利用PlantGSAD資料庫平台對農業基因體資料基因群富集分析之簡介

農試所作物組 賴牧謙 吳東鴻 李長沛

一、前言

作物遺傳學的進展，隨著基因定序技術的不斷突破，產生了與生物表現量有關的大量的體學資訊，如轉錄體學、蛋白質體學、微生物體學與代謝體學等重要性狀之遺傳調控數據。在轉錄體學方面，RNA定序 (RNA sequencing, RNA-seq) 分析為了解基因表現量的重要方式之一，透過次世代定序 (next generation sequencing, NGS) 技術的突破大量解序並量化生物體內所有基因或轉錄體的表現狀況，相較於過去的微陣列分析 (microarray) 具有較高的精確度與覆蓋度，其流程包含RNA萃取、互補DNA文庫 (cDNA library) 備製、上機定序與生物資訊分析。在成千上萬的RNA-seq資料中，研究者會根據基因的表現為上調或下調、P值顯著與否設定篩選的門檻，一般而言篩選的門檻為挑選P值小於0.05或倍數變化 (fold change) 大於2的DEGs，側重關注顯著上調或下調的差異性表現基因 (differential expression genes, DEGs)。然而，在生物體中，不同部位與組織對差異性表現量的敏感度不同，這樣的作法容易忽略差異性表現量不顯著，但卻具有重要生物功能的基因。例如轉錄因子，表現量少但一經開啟便可引發一連串下游基因的表達。

由於參與某些功能的基因，經常是一個接著一個序列啟動，因此參與該功能表達的類型或路徑就形成所謂的基因群 (gene-set)，從這些已知和某種功能有關的基因群中找出在試驗中存在差異表現的分析方法，就稱為基因群富集分析 (gene-set enrichment analysis, GSEA)。GSEA不需要特別指定差異性表現的閾值，而是利用統計檢定分析表現量資料的整體趨勢，銜接表現量數據與生物學意義，為分析基因表現量資料開闢出一條新的道路。本篇將介紹論基因富集分析的原理，並以水稻核糖核酸

通訊作者：李長沛副研究員
連絡電話：04-23317177

測序 (RNA sequencing, RNA-seq) 資料為例，利用PlantGSAD線上基因組功能註釋資料庫進行GSEA示範操作，導入視覺化圖表深度剖析複雜且龐大的表現量資料，有助於深化農業科研人員之研究成果。

二、基因群富集分析原理

基因群富集分析的定義，是根據已經建構好的基因組功能註釋資料庫對目標基因群進行功能性分類。當中，最常使用的基因組功能註釋資料庫為基因本體論 (gene ontology, GO) 以及京都基因與基因組百科全書 (Kyoto Encyclopedia of Genes and Genomes, KEGG)。GO 將註釋分為分子生物學功能、生物學過程與細胞學成分。透過此三大分類，對目標基因群的功能進行多層面的描述。KEGG則是一個整合基因組、化學、生理與系統生物學之綜合資料庫，詳細記錄各個基因與蛋白質所參與生物調控途徑。

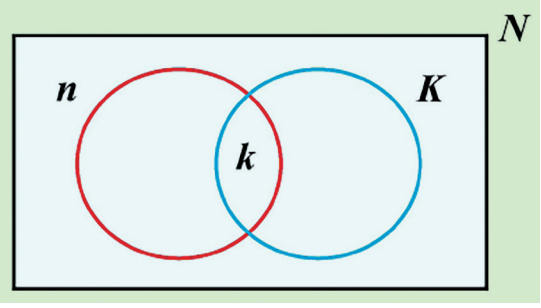
此外，除了挑選基因組功能註釋資料庫，還必須對目標基因群進行基因群富集檢定 (gene-set enrichment tests)，確認基因組功能註釋資料庫中預先定義之特定功能基因群，是否由目標基因群中的DEGs所富集。常見的檢定方法有超幾何檢定、費雪精確性檢定 (Fisher's exact test) 和卡方檢定。最常使用的費雪精確性檢定公式如下：

$$P(X=k)=\frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}}$$

其中，N為基因組功能註釋資料庫總基因個數；n為目標基因群的基因數目；K為特定功能基因群之基因數目；k為目標基因群與特定功能基因群重疊之基因數目。不同基因群之間的關係以文氏圖進行詳細說明，如圖一所示。最後，會利用false discovery rate (FDR) 對費雪精確性檢定所得到的P值進行校正。

三、PlantGSAD資料庫背景介紹

PlantGSAD (http://systemsbiology.cau.edu.cn/PlantGSEAv2/) 為 Ma 等人於2021年所架設之線上基因組功能註釋資料庫，前身為2013所架設的PlantGSEA (http://systemsbiology.cau.edu.cn/PlantGSEA/)。PlantGSAD 採用LAMP系統 (Linux + Apache + MySQL + PHP/Python) 建立線上資料庫，以Linux為底層作業系統，Apache架構網站伺服器，MySQL負責儲存與管理大量生物學資料， PHP/Python負責使用者互動介面與資料存取。由於其免費與開放原始碼之



圖一、基因群富集檢定各基因群間關係文氏圖。其中，N為基因組功能註釋資料庫總基因之集合；n為目標基因群之集合；K為特定功能基因群之集合；k為目標基因群n與特定功能基因群K之交集 (intersection)。

特性，近年來開始將這四個軟體組合在一起架構線上資料庫平台。

PlantGSAD 為近期新發表之高等植物整合型基因富集分析資料庫，其主要特色為彙整過去所有高等植物相關的基因組功能註釋資料庫，蒐集了涵蓋 44 個物種共 236,007 個功能基因集，並依據資料庫的屬性區分成9大類，如表一所示。其中，G1與G2主要為體學資料，如AgriGO(v2)、Phytozome(v13)。G3 主要收錄路徑調控分析資料庫，包含KEGG、MapMan (https://mapman.gabipd.org/imprint) 和 PlantCyc (https://plantcyc.org/)。G4 收錄數個基因家族的資料庫，例如 CAZy database，為醣類分子相關酵素資料庫。G5 則是收錄染色質狀態相關資料，主要資料來源為 PCSD database。前人研究顯示，細胞內的DNA序列會和組蛋白結合，進而構成不同的染色質狀態 (chromatin states)，主要區分成緊密與疏鬆兩種，不同的染色質狀態會影響後續的基因表達。因此，PlantGSAD 將其納入資料庫中並進行整合。G6 方面，收錄了轉錄因子與微核醣核酸相關之資料庫，諸如Plant Cistrome Database與 Plant ncRNA database (http://structuralbiology.cau.edu.cn/PNRD)。G7 收錄基因共表現網絡分析資料所建構之資料庫，如ATTED-II。G8 與 G9 則分別收錄了來自458篇期刊與液-液相分離之相關基因集與資料庫。液-液相分離 (liquid-liquid phase separation) 為2010年後興起的新學門，主要在探討細胞內不具有細胞膜的胞器 (例如核仁) 如何在細胞質內與其他物質區隔，獨立進行生化反應。PlantGSAD 亦將調控液-液相分離的基因資料庫納入，如DrLLPS。

四、PlantGSAD資料庫操作簡介

當使用者手上有一筆植物DEGs資料，便可以利用PlantGSAD對其進行基因富集分析。首先，在單一基因或基因集搜尋的部分，PlantGSAD提供使用者查特定基因在9大類基因註釋資料庫中的功能註解，亦可查詢指定功能基因集(圖二A、B、C)。在基因富集分析方面，選定物種、輸入目標 DEGs 後，可以選擇欲進行分析的富集類別，PlantGSAD允許使用者同時勾選9大類別的富集分析資料庫，以方便使用者比較不同類別間對同一目標差異表現基因集功能上的差異，此項功能是PlantGSAD的一大特色 (圖二D)。最後，PlantGSAD提供視覺化的圖表，包含有向無環樹狀圖 (direct acyclic graphical tree, DAG tree) 與基因-註解重疊矩陣 (圖二E)。DAG tree不同與網絡圖，各節點間具有方向性，但不會形成環狀之構造，適合用來描述註解與註解之間的上下游關係。基因-註解重疊矩陣則是能清楚呈現特定基因是否具有該註解之功能。

本文章以Wu等人於2016年所發表，蒐集抽穗後不同天數之高度休眠性水稻品系N22與輕度休眠性水稻品系Q4359、Q4646之種子進行RNA-seq分析，並利用活性氧化物 (ROS) 處理後

所篩選出的85個P值小於0.05的DEGs，其資料型態包含基因名稱、倍數變化(log2)、表現量上調 (up) 或下調 (down) 及P值，以這85個DEGs進行PlantGSAD基因富集分析之示範操作，探討這85個DEGs與水稻種子休眠性之間的關聯性。第一步，進入ANALYSIS頁面後，勾選欲進行分析的富集資料庫類別 (G1-G9)，選擇物種、匯入候選基因集後便可進行分析。以本次匯入的水稻基因為例，PlantGSAD能接受的基因名稱為LOC_Os01g01010，若基因名稱版本為Os01g0100100，可以利用rap-db水稻資料庫中的ID Converter (https://rapdb.dna.affrc.go.jp/tools/converter) 進行共用編號的轉換。結果方面，本文章以G3為範

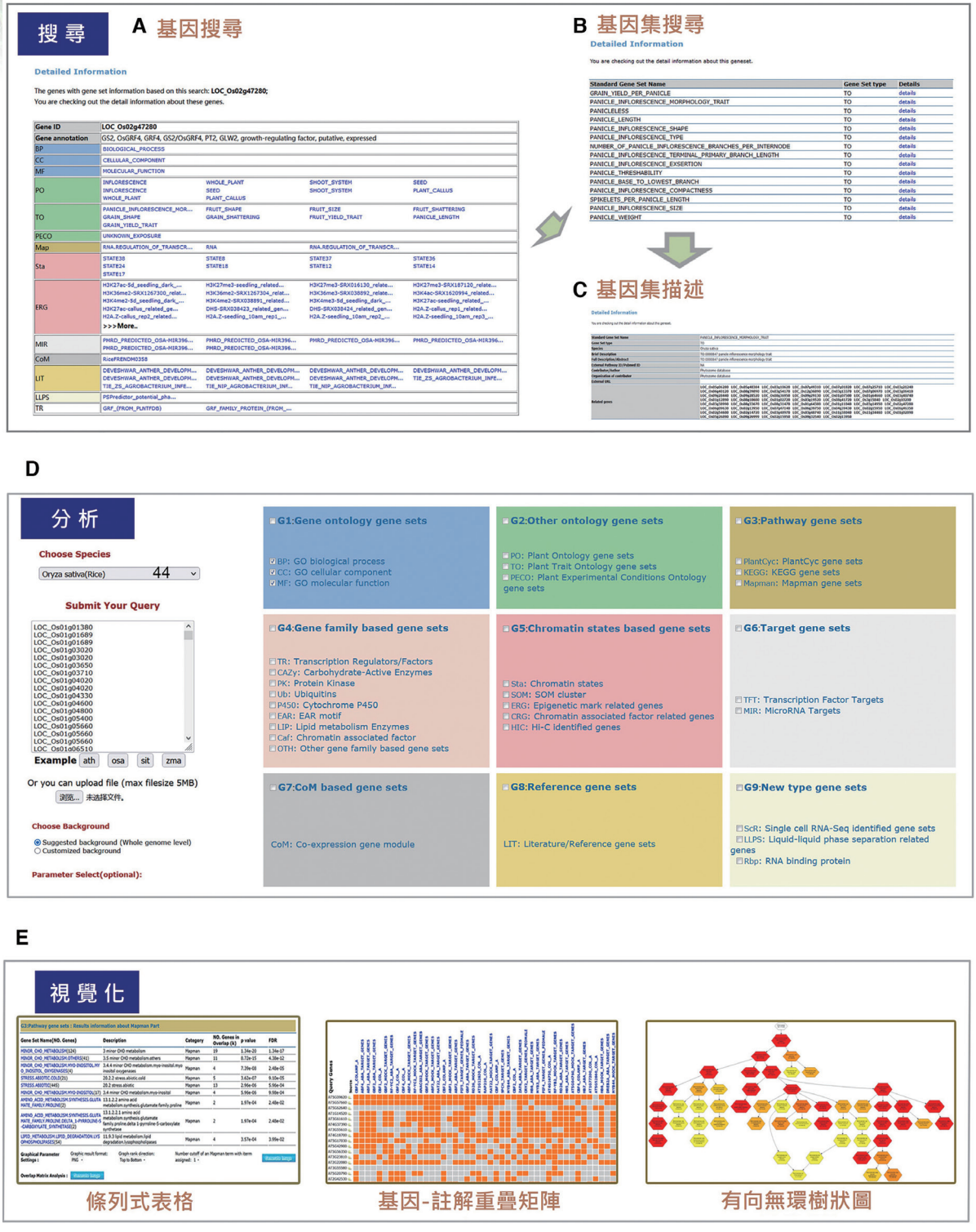
表一、PlantGSAD中9大類別 (G1-G9) 與其包含之物種、基因、功能基因集與主要資料庫來源。引用與修改自Ma *et al.* (2021)

類別	基因集描述	包含物種	包含基因集	包含基因	主要資料庫
G1	基因本體論	44	105,339	878,035	AgriGOv2/Phytozome
G2	其他本體論	4	9,444	65,958	Planteome
G3	路徑調控	39	90,526	1,174,307	KEGG
G4	基因家族	41	8,137	142,393	CAZy database
G5	染色質狀態	5	495	427,194	PCSD database
G6	轉錄因子與微核醣核酸	9	5,454	90,424	Plant Cistrome Database
G7	基因共表現網絡分析	10	14,220	82,020	ATTED-II
G8	期刊文獻	16	1,950	110,503	458 literatures
G9	液-液相分離	31	442	229,904	DrLLPS

表二、利用85個DEGs進行GSEA結果。由左至右詳細列出了85個DEGs參與了KEGG中的哪個功能基因集、該功能基因集共有幾個基因、功能基因集之描述、輸入DEGs與該功能基因集有多少重疊、P值與FDR。

G3:Pathway gene sets : Results information about KEGG Part

Gene Set Name(NO. Genes)	Description	Category	NO. Genes in Overlap (k)	p value	FDR
KEGG_GLUTATHIONE_METABOLISM(55)	ko00480 Glutathione metabolism	KEGG	6	3.56e-10	3.92e-08
KEGG_ARACHIDONIC_ACID_METABOLISM(12)	ko00590 Arachidonic acid metabolism	KEGG	4	6.80e-09	3.74e-07
KEGG_PEROXISOME(76)	ko04146 Peroxisome	KEGG	5	1.20e-07	4.42e-06
KEGG_TRYPTOPHAN_METABOLISM(17)	ko00380 Tryptophan metabolism	KEGG	3	3.10e-06	8.52e-05
KEGG_PHENYLALANINE_METABOLISM(75)	ko00360 Phenylalanine metabolism	KEGG	4	5.25e-06	1.15e-04
KEGG_PHENYLPROPANOID_BIOSYNTHESIS(86)	ko00940 Phenylpropanoid biosynthesis	KEGG	4	8.82e-06	1.62e-04
KEGG_ASCORBATE_AND_ALDARATE_METABOLISM(36)	ko00053 Ascorbate and aldarate metabolism	KEGG	3	2.44e-05	3.83e-04
KEGG_CITRATE_CYCLE_(TCA_CYCLE)(49)	ko00020 Citrate cycle (TCA cycle)	KEGG	3	5.81e-05	7.99e-04
KEGG_GLYOXYLATE_AND_DICARBOXYLATE_METABOLISM(53)	ko00630 Glyoxylate and dicarboxylate metabolism	KEGG	3	7.26e-05	8.87e-04
KEGG_PROTEIN_PROCESSING_IN_ENDOPLASMIC_RETICULUM(193)	ko04141 Protein processing in endoplasmic reticulum	KEGG	4	1.87e-04	2.06e-03
KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION(38)	ko00280 Valine, leucine and isoleucine degradation	KEGG	2	1.50e-03	1.50e-02
KEGG_BASE_EXCISION_REPAIR(42)	ko03410 Base excision repair	KEGG	2	1.82e-03	1.66e-02



圖二、PlantGSAD操作介面簡介。(A) 輸入基因進行搜尋 (B) 輸入基因集進行搜尋 (C) 功能基因集詳細資訊 (D) 輸入目標差異表現基因進行富集分析 (E) 視覺化圖表。引用自與修改自Ma et al. (2021)

例進行介紹。第一項結果如表二所示，85個DEGs顯著參與的調控路徑共有12個，P值由小到大排列前3名分別為穀胱甘肽代謝途徑 (glutathione metabolism)、花生油酸代謝途徑 (arachidonic acid metabolism) 和過氧化體 (peroxisome)。在種子發芽的過程中，會將脂肪分解以提供為呼吸作用的原料，這一連串的代謝途徑中會產生乙醛酸體，脂肪酸進入乙醛酸體進行ß-氧化作用與乙醛酸循環兩個主要生化途徑後，形成四碳酸進入細胞質中進行葡萄糖新生成。然而，當發芽的種子照射到陽光後，子葉逐漸轉變成綠色，此時乙醛酸體會逐漸轉換為過氧化體 (peroxisome)。85個DEGs中有5個基因參與過氧化體的形成 (P-value = 4.2×10-6)，4個基因參與花生油酸代謝途徑 (P-value = 3.74×10-7)，同時出現這兩項代謝途徑足以證明這85個DEGs與種子萌芽與休眠性有一定程度上的關聯。

五、結語

在基因表現量分析中，不論是微陣列或RNA-seq技術，都需要基因富集分析的輔助，對差異表現量基因進行功能性的分群。況且，隨著資訊設備軟硬體的提升，可提供基因富集分析的線上平台選擇與日俱增，挑選適合農業研究人員的基因富集分析平台更顯重要。PlantGSAD 平台為2021年架設之線上基因組功能註釋資料庫，整合過去其他資料庫之高等植物功能性基因注釋資料，並提供圖形化的操作介面與可輸出的視覺化圖表，不失為農業研究人員深化研究成果的一項新選擇。

六、參考文獻

Ma, X. L., H. Y. Yan, J. T. Yang, Y. Liu, Z. Q. Li, M. H. Sheng, Y. X. Cao, X. Y. Yu,
X. Yi, W. Y. Xu, Z. Su. 2021. PlantGSAD: a comprehensive gene set annotation database for plant species. Nucleic Acids Res. gkab794.

Wu, T., C. Y. Yang, B. X. Ding, Z. M. Feng, Q. Wang, J. He, J. H. Tong, L. T. Xiao, L. Jiang, J. M. Wan. 2016. Microarray-based gene expression analysis of strong seed dormancy in rice cv. N22 and less dormant mutant derivatives. Plant Physiol. Biochem. 99:27–38.

Yi, X., Z. Du, Z. Su. PlantGSEA: a gene set enrichment analysis toolkit for plant
community. 2013. Nucleic Acids Res. 41:W1, 98–103.