การวิเคราะห์แบ่งกลุ่มและการค้นหาโมดูลการทำงานของยีนจากข้อมูลการแสดงออกของยีนในรากมันสำปะหลัง

พรทิพย์ เดชพิชัย[1*] ฟารีดา พึ่งเพียร[2] สิริลักษณ์ สิทธิพูนปราชญา[2] ชื่นชม ศาลิคุปต[3]
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี แขวงบางมด เขตทุ่งครุ กรุงเทพฯ 10140
ตรีนุช สายทอง[4] และ เสาวลักษณ์ กัลปณุลักษณ์[5]
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี แขวงท่าข้าม เขต
บางขุนเทียน กรุงเทพฯ 10150
** Corresponding Author: porntip.dec@kmutt.ac.th*
[1] อาจารย์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
[2] นักศึกษา ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
[3] ผู้ช่วยศาสตราจารย์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
[4] ผู้ช่วยศาสตราจารย์ หลักสูตรชีวสารสนเทศและชีววิทยาระบบ คณะทรัพยากรชีวภาพและเทคโนโลยี
[5] รองศาสตราจารย์ หลักสูตรชีวสารสนเทศและชีววิทยาระบบ คณะทรัพยากรชีวภาพและเทคโนโลยี

ข้อมูลบทความ

ประวัติบทความ :
รับเพื่อพิจารณา : 1 มกราคม 2564
แก้ไข : 7 กรกฎาคม 2564
ตอบรับ : 9 สิงหาคม 2564
DOI : 10.14456/kmuttrd.2021.9

คำสำคัญ :
การวิเคราะห์แบ่งกลุ่มยีน /
การแสดงออกของยีน /
มันสำปะหลัง

บทคัดย่อ

มันสำปะหลังเป็นพืชเศรษฐกิจสำคัญทั้งในประเทศไทยและในโลก ความก้าวหน้าทางวิทยาศาสตร์ทำให้ทราบจีโนมของมันสำปะหลัง แต่การที่จะทราบหน้าที่ของยีนจากการวิเคราะห์ทางห้องปฏิบัติการทางชีววิทยาระดับโมเลกุลของพืชนั้นต้องใช้ระยะเวลาและทรัพยากรจำนวนมาก งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อทำนายหน้าที่การทำงานของยีนจากพฤติกรรมการแสดงออกของยีนด้วยการวิเคราะห์แบ่งกลุ่มด้วยวิธี K-means และอนุมานหน้าที่ให้แก่ยีนที่ไม่ทราบหน้าที่ภายในกลุ่มยีนที่โดดเด่นด้วย Gene Set Enrichment Analysis (GSEA) ข้อมูลการแสดงออกของยีนที่นำมาศึกษาเป็นข้อมูลการแสดงออกของราก ซึ่งประกอบด้วยเนื้อเยื่อ 3 ชนิด ได้แก่ เนื้อเยื่อรากสะสมอาหาร เนื้อเยื่อรากฝอย และเนื้อเยื่อเจริญปลายราก รวม 8 ตัวอย่าง โดยแบ่งชุดข้อมูลออกเป็นข้อมูลย่อย 2 ชุด ได้แก่ (1) เนื้อเยื่อรากฝอยและเนื้อเยื่อเจริญปลายราก และ (2) เนื้อเยื่อรากสะสมอาหาร เนื้อเยื่อรากฝอย และเนื้อเยื่อเจริญปลายราก พบว่า สามารถแบ่งกลุ่มรูปแบบการแสดงออกออกเป็น 21 และ 20 กลุ่มตามลำดับ ซึ่งมีเพียง 14 กลุ่มที่สามารถหาหน้าที่ที่โดดเด่นของยีนให้แต่ละกลุ่มได้ในชุดข้อมูลทั้ง 2 ชุด ทำให้สามารถทำนายหน้าที่ของยีนให้แก่ยีนที่ไม่ทราบหน้าที่ได้จำนวน 8,561 และ 8,727 ยีนในชุดข้อมูลที่ 1 และ 2 ตามลำดับ รวมแล้วสามารถทำนายหน้าที่ของยีนได้เพิ่มขึ้น 8,736 ยีน หรือคิดเป็นร้อยละ 26.45 ของยีนทั้งหมดในจีโนมมันสำปะหลัง ผลการทำนายหน้าที่ของยีนด้วยวิธีการดังกล่าวสามารถเพิ่มความสมบูรณ์ของหน้าที่ยีน คิดเป็นร้อยละ 75.38

Clustering and Exploring of Gene Functional Modules from Cassava Root Gene Expression Data

Porntip Dechpichai[1*], Fareeda Puengpien[2], Sirilak Sittipoonprachaya[2], Chunchom Salikupata[3],
King Mongkut's University of Technology Thonburi, Bangmod, Thungkhru, Bangkok 10140
Treenut Saithong[4] and Saowalak Kalapanulak[5]
King Mongkut's University of Technology Thonburi, Tha Kham, Bang Khun Thian, Bangkok 10150

** Corresponding Author: porntip.dec@kmutt.ac.th*
[1] Lecturer, Department of Mathematics, Faculty of Science.
[2] Student, Department of Mathematics, Faculty of Science.
[3] Assistant Professor, Department of Mathematics, Faculty of Science.
[4] Assistant Professor, Bioinformatics and Systems Biology Program, School of Bioresources and Technology.
[5] Associate Professor, Bioinformatics and Systems Biology Program, School of Bioresources and Technology.

Article Info

Article History:
Received: January 1, 2021
Revised: July 7, 2021
Accepted: August 9, 2021
DOI : 10.14456/kmuttrd.2021.9

Keywords:
Gene Clustering Analysis /
Gene Expression /
Cassava

Abstract

Cassava is an important economic crop, both in Thailand and internationally. Advances in sequencing technology have allowed cassava genome to be deciphered. However, identifying the functions of all genes in the cassava genome using plant molecular biology laboratory is a tedious and resource-extensive. The present research therefore aimed to predict gene functions based on their expression profiles using the K-means clustering method and to propose their functions to unknown genes via the use of Gene Set Enrichment Analysis (GSEA). Three tissues of cassava roots, including storage root, fibrous root and root apical meristem, were used in the study. The gene expression data were divided into 2 subsets, which are SET1: fibrous root and root apical meristem and SET2: storage root, fibrous root and root apical meristem. Cassava genes could be divided into 21 groups and 20 groups, respectively; however, only 14 groups can be assigned the significant functions in both subsets. 8,561 and 8,727 unknown genes can be assigned the functions in SET1 and SET2, respectively. Totally, putative related functions can be assigned to 8,736 cassava genes or 26.45 percent of all the genes in the cassava genome. The results allow 75.38 percent of the genes in the genome to be assigned with their related functions.

1. บทนำ

มันสำปะหลังเป็นพืชที่มีความสำคัญทางเศรษฐกิจของประเทศไทยโดยมีมูลค่าส่งออกเป็นอันดับสองรองจากข้าว [1] เนื่องจากมันสำปะหลังสามารถนำมาใช้ประโยชน์ได้ทุกส่วนตั้งแต่ยอดจนถึงราก และสามารถนำไปแปรรูปเป็นผลิตภัณฑ์ต่าง ๆ เพื่อเป็นอาหารของมนุษย์และสัตว์ จึงทำให้มันสำปะหลังเป็นอาหารหลักของประชากรโลกอีกด้วย และในปัจจุบันนี้ความต้องการอาหารนั้นมีอัตราการเพิ่มขึ้นอย่างต่อเนื่อง เนื่องจากประชากรโลกเพิ่มขึ้น แต่ผลผลิตทางการเกษตรมีน้อยลงอันเนื่องมาจากเปลี่ยนแปลงสภาพภูมิอากาศ [2] จึงทำให้เกิดการวิจัยเพื่อพัฒนาผลผลิตทางการเกษตรให้เพิ่มสูงขึ้น เพื่อเรียนรู้และเข้าใจกลไกทางชีววิทยาภายในเซลล์ของมันสำปะหลัง ซึ่งอาจนำไปสู่การพัฒนาสายพันธุ์ หรือการปรับเปลี่ยนวิธีการดูแลรักษาให้เหมาะสมเพื่อลดค่าใช้จ่ายในการเพาะปลูกของเกษตรกรและให้ผลผลิตอย่างยั่งยืน [3]

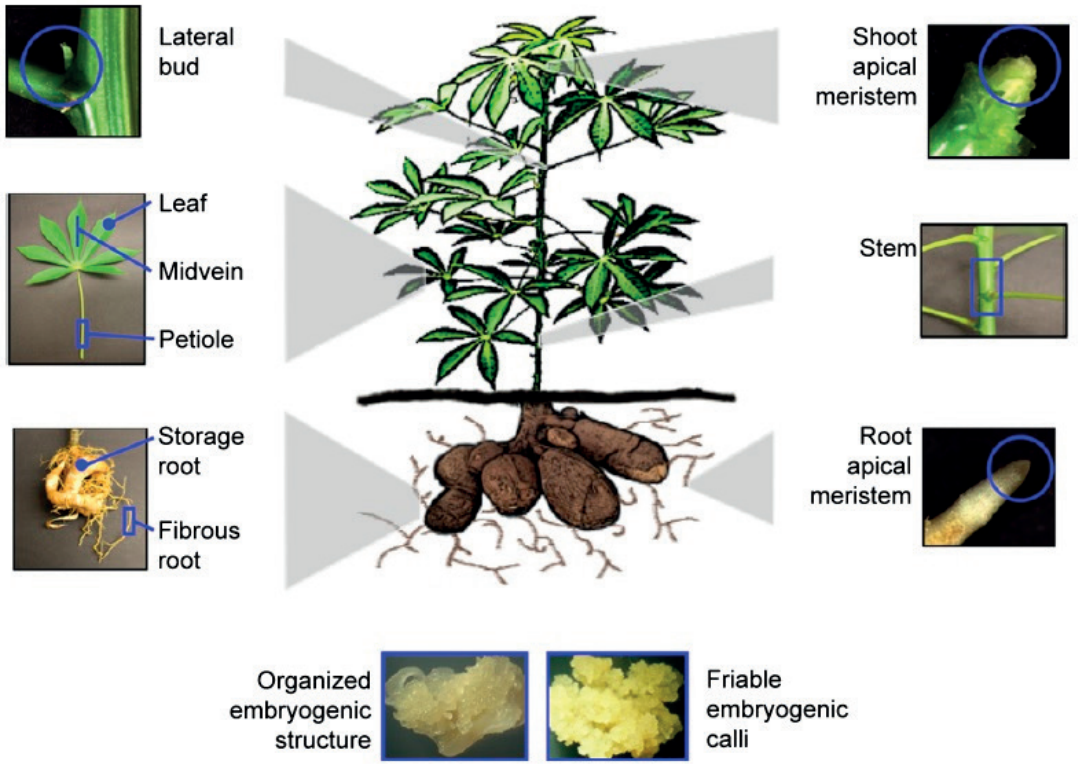
ในปัจจุบันได้มีเทคโนโลยีที่ใช้ศึกษาการทำงานในระดับเซลล์ โดยวัดการทำงานของยีนหรือการแสดงออกของยีนในระดับชีวโมเลกุล ซึ่งสามารถวัดการแสดงออกของยีนได้หลายยีนพร้อมกัน ทำให้ได้ข้อมูลการแสดงออกของยีนจำนวนมาก [4] เพื่อทำความเข้าใจกลไกการทำงานภายในของสิ่งที่กำลังศึกษา แต่อย่างไรก็ตามยังมียีนอีกจำนวนมากที่ยังไม่สามารถระบุหน้าที่การทำงานของมันได้ เช่น ในจีโนมของมันสำปะหลังนั้นพบยีนที่สามารถระบุหน้าที่ได้เพียงร้อยละ 49 [3, 5] และหน้าที่การทำงานของยีนที่ระบุได้ดังกล่าวนั้นไม่ได้มาจากต้นมันสำปะหลังเองแต่ถูกอนุมานจากพืชต้นแบบอื่นเป็นหลัก โดยการระบุหน้าที่การทำงานของยีนในห้องปฏิบัติการนั้นต้องใช้ระยะเวลาในการศึกษานาน เนื่องจากต้องใช้ทักษะ ความรู้ และระยะเวลาเพาะปลูกนาน [6] ดังนั้นการศึกษาหน้าที่การทำงานของยีนด้วยวิธีวิศวกรรมแบบย้อนกลับ (reverse engineering) จึงได้ถูกนำเสนอและประยุกต์ใช้กับงานทางด้านชีววิทยาเพื่อค้นหาหน้าที่การทำงานของยีนเหล่านั้นจากรูปแบบของการแสดงออกของยีน ณ สภาวะต่างๆ [7]

ดังนั้นงานวิจัยนี้นี้จึงสนใจที่จะค้นหาหน้าที่ของยีนจากการวิเคราะห์รูปแบบการแสดงออกที่สอดคล้องกันโดยการวิเคราะห์แบ่งกลุ่ม (clustering analysis) จากรูปแบบการทำงานของยีนที่จำเพาะต่อเนื้อเยื่อ โดยมีสมมติฐานว่า หากยีนที่มีรูปแบบการแสดงออกของยีนสอดคล้องกันจะทำหน้าที่สอดคล้องกัน ซึ่งจะทำให้สามารถอนุมานหน้าที่ของยีนภายในกลุ่มยีนให้แก่ยีนที่ไม่ทราบหน้าที่ภายในกลุ่มได้ การศึกษาวิเคราะห์ดังกล่าวจะช่วยให้สามารถระบุหน้าที่ของยีนในมันสำปะหลังได้เพิ่มมากขึ้น ซึ่งจะช่วยทำให้เข้าใจถึงกลไกการทำงานภายในของต้นมันสำปะหลัง และอาจนำไปสู่การปรับปรุงพันธุ์และพัฒนาสายพันธุ์มันสำปะหลังให้มีผลผลิตสูงต่อไปในอนาคตได้

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 มันสำปะหลัง

มันสำปะหลังเป็นพืชเศรษฐกิจสำคัญของโลก เนื่องจากเป็นพืชที่ให้ปริมาณแป้งต่อต้นค่อนข้างสูงเมื่อเทียบกับพืชชนิดอื่นๆ การปลูกมันสำปะหลังนั้นจะใช้การปักชำจากลำต้น (stem) ที่ตัดเป็นท่อนๆ หรือเรียกว่าท่อนพันธุ์ ท่อนปักอยู่ในดินจะออกรากเป็นรากฝอย (fibrous root) หลังจากปลูกได้ประมาณ 2 เดือนรากจะค่อยๆ สะสมแป้ง และมีขนาดใหญ่ขึ้น เรียกว่ารากสะสมอาหาร (storage root) หรือหัวมันสำปะหลัง มันสำปะหลังสามารถเก็บเกี่ยวได้หลังจากเพาะปลูกประมาณ 6 - 16 เดือน นอกจากนั้นลักษณะทางกายภาพของมันสำปะหลังยังประกอบไปด้วย เนื้อเยื่อเจริญปลายราก (root apical meristem; RAM) ใบ (leaf) เส้นกลางใบ (midvein) ก้านใบ (petiole) ตาข้าง (lateral bud) เนื้อเยื่อเจริญปลายยอด (shoot apical meristem) เนื้อเยื่อเจริญประเภท Organized Embryogenic Structure (OES) และเนื้อเยื่อเจริญประเภท Friable Embryognic Callus (FEC) (รูปที่ 1) [8] ซึ่งแต่ละเนื้อเยื่อนั้นจะประกอบไปด้วยเซลล์และสารพันธุกรรม



รูปที่ 1 ลักษณะทางกายภาพของต้นมันสำปะหลัง [8]

2.2 จีโนม (genome) และ ยีน (gene) ของมันสำปะหลัง

การวิเคราะห์ข้อมูลจีโนมของสิ่งมีชีวิต เป็นการศึกษาองค์ความรู้ด้านชีวโมเลกุล เพื่อเข้าใจเชิงลึกในระดับเซลล์ โดยจีโนมเป็นข้อมูลทางรหัสพันธุกรรมทั้งหมดที่จำเป็นต่อการดำรงชีวิต ขนาดจีโนมของสิ่งมีชีวิตแต่ละชนิดจะมีความแตกต่างกันขึ้นอยู่กับความซับซ้อนของสิ่งมีชีวิต จีโนมคือชุดของดีเอ็นเอ (DNA) ซึ่งถูกบรรจุอยู่ในนิวเคลียสของทุกเซลล์ ภายในจีโนมของสิ่งมีชีวิตจะประกอบไปด้วยยีนหลายยีน ซึ่งยีนนั้นคือหน่วยพันธุกรรมที่ทำหน้าที่ถ่ายทอดและควบคุมลักษณะทางพันธุกรรม และสามารถถ่ายทอดจากรุ่นสู่รุ่นได้ ยีนจะถูกถอดรหัสเป็นเอมอาร์เอนเอ (mRNA) และแปลรหัสให้เป็นกรดอะมิโนขดเรียงต่อกันเป็นสายยาวเรียกโปรตีน (protein) ซึ่งโปรตีนทำหน้าที่ได้หลายอย่าง ทั้งเป็นโครงสร้างของเซลล์ เป็นตัวรับส่งสัญญาณในการติดต่อสื่อสาร และเป็นเอนไซม์ที่ทำหน้าที่เร่งปฏิกิริยาภายในเซลล์ นอกจากนั้นโปรตีนยังเป็นตัวควบคุมการแสดงออกของยีนที่สามารถตอบสนองต่อสิ่งเร้า การทำงานของยีนหรือโปรตีนมักจะทำงานร่วมกันจะมีการแสดงออกร่วมกัน (gene expression) เพื่อให้เกิดการทำงานอย่างเป็นระบบและมีประสิทธิภาพสูงสุด [9]

การระบุหน้าที่ของยีนนั้นสามารถอธิบายอย่างเป็นระบบด้วย gene ontology (GO) ซึ่งสามารถแบ่งกลุ่ม (term) อย่างกว้างออกเป็น 3 กลุ่ม ได้แก่ (1) กระบวนการทางชีววิทยา (biological processes) (2) องค์ประกอบของเซลล์ (cellular components) (3) หน้าที่ระดับโมเลกุลภายในเซลล์ (molecular functions) [10] และการที่จะทราบหน้าที่ของยีนต่างๆ นั้นต้องใช้ข้อมูลทางห้องปฏิบัติการในการค้นหาหน้าที่ของยีน หรือใช้วิธีการอ้างอิงจากพืชต้นแบบ ได้แก่ พืชอะราบิดอพซิส (Arabidopsis thaliana) ข้าว และข้าวโพด ซึ่งเป็นพืชที่มีการศึกษาหน้าที่การทำงานของยีนมาค่อนข้างมาก โดยพิจารณา

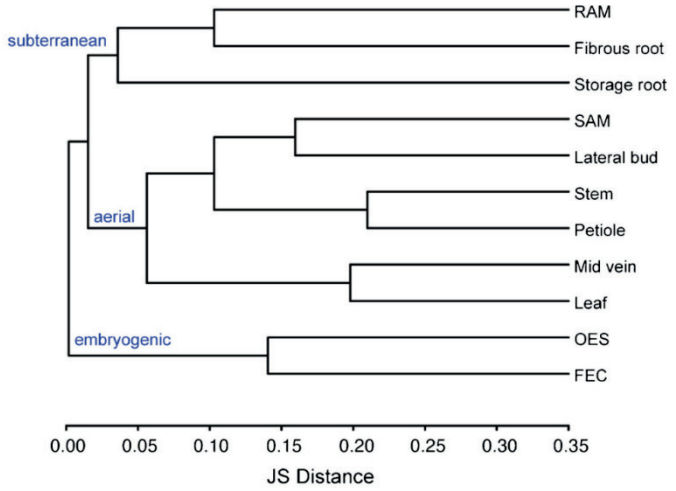
ความเหมือนกันของลำดับกรดอะมิโนระหว่างพืชต้นแบบและมันสำปะหลัง ซึ่งวิธีการนี้เรียกว่า comparative genomics

มันสำปะหลังประกอบไปด้วยโครโมโซมทั้งหมดจำนวน 18 โครโมโซม โดยมีขนาดจีโนมมันสำปะหลังรวม 495.48 ล้านคู่เบส และสามารถทำนายยีนทั้งหมดจำนวน 33,033 ยีนที่สามารถถอดรหัสเป็น mRNA และแปลรหัสเป็นโปรตีนได้ [3, 5] โดยในฐานข้อมูล Phytozome ได้ทำนายหน้าที่การทำงานของยีนผ่านการเปรียบเทียบจีโนมกับพืชต้นแบบ พบยีนที่ทราบหน้าที่มีจำนวน 16,164 ยีน หรือคิดเป็นร้อยละ 49 และยีนที่ไม่ทราบหน้าที่จำนวน 16,869 ยีน หรือคิดเป็นร้อยละ 51 [5]

2.3 การวัดการแสดงออกของยีนในพืชและมันสำปะหลังด้วยเทคโนโลยีการอ่านลำดับทางพันธุกรรม

การวัดค่าการแสดงออกของยีนผ่านเทคโนโลยี DNA sequencing นั้นจะวัดปริมาณของ RNA ของยีนที่มีการแสดงออกในแต่ละตัวอย่าง เรียกว่า RNA-seq ซึ่งสามารถตรวจวัดการแสดงออกในระดับ transcriptome ของ RNA ของทุกยีนภายในเซลล์ได้ [4]

การศึกษาการแสดงออกของยีนภายในแต่ละเนื้อเยื่อของมันสำปะหลังด้วยวิธี RNA-seq ซึ่งประกอบไปด้วยเนื้อเยื่อ 11 ชนิด ได้แก่ เนื้อเยื่อรากสะสมอาหาร เนื้อเยื่อรากฝอย เนื้อเยื่อเจริญปลายราก เนื้อเยื่อใบ เนื้อเยื่อเส้นกลางใบ เนื้อเยื่อก้านใบ เนื้อเยื่อตาข้าง เนื้อเยื่อลำต้น เนื้อเยื่อเจริญปลายยอด เนื้อเยื่อเจริญประเภท OES และเนื้อเยื่อเจริญประเภท FEC พบว่าเนื้อเยื่อแต่ละชนิดมีการแสดงออกที่แตกต่างกัน และเมื่อดูความคล้ายคลึงของลักษณะการแสดงออกของบริเวณใต้ดิน (subterranean) พบว่า เนื้อเยื่อรากฝอย และเนื้อเยื่อเจริญปลายรากนั้นมีความคล้ายคลึงกันมากกว่าเนื้อเยื่อรากสะสมอาหาร (รูปที่ 2) [8]



รูปที่ 2 เดนโดแกรมการวิเคราะห์แบ่งกลุ่มการแสดงออกในแต่ละเนื้อเยื่อโดยคำนวนหาระยะห่างแบบ Jensen–Shannon (JS) ระหว่างเนื้อเยื่อโดยใช้ค่าเฉลี่ยของทั้งสาม biological replicate [8]

2.4 การปรับฐาน (Normalization of gene expression)

เนื่องจากข้อมูลการแสดงออกของยีน ได้แก่ (1) ความยาวของยีน (gene length) แต่ละยีนมีขนาดไม่เท่ากัน (2) จำนวน read ทั้งหมดของแต่ละตัวอย่างไม่เท่ากัน (library sizes หรือ total reads) (3) สัดส่วนปริมาณการแสดงออกของยีนในแต่ละตัวอย่างที่ไม่เท่ากัน (RNA compositions) ทำให้ไม่สามารถนำข้อมูลการแสดงออกของยีนมาเปรียบเทียบ

ได้ทันที จึงต้องมีการปรับฐานให้สามารถเปรียบเทียบกันได้ โดยวิธี Gene length corrected Trimmed Mean of M-value (GeTMM) ซึ่งนำหลักการของการปรับค่ามาตรฐานด้วยวิธี Trimmed Mean of M-value (TMM) ใน EdgeR package มาปรับปรุงโดยการเพิ่มการลดความโน้มเอียงจากความยาวของยีน (reads per kilobase mapped reads; RPM) [11]

2.5 การวิเคราะห์แบ่งกลุ่ม (Cluster Analysis)

การวิเคราะห์แบ่งกลุ่มเป็นเทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning technique) เพื่อจำแนกกลุ่มข้อมูลที่มีคุณลักษณะคล้ายกันอยู่ในกลุ่มเดียวกัน [12] งานวิจัยนี้ได้ใช้การวิเคราะห์แบ่งกลุ่มแบบไม่เป็นขั้นตอน (Partitioning) ด้วยวิธี K-means โดยมีการกำหนดจำนวนกลุ่มด้วยวิธี Silhouette ดังสมการที่ (1-3)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad if \quad |c| > 1 \tag{1}$$

$$a(i) = \frac{1}{|c_i| - 1} \sum_{j \in c_i, i \neq j} d(i, j) \tag{2}$$

$$b(i) = \min \frac{1}{|c_k| - 1} \sum_{j \in c_k} d(i, j) \tag{3}$$

เมื่อ $s(i)$ คือ ค่าเฉลี่ยระยะห่างของ $a(i)$ และ $b(i)$ ซึ่ง $a(i)$ คือ ค่าเฉลี่ยระยะห่างภายในกลุ่มเดียวกัน และ $b(i)$ คือ ค่าเฉลี่ยระยะห่างระหว่างกลุ่ม และ $d(i, j)$ คือ ระยะห่างระหว่างข้อมูล ถ้าค่า $s(i)$ เข้าใกล้ 1 แสดงว่าจำนวนกลุ่มที่กำหนดมีความเหมาะสม แต่ถ้าค่า $s(i)$ เข้าใกล้ -1 มาก แสดงว่าจำนวนกลุ่มที่กำหนดไม่เหมาะสม

หลังจากกำหนดจำนวนกลุ่มแล้ว จะคำนวณหาจุดกึ่งกลางกลุ่มของแต่ละกลุ่ม โดยคำนวณค่า sse (Sum Square Error) ซึ่งคือระยะห่างกำลังสองของแต่ละหน่วยไปยังจุดกึ่งกลางกลุ่ม (c_i) ที่หน่วยนั้นอยู่ ดังสมการที่ (4) ต่อจากนั้นจะพิจารณาการย้ายกลุ่ม โดยจะย้ายหน่วยที่ i ไปยังกลุ่มที่ทำให้ sse มีค่าต่ำที่สุด เมื่อย้ายกลุ่มแล้วจะทำการคำนวณหาจุดกึ่งกลางกลุ่มใหม่ แล้วพิจารณาการย้ายกลุ่มอีกครั้ง แต่ถ้าไม่มีการย้ายกลุ่มอีกแล้ว แสดงว่ากลุ่มที่แบ่งได้นั้นเหมาะสมแล้ว

$$sse = \sum_{i=1}^{n} \left(x_i - \bar{x}_{c_i}\right)' \left(x_i - \bar{x}_{c_i}\right) \tag{4}$$

2.6 การระบุหน้าที่ของยีนที่โดดเด่นภายในกลุ่มยีนโดยวิธี Gene Set Enrichment Analysis (GSEA) ด้วยวิธีการทางสถิติ

การวิเคราะห์โดยวิธี GSEA หรือ functional enrichment analysis เป็นวิธีการหนึ่งที่ช่วยในการระบุหน้าที่ของยีนที่โดดเด่น (over-represented) ภายในชุดยีนหรือโปรตีน [13] โดยวิเคราะห์นัยสำคัญทางสถิติของกลุ่มยีนที่มีความสัมพันธ์กัน โดยความสัมพันธ์ของกลุ่มของยีนอาจเป็นความสัมพันธ์ในลักษณะของการเชื่อมโยงทางชีวภาพที่มีตั้งแต่เริ่มต้น (prior biological pathway) หรือการแสดงออกร่วมของยีน (co-expression) [14] งานวิจัยนี้จึงมีสมมติฐานว่า หากยีนที่มีรูปแบบการแสดงออกของยีนสอดคล้องกันจะทำหน้าที่สอดคล้องกัน ซึ่งจะทำให้สามารถอนุมานหน้าที่ของยีนภายในกลุ่มยีนให้แก่ยีนที่ไม่ทราบหน้าที่ภายในกลุ่มได้ วิธี GSEA ได้ถูกนำมาใช้ในงานวิจัยต่างๆ เช่น การวิเคราะห์ความแตกต่างของชุดยีนในเนื้อเยื่อมะเร็งระยะต่างๆ [15-17] นอกจากนี้ยังใช้ในการศึกษาชุดยีนที่มีลักษณะการแสดงออกที่แตกต่างกันในแต่ละเนื้อเยื่อของมันสำปะหลัง [8] วิธี GSEA จะใช้ค่าความถี่ของยีนที่พบ GO terms ในชุดยีนนั้นเพื่อเปรียบเทียบกับยีนทั้งหมดหรือ background โดยใช้การทดสอบทางสถิติด้วยสถิติทดสอบ

Fisher's exact test, Chi-square test, t-test, Z-score test และ Kolmogorov-Smirnov test [18-19]

หลังจากวิเคราะห์แบ่งกลุ่มแบบ K-means แล้ว จะทำให้ทราบสมาชิกของยีนภายในแต่ละกลุ่มของทั้งสองชุดข้อมูลย่อย ซึ่งจะนำมาวิเคราะห์เพื่อหาหน้าที่โดดเด่นของยีนภายในแต่ละกลุ่มยีน ด้วยวิธี GSEA โดย GO term ของแต่ละยีนภายในกลุ่มยีนจะถูกพิจารณาเพื่อหา GO terms ที่โดดเด่นเมื่อเทียบกับยีนทั้งหมดในจีโนมของมันสำปะหลัง โดยจะกำหนดให้ k เป็นตัวแปรสุ่มแบบไฮเปอร์จีออเมตริก (Hypergeometric random variable) ซึ่งแทนจำนวนครั้งความสำเร็จที่ได้จากการสุ่มหยิบของแบบไม่แทนที่ (without replacement) จำนวน n สิ่งจากของ N สิ่ง ซึ่งประกอบด้วยของ 2 ประเภท โดยมีจำนวน m หน่วย และ N-m หน่วยตามลำดับ โดยมีฟังก์ชันความน่าจะเป็นดังสมการที่ (5) และใช้ GOATOOLS ด้วยการทดสอบ Fisher's exact test โดยใช้การแจกแจงแบบ hypergeometric ซึ่งค่อนข้างจะมีความแม่นยำกว่าตัวทดสอบอื่นๆ และใช้ Benjamini-Hochberg ดังสมการที่ (6) ในการควบคุมความผิดพลาดที่เกิดจากการคำนวณค่า p-value จากการทดสอบหลายครั้ง [18]

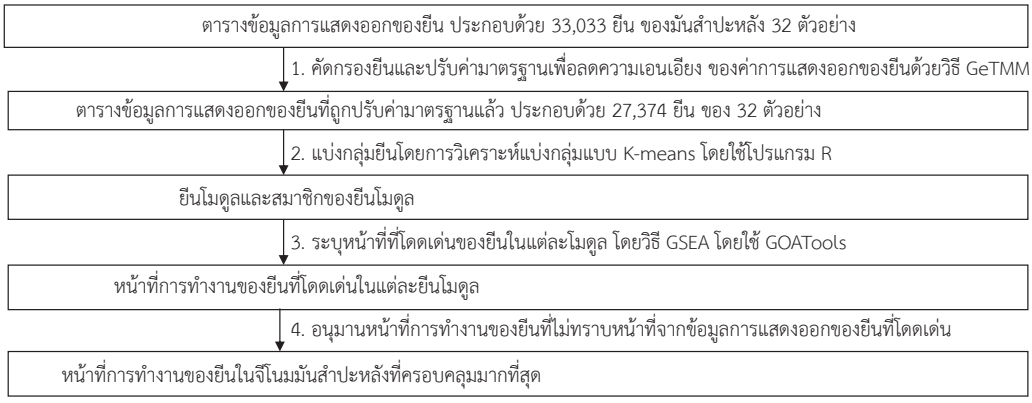
$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \quad (5)$$

$$p - value^{'}s \text{ Benjamini} - \text{Hochberg} = \frac{m}{i} \times Q \quad (6)$$

โดยที่ m = p-value rank (ลำดับค่า p-value ของทุกการทดสอบจากการเรียงแบบน้อยไปมาก) i = จำนวนข้อมูลทั้งหมด และ Q = ค่าความผิดพลาด (0.05)

3. วิธีการวิจัย

งานวิจัยนี้ใช้ข้อมูลการแสดงออกของยีนในมันสำปะหลังที่จำเพาะในแต่ละเนื้อเยื่อจากงานวิจัยของ Wilson et al. [8] ซึ่งศึกษาการแสดงออกของยีนในมันสำปะหลังสายพันธุ์ TME 204 โดยข้อมูลประกอบด้วยยีนทั้งหมด 33,033 ยีน ของเนื้อเยื่อ 11 ชนิด ได้แก่ เนื้อเยื่อรากสะสมอาหาร เนื้อเยื่อรากฝอย เนื้อเยื่อเจริญปลายราก เนื้อเยื่อใบ เนื้อเยื่อเส้นกลางใบ เนื้อเยื่อก้านใบ เนื้อเยื่อตาข้าง เนื้อเยื่อลำต้น เนื้อเยื่อเจริญปลายยอด เนื้อเยื่อเจริญประเภท OES และเนื้อเยื่อเจริญประเภท FEC ซึ่งวัดซ้ำจำนวน 3 ซ้ำ ยกเว้นรากสะสมอาหารที่วัด 2 ซ้ำ รวมทั้งหมด 32 ตัวอย่าง งานวิจัยนี้เลือกใช้เนื้อเยื่อบริเวณของราก 3 ชนิด ได้แก่ เนื้อเยื่อรากสะสมอาหาร เนื้อเยื่อรากฝอย และเนื้อเยื่อเจริญปลายราก รวมทั้งหมด 8 ตัวอย่าง เพื่อทำนายหน้าที่ของยีนจากรูปแบบการแสดงออกของยีน เนื่องจากรากเป็นส่วนที่ให้ผลผลิตและเป็นส่วนสำคัญในการสร้างแป้ง ซึ่งหากทราบหน้าที่บริเวณส่วนของรากจะทำให้สามารถช่วยเพิ่มผลผลิตได้ในอนาคต งานวิจัยนี้แบ่งขั้นตอนหลักออกเป็น 4 ขั้นตอน (รูปที่ 3) ดังนี้



รูปที่ 3 วิธีการศึกษา

3.1 คัดกรองยีนและปรับค่ามาตรฐานเพื่อลดความเอนเอียงของค่าการแสดงออกของยีนด้วยวิธี GeTMM

ข้อมูลตารางค่าการแสดงออกของยีนทั้งหมด 33,033 ยีน ในมันสำปะหลังจำนวน 32 ตัวอย่าง จะถูกคัดกรอง โดยจะตัดยีนที่มีจำนวน read น้อยกว่า 10 ทิ้ง แล้วเลือกยีนที่มีผลรวมทุกตัวอย่างมากกว่า 15 ขึ้นไป และยีนนั้นต้องแสดงออกอย่างน้อย 70% ของจำนวนตัวอย่างทั้งหมด หลังจากคัดกรองยีนในตารางค่าการแสดงออกของยีนแล้วจะนำค่าการแสดงออกของแต่ละยีนมาปรับค่ามาตรฐานเพื่อลดความโน้มเอียงของค่าการแสดงออกเพื่อให้แต่ละตัวอย่างสามารถเปรียบเทียบกันได้ โดยใช้วิธี GeTMM ด้วยฟังก์ชัน filterByExpr ใน Empirical analysis of digital gene expression data in R package (EdgeR) [20-21]

3.2 แบ่งกลุ่มยีนโมดูลโดยการวิเคราะห์แบ่งกลุ่มแบบ K-means

ข้อมูลจะถูกแบ่งออกเป็น 2 ชุด ได้แก่ ชุดที่ 1 ประกอบด้วย เนื้อเยื่อรากฝอย และเนื้อเยื่อเจริญปลายราก จำนวน 6 ตัวอย่าง ส่วนชุดที่ 2 ประกอบด้วย เนื้อเยื่อรากสะสมอาหาร เนื้อเยื่อรากฝอย และเนื้อเยื่อเจริญปลายราก จำนวน 8 ตัวอย่าง ข้อมูลทั้งสองชุดจะถูกนำไปแบ่งกลุ่มยีนจากลักษณะการแสดงออกของยีนแยกกันเพื่อค้นหาหน้าที่ที่โดดเด่นและอนุมานหน้าที่ให้แก่ยีนที่ไม่ทราบหน้าที่ต่อไป การค้นหากลุ่มยีนที่มีค่าการแสดงออกคล้ายคลึงกันด้วยการวิเคราะห์แบ่งกลุ่มแบบ K-means จะต้องมีการกำหนดจำนวนกลุ่มเพื่อทำการวิเคราะห์แบ่งกลุ่ม แต่เนื่องจากไม่ทราบจำนวนกลุ่มที่เหมาะสมได้ งานวิจัยนี้จึงค้นหาจำนวนกลุ่มที่เหมาะสมโดยการวิเคราะห์ Silhouette ด้วยฟังก์ชัน fviz_nbclust และกำหนดจำนวนกลุ่มที่คำนวณสูงสุดเป็น 100 กลุ่ม ซึ่งจะคำนวณค่าเฉลี่ย Silhouette ของแต่ละกลุ่มออกมา หลังจากนั้นจึงหาจำนวนกลุ่มที่เหมาะสมโดยหาจุดหักงอ (breaking point) ระหว่างค่าเฉลี่ย Silhouette เทียบกับจำนวนกลุ่ม ด้วยการลากเส้นตรง 2 เส้นและหาจุดตัด ซึ่งจะได้จำนวนกลุ่มที่เหมาะสมออกมา จากนั้นนำจำนวนกลุ่มที่เหมาะสมไปกำหนดจำนวนกลุ่มให้กับฟังก์ชัน K-means เพื่อหาสมาชิกภายในกลุ่ม

3.3 ระบุหน้าที่ที่โดดเด่นของยีนในแต่ละโมดูลโดยการวิเคราะห์ GSEA

หลังจากทราบสมาชิกของยีนภายในแต่ละกลุ่มของทั้งสองชุดข้อมูลย่อยแล้ว จะนำมาวิเคราะห์เพื่อหาหน้าที่โดดเด่นของยีนภายในแต่ละกลุ่ม ด้วยวิธี GSEA โดย GO term ของแต่ละยีนภายในกลุ่มจะถูกพิจารณาเพื่อหา GO terms ที่โดดเด่นเมื่อเทียบกับยีนทั้งหมดในจีโนมของมันสำปะหลัง โดยใช้หลักการของ Hypergeometric แบบ Fisher's exact test แล้วคำนวณ FDR โดย Benjamini – Hochberg ที่ใช้ควบคุมการผิดพลาดของค่า p-value เพื่อให้การวิเคราะห์หน้าที่มีความถูกต้องมากที่สุดด้วย GOAtools [18]

3.4 อนุมานหน้าที่การทำงานของยีนที่ไม่ทราบหน้าที่จากข้อมูลการแสดงออกของยีนที่โดดเด่น

โดย GO term ที่โดดเด่นทั้งหมดในแต่ละกลุ่มยีนจะถูกกำหนดให้แก่ยีนที่ไม่ทราบหน้าที่ทั้งหมดภายในแต่ละกลุ่ม งานวิจัยนี้จะทำการหาค่าความเชื่อมั่น (confidential score) ของสองชุดข้อมูลว่ามีความเหมือนหรือแตกต่างกันของ GO term ที่ถูกกำหนดให้ของแต่ละชุดข้อมูล โดยคำนวณค่า Jaccard similarity coefficient ดังสมการ (7) โดย set ของ GO term ในชุดข้อมูลที่ 1 แทนด้วย A และ set ของ GO term ในชุดข้อมูลที่ 2 แทนด้วย B ค่า J(A, B) คือค่าที่บ่งบอกจำนวนที่ซ้อนทับกันของชุด GO term ใน A และ B ซึ่งมีค่าอยู่ระหว่าง 0 – 100 โดยหากมีค่าเท่ากับ 0 หมายถึงทั้งสอง set ไม่มี GO term ที่ซ้อนทับกัน หรือหากมีค่าเท่ากับ 100 หมายถึง GO term ทั้งสอง set ซ้อนทับกัน

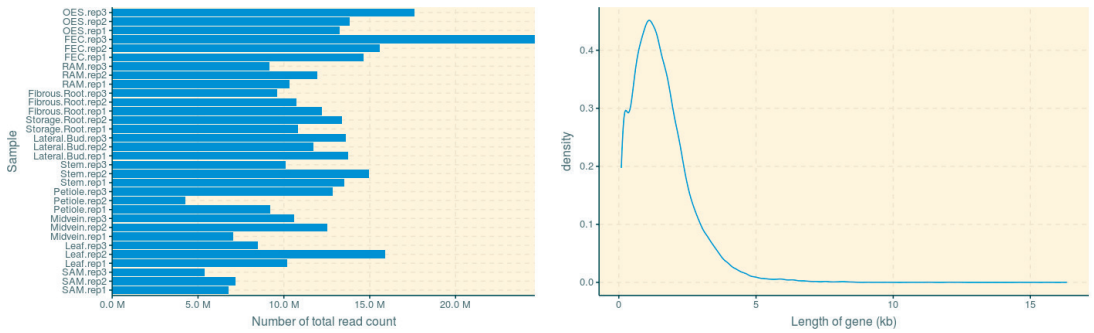
$$J(A,B) = \frac{A \cap B}{A \cup B} \times 100 \quad (7)$$

4. ผลการวิจัย

การศึกษาเพื่อวิเคราะห์แบ่งกลุ่มและค้นหาหน้าที่การทำงานของยีนบริเวณส่วนของรากในจีโนมมันสำปะหลังโดยวิธีการอ้างอิงจากรูปแบบที่จำเพาะของการแสดงออกของยีนในมันสำปะหลัง มีผลการศึกษาดังนี้

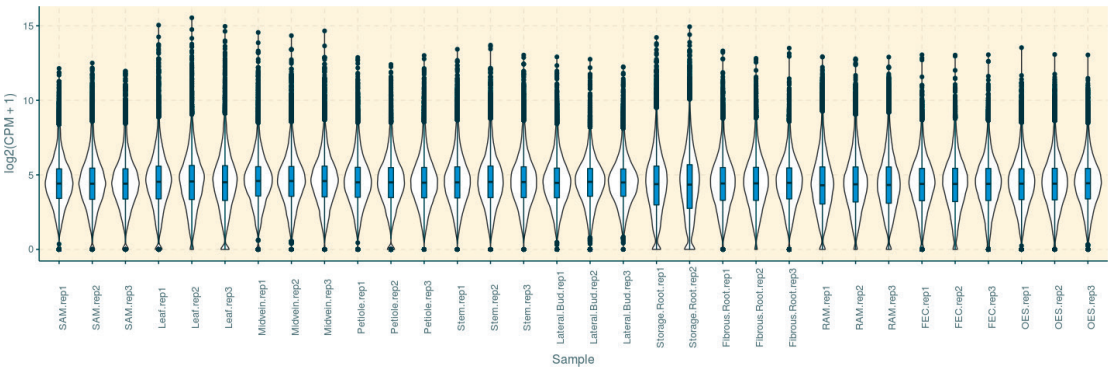
4.1 ลักษณะข้อมูลการแสดงออกของยีนในแต่ละเนื้อเยื่อของมันสำปะหลัง

ข้อมูลการแสดงออกของยีนในเนื้อเยื่อมันสำปะหลังจำนวน 32 ตัวอย่างของเนื้อเยื่อ 11 ชนิด จะมีจำนวน Reads ทั้งหมดในแต่ละตัวอย่างที่ถูกนับบนบริเวณของยีนในจีโนมมันสำปะหลังแตกต่างกัน เช่น เนื้อเยื่อเจริญประเภท friable embryogenic callus (FECrep3) มีจำนวน reads มากกว่าเนื้อเยื่ออื่นๆ (รูปที่ 4(1)) และจะเห็นได้ว่าเส้นโค้งความถี่ของข้อมูลความยาวของยีนเบ้ขวา และสัดส่วนการแสดงออกของยีนในแต่ละเนื้อเยื่อมีสัดส่วนที่แตกต่างกัน (รูปที่ 4(2)) ดังนั้นข้อมูลการแสดงออกของยีนด้วยเทคโนโลยี RNA-sequencing นั้น ไม่สามารถนำมาใช้ในการวิเคราะห์ได้โดยตรง จึงต้องมีการปรับค่ามาตรฐานเพื่อลดความเอนเอียงของค่าการแสดงออกของยีนด้วยวิธี GeTMM เพื่อให้สามารถเปรียบเทียบข้ามตัวอย่างได้ หลังจากปรับค่ามาตรฐานด้วยวิธี GeTMM พบว่าการแสดงออกของเนื้อเยื่อในมันสำปะหลัง สายพันธ์ TME204 จำนวน 33,033 ยีน จากเนื้อเยื่อ 11 ชนิด จำนวน 32 ตัวอย่าง จะมีจำนวนยีนลดลงเหลือ 27,374 ยีน เนื่องจากได้มีการตัดยีนที่ไม่มีค่าการแสดงออกของยีนออก ซึ่งการกระจายตัวของค่าการแสดงออกของยีนในมันสำปะหลังในแต่ละเนื้อเยื่อตัวอย่างหลังจากปรับค่ามาตรฐานด้วยวิธี GeTMM แสดงดังรูปที่ 5 และมีค่าการแสดงออกของยีนโดยเฉลี่ยในแต่ละตัวอย่างอยู่ในช่วง 4.35 – 4.55 log2-transformed ของค่า CPM+1 (รูปที่ 5)



(1) จำนวน Reads ของยีนในจีโนม (2) ความหนาแน่นของความยาวของยีน

รูปที่ 4 ข้อมูลการแสดงออกของยีนในเนื้อเยื่อมันสำปะหลังก่อนปรับค่าด้วยวิธี GeTMM



รูปที่ 5 ข้อมูลการแสดงออกของยีนในเนื้อเยื่อมันสำปะหลังหลังปรับค่าด้วยวิธี GeTMM

4.2 การวิเคราะห์แบ่งกลุ่มจากรูปแบบการแสดงออกของยีน

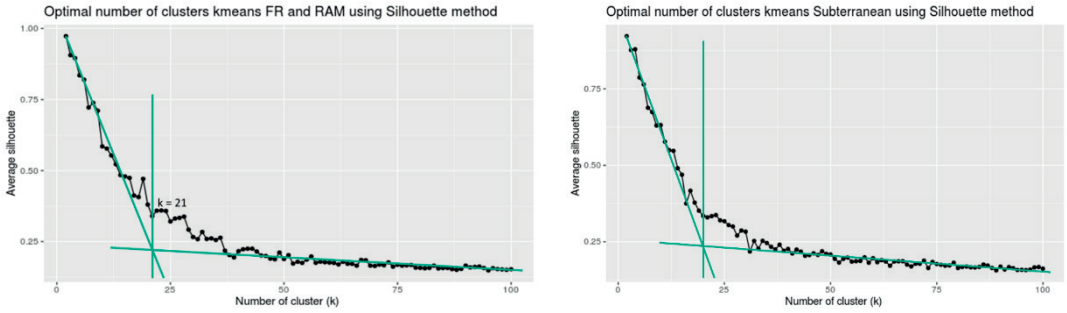
ข้อมูลที่ได้จากการปรับค่ามาตรฐาน คือ เนื้อเยื่อบริเวณของราก 3 ชนิด คือ เนื้อเยื่อรากสะสมอาหาร เนื้อเยื่อรากฝอย เนื้อเยื่อเจริญปลายราก จะถูกแบ่งออกเป็น 2 ชุด คือชุดที่ 1 ประกอบด้วย เนื้อเยื่อรากฝอย และเนื้อเยื่อเจริญปลายราก ส่วนชุดที่ 2 ประกอบด้วย เนื้อเยื่อรากสะสมอาหาร เนื้อเยื่อรากฝอย และเนื้อเยื่อเจริญปลายราก เพื่อนำมาแบ่งกลุ่มข้อมูลด้วยวิธี K-means โดยกำหนดจำนวนกลุ่มที่เหมาะสมด้วยค่า Silhouette พบว่า จำนวนกลุ่มที่เหมาะสมสำหรับข้อมูลชุดที่ 1 คือ 21 กลุ่ม และชุดที่ 2 คือ 20 กลุ่ม (รูปที่ 6) และผลการวิเคราะห์แบ่งกลุ่มสำหรับข้อมูลชุดที่ 1 และ ชุดที่ 2 แสดงดังตารางที่ 1 และ 2 ซึ่งแสดงจำนวนยีนในแต่ละกลุ่ม และรูปที่ 7 เป็นการแสดงลักษณะการแสดงออกของยีน โดยแกน y คือ ค่าการแสดงออก มีหน่วยเป็น CPM+1

ตารางที่ 1 จำนวนยีนสมาชิกภายในกลุ่มที่ถูกจัดโดยการวิเคราะห์แบ่งกลุ่มแบบ **K-means** ของข้อมูลชุดที่ 1

กลุ่มที่	จำนวนยีน	กลุ่มที่	จำนวนยีน	กลุ่มที่	จำนวนยีน	กลุ่มที่	จำนวนยีน	กลุ่มที่	จำนวนยีน
1	6	5	3	9	145	13	19	17	854
2	5,001	6	912	10	69	14	440	18	266
3	9,004	7	1	11	29	15	78	19	126
4	16	8	26	12	9	16	52	20	218
								21	1,930

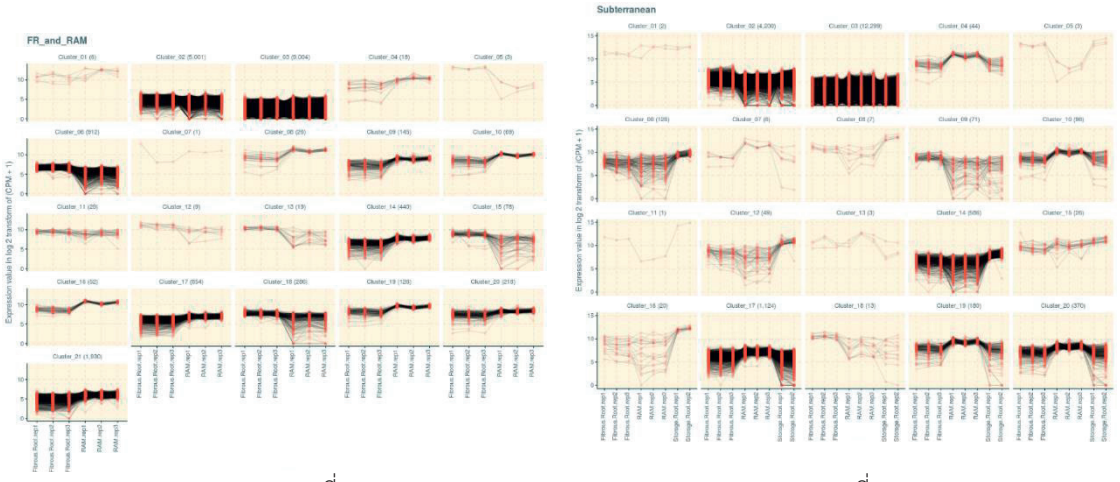
ตารางที่ 2 จำนวนยีนสมาชิกภายในกลุ่มที่ถูกจัดโดยการวิเคราะห์แบ่งกลุ่มแบบ **K-means** ของข้อมูลชุดที่ 2

กลุ่มที่	จำนวนยีน	กลุ่มที่	จำนวนยีน	กลุ่มที่	จำนวนยีน	กลุ่มที่	จำนวนยีน	กลุ่มที่	จำนวนยีน
1	2	5	3	9	71	13	3	17	1,124
2	4,200	6	126	10	96	14	586	18	13
3	12,299	7	6	11	1	15	26	19	180
4	44	8	7	12	49	16	20	20	370



(1) ชุดที่ 1 (1) ชุดที่ 2

รูปที่ 6 ความสัมพันธ์ระหว่างจำนวนกลุ่ม (k) และค่าเฉลี่ย Silhouette



(1) ชุดที่ 1 (1) ชุดที่ 2

รูปที่ 7 รูปแบบการแสดงออกของแต่ละยีน

4.3 การค้นหาหน้าที่โดดเด่นของยีนภายในกลุ่ม

งานวิจัยนี้มีสมมติฐานว่า หากยีนที่มีรูปแบบการแสดงออกของยีนสอดคล้องกันจะทำหน้าที่สอดคล้องกัน ซึ่งจะทำให้สามารถอนุมานหน้าที่ของยีนภายในกลุ่มยีนให้แก่ยีนที่ไม่ทราบหน้าที่ภายในกลุ่มได้ ดังนั้นการอนุมานหน้าที่ของยีนที่ยังไม่ทราบหน้าที่จึงสามารถทำได้ด้วยวิธีการ GSEA เพื่อค้นหา GO term หรือ หน้าที่ของยีนที่โดดเด่นภายในแต่ละกลุ่มยีนโมดูล ซึ่งพบว่ามีเพียง 14 กลุ่ม ในทั้งสองชุดที่สามารถหาหน้าที่ที่โดดเด่นของยีนให้แต่ละกลุ่มได้ ดังตารางที่ 3 และได้แสดงจำนวน GO ที่มีความ enrichment ของแต่ละกลุ่มในประเภทต่างๆ ดังนี้ กระบวนการทางชีววิทยา (biological processes) หน้าที่ของโมเลกุลในเซลล์ (molecular functions) และองค์ประกอบของเซลล์ (cellular components) เช่น ข้อมูลชุดที่ 1 กลุ่มที่ 19 มียีนทั้งหมด 126 ยีน แยกเป็นยีนไม่ทราบหน้าที่ 36 ยีน และทราบหน้าที่ 93 ยีน จาก GO term ที่โดดเด่นในกลุ่มนี้ พบว่าหน้าที่หลักนั้นเกี่ยวกับ sulfur metabolism และ carbohydrate metabolism เช่น GO:0006790: sulfur compound metabolic process, GO:1901137: carbo - hydrate derivative biosynthetic process, GO: 0015986:ATP synthesis coupled proton tran-sport,

GO:0015985: energy coupled proton transport, down electrochemical gradient, GO:0009152: purine ribonucleotide biosynthetic process, GO:0006754: ATP biosynthetic process เป็นต้น

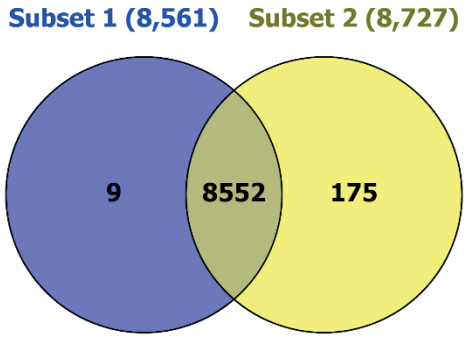
การวิเคราะห์เพื่อค้นหาหน้าที่การทำงานของยีนจากการอนุมานข้อมูลหน้าที่การทำงานของยีนที่โดดเด่นด้วยข้อมูลการแสดงออกของยีนในเนื้อเยื่อ พบว่า สามารถทำนายหน้าที่การทำงานของยีนได้โดยรวมคิดเป็น ร้อยละ 75.38 และพบว่าจำนวนของยีนที่สามารถหาหน้าที่ใหม่เพิ่มด้วยการวิเคราะห์แบ่งกลุ่มแบบ K-means ในชุดที่ 1 และชุดที่ 2 เท่ากับ 8,561 และ 8,727 ยีน ตามลำดับ รวมทั้งหมดเป็น 8,736 ยีน (ตารางที่ 4) และจากรูปที่ 8 พบว่ายีนที่หาหน้าที่เพิ่มใหม่ในสองชุดข้อมูลมีหน้าที่เหมือนกันจำนวน 8,552 ยีน

ตารางที่ 3 จำนวน GO term ของยีนที่โดดเด่นในแต่ละยีนโมดูลของข้อมูลชุดที่ 1 และ 2

No	ชุดที่ 1					ชุดที่ 2				
	กลุ่มที่	จำนวน GO term ที่โดดเด่น				กลุ่มที่	จำนวน GO term ที่โดดเด่น			
		รวม	Biological process	Molecular function	Cellular component		รวม	Biological process	Molecular function	Cellular component
1	2	155	93	27	35	2	240	156	50	34
2	3	150	57	69	24	3	155	60	71	24
3	4	5	0	5	0	4	10	0	5	5
4	6	4	2	0	2	6	34	29	0	5
5	8	36	26	2	8	8	35	25	2	8
6	9	50	27	4	19	9	55	27	6	22
7	10	43	26	3	14	10	43	26	3	14
8	13	1	1	0	0	13	1	1	0	0
9	14	154	80	25	49	14	146	73	34	36
10	16	39	25	3	11	15	1	0	1	0
11	17	182	117	37	28	16	39	25	3	11
12	19	93	68	6	19	17	201	123	41	37
13	20	143	89	27	27	19	93	68	6	19
14	21	146	95	20	31	20	172	102	40	30

ตารางที่ 4 จำนวนและร้อยละของยีนสมาชิกจำแนกตามการทราบหน้าที่

	ชุดที่ 1	ชุดที่ 2	Union sets
จำนวนยีนที่ทราบหน้าที่เดิม	16,164 (48.93 %)		
จำนวนยีนที่สามารถหาหน้าที่ใหม่เพิ่มด้วยการวิเคราะห์แบ่งกลุ่มแบบ K-means	8,561 (25.92 %)	8,727 (26.42 %)	8,736 (26.45 %)
จำนวนยีนที่ทราบหน้าที่ทั้งหมด	24,725 (74.85 %)	24,891 (75.35 %)	24,900 (75.38 %)



รูปที่ 8 แผนภาพเวนน์-ออยเลอร์ ของจำนวนยีนที่สามารถหาหน้าที่ใหม่เพิ่มในชุดที่ 1 และ 2

5. สรุปผลการวิจัย

การอนุมานหน้าที่ให้ยีนที่ไม่ทราบหน้าที่ภายในกลุ่มยีนโดยแบ่งชุดข้อมูลออกเป็น 2 ชุดข้อมูลย่อย ได้แก่ (1) รากฝอยและเนื้อเยื่อเจริญ และ (2) รากสะสมอาหาร รากฝอย และเนื้อเยื่อเจริญปลายราก พบว่า สามารถแบ่งกลุ่มรูปแบบการแสดงออกด้วยการวิเคราะห์แบ่งกลุ่มแบบ K-means ได้ 21 และ 20 กลุ่มตามลำดับ ซึ่งมีเพียง 14 กลุ่ม ในทั้งสองชุดที่สามารถหาหน้าที่ที่โดดเด่นของยีนให้แต่ละกลุ่มได้ ทำให้สามารถทำนายหน้าที่ของยีนให้แก่ยีนที่ไม่ทราบหน้าที่ได้เพิ่มขึ้นจำนวน 8,561 และ 8,727 ยีน ในชุดข้อมูลที่ 1 และ 2 ตามลำดับ ซึ่งรวมแล้วสามารถทำนายหน้าที่ของยีนได้เพิ่มขึ้น 8,736 หรือคิดเป็นร้อยละ 26.45 ของยีนทั้งหมดในจีโนมมันสำปะหลัง ผลการทำนายหน้าที่ของยีนด้วยวิธีการดังกล่าวสามารถทำให้ทราบหน้าที่ของยีนคิดเป็นร้อยละ 75.38 จะเห็นได้ว่าผลการวิจัยนี้แสดงให้เห็นว่าเทคนิคการจัดการข้อมูลด้วยการวิเคราะห์แบ่งกลุ่มยีนสามารถนำมาใช้เพื่อที่จะทำนายหน้าที่ที่เกี่ยวข้องให้ยีนที่ไม่ทราบหน้าที่ได้ซึ่งจะช่วยลดเวลาและค่าใช้จ่ายเมื่อเปรียบเทียบกับการระบุหน้าที่การทำงานของยีนในห้องปฏิบัติการ [21]

6. กิตติกรรมประกาศ

ขอขอบคุณห้องปฏิบัติการชีววิทยาระบบและชีวสารสนเทศและศูนย์นวัตกรรมซอฟต์แวร์และการประมวลผล (Innosoft) มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี สำหรับคำแนะนำและการสนับสนุนการดำเนินการวิจัยครั้งนี้

7. เอกสารอ้างอิง

1. Office of Agricultural Economics, 2020, Thailand Foreign Agricultural Trade Statistics 2019 [Online], Available: http://www.oae.go.th/ assets/portals/1/ files/journal/ 2563/trade st-at62.pdf. (In Thai)

2. Food and Agriculture Organization of the United Nations (FAO), 2017, The Future of Food and Agriculture: Trends and Challenges [Online], Available: http://www.fao.org/3/-a-i6881e.pdf.

3. Bredeson, J.V., Lyons, J.B., Prochnik, S., Wu, G.A., Ha, C.M., Ha, C.M., Edsinger-Gonzales, E., Edsinger-Gonzales, E., Grimwood, J., Schmutz, J., Rabbi, I.Y., Egesi, C., Nauluvula, P., Lebot, V., Ndunguru, J., Mkamilo, G.S., Bart, R., Setter, T.L., Gleadow, R. M., Kulakow, P., Ferguson, M., Rounsley, S., Rokhsar, D.S., Rokhsar, D.S. and Rokhsar, D.S., 2016, "Sequencing Wild and Cultivated Cassava and Related Species Reveals Extensive Interspecific Hybridization and Genetic Diversity," *Nature Biotechnology*, 34 (5), pp. 562-570.

4. Mackenzie, R., 2018, RNA-seq: Basics, Applications and Protocol [Online], Available: https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applica-tions-and-protocol-299461.

5. Goodstein, D., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N.H. and Rokhsar, D.S., 2012, "Phytozome: a Comparative Platform for Green Plant Genomics," *Nucleic Acids Research*, 40, pp. 1178-1186. https://doi.org/10.1093/nar/gkr944

6. Wong, D.C., Sweetman, C. and Ford, C.M., 2014, "Annotation of Gene Function in Citrus Using Gene Expression Information and Co-expression Networks," *BMC Plant Biology*, 14 (1): 186. https://doi.org/10.1186/1471-2229-14-186

7. Villaverde, A.F. and Banga, J.R., 2014, "Reverse Engineering and Identification in Systems Biology: Strategies, Perspectives and Challenges," *Journal of the Royal Society Interface*, 11: 20130505. https://doi.org/10.1098/rsif.2013.0505

8. Wilson, M.C., Mutka, A.M., Hummel, A.W., Berry, J., Chauhan, R.D., Vijayaragha-van, A., Taylor, N.J., Voytas, D.F., Chitwood, D.H. and Bart, R.S., 2017, "Gene Expression Atlas for the Food Security Crop Cassava," *New Phytologist*, 213 (4), pp. 1632-1641. https://doi.org/10.1111/nph.14443

9. Brown, T.A., 2002, Genomes [Online], Available: https://www.ncbi.nlm.nih.gov/ books/NBK21130/.

10. The Gene Ontology Consortium, Gene Ontolog Overview [Online], Available: http://geneontology.org/docs/ontology-documentation/.

11. Smid, M., Coebergh van den Braak, R.R.J., van de Werken, H.J.G., van Riet, J., van Galen, A., de Weerd V., van der Vlugt-Daane, M., Bril, S.I., Lalmahomed, Z.S., Kloosterman, W.P., Wilting, S.M., Foekens, J.A., IJzermans, J.N.M., Martens, J.W.M. and Sieuwerts, A.M., 2018, "Gene Length Corrected Trimmed Mean of M-values (GeTMM) Processing of RNA-seq Data Performs Similarly in Intersample Analyses while Improving Intrasample Comparisons," *BMC Bioinformatics*, 19: 236. https://doi.org/10.1186/s12859-018-2246-7

12. James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013 An Introduction to Statistical Learning: with Applications in R, Springer, New York.

13. Shi, J. and Walker, M.G., 2007, "Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles," *Current Bioinformatics*, 2 (2), pp. 133-137. https://doi.org/10.2174/157489307780618231

14. Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S. and Mesirov, J.P., 2005, "Gene Set Enrichment Analysis: a Knowledge-based Approach for Interpreting Genome-wide Expression Profiles," *Proceedings of the National Academy of Sciences of the United States of America*, 102 (43), pp. 15545-15550. https://doi.org/10.1073/pnas.0506580102

15. Maruschke, M., Hakenberg, O.W., Koczan, D., Zimmermann, W., Stief, C.G. and Buchner, A., 2014, "Expression Profiling of Metastatic Renal Cell Carcinoma Using Gene Set Enrichment Analysis," *International Journal of Urology*, 21 (1), pp. 46-51. https://doi.org/10.1111/iju.12183

16. Wu, B., Li, C., Xie, J., Du, Z., Luo, L., Wu, J., Zhang, P., Xu, L. and Li, E., 2014, "Bioinformatics Analyses of m-RNA Profiling Following Ezrin Knockdown in Esophageal Squamous Cell Carcinoma," *Journal of Cancer Science and Therapy*, 6 (9), pp. 314-321. https://doi.org/10.4172/1948-5956.1000287

17. Yu, Y., Blokhuis, B.R., Garssen, J. and Redegeld, F.A., 2019, "A Transcriptomic Insight into the Impact of Colon Cancer Cells on Mast Cells," *International Journal of Molecular Sciences*, 20 (7): 1689. https://doi.org/10.3390/ijms20071689

18. Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramírez, F., Vesztrocy, A.W., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., Dampier, W.,

Dessimoz, C., Flick, P. and Tang, H., 2018, "GOATOOLS: A Python Library for Gene Ontology Analyses," *Scientific Reports*, 8: 10872. https://doi.org/10.1038/s41598-018-28948-z

19. Huang, D.W., Sherman, B.T. and Lempicki, R.A., 2009, "Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists," *Nucleic Acids Research*, 37 (1), pp. 1-13. https://doi.org/10.1093/nar/gkn923

20. Robinson, M.D., McCarthy, D.J. and Smyth, G.K., 2010, edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data, *Bioinformatics*, 26 (1), pp. 139-140. https://doi.org/10.1093/bioinformatics/btp616

21. Chen, Y., McCarthy, D., Robinson, M. and Smyth, G.K, 2008, edgeR: Differential Expression Analysis of Digital Gene Expression Data User's Guide [Online], Available: http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf.

22. Williams, A. and Halappanavar, S., 2017, "Application of Bi-clustering of Gene Expression Data and Gene Set Enrichment Analysis Methods to Identify Potentially Disease Causing Nanomaterials," *Data in Brief*, 15, pp. 933-940. https://doi.org/10.1016/j.dib.2017.10.060