การเปรียบเทียบการจัดเรียงลำดับอาร์เอ็นเอนิวคลีโอไทด์ของเซลล์มะเร็งเม็ดเลือดขาวในแมว ระหว่างโปรแกรมแพคเกจ QuasR และโปรแกรมแพคเกจ Rsubread

Comparison of Feline Lymphoma RNA Sequencing Data Alignment Performed by QuasR and Rsubread Packages

กาจ โชคชัยอุสาหะ* และ ธนิดา สนั่นเมือง

Kaj Chokeshai-u-saha and Thanida Sananmuang

คณะสัตวแพทยศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก วิทยาเขตบางพระ ศรีราชา ชลบุรี

**ไปรษณีย์อิเล็กทรอนิกส์ : kaj.chk@gmail.com หมายเลขโทรศัพท์ : 038-358137 ต่อ 102*

บทคัดย่อ

การจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมเป็นขั้นตอนในการระบุระดับการแสดงออกของยีนในเทคโนโลยีอาร์เอ็นเอซีเควนซิ่ง (RNA-sequencing) เนื่องจาก QuasR และ Rsubread เป็นโปรแกรมแพคเกจสาธารณะในไบโอคอนดัคเตอร์ซึ่งนิยมใช้ในการวิเคราะห์จัดเรียงลำดับนิวคลีโอไทด์ตัวอย่างกับจีโนมของมนุษย์และสัตว์ ผู้วิจัยทางด้านสุขภาพและการจัดการสัตว์จึงควรรู้จักและสามารถใช้งานโปรแกรมแพคเกจทั้งสองได้ การศึกษาครั้งนี้เป็นการสาธิตการใช้งานและเปรียบเทียบโปรแกรมแพคเกจทั้งสองในการจัดเรียงลำดับนิวคลีโอไทด์ข้อมูลอาร์เอ็นเอซีเควนของเซลล์มะเร็งเม็ดเลือดขาวชนิดทีลิมโฟไซต์ของแมว โดยเปรียบเทียบขั้นตอนของการจัดเรียงลำดับนิวคลีโอไทด์ เวลาที่ใช้ในการวิเคราะห์เปอร์เซ็นต์การจัดเรียงนิวคลีโอไทด์สำเร็จ และคุณภาพของข้อมูลหลังการจัดเรียงลำดับนิวคลีโอไทด์ ผลจากการศึกษาแสดงให้เห็นว่า QuasR ต้องการชุดคำสั่งในการปฏิบัติการ (2 ชุดคำสั่ง) น้อยกว่า Rsubread (4 ชุดคำสั่ง) ในขณะที่ Rsubread ใช้การเวลาในการจัดเรียงนิวคลีโอไทด์ข้อมูลทั้งหมด (6 ชั่วโมง 22 นาที) น้อยกว่า QuasR (12 ชั่วโมง 56 นาที) และให้เปอร์เซ็นต์การจัดเรียงนิวคลีโอไทด์สำเร็จสูงกว่า (Rsubread – 89.21±0.20% และ QuasR - 76.72±0.21%) ($p<0.001$, Student-t-test) Rsubread จึงแสดงประสิทธิภาพสูงกว่า QuasR อย่างไรก็ตามผู้ใช้งานโปรแกรมควรคำนึงถึงลักษณะข้อมูลและการตั้งค่าพารามิเตอร์ในการวิเคราะห์ ซึ่งสามารถส่งผลต่อเวลาและสัดส่วนการจัดเรียงนิวคลีโอไทด์สำเร็จด้วย

คำสำคัญ : การจัดเรียงลำดับนิวคลีโอไทด์กับจีโนม อาร์เอ็นเอซีเควนซิ่ง QuasR Rsubread

Abstract

Alignment of RNA-sequencing data is the process to map short sequences to genome in order to incicate expressions of target genes or exons. Since QuasR and Rsubread are well-known Bioconductor packages for animal RNA-sequencing alignment, all molecular researchers in animal health science should acknowledge the use of both packages. In this study, RNA-sequencing data of feline lymphoma cell lines were aligned with cat genome using QuasR and Rsubread packages with default settings. QuasR and Rsubread were compared in terms of analysis process, time requirement, percent alignment and processed data quality. The results implied QuasR as more user-friendly package than Rsubread due to its lesser instruction sets to process (2 instruction sets for QuasR and 3 instruction sets for Rsubread). On the contrary, Rsubread performed faster alignment (12 hours 56 minutes for QuasR and 6 hours 22 minutes for Rsubread) yet provide higher percent alignment

(p<0.001, Student-t-test) (89.21±0.20% for Rsubread and 76.72±0.21% for QuasR). Though Rsubread outstanded QuasR in term of time and percent alignment, changes of data and parameter adjustment could greatly affect the alignment outcome. These should be always considered for further application of both packages.

Keywords : Alignment, RNA-sequencing, QuasR, Rsubread

1. บทนำ

นับตั้งแต่ความสำเร็จในการระบุลำดับนิวคลีโอไทด์จีโนมของมนุษย์ในปี ค.ศ. 2000 (Yamey, 2000) การศึกษาศาสตร์ด้านจีโนมิกส์ (Genomics) เอ็กโซมิกส์ (Exomics) และทรานคริปโตมิกส์ (Transcriptomics) ได้รับการพัฒนาสู่ยุคของการหาลำดับนิวคลีโอไทด์แบบไฮทรูพุท (high throughput sequencing) อย่างเต็มรูปแบบ เทคโนโลยี Next generation sequencing เป็นเทคโนโลยีการหาลำดับนิวคลีโอไทด์แบบไฮทรูพุท ที่ได้รับความนิยมเพิ่มขึ้นอย่างต่อเนื่อง โดยการวิเคราะห์ลำดับนิวคลีโอไทด์ด้วยเทคโนโลยี Next generation sequencing เริ่มต้นจากการตัดสายดีเอ็นเอต้นแบบที่สนใจออกเป็นชิ้นส่วนดีเอ็นเอสายสั้นๆ ก่อนเชื่อมต่อชิ้นส่วนดีเอ็นเอชิ้นที่ได้กับนิวคลีโอไทด์ที่ทราบลำดับ (adaptor) เพื่อให้สามารถสังเคราะห์เพิ่มจำนวนดีเอ็นเอเหล่านั้นด้วยไพรเมอร์ที่จำเพาะกับ adapter ได้โดยไม่ต้องทราบลำดับนิวคลีโอไทด์ในการออกแบบไพรเมอร์ ทำให้สามารถเพิ่มจำนวนนิวคลีโอไทด์ต้นแบบทั้งหมดสำหรับการวิเคราะห์หาลำดับได้พร้อมกัน ด้วยเหตุนี้เทคโนโลยี Next generation sequencing จึงสามารถให้ข้อมูลลำดับนิวคลีโอไทด์ปริมาณมากได้อย่างรวดเร็ว (Zhang, *et al.*, 2011)

ในปัจจุบันเทคโนโลยี Next generation sequencing ได้รับการประยุกต์ใช้สำหรับศึกษาอาร์เอ็นเอผ่านการจัดเรียงลำดับนิวคลีโอไทด์ของซีดีเอ็นเอ (cDNA) ซึ่งสังเคราะห์จากอาร์เอ็นเอ (RNA) ต้นแบบกับจีโนมเพื่อใช้ในการศึกษาการแสดงออกของยีนและเอกซอน (exon) ในสิ่งมีชีวิตหลากหลายชนิด การประยุกต์ใช้ เทคโนโลยี Next generation sequencing กับอาร์เอ็นเอก่อให้เกิดแขนงของการศึกษาอาร์เอ็นเอที่เรียกว่าอาร์เอ็นเอซีเควนซิ่ง (RNA-sequencing) หรืออาร์เอ็นเอเซค (RNA-seq) จากหลักการที่กล่าวมาข้างต้นทำให้การจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมเป็นหัวใจสำคัญของการวิเคราะห์ข้อมูลที่ได้จากอาร์เอ็นเอซีเควน เนื่องจากปริมาณสายซีดีเอ็นเอสายสั้นที่ได้รับการจัดเรียงกับยีน/เอกซอน จะแสดงถึงระดับการแสดงออกของแต่ละยีน/เอกซอนเพื่อนำใช้ในการวิเคราะห์ในลำดับต่อๆ ไป (Guan, *et al.*, 2011; Xiao, et al., 2011; Zhou, *et al.*, 2010)

การจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมอาศัยการนำลำดับนิวคลีโอไทด์สายสั้นที่ได้จากการเพิ่มจำนวนหรือ reads ที่มีความยาวเฉลี่ย 30 ถึง 400 นิวคลีโอไทด์ ซึ่งมีความยาวแตกต่างกันตามชนิดของเทคโนโลยีที่ใช้มาจัดเรียง (Aligning sequencing) ให้ตรงกับลำดับนิวคลีโอไทด์ของจีโนมสิ่งมีชีวิตโดยอาศัยเครื่องมือทางชีวสารสนเทศที่เหมาะสม จากข้อจำกัดของเทคโนโลยีในการจัดวางลำดับนิวคลีโอไทด์สายสั้นดังกล่าวกับจีโนมของสิ่งมีชีวิต และความซับซ้อนของกลไกที่สิ่งมีชีวิตใช้ในการควบคุมการแสดงออกของยีนผ่านการเชื่อมต่อเอกซอนซึ่งเป็นลำดับนิวคลีโอไทด์ที่ไม่ปรากฏต่อเนื่องกันบนจีโนมและยังมีรูปแบบที่หลากหลายในการประกอบกันเป็นสายอาร์เอ็นเอของยีนแต่ละยีน ทำให้เครื่องมือทางชีวสารสนเทศสำหรับการจัดเรียงและเชื่อมลำดับนิวคลีโอไทด์ได้รับการพัฒนาขึ้นอย่างต่อเนื่องด้วยหลักการที่แตกต่างกันเพื่อให้ผลการนับลำดับนิวคลีโอไทด์ที่ได้รับการจัดเรียงกับจีโนมถูกต้องและได้สัดส่วนการจัดเรียงสำเร็จมากที่สุด

ปริมาณข้อมูลอาร์เอ็นเอเซคของสัตว์หลากหลายชนิดในฐานข้อมูลสาธารณะมีอัตราการเติบโตเพิ่มขึ้นอย่างต่อเนื่อง ส่งผลให้การวิเคราะห์ข้อมูลอาร์เอ็นเอเซคในสัตว์เข้ามามีบทบาทต่อการพัฒนางานวิจัยทางด้านสุขภาพและการจัดการสัตว์มากขึ้นตามลำดับ เนื่องจากการวิเคราะห์ข้อมูลอาร์เอ็นเอเซคที่เหมาะสมสามารถสร้างองค์ความรู้ใหม่ที่มีประโยชน์และยั่งยืน จึงมีความจำเป็นอย่างยิ่งที่ผู้วิจัยทางด้านสุขภาพและการจัดการสัตว์ควรเข้าใจหลักการจัดเรียง

ลำดับนิวคลีโอไทด์พื้นฐานและสามารถพิจารณาการใช้โปรแกรมสาธารณะสำเร็จรูปเพื่อการวิเคราะห์ดังกล่าวได้อย่างเหมาะสม เนื่องจาก ไบโอคอนดัคเตอร์เป็นโครงการทางเทคโนโลยีชีวสารสนเทศที่มีจุดมุ่งหมายหลักเพื่อพัฒนาโปรแกรมและฐานข้อมูลสำหรับวิเคราะห์ข้อมูลต่างๆ ที่เกี่ยวข้องกับยีนโดยใช้ภาษาคอมพิวเตอร์อาร์ (R) ซึ่งได้รับการพัฒนาขึ้นสำหรับการวิเคราะห์ทางสถิติ (Gentleman, *et al.*, 2004) ด้วยเหตุนี้โปรแกรมวิเคราะห์ข้อมูลอาร์เอ็นเอเซคที่ได้รับการพัฒนาจากไบโอคอนดัคเตอร์จึงมีจุดเด่นที่ใช้งานได้ง่าย เหมาะกับผู้ใช้งานที่มิใช่นักชีวสารสนเทศและมีพื้นฐานด้านการใช้ภาษาคอมพิวเตอร์จำกัด

ในปัจจุบันโปรแกรมแพคเกจสำหรับจัดเรียงลำดับนิวคลีโอไทด์ที่ได้จากข้อมูลอาร์เอ็นเซคที่ได้รับความนิยมของไบโอคอนดัคเตอร์ ได้แก่ QuasR (Gaidatzis, et al., 2014) และ Rsubread (Liao, et al., 2013) อย่างไรก็ตามตัวอย่างการประยุกต์ใช้โปรแกรมทั้งสองในงานวิจัยทางด้านสุขภาพและการจัดการในสัตว์ยังมีอยู่จำกัด ทำให้ผู้วิจัยงานด้านวิทยาศาสตร์เกี่ยวกับสัตว์สูญเสียโอกาสในการใช้โปรแกรมสาธารณะดังกล่าว การศึกษาครั้งนี้เป็นการเปรียบเทียบการทำงานของโปรแกรม QuasR และ Rsubread ในการจัดเรียงข้อมูลลำดับนิวคลีโอไทด์สายสั้นของข้อมูลอาร์เอ็นเซคของเซลล์มะเร็งเม็ดเลือดขาวชนิดทีลิมโฟไซต์ในแมว (FeT-J cells) (Ertl and Klein, 2014) เพื่อสาธิตการประยุกต์ใช้โปรแกรมทั้งสองกับข้อมูลอาร์เอ็นเอเซคในแมวและเปรียบเทียบผลการวิเคราะห์ข้อมูลที่ได้จากแต่ละโปรแกรมในแง่มุมต่างๆ อันจะก่อประโยชน์ต่อการพิจารณาเลือกใช้โปรแกรมในการวิเคราะห์ข้อมูลอาร์เอ็นเอเซคต่อไปอนาคต

2. วิธีการทดลอง

2.1 ระบบคอมพิวเตอร์ที่ใช้ในการวิเคราะห์

คอมพิวเตอร์ที่ใช้การวิเคราะห์มีหน่วยความจำประมวลผล (RAM) ขนาด 8GB และมีพื้นที่ว่างสำหรับบันทึกข้อมูล 10 GB ระบบไมโครโปรเซสเซอร์ที่ใช้คือ Intel Pentium V processor โดยทำงานภายใต้ระบบปฏิบัติการ Bio-Linux-7 operating systems (OS) (Field, *et al.*, 2006)

2.2 ข้อมูลอาร์เอ็นเอเซคของเซลล์มะเร็งเม็ดเลือดขาวชนิดทีลิมโฟไซต์ในแมว

ข้อมูลอาร์ซีเควนซิงของเซลล์มะเร็งเม็ดเลือดขาวชนิดทีลิมโฟไซต์ของแมวบ้าน (FeT-J cells) ที่ติดเชื้อไวรัสเอดส์และของเซลล์มะเร็งเม็ดเลือดขาวกลุ่มควบคุม (Ertl and Klein, 2014) สามารถดาวน์โหลดได้จาก http://www.ebi.ac.uk/arrayexpress ภายใต้ accession number E-MTAB-2083 โดยรายละเอียดของการเตรียมเซลล์และข้อมูลระบุอยู่ใน Ertl and Klein (2014) โดยข้อมูลดังกล่าวได้จากการใช้เทคโนโลยี Genome Analyzer II (Illumina) ในการอ่านลำดับนิวคลีโอไทด์สายสั้นความยาว 37 bp โดยข้อมูลลำดับนิวคลีโอไทด์สายสั้นของแต่ละตัวอย่างถูกดาวน์โหลดและเก็บไว้ด้วยรูปแบบไฟล์ชนิด fastq.gz (ตารางที่ 1)

ตารางที่ 1 รายละเอียดของตัวอย่างข้อมูลลำดับนิวคลีโอไทด์สายสั้นที่ใช้ในการทดลอง

รหัสประจำตัวอย่าง	ชื่อไฟล์ที่จัดเก็บ	การเตรียมเซลล์ก่อนสกัดอาร์เอ็นเอ
ERR371789	ERR371789.fastq.gz	เลี้ยง FeT-J cells 24 ชั่วโมง ในตัวทำละลายไวรัส FIV
ERR 371790	ERR 371790.fastq.gz	เลี้ยง FeT-J cells 24 ชั่วโมง ในตัวทำละลายไวรัส FIV
ERR 371791	ERR 371791.fastq.gz	เลี้ยง FeT-J cells 24 ชั่วโมง ในตัวทำละลายไวรัส FIV
ERR 371792	ERR 371792.fastq.gz	เลี้ยง FeT-J cells 24 ชั่วโมง ในตัวทำละลายไวรัส FIV
ERR 371793	ERR 371793.fastq.gz	เลี้ยง FeT-J cells 24 ชั่วโมง หลังติดเชื้อไวรัส FIV
ERR 371794	ERR 371794.fastq.gz	เลี้ยง FeT-J cells 24 ชั่วโมง หลังติดเชื้อไวรัส FIV
ERR 371795	ERR 371795.fastq.gz	เลี้ยง FeT-J cells 24 ชั่วโมง หลังติดเชื้อไวรัส FIV
ERR 371796	ERR 371796.fastq.gz	เลี้ยง FeT-J cells 24 ชั่วโมง หลังติดเชื้อไวรัส FIV

2.3 จีโนมของแมวบ้าน

จีโนมที่ใช้ในการระบุตำแหน่งในการทดลองครั้งนี้ คือ ซอฟท์ฟอร์แมทของจีโนมแมวบ้านในรูปของฟาสตาไฟล์ (fasta file) (Felis_catus.Felis_catus_6.2.dna.sm.toplevel.fa) ซึ่งได้จาก http://www.ensembl.org/ ซึ่งสร้างโดย Ensembl gene annotation project (e!70) Felis catus (cat, Felis_catus-6.2) โดยไฟล์ดังกล่าวสามารถดาวน์โหลดได้จาก http://www.ensembl.org/

2.4 โปรแกรม R โปรแกรมแพคเกจ QuasR โปรแกรมแพคเกจ Rsubread และโปรแกรม FastQC

โปรแกรม R เวอร์ชัน 2.15.1 ถูกใช้ในการศึกษาครั้งนี้ โดยสามารถดาวน์โหลดและติดตั้งโปรแกรม R ผ่าน CRAN (Comprehensive R Archive Network) ได้จาก http://www.r-project.org/ แพคเกจ GenomicFeatures แพคเกจ QuasR และแพคเกจ Rsubread เป็นแพคเกจสาธารณะที่สามารถดาวน์โหลดได้จากไบโอคอนดัคเตอร์ (http://www.bioconductor.org/) (Gentleman, *et al.*, 2004) สำหรับโปรแกรม FastQC สามารถดาวน์โหลดได้จากเว็บไซต์ของสถาบัน Babraha Bioinformatics Institutes (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

2.5 การจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมแมวโดย QuasR และ Rsubread

ในการศึกษาครั้งนี้กำหนดขั้นตอนของการจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมออกเป็น 4 ขั้นตอน ได้แก่ การจัดเตรียมข้อมูลให้อยู่ในรูปแบบที่โปรแกรมแพคเกจสามารถใช้งานได้ การจัดเตรียมไฟล์จีโนมของแมวเพื่อใช้จัดเรียงลำดับนิวคลีโอไทด์ การจัดเรียงลำดับนิวคลีโอไทด์ตัวอย่างกับจีโนม และการเขียนชุดคำสั่งเพื่อตรวจสอบเปอร์เซนต์ลำดับนิวคลีโอไทด์ที่จัดเรียงกับจีโนมสำเร็จในแต่ละตัวอย่าง โดยขั้นตอนในการปฏิบัติการทั้งหมดแสดงในวิธีการใช้งานแพคเกจฉบับสั้น โดยคู่มือฉบับสั้นของ QuasR ดาวน์โหลดได้จาก http://www.bioconductor.org/packages/release/bioc/vignettes/QuasR/inst/doc/QuasR.pdf และคู่มือฉบับสั้นของRsubread ดาวน์โหลดจาก http://www.bioconductor.org/packages/release/bioc/vignettes/Rsubread/inst/doc/Rsubread.pdf โดย arguments ของคำสั่งที่ใช้ในแพคเกจต่างๆให้ใช้ค่าพื้นฐานตามที่แพคเกจแต่ละชนิดระบุทั้งหมด ซึ่งสามารถตรวจสอบได้โดยการพิมพ์ '?' ตามด้วยคำสั่งที่ต้องการตรวจสอบ arguments เพื่อระบุค่าพารามิเตอร์ที่ใช้ในขั้นตอนต่างๆ

2.6 การตรวจสอบคุณภาพข้อมูลที่ได้จากการจัดเรียงนิวคลีโอไทด์กับจีโนม

ข้อมูลที่ได้จากการจัดเรียงนิวคลีโอไทด์กับจีโนมด้วยโปรแกรม QuasR และ Rsubread จะถูกบันทึกในรูปแบบ BAM (Binary version of Sequence Alignment/Map) BAM ไฟล์ที่ได้จากการจัดเรียงนิวคลีโอไทด์แต่ละตัวอย่างจะได้รับการประเมินคุณภาพด้วยโปรแกรม FastQC โดยโปรแกรมจะประเมินข้อมูลใน 12 ประเด็น ได้แก่ Basic Statistics, Per Base Sequence Quality, Per Sequence Quality Scores, Per Base Sequence Content, Per Sequence GC Content, Per Base N Content, Sequence Length, Distribution Duplicate Sequences, Overrepresented Sequences, Adapter Content, Kmer Content และ Per Tile Sequence Quality โดยสามารถอ่านวิธีการแปรผลการประเมินในแต่ละประเด็นได้เพิ่มเติมจาก http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

3. ผลการทดลอง

3.1 ขั้นตอนการสั่งงานโปรแกรมรวมถึงระยะเวลาในการประมวลผลการจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมระหว่างแพคเกจ QuasR และ Rsubread มีความแตกต่างกัน

เมื่อเปรียบเทียบขั้นตอนและระยะเวลาที่ใช้ทั้งหมดในการประมวลผลการจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมระหว่างแพคเกจ QuasR และ Rsubread พบว่าโปรแกรมแพคเกจ QuasR ต้องการชุดคำสั่งเพื่อครอบคลุมขั้นตอนการประมวลผลทั้งหมดน้อยกว่าโปรแกรมแพคเกจ Rsubread อย่างไรก็ตามระยะเวลาในการประมวลผลของ QuasR มากกว่า Rsubread ถึง 6 ชั่วโมง 34 นาที (ตารางที่ 2)

3.2 โปรแกรมแพคเกจ Rsubread ให้เปอร์เซ็นต์ลำดับนิวคลีโอไทด์ที่จัดเรียงกับจีโนมสำเร็จสูงกว่าแพคเกจ QuasR โดยที่คุณภาพข้อมูลหลังการจัดเรียงลำดับนิวคลีโอไทด์ด้วยแพคเกจทั้งสองไม่แตกต่างกัน

เมื่อเปรียบเทียบเครื่องมือ ค่าพารามิเตอร์ และช่วงเปอร์เซ็นต์ที่ได้จากการจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมสำเร็จระหว่างโปรแกรมแพคเกจ QuasR และ Rsubread พบว่าโปรแกรมแพคเกจทั้งสองใช้เครื่องมือทางชีวสารสนเทศที่แตกต่าง โดย QuasR ใช้ Bowtie ในขณะที่ Rsubread ใช้ subread อย่างไรก็ตามโปรแกรมทั้งสองมีการกำหนดค่าพารามิเตอร์พื้นฐาน ได้แก่ Phred score cut-off ปริมาณตำแหน่งสูงสุดที่นิวคลีโอไทด์แต่ละเส้นจับกับจีโนมที่ถูกรายงาน ปริมาณนิวคลีโอไทด์ไม่เข้าคู่สูงสุดที่อนุญาตให้มีได้ และปริมาณนิวคลีโอไทด์ที่แทรกหรือหายไปสูงสุดที่อนุญาตให้มีได้เหมือนกันในโปรแกรมทั้งสอง (ตารางที่ 3) อย่างไรก็ตามแพคเกจ Rsubread ให้เปอร์เซ็นต์ลำดับนิวคลีโอไทด์ที่จัดเรียงกับจีโนมสำเร็จสูงกว่าแพคเกจ QuasR ($p<0.001$, Student-t-test) (ตารางที่ 4) โดยที่ให้คุณภาพของข้อมูลหลังจัดเรียงสำเร็จไม่แตกต่างกัน เนื่องจากผลการตรวจสอบคุณภาพข้อมูลทั้งหมดมีปริมาณมาก ณ ที่นี้จึงขอยกผลการตรวจสอบคุณภาพของตัวอย่างรหัส ERR371789 เพียงบางส่วน ดังแสดงในภาพที่ 1

ตารางที่ 2 เปรียบเทียบขั้นตอนการสั่งงานโปรแกรมและระยะเวลาในการประมวลผลระหว่างแพคเกจ QuasR และ Rsubread

ข้อพิจารณาเปรียบเทียบ	ชนิดโปรแกรมแพคเกจ	
	QuasR	Rsubread
ความต้องการการจัดเตรียมข้อมูลสำหรับโปรแกรมแพคเกจ	ต้องการ	ต้องการ
ความต้องการการจัดเตรียมไฟล์ genome index เฉพาะสำหรับโปรแกรมแพคเกจ จากไฟล์จีโนม	ไม่ต้องการ	ต้องการ
ความต้องการการเขียนชุดคำสั่งวน (loop) ให้ดำเนินการจัดเรียงลำดับนิวคลีโอไทด์ของแต่ละตัวอย่างกับจีโนม	ไม่ต้องการ	ต้องการ
ความต้องการการเขียนชุดคำสั่งเพื่อตรวจสอบเปอร์เซนต์ลำดับนิวคลีโอไทด์ที่จัดเรียงกับจีโนมสำเร็จในแต่ละตัวอย่างภายหลัง	ไม่ต้องการ	ต้องการ
ความต้องการจำนวนขั้นตอนทั้งหมด	2	4
ระยะเวลาที่ใช้ในการประมวลผล	12 ชั่วโมง 56 นาที	6 ชั่วโมง 22 นาที

ตารางที่ 3 แสดงข้อเปรียบเทียบเครื่องมือและค่าพารามิเตอร์ระหว่างโปรแกรมแพคเกจ QuasR และ Rsubread

ข้อพิจารณาเปรียบเทียบ	ชนิดโปรแกรมแพคเกจ	
	QuasR	Rsubread
เครื่องมือทางชีวสารสนเทศที่ใช้	Bowtie	subread
Phred score cut-off	33	33
ปริมาณตำแหน่งสูงสุดที่นิวคลีโอไทด์แต่ละเส้นจับกับจีโนมที่ถูกรายงาน (maximal number of equally-best mapping locations)	1	1
ปริมาณนิวคลีโอไทด์ไม่เข้าคู่สูงสุดที่อนุญาตให้มีได้ (maximal mismatched bases)	3	3
ปริมาณนิวคลีโอไทด์ที่แทรกหรือหายไปสูงสุดที่อนุญาตให้มีได้ (maximal insertions/deletions)	5	5

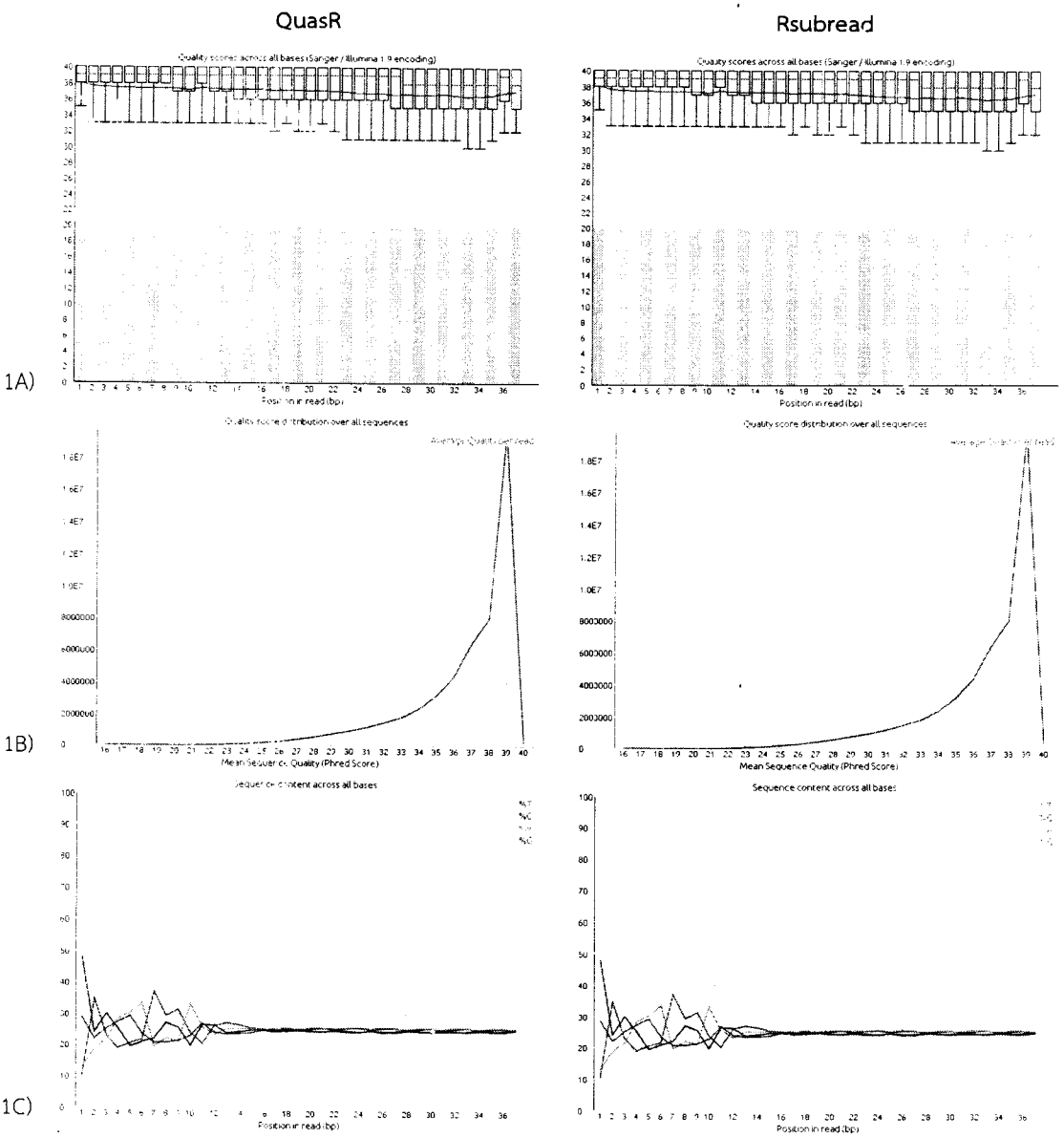
ตารางที่ 4 แสดงปริมาณลำดับนิวคลีโอไทด์ที่จัดเรียงกับจีโนมที่ได้จากโปรแกรมแพคเกจ QuasR และ Rsubread

ไฟล์ตัวอย่าง	จำนวนลำดับนิวคลีโอไทด์ทั้งหมด	จำนวนลำดับนิวคลีโอไทด์ที่จัดเรียงกับจีโนมสำเร็จ		เปอร์เซนต์ลำดับนิวคลีโอไทด์ที่จัดเรียงกับจีโนมสำเร็จ	
		QuasR	Rsubread	QuasR	Rsubread
ERR371789.fastq.gz	52542136	40366252	46950750	76.82644	89.35828
ERR 371790.fastq.gz	45147168	34615820	40326420	76.67329	89.32215
ERR 371791.fastq.gz	42296304	32478784	37557638	76.7887	88.7965
ERR 371792.fastq.gz	45552109	34885861	40644624	76.58451	89.22666
ERR 371793.fastq.gz	47351395	36548783	42323868	77.18629	89.38252
ERR 371794.fastq.gz	56196638	43040040	50181604	76.58828	89.29645
ERR 371795.fastq.gz	40486522	31002878	36041125	76.5758	89.02006
ERR 371796.fastq.gz	51444090	39389960	45925495	76.56848	89.27264
เฉลี่ยช่วงเปอร์เซ็นต์ลำดับนิวคลีโอไทด์ที่จัดเรียงกับจีโนม (Percent alignment) (mean±SD)				76.72±0.21	39.21±0.20

4. วิจารณ์ผลการทดลอง

แพคเกจ QuasR ใช้ Bowtie (Langmead, *et al.*, 2009) ซึ่งเป็นเครื่องมือทางชีวสารสนเทศในการจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมโดยอาศัยการบ่งชี้ตำแหน่งในจีโนมด้วยวิธี BWT (Burrows-Wheeler space transformation) เพื่อลดพื้นที่หน่วยความจำที่ต้องใช้ในการปฏิบัติงาน ในขณะที่แพคเกจ Rsubread ใช้ Subread เป็นเครื่องมือทางชีวสารสนเทศในการจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมโดยอาศัยวิธี seed-and-vote เพื่อให้สามารถจัดเรียงลำดับนิวคลีโอไทด์ได้อย่างรวดเร็ว (Liao, *et al.*, 2013) ผลจากการศึกษาแสดงให้เห็นว่าแพคเกจ QuasR มีจุดเด่นที่ความสะดวกในการใช้งานมากกว่า Rsubread เนื่องจากสามารถเขียนคำสั่งให้โปรแกรมสามารถจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมครบทุกขั้นตอน โดยใช้คำสั่ง qAlign และไม่ต้องการการเขียนชุดคำสั่งวนเพื่อใช้งาน (ตารางที่ 2) นอกจากนี้ผลที่ได้จากคำสั่ง qAlign ยังจัดเก็บผลการวิเคราะห์ในรูปแบบโปรเจค (qProject) ซึ่งสะดวกต่อการ ตรวจเปอร์เซ็นต์ของลำดับนิวคลีโอไทด์ที่จัดวางกับจีโนมสำเร็จ (Percent alignment) และการตรวจสอบคุณภาพ รวมถึงการวิเคราะห์ความแตกต่างทางสถิติในลำดับต่อไป

ผลจากการการจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมแสดงให้เห็นว่าโปรแกรมแพคเกจ Rsubread ใช้เวลาในการจัดเรียงลำดับนิวคลีโอไทด์น้อยกว่า QuasR (ตารางที่ 2) และให้ปริมาณลำดับนิวคลีโอไทด์ที่สามารถจัดเรียงกับจีโนมได้สำเร็จสูงกว่า QuasR อย่างมีนัยสำคัญทางสถิติ (ตารางที่ 4) โดยที่คุณภาพข้อมูลที่ได้หลังจากการจัดเรียงไม่แตกต่างกัน (ภาพที่ 1) ด้วยเหตุนี้โปรแกรมแพคเกจ Rsubread จึงมีประสิทธิภาพในการจัดเรียงลำดับนิวคลีโอไทด์มากกว่า QuasR ในการศึกษาครั้งนี้ อย่างไรก็ตามการเปลี่ยนแปลงข้อมูลลำดับนิวคลีโอไทด์ที่นำมาวิเคราะห์และค่าพารามิเตอร์ต่างๆ ในกระบวนการจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมสามารถส่งผลต่อเวลาและปริมาณลำดับนิวคลีโอไทด์ที่สามารถจัดเรียงกับจีโนมได้สำเร็จได้ ดังนั้นความแตกต่างในประสิทธิภาพของโปรแกรมแพคเกจทั้งสองที่เกิดขึ้นในการศึกษาครั้งนี้จึงอยู่ภายใต้ข้อสรุปการใช้ค่าพารามิเตอร์พื้นฐานตามที่แพคเกจแต่ละชนิดระบุ ซึ่งเหมาะสมกับผู้ใช้งานโปรแกรมในระดับทั่วไป อย่างไรก็ตามผู้ใช้งานโปรแกรมจึงควรคำนึงถึงปัจจัยดังกล่าวประกอบการเลือกใช้งานโปรแกรมอย่างเหมาะสม และควรหมั่นตรวจสอบการเปลี่ยนแปลงของแพคเกจแต่ละชนิดอย่างสม่ำเสมอ เพื่อการใช้งานให้เกิดประสิทธิภาพสูงสุด



ภาพที่ 1 แสดงผลการประเมินคุณภาพของไฟล์ ERR371789 หลังผ่านการจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมด้วยโปรแกรมแพคเกจ QuasR และ Rsubread แล้วบันทึกในรูปของ Binary version of SAM file (BAM) โดยทำการเปรียบคุณภาพของนิวคลีโอไทด์แต่ละลำดับจากลำดับที่ 1 ถึงลำดับที่ 36 (1A) คุณภาพของลำดับนิวคลีโอไทด์ทุกเส้นที่ได้รับการเข้าคู่กับจีโนมในรูป Phred-score (1B) และเปอร์เซ็นต์นิวคลีโอไทด์แต่ละชนิดจากลำดับที่ 1 ถึงลำดับที่ 36 (1C)

5. สรุปผลการทดลอง

โปรแกรมแพคเกจ QuasR มีจุดเด่นในเรื่องความสะดวกในการใช้งาน เนื่องจากต้องการชุดคำสั่งในการปฏิบัติการน้อยและไม่ต้องการการเขียนชุดคำสั่งวนในขณะที่โปรแกรมแพคเกจ Rsubread มีประสิทธิภาพในการจัดเรียงลำดับนิวคลีโอไทด์กับจีโนมได้รวดเร็วและให้เปอร์เซ็นต์การจัดเรียงนิวคลีโอไทด์สำเร็จสูงกว่า QuasR อย่างไรก็ตามผู้ใช้งานโปรแกรมพึงคำนึงถึงข้อมูลดิบที่นำมาวิเคราะห์และการตั้งค่าพารามิเตอร์ประกอบด้วยเสมอ

5. กิตติกรรมประกาศ

งานวิจัยครั้งนี้ได้รับการสนับสนุนจากทุนอุดหนุนการวิจัยประจำปี 2558 จากเงินงบประมาณรายจ่าย (เงินรายได้) ของมหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก เพื่อดำเนินงานโครงการวิจัยเรื่อง การเทียบเคียงคุณสมบัติของเซลล์เม็ดเลือดขาวชนิดนิวเคลียสเดี่ยวระหว่างมนุษย์และสุกรผ่านการแสดงออกของจีโนมเพื่อประเมินการใช้สุกรสำหรับการศึกษาภูมิคุ้มกันทางการแพทย์

6. เอกสารอ้างอิง

Ertl, R., and D. Klein. 2014. Transcriptional profiling of the host cell response to feline immunodeficiency virus infection. **Virol J.** 11: 52.

Field, D., B. Tiwari, T. Booth, S. Houten, D. Swan, N. Bertrand, and M. Thurston. 2006. Open software for biologists: from famine to feast. **Nat Biotechnol** 24: 801-803.

Gaidatzis, D., A. Lerch, F. Hahne and M.B. Stadler. 2014. QuasR: quantification and annotation of short reads in R. **Bioinformatics**.

Gentleman, R.C., V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge and Gentry. 2004. Bioconductor: open software development for computational biology and bioinformatics. **Genome Biol.** 5: R80.

Guan, D.G., J.Y. Liao, Z.H. Qu, Y. Zhang and L.H. Qu. 2011. mirExplorer: detecting microRNAs from genome and next generation sequencing data using the AdaBoost method with transition probability matrix and combined features. **RNA Biol.** 8: 922-934.

Langmead, B., C. Trapnell, M Pop and S.L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Biol.** 10: R25.

Liao, Y., G.K. Smyth and W. Shi. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. **Nucleic Acids Res.** 41: e108.

Xiao, L., J. Zhang, P. Sirois, N. He and K. Li. 2011. A new strategy for next generation sequencing: merging the Sanger's method and the sequencing by synthesis through replacing extension. **J. Biomed Nanotechnol.** 7: 568-571.

Yamey, G. 2000. Scientists unveil first draft of human genome. **BMJ** 321: 7.

Zhang, W., J. Chen, Y. Yang, Y. Tang, J. Shang and B. Shen. 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. **PLoS One** 6: e17915.

Zhou, X., L. Ren, Y. Li, M. Zhang, Y. Yu and J. Yu. 2010. The next-generation sequencing technology: a technology review and future perspective. **Sci China Life Sci.** 53: 44-57.