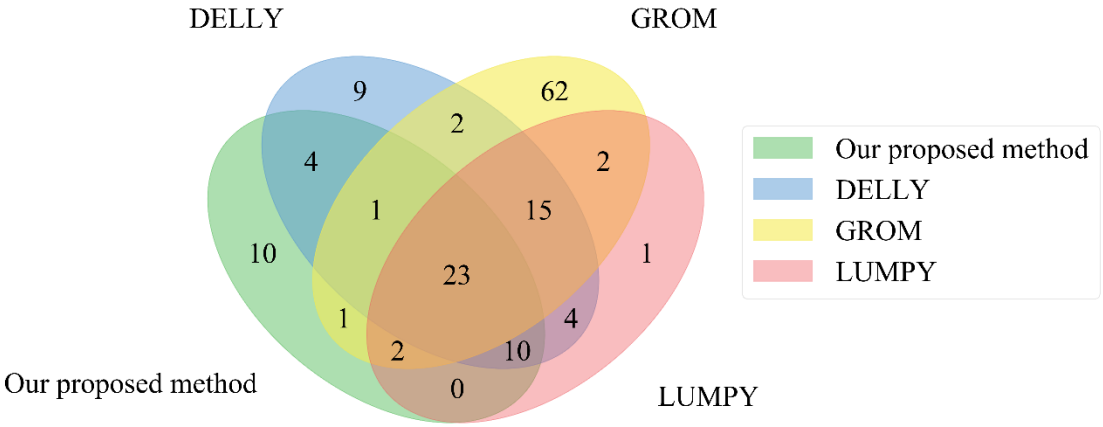
รูปที่ ข.24 แผนภาพเวนน์ของ ERR091572 ประเภท tandem duplication



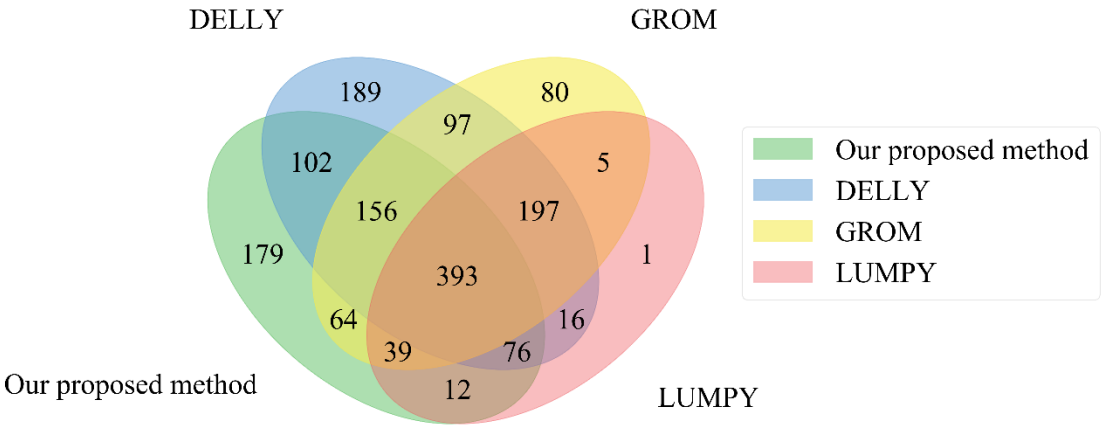
รูปที่ ข.25 แผนภาพเวนน์ของ ERR091573 ประเภท deletion



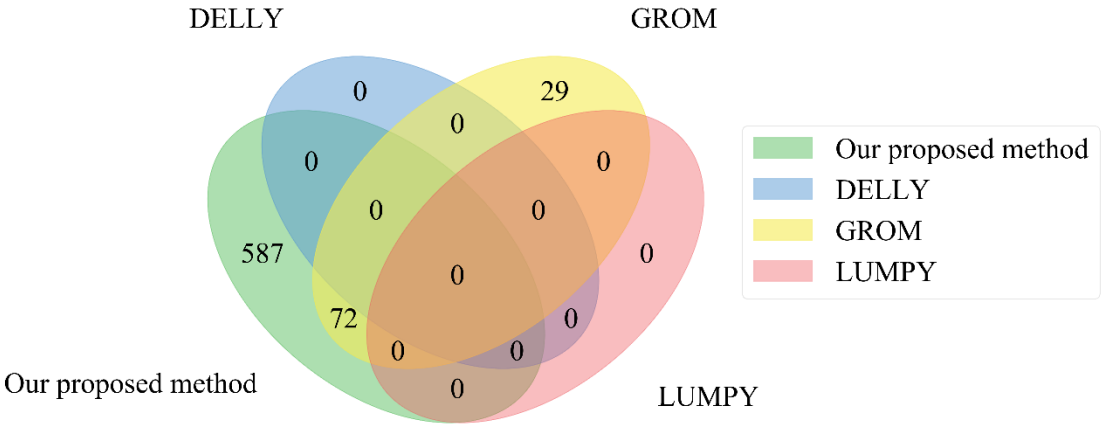
รูปที่ ข.26 แผนภาพเวนน์ของ ERR091573 ประเภท insertion



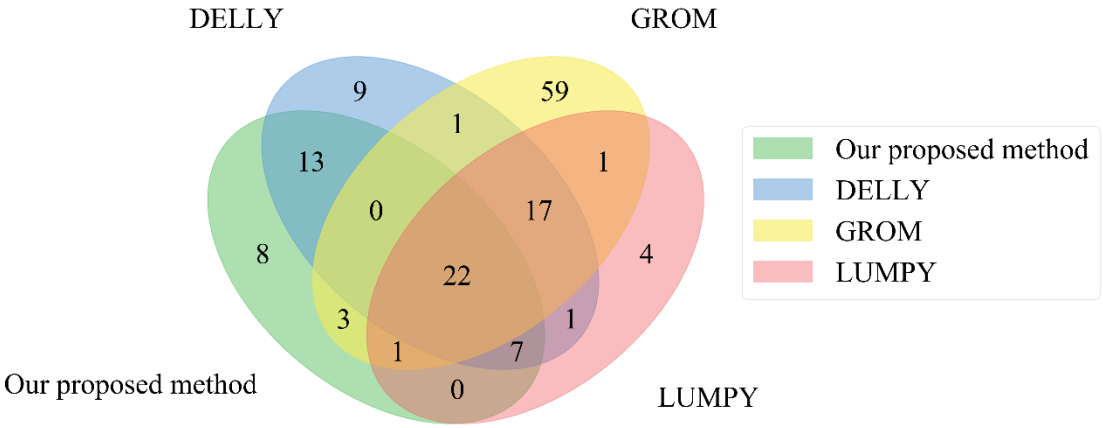
รูปที่ ข.27 แผนภาพเวนน์ของ ERR091573 ประเภท tandem duplication



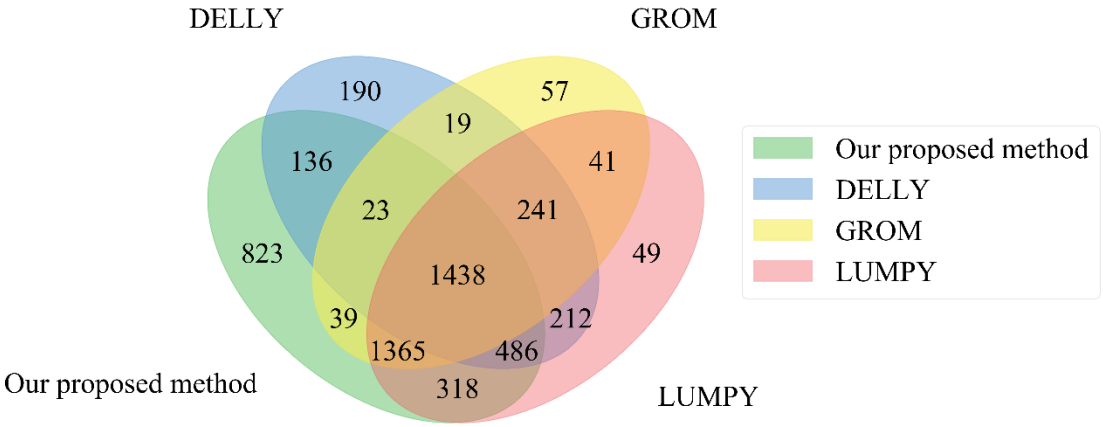
รูปที่ ข.28 แผนภาพเวนน์ของ ERR091574 ประเภท deletion



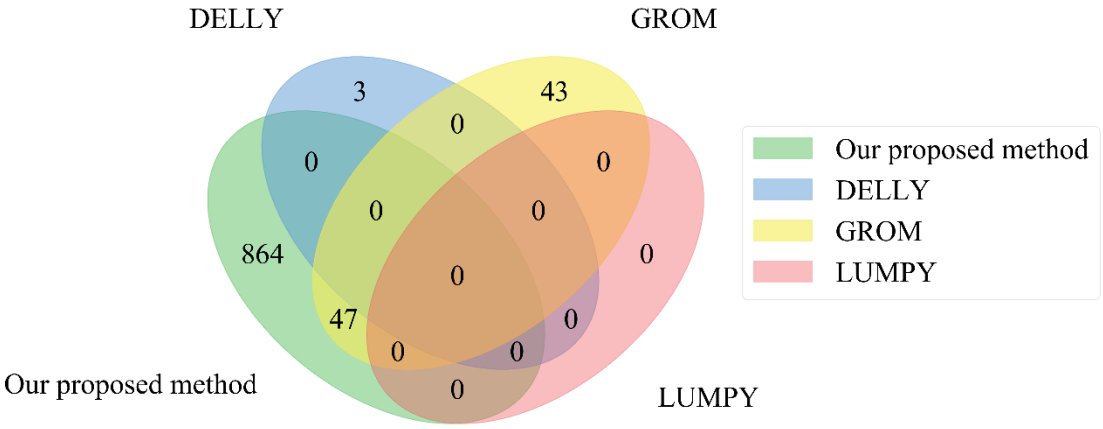
รูปที่ ข.29 แผนภาพเวนน์ของ ERR091574 ประเภท insertion



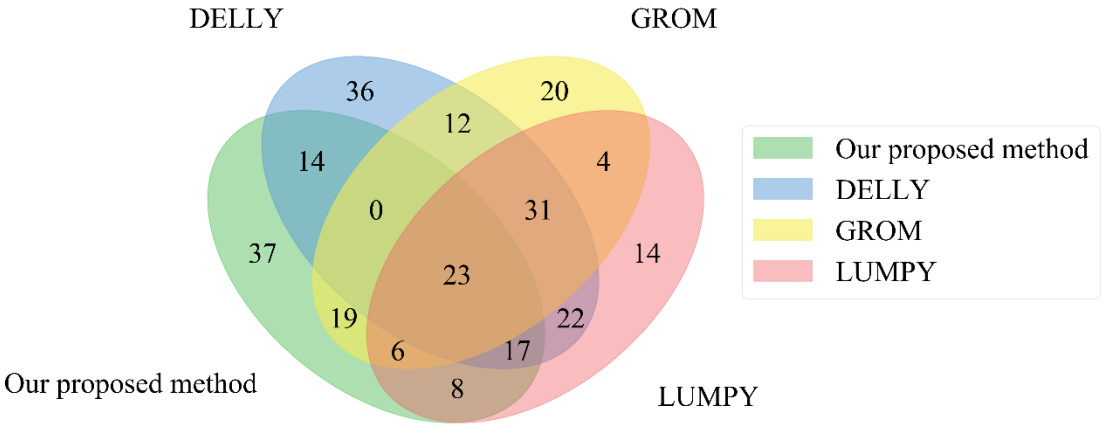
รูปที่ ข.30 แผนภาพเวนน์ของ ERR091574 ประเภท tandem duplication



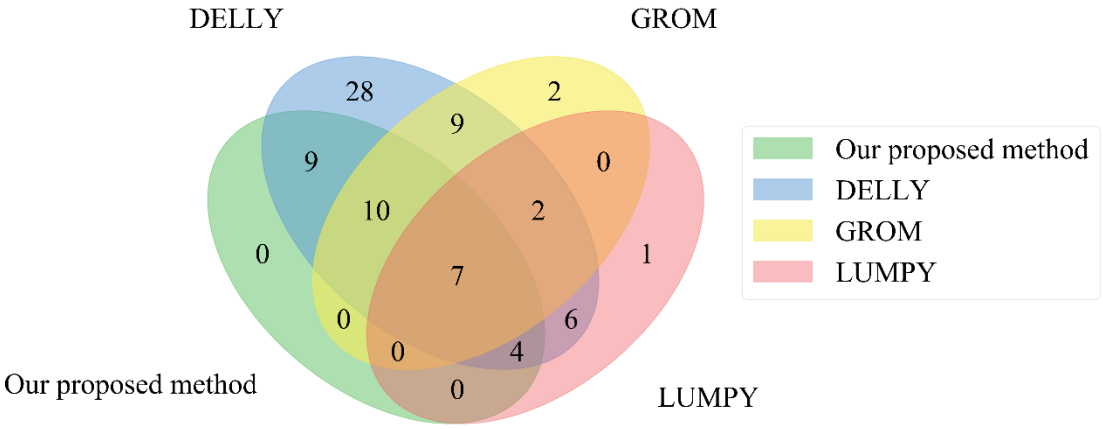
รูปที่ ข.31 แผนภาพเวนน์ของ ERR894729 ประเภท deletion



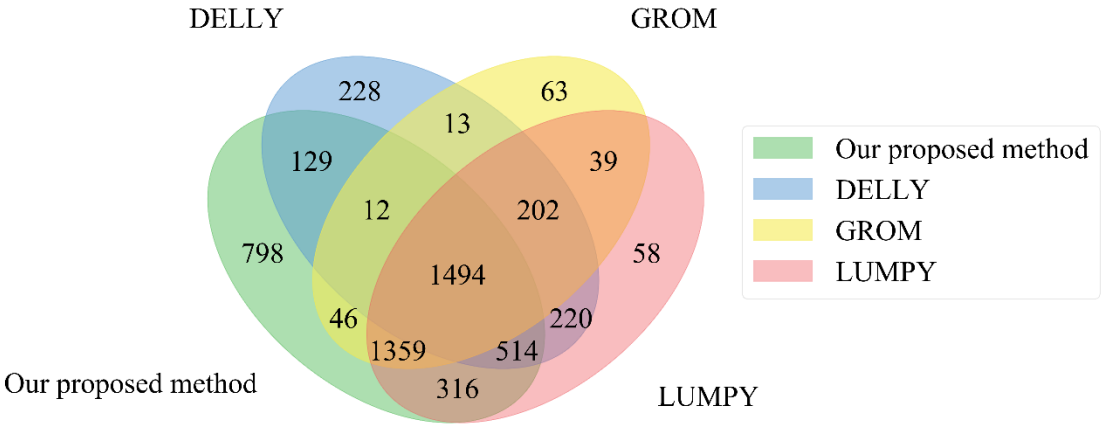
รูปที่ ข.32 แผนภาพเวนน์ของ ERR894729 ประเภท insertion



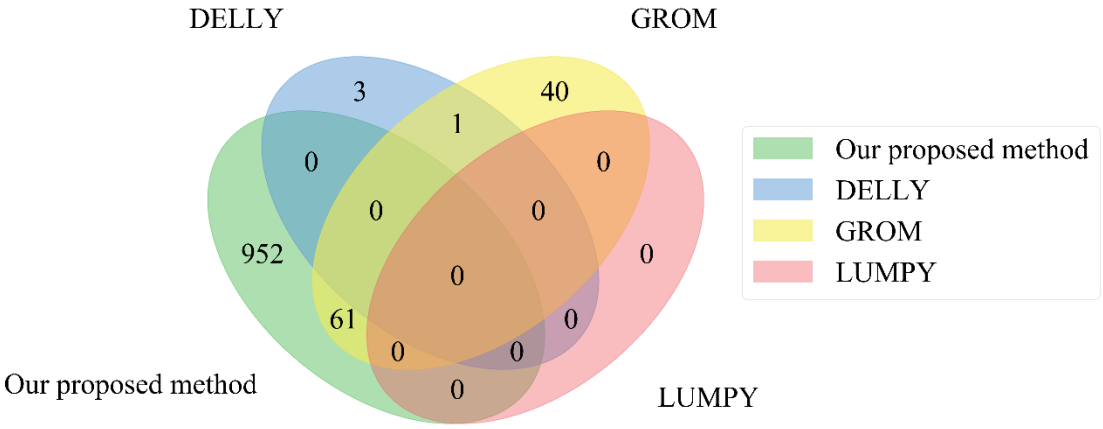
รูปที่ ข.33 แผนภาพเวนน์ของ ERR894729 ประเภท tandem duplication



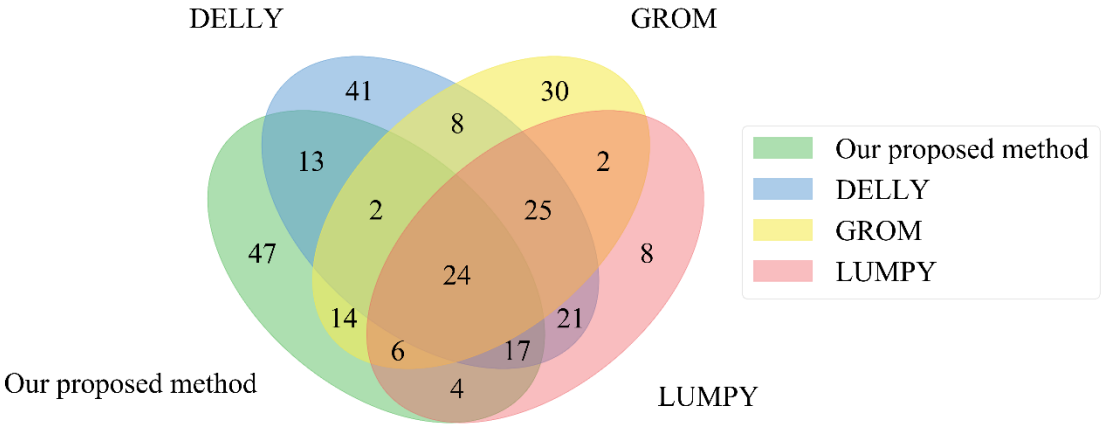
รูปที่ ข.34 แผนภาพเวนน์ของ ERR894729 ประเภท inversion



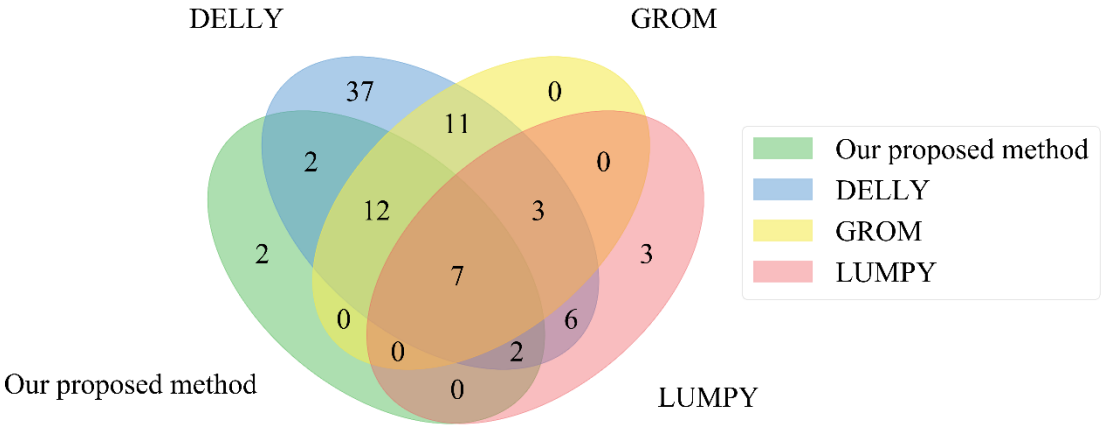
รูปที่ ข.35 แผนภาพเวนน์ของ ERR903030 ประเภท deletion



รูปที่ ข.36 แผนภาพเวนน์ของ ERR903030 ประเภท insertion



รูปที่ ข.37 แผนภาพเวนน์ของ ERR903030 ประเภท tandem duplication



รูปที่ ข.38 แผนภาพเวนน์ของ ERR903030 ประเภท inversion

บรรณานุกรม

1. Stankiewicz, P. and J.R. Lupski, *Structural variation in the human genome and its role in disease.* Annu Rev Med, 2010. **61**: p. 437-55.
2. Nazarenko, M.S., et al., *Genomic structural variations for cardiovascular and metabolic comorbidity.* Sci Rep, 2017. **7**: p. 41268.
3. Pollex, R.L. and R.A. Hegele, *Copy number variation in the human genome and its implications for cardiovascular disease.* Circulation, 2007. **115**(24): p. 3130-8.
4. Smith, J.G. and C. Newton-Cheh, *Genome-wide association studies of late-onset cardiovascular disease.* J Mol Cell Cardiol, 2015. **83**: p. 131-41.
5. Le Marechal, C., et al., *Hereditary pancreatitis caused by triplication of the trypsinogen locus.* Nat Genet, 2006. **38**(12): p. 1372-4.
6. Helbig, I., et al., *15q13.3 microdeletions increase risk of idiopathic generalized epilepsy.* Nat Genet, 2009. **41**(2): p. 160-2.
7. Greenman, C., et al., *Patterns of somatic mutation in human cancer genomes.* Nature, 2007. **446**(7132): p. 153-8.
8. Lee, W., et al., *The mutation spectrum revealed by paired genome sequences from a lung cancer patient.* Nature, 2010. **465**(7297): p. 473-7.
9. Pleasance, E.D., et al., *A comprehensive catalogue of somatic mutations from a human cancer genome.* Nature, 2010. **463**(7278): p. 191-6.
10. Autism Genome Project, C., et al., *Mapping autism risk loci using genetic linkage and chromosomal rearrangements.* Nat Genet, 2007. **39**(3): p. 319-28.
11. Marshall, C.R., et al., *Structural variation of chromosomes in autism spectrum disorder.* Am J Hum Genet, 2008. **82**(2): p. 477-88.
12. International Schizophrenia, C., *Rare chromosomal deletions and duplications increase risk of schizophrenia.* Nature, 2008. **455**(7210): p. 237-41.
13. Stefansson, H., et al., *Large recurrent microdeletions associated with schizophrenia.* Nature, 2008. **455**(7210): p. 232-6.
14. Singleton, A.B., et al., *alpha-Synuclein locus triplication causes Parkinson's disease.* Science, 2003. **302**(5646): p. 841.

15. Rovelet-Lecrux, A., et al., *APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy.* Nat Genet, 2006. **38**(1): p. 24-6.
16. Sboner, A., et al., *The real cost of sequencing: higher than you think!* Genome Biol, 2011. **12**(8): p. 125.
17. Jeffares, D.C., et al., *Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast.* Nat Commun, 2017. **8**: p. 14061.
18. Wgsim, *https://github.com/lh3/wgsim*.
19. Ogasawara, O., et al., *DDBJ Database updates and computational infrastructure enhancement.* Nucleic Acids Res, 2020. **48**(D1): p. D45-D50.
20. Leinonen, R., et al., *The sequence read archive.* Nucleic Acids Res, 2011. **39**(Database issue): p. D19-21.
21. MacDonald, J.R., et al., *The Database of Genomic Variants: a curated collection of structural variation in the human genome.* Nucleic Acids Res, 2014. **42**(Database issue): p. D986-92.
22. Chaisson, M.J.P., et al., *Multi-platform discovery of haplotype-resolved structural variation in human genomes.* Nat Commun, 2019. **10**(1): p. 1784.
23. Park, S.T. and J. Kim, *Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing.* Int Neurourol J, 2016. **20**(Suppl 2): p. S76-83.
24. Turner, F.S., *Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries.* Front Genet, 2014. **5**: p. 5.
25. Yu, X., et al., *How do alignment programs perform on sequencing data with varying qualities and from repetitive regions?* BioData Min, 2012. **5**(1): p. 6.
26. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
27. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.
28. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
29. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA*

sequences to the human genome. Genome Biol, 2009. **10**(3): p. R25.

30. Li, R., et al., *SOAP: short oligonucleotide alignment program.* Bioinformatics, 2008. **24**(5): p. 713-4.
31. Li, R., et al., *SOAP2: an improved ultrafast tool for short read alignment.* Bioinformatics, 2009. **25**(15): p. 1966-7.
32. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
33. Huddleston, J., et al., *Discovery and genotyping of structural variation from long-read haploid genome sequence data.* Genome Res, 2017. **27**(5): p. 677-685.
34. Chiang, C., et al., *The impact of structural variation on human gene expression.* Nat Genet, 2017. **49**(5): p. 692-699.
35. Shigemizu, D., et al., *A practical method to detect SNVs and indels from whole genome and exome sequencing data.* Sci Rep, 2013. **3**: p. 2161.
36. Tattini, L., R. D'Aurizio, and A. Magi, *Detection of Genomic Structural Variants from Next-Generation Sequencing Data.* Front Bioeng Biotechnol, 2015. **3**: p. 92.
37. Hickey, L., *Hunting Structural Variants: Population by Population.* Front Line Genomics Magazine, 2017(15): p. 43-45.
38. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping.* Nat Rev Genet, 2011. **12**(5): p. 363-76.
39. SAMtools. *http://samtools.github.io/hts-specs/SAMv1.pdf*.
40. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
41. Rausch, T., et al., *DELLY: structural variant discovery by integrated paired-end and split-read analysis.* Bioinformatics, 2012. **28**(18): p. i333-i339.
42. Smith, S.D., J.K. Kawash, and A. Grigoriev, *Lightning-fast genome variant detection with GROM.* Gigascience, 2017. **6**(10): p. 1-7.
43. Layer, R.M., et al., *LUMPY: a probabilistic framework for structural variant discovery.* Genome Biol, 2014. **15**(6): p. R84.
44. Wala, J.A., et al., *SvABA: genome-wide detection of structural variants and indels by local assembly.* Genome Res, 2018. **28**(4): p. 581-591.

45. Kronenberg, Z.N., et al., *Wham: Identifying Structural Variants of Biological Consequence.* PLoS Comput Biol, 2015. **11**(12): p. e1004572.

46. Karp, R.M. and M.O. Rabin, *Efficient randomized pattern-matching algorithms.* IBM Journal of Research and Development, 1987. **31**(2): p. 249-260.

47. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* Journal of Molecular Biology, 1981. **147**(1): p. 195-197.

48. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.

49. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

ประวัติผู้เขียน

ชื่อ-สกุล	ศักยภาพ ผิวเหลือง
วัน เดือน ปี เกิด	25 มกราคม 2535
วุฒิการศึกษา	วิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมโยธา สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ