Version of Record: https://www.sciencedirect.com/science/article/pii/S1369526622000516
Manuscript_94266c35615fea4b3353179f9e455587

# Omics-guided metabolic pathway discovery in plants: resources, approaches, and opportunities

Kangmei Zhao and Seung Y. Rhee

Carnegie Institution for Science, Department of Plant Biology, Stanford, CA, USA

Corresponding: K.Z. (kzhao@carnegiescience.edu)

Keywords: omics, metabolic pathway databases, metabolic gene clusters, novel pathway discovery

## Abstract
Plants produce a vast array of metabolites, the biosynthetic routes of which remain largely undetermined. Genome-scale enzyme and pathway annotations and omics technologies have revolutionized research to decrypt plant metabolism and produced a growing list of functionally characterized metabolic genes and pathways. However, what is known is still a tiny fraction of the metabolic capacity harbored by plants. Here, we review plant enzyme and pathway annotation resources and cutting-edge omics approaches to guide discovery and characterization of plant metabolic pathways. We also discuss strategies for improving enzyme function prediction by integrating protein 3D structure information and single cell omics. This review aims to serve as a primer for plant biologists to leverage omics datasets to facilitate understanding and engineering plant metabolism.

## Introduction
Plants harbor tremendous metabolic diversity, which is essential to cope with many of the world's challenges, including food security, drug development, and ecosystem functioning [1–3]. The social and economic importance of plant metabolites has motivated research to elucidate how they are biosynthesized. Yet, the overall understanding of the genetic basis for plant metabolism is still limited [4]. Plants are predicted to synthesize over 1 million metabolites [5], but only about 0.1% of the biosynthetic pathways have been functionally elucidated [6]. Molecular genetics and analytical biochemistry are classic approaches to uncover the biochemical functions of individual metabolic genes and pathways [7,8]. However, this process tends to be labor intensive and limits the functional study of enzymatic genes to species that are amenable for genetic analysis [8].

Advances in metabolic pathway mapping infrastructure and omics approaches facilitate the discovery of novel pathways in diverse plant species. Several publicly available knowledge bases provide annotations for enzymes and pathways. The widely cited resources for enzyme annotation include UniProt, BRENDA, and Rhea [9–11] (**Table 1**).The infrastructure for metabolic pathway mapping is available at MetaCyc, Kyoto Encyclopedia of Genes and Genomes (KEGG), PlantReactome, and the Plant Metabolic Network (PMN) [12–16] (**Table 1**). Besides resources for genome-scale enzyme annotation and gene-to-pathway mapping, tools are available to predict metabolic gene clusters (MGCs), which are formed by physically co-

localized genes catalyzing reactions in the same biosynthetic pathway [6,17,18]. These resources and tools enable the initial identification of metabolic pathways and gene clusters in the species of interest. Omics technologies (e.g. genomics, transcriptomics, proteomics, and metabolomics) can further enhance characterization of metabolic pathways by prioritizing candidate genes responsible for producing metabolites of interest in certain tissue types and conditions (**Fig. 1**). Here, we discuss recent advances in enzyme and pathway annotation resources and omics approaches to guide novel pathway characterization. First, we summarize widely applied pathway annotation resources and tools that can systematically identify metabolic genes, pathways, and metabolic gene clusters. Then, we discuss cutting-edge approaches for leveraging omics datasets to guide novel pathway discovery. We also provide perspectives on future directions for metabolic gene function prediction to enhance high throughput discovery of pathways important for agriculture, healthcare, and the environment.

**Plant enzyme function annotation and pathway mapping**

Effective omics-guided metabolic pathway discovery requires accessing enzyme annotations and gene-to-pathway mapping infrastructure. The growing list of experimentally characterized enzymes expands the understanding of the genetic basis underlying the biosynthesis of diverse groups of metabolites. Several public databases continue to expand the inventory of enzymes with experimental evidence by curating the data from the literature (**Table 1**) [6,9,11,13,15,16]. These resources are valuable for inferring the function of unknown genes based on sequence homology to generate genome-scale annotations for all enzymes in a species [16].

Once enzyme annotations are established, pathways can be inferred by connecting enzymes involved in making the same metabolite. A widely applied strategy is to assemble enzymes into pathways by searching against “gold standard” databases that contain all experimentally characterized pathways, such as MetaCyc and PlantCyc [4,6,15,16,19]. Then a validation step is followed to compute whether a pathway is present in each species based on the fraction of its constituent reactions predicted in that species. This approach can systematically map enzymes into pathways in sequenced genomes (**Fig. 1A**). The prediction of metabolic gene clusters requires a different strategy to capture physically co-localized metabolic genes that might be involved in a biosynthetic pathway [17,18,20*–22*]. Several tools have been developed to predict MGCs in sequenced plant genomes. Most predict MGCs based on sequence homology to known clusters and local metabolic gene density [20*,22*]. PlantiSMASH represents a homology-based prediction tool, which compares protein sequence similarity for the co-localized metabolic genes within a genomic region to the enzyme families associated with all characterized clusters [23]. In contrast, PlantClusterFinder (PCF) is a *de novo* prediction method based on local enzyme density [6]. PCF scans the genome to find regions that are highly enriched with metabolic genes and the loci containing at least three enzymes catalyzing two different reactions are identified as MGCs. The inferred metabolic pathway and gene clusters can be integrated with omics datasets to further dissect functional patterns of metabolism and accelerate novel pathway discovery.

**Leveraging co-expression and gene-metabolite correlation**

Genes participating in the same biological process often show coordinated expression patterns in response to genetic or environmental perturbations [21,24,25]. Built upon this foundation, co-expression analysis has been widely applied to identify candidate genes associated with pathways of interest (**Fig. 1B**). For example, co-expression analysis facilitates identifying a triterpenoid metabolic network derived from the metabolites produced by known MGCs [26**]. In *Arabidopsis,* four MGCs have been characterized to produce triterpenoids and genes within each cluster show high co-expression [27,28]. Interestingly, several metabolic genes scattered in the genome are also highly co-expressed with terpene synthases in the clusters, but their function remains characterized [26**]. Co-transformation assays in *N. benthamiana* showed that these metabolic genes can use terpenoids produced by the gene clusters as substrates to synthesize new metabolites [26**]. Disrupting the biosynthesis of terpenoid-derived metabolites using *Arabidopsis* mutants resulted in shifted microbe communities in the rhizosphere, which may affect plant-microbe interactions [26**]. This study shows how genes catalyzing novel reactions can be discovered based on co-expression analysis with known pathways.

Besides co-expression, correlation between gene expression and metabolite accumulation has been used to prioritize candidate genes involved in synthesizing metabolites of interest (**Fig. 1B**) [29*]. A prominent example is the discovery of the falcarindiol biosynthesis pathway in tomato. Falcarindiol is a highly modified lipid, which can be induced by different biotic stresses in tomato to promote resistance against fungal and bacterial pathogens [30**]. Based on the chemical structure of falcarindiol, an acetylenase was hypothesized to catalyze early steps of the biosynthesis using linoleic acid as the precursor [31]. To identify the acetylenase involved in falcarindiol biosynthesis, correlation analysis was applied to identify the candidate enzyme whose gene expression pattern showed the highest similarity to falcarindiol accumulation under diverse biotic stress conditions [30**]. The top candidate identified from this analysis showed expected enzymatic activity based on experimental validation in a heterologous system and native plants [30**]. This study demonstrates the advantage of correlation analysis between gene expression and metabolite accumulation in elucidating previously unknown pathways.

**Exploiting metabolic diversity of natural populations**

Genome-wide association studies (GWAS) have been widely applied to dissect genetic architectures underlying phenotypes of interest [32]. Combining GWAS with metabolic profiling facilitates identifying genes underpinning metabolic diversity using the content of metabolites as phenotypic traits [33,34]. Metabolite GWAS (mGWAS) can discover various types of genes, such as transcription factors and biosynthetic genes, associated with a metabolic trait (**Fig. 1C**) [34–38]. For example, a mGWAS was conducted using kernels of seven hundred maize genotypes grown at multiple locations and resulted in over 1,000 associations between genomic loci and metabolic traits. Two candidate genes highly associated with phenolamides, *PHT* (*Putrescine Hydroxycinnamoyl-Transferase*) and *CCoAOMT* (*Caffeoyl-CoA O-MethylTransferase*), were tested for function using mutants in rice and maize. Metabolite composition analysis confirmed that these two enzymes catalyzing early reactions involved in phenolamide biosynthesis using both arginine and putrescine as substrates [35**]. Despite the usefulness of mGWAS, it can be challenging to functionally validate causal genes for the trait of interest as multiple SNPs can be associated with the same metabolic feature [38]. Co-

expression and gene-metabolite correlation analysis can serve as orthogonal approaches to mGWAS to prioritize causal genes and eliminate false positives, which facilitates downstream functional validation using molecular genetics [29*,39].

**Harnessing evolutionary diversification of metabolism**

Plant metabolism diversification arises from gradual modification of existing enzymes, which is mainly caused by gene duplication followed by sub- or neo-functionalization. This can lead to the emergence of new metabolic pathways and metabolites in specific lineages [40–42]. Combining the taxonomic distribution of a metabolite and evolutionary patterns of enzymes help prioritize candidates involved in synthesizing the compound of interest in specific lineages (**Fig. 1D**). This approach was used to discover the biosynthesis of various economically important metabolites showing lineage specific distribution [43,44**]. A prominent example is the identification of the enzyme catalyzing the first committed step of anthraquinone biosynthesis in a medicinal plant *Senna tora*. *S. tora* belongs to the Fabaceae family and accumulates high levels of anthraquinones, which are a group of aromatic polyketides that have been used as a traditional herbal medicine to treat various diseases [44**]. To elucidate anthraquinone biosynthesis in *S. tora*, phylogenomics was used to distinguish the two hypothesized routes to produce this compound in plants. This approach identified a group of chalcone synthase-like proteins that have the catalytic capacity to generate anthraquinone scaffold and showed lineage-specific expansion in *S. tora*. Combining transcriptomics and *in vitro* enzymatic assays, CHS-L9 was identified as the enzyme candidate for catalyzing the initial step of anthraquinone biosynthesis via the polyketide pathway [44**]. This study demonstrates the power of leveraging evolutionary diversification to discover pathways.

**Future directions**

Fruitful progress has been made in discovering metabolic pathways, yet, this only represents a small fraction of the metabolic capacity in plants [16]. To accelerate metabolic pathway characterization, high quality enzyme function annotation is essential. Sequence similarity serves as the major criterion to propagate function annotation between homologs. This strategy has limited power to distinguish enzymes with high sequence similarity in large protein families or polyploid genomes. To establish accurate homologous relationships between species, gene expression patterns can serve as an additional feature besides sequence similarity [45]. With the advances of single cell technologies in plants, high resolution gene expression maps are becoming available across diverse cell types, which can further enhance ortholog prediction and enzyme function annotation [45,46*]. Besides sequence homology, developing a holistic understanding of enzymes can improve function prediction, especially for orphan genes with limited prior knowledge. This goal can be achieved by integrating diverse types of information to infer function. For example, protein 3D structures generated by AlphaFold provide new opportunities to infer substrate-enzyme pairs using protein sequences [47]. Taken together, integrating various types of omics resources facilitates toolset innovation and functional characterization of novel metabolic pathways.

**Conclusion**

Massive amounts of omics datasets can help delineate novel metabolic pathways in non-model species, which represent major sources for economically significant metabolites [8,30**,44**]. In this review, we summarized publicly available knowledge bases that enable the identification of metabolic genes and pathways and discussed cutting-edge omics-guided approaches for elucidating the biosynthetic routes for metabolites of interest. The growing list of genome-scale resources can help guide traditional molecular genetics and biochemical studies. Future advances in enzyme annotation and integrated omics will inform the characterization and engineering of plant metabolism, which promotes sustainable agriculture, healthcare innovation, and climate stabilization.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgement**

We thank Olivia MacDonald and Dr. Charles Hawkins for their help in preparing this manuscript. This work was supported, in part, by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program grant nos. DE-SC0018277, DE-SC0008769, DE-SC0020366 and DE-SC0021286 and the U.S. National Science Foundation grants MCB-1617020 and IOS-1546838. This work was done on the ancestral land of the Muwekma Ohlone Tribe, which was and continues to be of great importance to the Ohlone people.

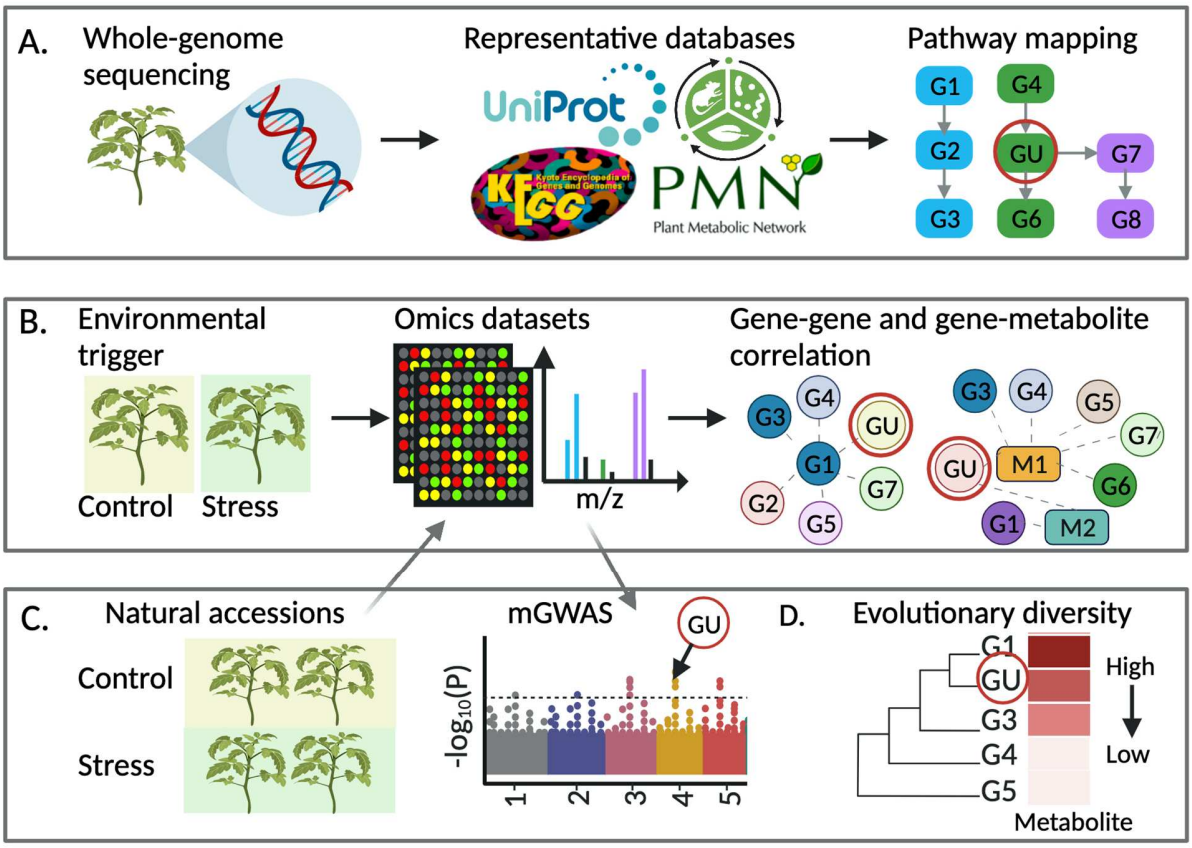


**Figure 1.** Omics-guided strategies to elucidate novel metabolic pathways: A. Plant metabolic pathway databases, B. Correlation analysis, C. metabolic genome wide association studies (mGWAS), D. phylogenomics. G1 to G8 represent metabolic genes. Red circle highlights the gene of unknown function (GU) discovered by each omics approach. M1 and M2 represent metabolites.

**Table 1. Summary of publicly available knowledge bases that provide enzyme annotation and gene to pathway mapping infrastructure.**

| Database | Resources available | Number of curated enzymes | Number of tested pathways | Plant specific? | Metabolism specific? |
|---|---|---|---|---|---|
| UNIPROT[9] | Enzyme annotation | 107,868 | NA | No | No |
| BRENDA[11] | Enzyme annotation | 113,179 | NA | No | No |
| Rhea[10] | Enzyme annotation | Unavailable | NA | No | Yes |
| MetaCyc[15] | Enzyme annotation and pathway mapping | 13,540 | 2,980 | No | No |
| KEGG[12] | Enzyme annotation and pathway mapping | Unavailable | 543 | No | No |
| Plant Reactome[14] | Enzyme annotation and pathway mapping | 1,824 | 298 | Yes | No |
| PMN[16] | Enzyme annotation and pathway mapping | 3,769 | 1,163 | Yes | Yes |

**Box 1. Outstanding questions that can be addressed to further leverage omics resources to investigate plant metabolism**

1. What new infrastructures are required to provide high-resolution annotation for genes and pathways from leveraging single cell or single-molecule level omics datasets?
2. What resources need to be developed (e.g. data generation standards, data processing tools, repository databases, benchmarking data) to make publicly available omics datasets Findable, Accessible, Interoperable, and Reusable (FAIR) [48]?
3. What are the best methods of integrating different types of omics datasets (e.g. transcription factor binding, epigenomics, chromatin accessibility assays) to provide a holistic view of information flow from genes to phenotypes in the context of metabolism?

## References


1. Weng J-K, Lynch JH, Matos JO, Dudareva N: **Adaptive mechanisms of plant specialized metabolism connecting chemistry to function**. *Nat Chem Biol* 2021, **17**:1037–1045.

2. Fang C, Fernie AR, Luo J: **Exploring the Diversity of Plant Metabolism**. *Trends Plant Sci* 2019, **24**:83–98.

3. Wurtzel ET, Kutchan TM: **Plant metabolism, the diverse chemistry set of the future**. *Science* 2016, **353**:1232–1236.

4. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, et al.: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases**. *Nucleic Acids Res* 2014, **42**:D459–71.

5. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin, Darusman LK, et al.: **KNApSAcK Family Databases: Integrated Metabolite–Plant Species Databases for Multifaceted Plant Research**. *Plant Cell Physiol* 2011, **53**:e1–e1.

6. Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T, et al.: **Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants**. *Plant Physiol* 2017, **173**:2041–2059.

7. Kim J, Buell CR: **A Revolution in Plant Metabolism: Genome-Enabled Pathway Discovery**. *Plant Physiol* 2015, **169**:1532–1539.

8. Jacobowitz JR, Weng J-K: **Exploring Uncharted Territories of Plant Specialized Metabolism in the Postgenomic Era**. *Annu Rev Plant Biol* 2020, **71**:631–658.

9. UniProt Consortium: **UniProt: the universal protein knowledgebase in 2021**. *Nucleic Acids Res* 2021, **49**:D480–D489.

10. Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, Hyka-Nouspikel N, Gasteiger E, Kerhornou A, Neto TB, et al.: **Rhea, the reaction knowledgebase in 2022**. *Nucleic Acids Res* 2022, **50**:D693–D700.

11. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, Schomburg D: **BRENDA, the ELIXIR core data resource in 2021: new developments and updates**. *Nucleic Acids Res* 2021, **49**:D498–D508.

12. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K: **KEGG: new perspectives on genomes, pathways, diseases and drugs**. *Nucleic Acids Res* 2017, **45**:D353–D361.

13. Kanehisa M: **KEGG Bioinformatics Resource for Plant Genomics and Metabolomics**. *Methods Mol Biol* 2016, **1374**:55–70.

14. Naithani S, Gupta P, Preece J, D'Eustachio P, Elser JL, Garg P, Dikeman DA, Kiff J, Cook J, Olson A, et al.: **Plant Reactome: a knowledgebase and resource for comparative pathway analysis**. *Nucleic Acids Res* 2020, **48**:D1093–D1103.

15. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley

S, Subhraveti P, Karp PD: **The MetaCyc database of metabolic pathways and enzymes - a 2019 update**. *Nucleic Acids Res* 2020, **48**:D445–D453.

16. Hawkins C, Ginzburg D, Zhao K: **Plant Metabolic Network 15: A resource of genome wide metabolism databases for 126 plants and algae**. *Journal of Integrative* 2021,

17. Nützmann H-W, Huang A, Osbourn A: **Plant metabolic clusters - from genetics to genomics**. *New Phytol* 2016, **211**:771–789.

18. Polturak G, Osbourn A: **The emerging role of biosynthetic gene clusters in plant defense and plant interactions**. *PLoS Pathog* 2021, **17**:e1009698.

19. Karp PD, Latendresse M, Caspi R: **The pathway tools pathway prediction algorithm**. *Stand Genomic Sci* 2011, **5**:424–429.

20*. Nützmann H-W, Scazzocchio C, Osbourn A: **Metabolic Gene Clusters in Eukaryotes**. *Annu Rev Genet* 2018, **52**:159–183. This review summarizes functionally characterized metabolic gene clusters in plants.

21. Banf M, Zhao K, Rhee SY: **METACLUSTER—an R package for context-specific expression analysis of metabolic gene clusters**. *Bioinformatics* 2019,

22*. Chavali AK, Rhee SY: **Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites**. *Brief Bioinform* 2018, **19**:1022–1034. This review summarizes toolsets available to predict metabolic gene clusters in plants.

23. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH: **plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters**. *Nucleic Acids Res* 2017, **45**:W55–W63.

24. Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ: **Co-expression tools for plant biology: opportunities for hypothesis generation and caveats**. *Plant Cell Environ* 2009, **32**:1633–1651.

25. Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W: **Learning from Co-expression Networks: Possibilities and Challenges**. *Front Plant Sci* 2016, **7**:444.

26**. Huang AC, Jiang T, Liu Y-X, Bai Y-C, Reed J, Qu B, Goossens A, Nützmann H-W, Bai Y, Osbourn A: **A specialized metabolic network selectively modulates Arabidopsis root microbiota**. *Science* 2019, **364**. This studies demonstrates the usefulness of co-expression in characterizing unknown metabolic genes.

27. Field B, Osbourn AE: **Metabolic diversification--independent assembly of operon-like gene clusters in different plants**. *Science* 2008, **320**:543–547.

28. Field B, Fiston-Lavier A-S, Kemen A, Geisler K, Quesneville H, Osbourn AE: **Formation of plant metabolic gene clusters within dynamic chromosomal regions**. *Proc Natl Acad Sci U S A* 2011, **108**:16116–16121.

29*. Wu S, Alseekh S, Cuadros-Inostroza Á, Fusari CM, Mutwil M, Kooke R, Keurentjes JB, Fernie AR, Willmitzer L, Brotman Y: **Combined Use of Genome-Wide Association Data and Correlation Networks Unravels Key Regulators of Primary Metabolism in Arabidopsis thaliana**. *PLoS Genet* 2016, **12**:e1006363. This study identifies heretability

and polygenicity associated with primary and specialized metabolism.

30**. Jeon JE, Kim J-G, Fischer CR, Mehta N, Dufour-Schroif C, Wemmer K, Mudgett MB, Sattely E: **A Pathogen-Responsive Gene Cluster for Highly Modified Fatty Acids in Tomato**. *Cell* 2020, **180**:176–187.e19. This study employs correlation analysis between genes and metabolites to prioritize candidates involved in producing falcarindiol in tomato.

31. Li-Beisson Y, Shorrosh B, Beisson F, Andersson MX, Arondel V, Bates PD, Baud S, Bird D, Debono A, Durrett TP, et al.: **Acyl-lipid metabolism**. *Arabidopsis Book* 2013, **11**:e0161.

32. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D: **Benefits and limitations of genome-wide association studies**. *Nat Rev Genet* 2019, **20**:467–484.

33. Chen J, Xue M, Liu H, Fernie AR, Chen W: **Exploring the genic resources underlying metabolites through mGWAS and mQTL in wheat: From large-scale gene identification and pathway elucidation to crop improvement**. *Plant Commun* 2021, **2**:100216.

34. Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W, Li Y, Liu X, Zhang H, Dong H, et al.: **Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism**. *Nat Genet* 2014, **46**:714–721.

35**. Wen W, Li D, Li X, Gao Y, Li W, Li H, Liu J, Liu H, Chen W, Luo J, et al.: **Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights**. *Nat Commun* 2014, **5**:3438. This study performs metabolic GWAS to identify genes associated with a variety of metabolic traits in maize.

36. Wen CC, Yee SW, Liang X, Hoffmann TJ, Kvale MN, Banda Y, Jorgenson E, Schaefer C, Risch N, Giacomini KM: **Genome-wide association study identifies ABCG2 (BCRP) as an allopurinol transporter and a determinant of drug response**. *Clin Pharmacol Ther* 2015, **97**:518–525.

37. Peng S, Deyssenroth MA, Di Narzo AF, Lambertini L, Marsit CJ, Chen J, Hao K: **Expression quantitative trait loci (eQTLs) in human placentas suggest developmental origins of complex diseases**. *Hum Mol Genet* 2017, **26**:3432–3441.

38. Soltis NE, Kliebenstein DJ: **Natural Variation of Plant Metabolism: Genetic Mechanisms, Interpretive Caveats, and Evolutionary and Mechanistic Insights**. *Plant Physiol* 2015, **169**:1456–1468.

39. Chan EKF, Rowe HC, Corwin JA, Joseph B, Kliebenstein DJ: **Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana**. *PLoS Biol* 2011, **9**:e1001125.

40. Weng J-K, Philippe RN, Noel JP: **The rise of chemodiversity in plants**. *Science* 2012, **336**:1667–1670.

41. Moghe GD, Last RL: **Something Old, Something New: Conserved Enzymes and the Evolution of Novelty in Plant Specialized Metabolism**. *Plant Physiol* 2015, **169**:1512–1523.

42. Torrens-Spence MP, Fallon TR, Weng JK: **A Workflow for Studying Specialized**

**Metabolism in Nonmodel Eukaryotic Organisms**. *Methods Enzymol* 2016, **576**:69–97.

43. Hodgson H, De La Peña R, Stephenson MJ, Thimmappa R, Vincent JL, Sattely ES, Osbourn A: **Identification of key enzymes responsible for protolimonoid biosynthesis in plants: Opening the door to azadirachtin production**. *Proc Natl Acad Sci U S A* 2019, **116**:17096–17104.

44**. Kang S-H, Pandey RP, Lee C-M, Sim J-S, Jeong J-T, Choi B-S, Jung M, Ginzburg D, Zhao K, Won SY, et al.: **Genome-enabled discovery of anthraquinone biosynthesis in Senna tora**. *Nat Commun* 2020, **11**:1–11. This paper illustrates how to integrate omics approaches and biochemistry to discover novel pathways in a medicinal plant.

45. Coate JE, Farmer AD, Schiefelbein JW, Doyle JJ: **Expression Partitioning of Duplicate Genes at Single Cell Resolution in Arabidopsis Roots**. *Front Genet* 2020, **11**:596150.

46.* Cole B, Bergmann D, Blaby-Haas CE, Blaby IK, Bouchard KE, Brady SM, Ciobanu D, Coleman-Derr D, Leiboff S, Mortimer JC, et al.: **Plant single-cell solutions for energy and the environment**. *Commun Biol* 2021, **4**:962. * This review describes how single cell omics helps gene function annotation.

47. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al.: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**:583–589.

48. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al.: **The FAIR Guiding Principles for scientific data management and stewardship**. *Sci Data* 2016, **3**:160018.