Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in Biotechnology

Beyond co-expression: pathway discovery for plant pharmaceuticals

Sandesh H Swamidatta and Benjamin R Lichman

Plant natural products have been an important source of medicinal molecules since ancient times. To gain access to the whole diversity of these molecules for pharmaceutical applications, it is important to understand their biosynthetic origins. Whilst co-expression is a reliable tool for identifying gene candidates, a variety of complementary methods can aid in screening or refining candidate selection. Here, we review recently employed plant biosynthetic pathway discovery approaches, and highlight future directions in the field.

Address
Centre for Novel Agricultural Products, Department of Biology, University of York, York YO10 5DD, UK

Corresponding author:
Lichman, Benjamin R (benjamin.lichman@york.ac.uk)

Current Opinion in Biotechnology 2024, **88**:103147

This review comes from a themed issue on **Pharmaceutical Biotechnology**

Edited by **Michael Krogh Jensen**

Available online xxxx

https://doi.org/10.1016/j.copbio.2024.103147

Introduction

Plant natural products (NPs) can be used as pharmaceuticals or as the inspiration for synthetic drugs [1]. Discovering the genetic basis of the biosynthetic pathways to these compounds enables their reconstruction and/or modification in heterologous hosts [2]. It also provides access to biocatalysts that can be used in chemoenzymatic or synthetic biology routes to high-value pharmaceuticals [3]. Pathway gene discovery is dominated by RNA-seq-based co-expression analysis, exemplified by field-defining discoveries of enzymes [4] and pathways [5]. Here, we attempt to look beyond co-expression, examining relevant classical, state-of-the-art and emerging methods for plant biosynthetic gene discovery. We examine examples beyond pharmaceuticals as the methods described can be applied broadly across plant NPs.

The genetic basis

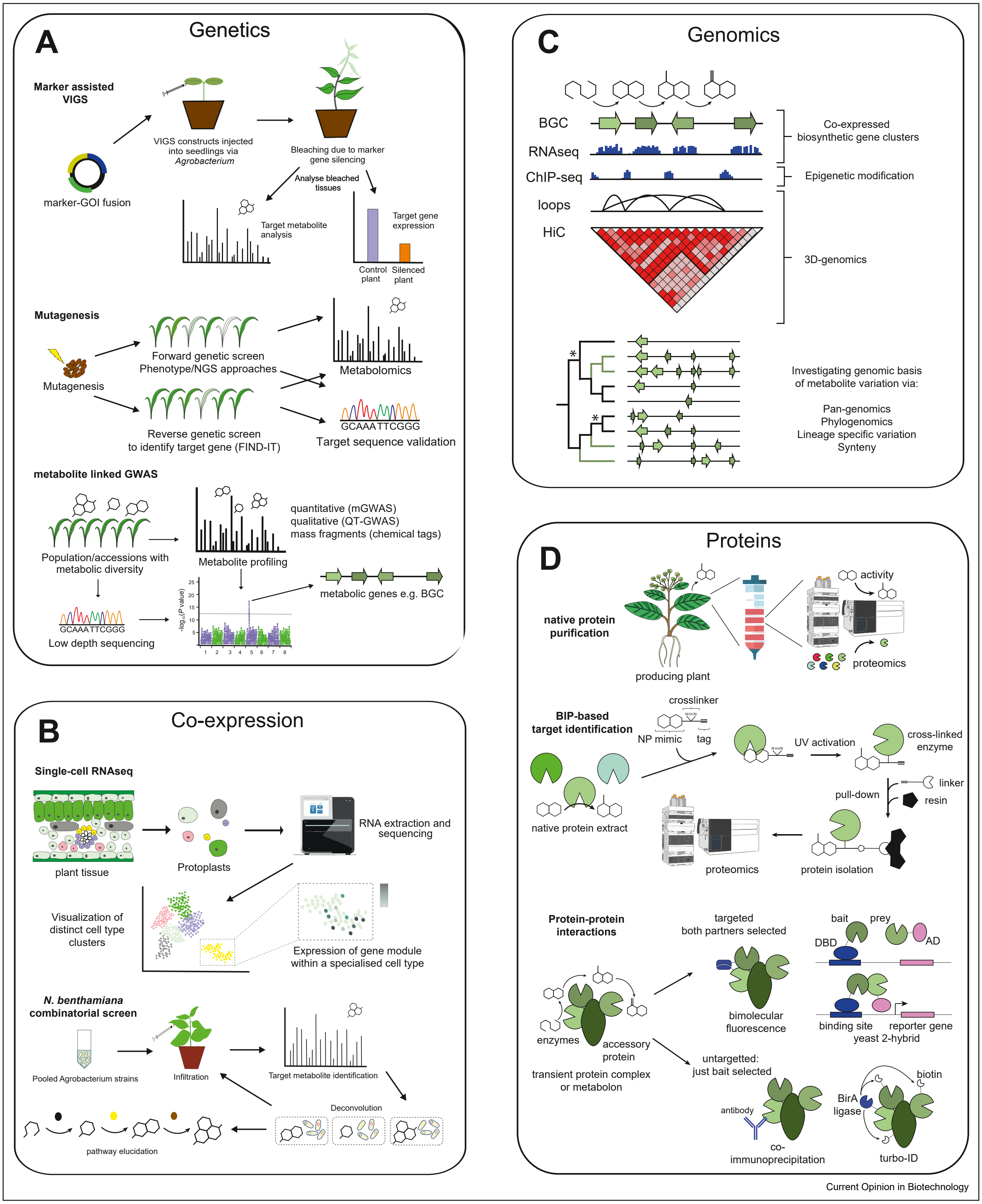
Currently, co-expression and omics-based technologies are the main route for understanding and characterising biosynthetic pathways. However, classical forward and reverse genetics remain valuable tools (Figure 1a). A modified virus-induced gene-silencing (VIGS) system was used to identify a serpentine synthase enzyme in *Catharanthus roseus* wherein the silencing construct also included a marker gene, phytoene desaturase, to pinpoint the silenced tissue [6]. This fusion method was also successfully used in elucidating a branchpoint enzyme in the bioactive diterpenoid pathway for jatrophanes and ingenanes in *Euphorbia peplus* [7]. VIGS can be used to quickly validate related pathway enzymes within plant families [8] to understand the pathways that could help to reconstruct pathways to produce relevant molecules for pharmaceutical applications.

Classical mutagenesis remains valuable where creating genetic variants is feasible. A recently developed approach for screening variant populations, FIND-IT [9], was used to characterise an unexpected acetyltransferase involved in biosynthesis of lupin quinolizidine-type alkaloids [10]. Genome-wide association studies (GWAS), although limited by the availability of population-level data, can be useful for determining the genetic basis of plant-specialized metabolism [11]. In a recent study, a modified metabolic GWAS (mGWAS) was developed where the metabolic fragment enrichment features were used as chemical tags and applied to a natural population of 391 diverse wheat accessions resulting in identification of around 500 potential metabolism-related genes for future analysis [12]. Whilst mGWAS focussed on quantitative metabolite traits, a new method QT-GWAS can detect associations with qualitative metabolite traits [13]. An intriguing 'genetical metabolomics' method has been proposed, which further combines biosynthetic gene cluster (BGC) prediction with metabolite-QTLs, looking for overlaps in signals, which can lead to prioritisation of BGCs for further investigation; though this method is yet to be experimentally validated [14].

Co-expression reimagined

Gene co-expression analysis remains the 'go to' approach for gene mining in biosynthetic gene discovery [5]. The

Figure 1



Current Opinion in Biotechnology

Schematic representation of methods for pathway discovery. **(a)** Genetics-based methods. **(b)** Emerging trends in co-expression. **(c)** Methods for aiding gene discovery within a single-genome assembly and across multiple genomes. **(d)** Direct interactions with proteins and metabolites for gene discovery.

spatial organisation of specialised metabolism means pathways may be localised in specific tissues, cells or specialised structures. Therefore, careful assessment of metabolite distribution followed by precision sampling can greatly increase the power of co-expression analysis. This has been demonstrated across plant taxonomic groups, with precision sampling in faba bean [15], daffodil [16] and clubmoss [17,18] being crucial for alkaloid pathway elucidation.

Cellular localisation represents an extreme of metabolite spatial organisation. Single-cell (sc) methods are revolutionising this approach, with expression patterns of specific cell types now available (Figure 1b). scRNA sequencing has been used to reveal the spatial distribution of monoterpene indole alkaloid (MIA) biosynthesis genes in *C. roseus* leaves [19], a pathway in *Nicotiana attenuata* corolla cells [20] and a transcriptional regulatory network of terpenoid biosynthesis in cotton-secretory glandular cells [21]. In *Hypericum perforatum*, scRNA-seq led to the identification of distinct cells responsible for hyperforin biosynthesis, and identification of the pathway genes [22]. A multi-omics investigation into *C. roseus* combined chromosome-scale genome assembly with scRNA sequencing alongside sc metabolomics [23]. Such integrated approaches will greatly improve the precision in identifying candidate genes for bioactive compounds.

An efficient screening system is equally important for functionally characterising the predicted plant biosynthetic genes. *Nicotiana benthamiana* is the current system of choice [24] due to the availability of a highly efficient transient expression toolbox and ease of expression of complex plant enzymes (i.e. membrane-bound, glycosylated). Moreover, *N. benthamiana* can be used for combinatorial screening of large pools of genes rapidly (Figure 1b) [25]. This is particularly useful for when computational prediction methods are unable to narrow down the candidate gene set. The combinatorial approach was used in the solution of the 20-step biosynthetic pathway for QS-21, a potent vaccine adjuvant, with 68 candidate genes pooled in one experiment [26]. Furthermore, the entire 20-step QS-21 pathway was reconstituted in *N. benthamiana* highlighting its use as a production system [27].

Genomics-based approaches

Thanks to the falling costs and increasing availability of long-read and scaffolding sequencing technologies, obtaining chromosome-level genome assemblies for medicinal plants is becoming a routine part of biosynthetic elucidation (e.g. [26]). Selected recent progress here includes the 10-Gb sequences of taxol-producing *Taxus* species [28,29], new or improved assemblies for MIA-producing species [23,30], a genome for *Ginkgo biloba* [31] and improved assemblies for the chassis *N. benthamiana* [32].

The major benefit in obtaining a genome sequence for biosynthetic pathway elucidation is in the chance of identifying a BGC for the pathway of interest (Figure 1c). BGCs are defined as the close genomic association of three or more independently evolved biosynthetic genes [33], though other genomic features can also have biosynthetic importance, including gene pairs and tandem arrays. BGCs can be identified by identifying the location of genes already known to be involved in the pathway. For example, in *G. biloba*, five CYPs proximal to a known diterpene synthase were characterised and found to catalyse the early oxidative steps in ginkgolide biosynthesis [34]. BGCs are especially useful for identifying genes that might not be easily identifiable via homology or functional annotation, for example, the clustered isopiperitenone reductase in *Schizonepeta tenuifolia* monoterpenoid biosynthesis, which is unrelated to the equivalent enzyme in *Mentha* [35].

Bioinformatic tools for mining genomes can identify BGCs [36,37]. For example, PlantiSmash was used to uncover a BGC of TPSs and P450s involved in paclitaxel biosynthesis [28]. Such tools enable a cluster-first approach where predicted but uncharacterised BGCs can be screened to discover new chemistry, akin to methods in prokaryotic systems [38]. This approach has been employed to discover the unusual amide bond-forming capability of chalcone synthase through heterologous expression of a predicted tomato BGC in yeast [39]. However, a proportion of predicted BGCs are likely to be non-functional [40], and the inclusion of data such as co-expression or phylogenomic context can prioritise BGC candidates or identify non-clustered gene partners.

Active and repressed BGCs may have distinct chromosomal conformations, which are also associated with epigenetic markers (i.e. H2A.Z for activation and H3K27me3 for repression) (Figure 1c) [41]. Furthermore, chromatin accessibility, which is related to both DNA conformation and epigenetic markers, may also be modulated in biosynthetic genes [42]. However, the connections between the spatial genome, epigenetics and plant biosynthesis are still emerging and have yet to be widely used for biosynthetic gene discovery, with the exception of a *C. roseus* secologanin transporter that was studied in part as it was present in a topologically associated domain with other biosynthetic genes [23].

Natural variation of metabolism within a species can be used to identify genes and genomic regions responsible for biosynthetic pathways (Figure 1c) [43]. When many genomes are available, it becomes possible to take a pangenomic approach, identifying groups of genes only present in a proportion of a population. For example, in a rice diterpenoid investigation, pangenome (three assemblies) analysis was coupled to the mGWAS (424 samples) to identify a BGC present in *japonica* but absent in *indica* [44]. Pangenomic approaches have been used to investigate evolution and discover variation in *Arabidopsis* triterpene biosynthesis gene clusters [45,46].

As specialised metabolism is taxonomically restricted, comparisons of genetic and genomic content between taxa with different metabolic content can be useful for identifying genes involved in biosynthetic pathways. A phylogenetic approach was used to identify the lineage-specific expansion of *Taxus* CYP725As that are functional in paclitaxel biosynthesis [28,47]. Lineage-specific variation in *Salvia* CYP76AKs has also been explored to investigate diterpenoid diversity [48]. Phylogenetics integrated with metabolite data and computational pathway reconstruction led to the identification of a CYP involved in iridoid biosynthesis [49]. Phylogenomic approaches, looking at genome synteny across families, have identified variations in terpenoid BGCs in Brassicaceae [50] and the mint family [51,52], with implications for gene discovery and exploring chemical variation.

Protein-directed

The methods discussed so far are genetics-based. However, approaches that target the proteins and metabolites can provide a more direct route to the activities of interest (Figure 1d). The classic method of activity-guided isolation of enzymes from plant extracts still has considerable value, especially for isolating enzymes of unknown family type or so-called auxiliary proteins that influence metabolic flux [53]. This approach led to the isolation of thebaine synthase (THS) from *Papaver somniferum*, a PR10 protein that catalyses a reaction that can also occur without the enzyme catalyst present [54]. The identification of THS led to a further to the identification of neopinone isomerase, which also catalyses a so-called spontaneous reaction [55], alongside related proteins that interact with alkaloid metabolites and influence metabolic flux [56].

The binding of metabolites to proteins can be exploited for protein identification through the development of probes. In such chemoproteomics approaches, chemical probes that mimic the enzyme substrate or product can be used to enrich proteins of interest, which can then be identified and quantified via proteomics [57]. This method, biosynthetic intermediate probe (BIP)-based target identification, can be used for identifying the targets of bioactive NPs [58]. A key demonstration of its utility was shown in its role in identifying Diels–Alderases from *Morus alba* that are responsible for catalysing intermolecular cycloadditions to produce complex flavonoids [59]. An equivalent approach was used in *Ophiorrhiza pumila* to discover a CYP able to produce strictosamide epoxide from strictosamide [60].

In some specialised metabolic pathways, biosynthetic enzymes in a single pathway have been shown to physically interact in enzyme–enzyme assemblies. In some specific cases, such multi-protein interactions are termed metabolons, wherein the interactions facilitate substrate channelling [61]. In either case, it may be possible to identify biosynthetic enzymes by probing protein–protein interactions (PPI). Bimolecular fluorescence (biFC), a method that demonstrates whether two labelled proteins are proximal, was used in kratom to prioritise medium-chain dehydrogenase/reductase candidates that interact with strictosidine glucosidase [62]. However, the requirement to clone fusion tags onto candidate genes makes this method less widely applicable than untargeted approaches, which include chemical cross-linking and immunoprecipitation as described in approaches to elucidate indican biosynthesis in *Persicaria tinctoria* [63]. Last, three complementary approaches to assess PPIs in *Arabidopsis thaliana* glucosinolate biosynthesis have been applied: yeast two-hybrid (Y2H), coimmunoprecipitation and BiFC [64].

Protein interactions may be important to identify auxiliary proteins that are not directly responsible for key pathway steps but could aid flux through pathways in heterologous systems. For example, membrane steroid-binding proteins that aid Arabidopsis lignin biosynthesis were discovered through a Y2H system [65]. Similarly, Y2H revealed interactions between cytochrome P450 (CYP)s and distinct electron transfer chain proteins in the biosynthesis of phenolics [66]. Newer methods for identifying PPIs, such as turbo-ID, have yet to be applied for plant biosynthetic pathways but hold much promise [67]. Use of PPIs to detect similar interactions in medicinal plant systems could lead to identification of valuable auxiliary proteins.

AI ahead

Computational approaches for specialised metabolite pathways and gene identification are improving in sophistication and utility [68], and we anticipate machine learning methods will replace classical co-expression approaches for gene candidate identification. Prediction of specific specialised metabolic pathway membership was developed using transcription co-expression datasets, looking at three common strategies (naive, unsupervised and supervised predictions) [69]. The work focussed on tomato, and found that multiple datasets, especially those that are pathway-function-associated, are particularly beneficial for accurate predictions. More specifically, these are datasets with information about biological processes related to the metabolites of interest such as hormone elicitation or pathogen treatment.

Recently, computational/AI methods for gene identification have been used to identify genes that have been subsequently validated. An *Atropa belladonna* alkaloid transporter was discovered by first selecting gene candidates from a transcriptome using supervised classification strategies, with a neural network performing best [70]. This method yielded just three candidates, of which two were functional, compared with over 100 candidates in a typical co-expression and protein homology-determined list. A similar method was used to identify alkaloid alcohol dehydrogenases from the *Rauvolfia tetraphylla* genome, through the classification of sequences as alkaloid-related [30]. The machine learning method added nine candidates that classical co-expression did not identify, including an enzyme later characterised as an ajmalicine/mayumbine synthase.

Tools that can predict gene function based on sequence alone have been used to identify alternative entry points to alkaloid biosynthesis in *Papaver somniferum* [71]. Support-vector-machine-based algorithms were developed, based on sequences, to help find the elusive aromatic aldehyde synthases and phenylpyruvate decarboxylases that contribute to alkaloid production. The inclusion of chemistry into machine learning methods could be useful for biosynthetic pathway discovery and engineering. Pathways for NPs and NP-like compounds can be predicted or designed using chemical logic [72,73]. Enzyme activity prediction is also improving, with some methods providing broad EC classifications that could be used to prioritise candidates [72], or even predict protein ligand interactions that could be used to identify enzyme-substrate pairs [74,75].

Conclusion

As described above, an assortment of tools providing gene-level resolution can be used to aid gene discovery. Improved tools in this regard either enable higher-throughput screening (i.e. combinatorial *N. benthamiana*) or reducing candidate choice through experiment (i.e. PPIs or scRNA-seq) or computation (i.e. sequence-based AI algorithms). Approaches that require high effort or cost for only slight improvements in candidate choice remain more suited to mechanistic investigations of already-identified biosynthetic genes. Within the next five years, we anticipate AI-based computational methods that integrate phylogenetics, genomics, structure/activity prediction and co-expression to become standard. There is rapid progress in methods for untargeted metabolomics that can be applied to plant-specialised metabolism [76]. An integration of metabolomics with plant biosynthetic pathway discovery would unlock the full complexity and diversity of plant NPs and their potential as therapeutic agents.

Author contributions

Both authors conceptualized and wrote the manuscript.

Declaration of Competing Interest

The authors declare no conflict of interest.

Data Availability

No data were used for the research described in the article.

Acknowledgements

We acknowledge funding from UK Research and Innovation (MR/S01862X/1 and MR/X010260/1).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Chaachouay N, Zidane L: **Plant-derived natural products: a source for drug discovery and development**. *Drugs Drug Candidates* 2024, **3**:184-207.

2. Zhang J, Hansen LG, Gudich O, Viehrig K, Lassen LMM, Schrübbers L, Adhikari KB, Rubaszka P, Carrasquer-Alvarez E, Chen L, *et al*.: **A microbial supply chain for production of the anti-cancer drug vinblastine**. *Nature* 2022, **609**:341-347.

3. Roddan R, Carter EM, Thair B, Hailes HC: **Chemoenzymatic approaches to plant natural product inspired compounds**. *Nat Prod Rep* 2022, **39**:1375-1382.

4. Geu-Flores F, Sherden NH, Courdavault V, Burlat V, Glenn WS, Wu C, Nims E, Cui Y, O'Connor SE: **An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis**. *Nature* 2012, **492**:138-142.

5. Lau W, Sattely ES: **Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone**. *Science* 2015, **349**:1224-1228.

6. • Yamamoto K, Grzech D, Koudounas K, Stander EA, Caputi L, Mimura T, Courdavault V, O'Connor SE: **Improved virus-induced gene silencing allows discovery of a serpentine synthase gene in *Catharanthus roseus***. *Plant Physiol* 2021, **187**:846-857.
The authors slightly modified the original VIGS construct, which greatly helps to pinpoint the tissues that have undergone silencing. This is particularly useful where one or more steps of biosynthesis take place in specialised cells.

7. Czechowski T, Forestier E, Swamidatta SH, Gilday AD, Cording A, Larson TR, Harvey D, Li Y, He Z, King AJ, *et al*.: **Gene discovery and virus-induced gene silencing reveal branched pathways to major classes of bioactive diterpenoids in *Euphorbia peplus***. *Proc Natl Acad Sci USA* 2022, **119**:e2203890119.

8. Yang J, Wu Y, Zhang P, Ma J, Yao YJ, Ma YL, Zhang L, Yang Y, Zhao C, Wu J, *et al*.: **Multiple independent losses of the**

biosynthetic pathway for two tropane alkaloids in the Solanaceae family. *Nat Commun* 2023, **14**:8457.

9. Knudsen S, Wendt T, Dockter C, Thomsen HC, Rasmussen M, Egevang Jørgensen M, Lu Q, Voss C, Murozuka E, Østerberg JT, *et al*.: **FIND-IT: accelerated trait development for a green evolution**. *Sci Adv* 2022, **8**:eabq2266.

10. Mancinotti D, Czepiel K, Taylor JL, Golshadi Galehshahi H, Møller LA, Jensen MK, Motawia MS, Hufnagel B, Soriano A, Yeheyis L, *et al*.: **The causal mutation leading to sweetness in modern white lupin cultivars**. *Sci Adv* 2023, **9**:eadg8866.

11. Burgos E, Belen De Luca M, Diouf I, de Haro LA, Albert E, Sauvage C, Tao ZJ, Bermudez L, Asís R, Nesi AN, *et al*.: **Validated MAGIC and GWAS population mapping reveals the link between vitamin E content and natural variation in chorismate metabolism in tomato**. *Plant J* 2021, **105**:907-923.

12. •• Zhu A, Liu M, Tian Z, Liu W, Hu X, Ao M, Jia J, Shi T, Liu H, Li D, *et al*.: **Chemical-tag-based semi-annotated metabolomics facilitates gene identification and specialized metabolic pathway elucidation in wheat**. *Plant Cell* 2023, **36**:540-558, https://doi.org/10.1093/plcell/koad286.

A novel approach using mass fragments as tags for genome association studies. Potential for associating specific tags with gene annotations for the discovery of tailoring enzymes.

13. • Brouckaert M, Peng M, Höfer R, El Houari I, Darrah C, Storme V, Saeys Y, Vanholme R, Goeminne G, Timokhin VI, *et al*.: **QT-GWAS: a novel method for unveiling biosynthetic loci affecting qualitative metabolic traits**. *Mol Plant* 2023, **16**:1212-1227.

Introduces a method for integrating qualitative metabolic traits into genome wide association studies.

14. Witjes L, Kooke R, van der Hooft JJJ, de Vos RCH, Keurentjes JJB, Medema MH, Nijveen H: **A genetical metabolomics approach for bioprospecting plant biosynthetic gene clusters**. *BMC Res Notes* 2019, **12**:194.

15. Björnsdotter E, Nadzieja M, Chang W, Escobar-Herrera L, Mancinotti D, Angra D, Xia X, Tacke R, Khazaei H, Crocoll C, *et al*.: **VC1 catalyses a key step in the biosynthesis of vicine in faba bean**. *Nat Plants* 2021, **7**:923-931.

16. •• Mehta N, Meng Y, Zare R, Kamenetsky-Goldstein R, Sattely E: **A developmental gradient reveals biosynthetic pathways to eukaryotic toxins in monocot geophytes**. *bioRxiv* 2023, https://doi.org/10.1101/2023.05.12.540595.

This study integrates precision sampling and labelling studies to identify the site of active biosynthesis, leading to the discovery of a complete set of biosynthetic genes for Amaryllidaceae alkaloids.

17. Nett RS, Dho Y, Low Y-Y, Sattely ES: **A metabolic regulon reveals early and late acting enzymes in neuroactive Lycopodium alkaloid biosynthesis**. *Proc Natl Acad Sci USA* 2021, **118**:e2102949118.

18. Nett RS, Dho Y, Tsai C, Passow D, Martinez Grundman J, Low Y-Y, Sattely ES: **Plant carbonic anhydrase-like enzymes in neuroactive alkaloid biosynthesis**. *Nature* 2023, **624**:182-191.

19. Sun S, Shen X, Li Y, Li Y, Wang S, Li R, Zhang H, Shen G, Guo B, Wei J, *et al*.: **Single-cell RNA sequencing provides a high-resolution roadmap for understanding the multicellular compartmentation of specialized metabolism**. *Nat Plants* 2023, **9**:179-190.

20. Kang M, Choi Y, Kim H, Kim S-G: **Single-cell RNA-sequencing of *Nicotiana attenuata* corolla cells reveals the biosynthetic pathway of a floral scent**. *New Phytol* 2022, **234**:527-544.

21. Lin J-L, Chen L, Wu W-K, Guo X-X, Yu C-H, Xu M, Nie G-B, Dun J-L, Li Y, Xu B, *et al*.: **Single-cell RNA sequencing reveals a hierarchical transcriptional regulatory network of terpenoid biosynthesis in cotton secretory glandular cells**. *Mol Plant* 2023, **16**:1990-2003.

22. Wu S, Morotti ALM, Yang J, Wang E, Tatsis EC: **Single-cell RNA-seq based elucidation of the antidepressant hyperforin biosynthesis de novo in St. John's wort**. *bioRxiv* 2024, https://doi.org/10.1101/2024.01.24.577018

23. •• Li C, Wood JC, Vu AH, Hamilton JP, Rodriguez Lopez CE, Payne RME, Serna Guerrero DA, Gase K, Yamamoto K, Vaillancourt B, *et al*.: **Single-cell multi-omics in the medicinal plant *Catharanthus roseus***. *Nat Chem Biol* 2023, **19**:1031-1041.

A comprehensive analysis of the alkaloid biosynthesis pathway in *C. roseus* by integrating genome sequencing, chromatin interaction data, single-cell transcriptomics, and single-cell metabolomics.

24. Reed J, Stephenson MJ, Miettinen K, Brouwer B, Leveau A, Brett P, Goss RJM, Goossens A, O'Connell MA, Osbourn A: **A translational synthetic biology platform for rapid access to gram-scale quantities of novel drug-like molecules**. *Metab Eng* 2017, **42**:185-193.

25. Carlson ED, Rajniak J, Sattely ES: **Multiplicity of the Agrobacterium infection of *Nicotiana benthamiana* for transient DNA delivery**. *ACS Synth Biol* 2023, **12**:2329-2338.

26. Reed J, Orme A, El-Demerdash A, Owen C, Martin LBB, Misra RC, Kikuchi S, Rejzek M, Martin AC, Harkess A, *et al*.: **Elucidation of the pathway for biosynthesis of saponin adjuvants from the soapbark tree**. *Science* 2023, **379**:1252-1264.

27. •• Martin LBB, Kikuchi S, Rejzek M, Owen C, Reed J, Orme A, Misra RC, El-Demerdash A, Hill L, Hodgson H, *et al*.: **Complete biosynthesis of the potent vaccine adjuvant QS-21**. *Nat Chem Biol* 2024, **20**:493-502, https://doi.org/10.1038/s41589-023-01538-5.

The authors identify the downstream steps that complete the biosynthetic pathway for the vaccine adjuvant QS-21. The work also shows the effectiveness of combinatorial high-throughput screening and pathway reconstruction in *N. benthamiana*.

28. Xiong X, Gou J, Liao Q, Li Y, Zhou Q, Bi G, Li C, Du R, Wang X, Sun T, *et al*.: **The *Taxus* genome provides insights into paclitaxel biosynthesis**. *Nat Plants* 2021, **7**:1026-1036.

29. Song C, Fu F, Yang L, Niu Y, Tian Z, He X, Yang X, Chen J, Sun W, Wan T, *et al*.: ***Taxus yunnanensis* genome offers insights into gymnosperm phylogeny and taxol production**. *Commun Biol* 2021, **4**:1203.

30. • Stander EA, Lehka B, Carqueijeiro I, Cuello C, Hansson FG, Jansen HJ, Dugé De Bernonville T, Birer Williams C, Vergès V, Lezin E, *et al*.: **The *Rauvolfia tetraphylla* genome suggests multiple distinct biosynthetic routes for yohimbane monoterpene indole alkaloids**. *Commun Biol* 2023, **6**:1197.

Demonstration of the integration of AI methods for gene candidate classification as part of the candidate selection workflow.

31. Liu H, Wang X, Wang G, Cui P, Wu S, Ai C, Hu N, Li A, He B, Shao X, *et al*.: **The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution**. *Nat Plants* 2021, **7**:748-756.

32. Ranawaka B, An J, Lorenc MT, Jung H, Sulli M, Aprea G, Roden S, Llaca V, Hayashi S, Asadyar L, *et al*.: **A multi-omic *Nicotiana benthamiana* resource for fundamental research and biotechnology**. *Nat Plants* 2023, **9**:1558-1571.

33. Smit SJ, Lichman BR: **Plant biosynthetic gene clusters in the context of metabolic evolution**. *Nat Prod Rep* 2022, **39**:1465-1482.

34. Forman V, Luo D, Geu-Flores F, Lemcke R, Nelson DR, Kampranis SC, Staerk D, Møller BL, Pateraki I: **A gene cluster in *Ginkgo biloba* encodes unique multifunctional cytochrome P450s that initiate ginkgolide biosynthesis**. *Nat Commun* 2022, **13**:5143.

35. Liu C, Smit SJ, Dang J, Zhou P, Godden GT, Jiang Z, Liu W, Liu L, Lin W, Duan J, *et al*.: **A chromosome-level genome assembly reveals that a bipartite gene cluster formed via an inverted duplication controls monoterpenoid biosynthesis in *Schizonepeta tenuifolia***. *Mol Plant* 2023, **16**:533-548.

36. 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.

37. 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.

38. Lin Z, Nielsen J, Liu Z: **Bioprospecting through cloning of whole natural product biosynthetic gene clusters**. *Front Bioeng Biotechnol* 2020, **8**:526.

39. Kong D, Li S, Smolke CD: **Discovery of a previously unknown biosynthetic capacity of naringenin chalcone synthase by heterologous expression of a tomato gene cluster in yeast**. *Sci Adv* 2020, **6**:eabd1143.

40. Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A: **A global coexpression network approach for connecting genes to specialized metabolic pathways in plants**. *Plant Cell* 2017, **29**:944-959.

41. Nützmann H-W, Doerr D, Ramírez-Colmenero A, Sotelo-Fonseca JE, Wegel E, Di Stefano M, Wingett SW, Fraser P, Hurst L, Fernandez-Valverde SL, *et al*.: **Active and repressed biosynthetic gene clusters have spatially distinct chromosome states**. *Proc Natl Acad Sci USA* 2020, **117**:13800-13809.

42. Zhou L, Huang Y, Wang Q, Guo D: **Chromatin accessibility is associated with artemisinin biosynthesis regulation in *Artemisia annua***. *Molecules* 2021, **26**:1194.

43. Zhou X, Liu Z: **Unlocking plant metabolic diversity: a (pan)-genomic view**. *Plant Commun* 2022, **3**:100300.

44. Zhan C, Lei L, Liu Z, Zhou S, Yang C, Zhu X, Guo H, Zhang F, Peng M, Zhang M, *et al*.: **Selection of a subspecies-specific diterpene gene cluster implicated in rice disease resistance**. *Nat Plants* 2020, **6**:1447-1454.

45. Liu Z, Cheema J, Vigouroux M, Hill L, Reed J, Paajanen P, Yant L, Osbourn A: **Formation and diversification of a paradigm biosynthetic gene cluster in plants**. *Nat Commun* 2020, **11**:5354.

46. Marszalek-Zenczak M, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P, Brzezinski K, Iefimenko T, Figlerowicz M, Zmienko A: **Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members**. *Front Plant Sci* 2023, **14**:1104303.

47. Jiang B, Gao L, Wang H, Sun Y, Zhang X, Ke H, Liu S, Ma P, Liao Q, Wang Y, *et al*.: **Characterization and heterologous reconstitution of *Taxus* biosynthetic enzymes leading to baccatin III**. *Science* 2024, **383**:622-629.

48. Hu J, Qiu S, Wang F, Li Q, Xiang C-L, Di P, Wu Z, Jiang R, Li J, Zeng Z, *et al*.: **Functional divergence of CYP76AKs shapes the chemodiversity of abietane-type diterpenoids in genus Salvia**. *Nat Commun* 2023, **14**:4696.

49. Rodríguez-López CE, Jiang Y, Kamileen MO, Lichman BR, Hong B, Vaillancourt B, Buell CR, O'Connor SE: **Phylogeny-aware chemoinformatic analysis of chemical diversity in Lamiaceae enables iridoid pathway assembly and discovery of aucubin synthase**. *Mol Biol Evol* 2022, **39**:msac057.

50. Liu Z, Suarez Duran HG, Harnvanichvech Y, Stephenson MJ, Schranz ME, Nelson D, Medema MH, Osbourn A: **Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae**. *New Phytol* 2020, **227**:1109-1123.

51. Bryson AE, Lanier ER, Lau KH, Hamilton JP, Vaillancourt B, Mathieu D, Yocca AE, Miller GP, Edger PP, Buell CR, *et al*.: **Uncovering a miltiradiene biosynthetic gene cluster in the Lamiaceae reveals a dynamic evolutionary trajectory**. *Nat Commun* 2023, **14**:343.

52. Li H, Wu S, Lin R, Xiao Y, Malaco Morotti AL, Wang Y, Galilee M, Qin H, Huang T, Zhao Y, *et al*.: **The genomes of medicinal skullcaps reveal the polyphyletic origins of clerodane diterpene biosynthesis in the family Lamiaceae**. *Mol Plant* 2023, **16**:549-570.

53. Dastmalchi M: **Elusive partners: a review of the auxiliary proteins guiding metabolic flux in flavonoid biosynthesis**. *Plant J* 2021, **108**:314-329.

54. Chen X, Hagel JM, Chang L, Tucker JE, Shiigi SA, Yelpaala Y, Chen H-Y, Estrada R, Colbeck J, Enquist-Newman M, *et al*.: **A pathogenesis-related 10 protein catalyzes the final step in thebaine biosynthesis**. *Nat Chem Biol* 2018, **14**:738-743.

55. Dastmalchi M, Chen X, Hagel JM, Chang L, Chen R, Ramasamy S, Yeaman S, Facchini PJ: **Neopinone isomerase is involved in codeine and morphine biosynthesis in opium poppy**. *Nat Chem Biol* 2019, **15**:384-390.

56. Ozber N, Carr SC, Morris JS, Liang S, Watkins JL, Caldo KM, Hagel JM, Ng KKS, Facchini PJ: **Alkaloid binding to opium poppy major latex proteins triggers structural modification and functional aggregation**. *Nat Commun* 2022, **13**:6768.

57. Gao Y, Ma M, Li W, Lei X, Chemoproteomics A: **Broad avenue to target deconvolution**. *Adv Sci* 2024, **11**:2305608.

58. Wang D, Cao Y, Zheng L, Lv D, Chen L, Xing X, Zhu Z, Li X, Chai Y: **Identification of Annexin A2 as a target protein for plant alkaloid matrine**. *Chem Commun* 2017, **53**:5020-5023.

59. •• Gao L, Su C, Du X, Wang R, Chen S, Zhou Y, Liu C, Liu X, Tian R, Zhang L, *et al*.: **FAD-dependent enzyme-catalysed intermolecular [4+2] cycloaddition in natural product biosynthesis**. *Nat Chem* 2020, **12**:620-628.
The authors synthesised a BIP that aided in the identification of an enzyme involved in intermolecular cycloaddition reactions in chalcomoracin biosynthesis.

60. Zhang T, Wang Y, Wu S, Tian E, Yang C, Zhou Z, Yan X, Wang P: **Chemoproteomics reveals the epoxidase enzyme for the biosynthesis of camptothecin in *Ophiorrhiza pumila***. *J Integr Plant Biol* 2023, https://doi.org/10.1111/jipb.13594

61. Zhang Y, Fernie AR: **Metabolons, enzyme-enzyme assemblies that mediate substrate channeling, and their roles in plant metabolism**. *Plant Commun* 2021, **2**:100081.

62. Wu Y, Liu C, Koganitsky A, Gong FL, Li S: **Discovering dynamic plant enzyme complexes in yeast for kratom alkaloid pathway identification**. *Angew Chem Int Ed Engl* 2023, **62**:e202307995.

63. Inoue S, Morita R, Kuwata K, Ishii K, Minami Y: **Detection of candidate proteins in the indican biosynthetic pathway of *Persicaria tinctoria* (Polygonum tinctorium) using protein-protein interactions and transcriptome analyses**. *Phytochemistry* 2020, **179**:112507.

64. Chen L-Q, Chhajed S, Zhang T, Collins JM, Pang Q, Song W, He Y, Chen S: **Protein complex formation in methionine chain-elongation and leucine biosynthesis**. *Sci Rep* 2021, **11**:3524.

65. Gou M, Ran X, Martin DW, Liu C-J: **The scaffold proteins of lignin biosynthetic cytochrome P450 enzymes**. *Nat Plants* 2018, **4**:299-310.

66. Zhao X, Zhao Y, Gou M, Liu C-J: **Tissue-preferential recruitment of electron transfer chains for cytochrome P450-catalyzed phenolic biosynthesis**. *Sci Adv* 2023, **9**:eade4389.

67. Xu S-L, Shrestha R, Karunadasa SS, Xie P-Q: **Proximity labeling in plants**. *Annu Rev Plant Biol* 2023, **74**:285-312.

68. Wang P, Schumacher AM, Shiu S-H: **Computational prediction of plant metabolic pathways**. *Curr Opin Plant Biol* 2022, **66**:102171.

69. Wang P, Moore BM, Uygun S, Lehti-Shiu MD, Barry CS, Shiu S-H: **Optimising the use of gene expression data to predict plant metabolic pathway memberships**. *New Phytol* 2021, **231**:475-489.

70. Srinivasan P, Smolke CD: **Engineering cellular metabolite**
•• **transport for biosynthesis of computationally predicted tropane alkaloid derivatives in yeast**. *Proc Natl Acad Sci USA* 2021, **118**:e2104460118.

The authors tested supervised learning models to predict genes involved in TA biosynthesis and identify two transporters through a neural network model.

71. Vavricka CJ, Takahashi S, Watanabe N, Takenaka M, Matsuda M, Yoshida T, Suzuki R, Kiyota H, Li J, Minami H, *et al*.: **Machine learning discovery of missing links that mediate alternative branches to plant alkaloids**. *Nat Commun* 2022, **13**:1405.

72. Kim GB, Kim JY, Lee JA, Norsigian CJ, Palsson BO, Lee SY: **Functional annotation of enzyme-encoding genes using deep learning with transformer layers**. *Nat Commun* 2023, **14**:7370.

73. Hafner J, Payne J, MohammadiPeyhani H, Hatzimanikatis V, Smolke C: **A computational workflow for the expansion of heterologous biosynthetic pathways to natural product derivatives**. *Nat Commun* 1760, **2021**:12.

74. Kroll A, Ranjan S, Engqvist MKM, Lercher MJ: **A general model to predict small molecule substrates of enzymes based on machine and deep learning**. *Nat Commun* 2023, **14**:2787.

75. Qiao Z, Nie W, Vahdat A, Miller TF, Anandkumar A: **State-specific protein–ligand complex structure prediction with a multiscale deep generative model**. *Nat Mach Intell* 2024, **6**:195-208.

76. van der Hooft JJJ, Ernst M, Papenberg D, Kang KB, Kappers IF, Medema MH, Dorrestein PC, Rogers S: **Deciphering complex natural mixtures through metabolome mining of mass spectrometry data**. In *Recent Adv Polyphen Res*. Edited by Salminen J-P, Wähälä K, de Freitas V, Quideau S. 8 John Wiley & Sons Ltd; 2023:139-168, https://doi.org/10.1002/9781119844792©