

# Multi-omics approaches for biosynthetic pathway prediction in plants



Hernando Suarez

## Propositions

1. Although guilt-by-association (the principle behind many gene function prediction approaches) promotes “throwing data against the wall to see what sticks”, there is value in interpreting what did not “stick”.  
(this thesis)
2. To aid data interoperability and re-use, the design of new -omics experiments should always consider the design of previous ones with similar objectives or parameters.  
(this thesis)
3. Skilled computational scientists make themselves irrelevant for routine tasks, and indispensable for novel ones.
4. Within Academia excess attention is given to the name and prestige of journals, laboratories and authors.
5. The basics of “programming” and “computer science” will soon be seen as analogous to “literacy” and “numeracy”.
6. Children should be taught about life that most adults, if not all, are making it up as they live.

Propositions belonging to the thesis, entitled  
Multi-omics Approaches for Biosynthetic Pathway Prediction in Plants  
Hernando Suarez  
Wageningen, 8 February 2021

# **Multi-omics approaches for biosynthetic pathway prediction in plants**

**Hernando Suarez**

**Thesis committee****Promotors**

Prof. Dr. Ir. D. de Ridder  
Professor of Bioinformatics  
Wageningen University & Research

**Co-promotor**

Dr. M.H. Medema  
Assistant Professor of Bioinformatics  
Wageningen University & Research

**Other members**

Prof. Dr. C.S. Testerink, Wageningen University & Research  
Dr. M.F. Seidl, Utrecht University  
Dr. S.E. O'Connor, Max Planck Institute of Chemical Ecology, Jena, Germany  
Dr. R. Cavill, Maastricht University

This research was conducted under the auspices of the Graduate School  
Experimental Plant Sciences

# **Multi-omics approaches for biosynthetic pathway prediction in plants**

## **Thesis**

submitted in fulfilment of the requirements for the degree of doctor  
at Wageningen University  
by the authority of the Rector Magnificus,  
Prof. Dr A.P.J. Mol,  
in the presence of the  
Thesis Committee appointed by the Academic Board  
to be defended in public  
on Monday 8 February 2021  
at 4 p.m. in the Aula.

Hernando Suarez

Multi-omics approaches for biosynthetic pathway prediction in plants, 104 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2021)

With references, with summary in English

ISBN 978-94-6395-669-7

DOI <https://doi.org/10.18174/538330>

# Table of Contents

<b>Chapter 1:</b> Introduction	6
<b>Chapter 2:</b> plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters <i>Published in Nucleic Acids Research 45(W1):55-63, 2017</i>	19
<b>Chapter 3:</b> Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae <i>Published in New Phytologist 227:1109-1123</i>	35
<b>Chapter 4:</b> Comparative transcriptomics reveals a conserved core of the phosphate starvation response across monocots and eudicots	62
<b>Chapter 5:</b> MEANtools: An Integrative Multi-omics Approach for Metabolic Pathway Prediction	77
<b>Chapter 6:</b> General Discussion	96
Summary	107
Envoi	108

# **Chapter 1**

## **Introduction**



## 1.1. The world's finest chemists

Intrinsically tied to our history as omnivores, humans have consumed plants for more than just nutrients since the dawn of our species, and likely even before. Although the oldest suspected instance of medicinal plant usage dates to 60,000 BC, attributed to Neanderthals in the Shanidar Cave<sup>1</sup>, the practice of ingesting plants to self-medicate has been documented in many non-human animals, who use them for diverse purposes such as regulating gut acidity, purging intestinal parasites and infections, or even as prophylaxis to prevent those and many other conditions<sup>2</sup>. In fact, the discovery of many plants' beneficial usages can be attributed to observing animal behaviors<sup>2</sup>, tracing the origins of the plant natural product (NP) discovery field to pharmacognosy, zoo-pharmacognosy and even ethology.

One reason for this may be the ubiquity of plant life throughout the globe: with over 390,000 species<sup>3</sup>, it is estimated that ~80% of Earth's biomass is composed of plants<sup>4</sup>. Unaided by sessility, their effective proliferation and survival in the diverse scenarios posed by Earth's ecosystems can, at least in part, be attributed to plants' specialized metabolism (SM)<sup>5</sup>. This part of plant metabolism involves the production of a diverse range of metabolites of various degrees of complexity, which help plants combat abiotic stresses, and mediate interactions with other organisms, be it antagonistic (like herbivores) or mutualistic (like pollinators)<sup>6,7</sup>. As humans, we have learned how to use these metabolites for our benefit, and have documented approximately ~10,000 plant species for medicinal use, a stark contrast to the ~3,000 species that have been cultivated for food<sup>8</sup>.

Because of their abundance and diversity, it is no surprise that plant NPs have found their way into multiple facets of human industry, from pharmaceutical applications to flavors and fragrances<sup>9</sup>. Morphine, a natural alkaloid of opium poppy, is perhaps the most famous of NPs with pharmaceutical application, but many others have gained a foothold in medicine. One such example is artemisinin, a sesquiterpene produced by sweet wormwood that is firmly established in combination therapy to treat malaria and is also currently being studied for other medicinal applications<sup>10</sup>. Another plant terpenoid that has been interwoven into modern civilization is rubber: from gloves to tires, rubber is essential to all manufacturing sectors, and while the majority of the rubber in the market is synthetic, 35% to 45% is natural rubber harvested almost exclusively from a single species<sup>11</sup>. One plant NP is key to a ~\$814 billion industry: nicotine, the highly addictive stimulant behind the 5,300 billion cigarettes consumed yearly over the world<sup>12</sup>.

Taming the wildly diverse world of plant specialized metabolism has evidently pushed humanity forward. As we continue to face new challenges, it is undeniable we must also continue learning from the world's finest chemists.

## 1.2. A short timeline of plant specialized metabolism research

The history of plant specialized metabolism research is best described by T. Hartmann in his 2007 review of the history of the field<sup>13</sup>:

"Within 50 years a bewildering array of waste products turned into a classified selection of chemical structures with indispensable ecological functions."

The history of plant NP discovery, however, is certainly older than fifty years: as posited in 1.1., humans have likely been discovering and using beneficial plant metabolites for most of the history of our species. The oldest documented evidence of medicinal plants usage to prepare drugs is ~5,000 years old with 12 distinct plant-based drug recipes, and a few centuries later, evidence shows knowledge of 365 recipes<sup>14</sup>. Arguably, these and the many other recipe preparations were informal or unknowing attempts of isolating and identifying the plants' beneficial NPs: "plant secondary metabolism research pre-history", as Hartmann calls it. The most acknowledged success of these informal attempts happened in 1806, when F. Sertürner isolated morphine from opium poppy, which sparked an era of new NP discovery and isolation from a plethora of plants, including ipecacuanha, strychnos, quinine and pomegranate<sup>14</sup>. To Hartmann, though, plant secondary metabolism research formally starts in the 1950s, when radioactively-labeled molecules allowed us to use biochemical evidence to trace back the biosynthetic pathways behind the production of many plant NPs<sup>13</sup>.

In the world of bacteria and fungi, NP discovery around the same period revolved around phenotypic screenings, bioassays in which antibacterial, antifungal or anti-cancer activity was detected visually by measuring growth inhibition induced by extracted metabolites. Between 1930 and 1970, thousands of new chemical entities were discovered with this methodology<sup>15</sup>. The following decades introduced genetics to the field in a more prevalent manner: by the 1970s, recombinant DNA technology was developed for *Escherichia coli* (and later *Streptomyces*), which allowed for the characterization of genes involved in important antibiotics like penicillin and kanamycin<sup>15</sup>. Around a decade afterwards, by the mid-1980s, the first enzyme-coding genes of plant secondary metabolism were cloned and functionally expressed; shortly after cDNA for phenylpropanoid, alkaloid and terpene biosynthesis from plant sources had been successfully cloned<sup>13</sup>.

It was in the early 2000s when genomics took the forefront of the field. Improvements in sequencing technologies led to obtaining a 99% complete genome sequence for *Streptomyces coelicolor* in 2002, the largest bacterial genome sequence at the time. This study uncovered over 20 previously unidentified groups of chromosomally clustered genes involved in the biosynthesis of specialized metabolites<sup>16</sup>, and several biosynthetic gene clusters (BGCs) were subsequently associated with distinct NPs. This quickly became a popular methodology to identify SM pathways and their associated NP in microorganisms: within 10 years, hundreds of gene clusters had been characterized, and this decade culminated in the release of antiSMASH in 2011, a computational tool to quickly identify and annotate SM gene clusters in bacteria and fungi with high confidence<sup>17</sup>. In the almost ten years since then, around ~2,000 BGCs from microorganisms have been characterized experimentally to link them to specific metabolic products<sup>18</sup>.

Although knowledge of chromosomally clustered SM genes in plants dates back to 1997, when five genes required for DIBOA biosynthesis in maize (Bx1 through Bx5) were found to be clustered on chromosome 4<sup>19</sup>, it took several more years to discover how widespread this phenomenon was in the plant kingdom. The next plant BGC was identified in oat in 2004: Qi *et al.* identified multiple genes required for the biosynthesis of the antimicrobial triterpenoid avenacin that were physically clustered<sup>20</sup>. Similar to what was the case for bacteria and fungi, this discovery led to ~10 years of BGC discovery in plants, but at a slower pace: by 2016, over twenty BGCs had been identified in plants<sup>21</sup>. The majority of plant NPs characterized during

that time, however, were not synthesized by BGCs, instead by un-clustered or partially clustered pathways and single-gene pathways; nevertheless, this effectively demonstrated that analyzing these genomic structures was a valid method for NP discovery in plants too, and the ongoing evolution of the NP discovery field brought high expectation of newer, more complex, methods to identify novel metabolic pathways by integrating multiple sources of data<sup>22</sup>.

Whereas the 2000s brought genomics to the forefront of the field, midway through the 2010s one word had been added; now, computational genomics was the name of the game. By 2015, using computational tools to mine -omics data was not restricted to BGC identification. Similar efforts were being made using transcriptomic data: the biosynthetic pathways for podophyllotoxin in mayapple<sup>23</sup> and 4-hydroxyindole-3-carbonyl nitrile (4-OH-ICN) in *Arabidopsis thaliana*<sup>24</sup> were elucidated by mining publicly available transcriptomic datasets, based on the principle that genes that encode enzymes belonging to the same pathway have correlated expression levels across samples. Indeed, large-scale analysis of transcripts from several plants demonstrated that levels of coexpression among genes involved in SM are larger than among other genes<sup>25</sup>, supporting this guilt-by-association principle. Moreover, both studies combined transcriptomic with metabolomic analyses to some extent: by analyzing the transcriptomes, they identified genes likely involved in targeted pathways of interest; their hypotheses were later reinforced through observed changes in the metabolome of mutants, or in tobacco after heterologous combinatorial expression of the targeted enzymes. A similar approach had been used to identify the final step in the biosynthesis of noscapine a year before: analyses of the transcriptomes of multiple opium poppy chemotypes were used to identify the noscapine synthase (NOS) before validating their results with LC-MS/MS<sup>26</sup>. These informed sequential sets of multiple -omics analyses illuminated the advantages that a fully integrated multi-omic analysis could eventually bring to the table: gene expression and metabolite abundance are two entry points to study the same metabolic pathways; measuring them simultaneously and integrating these data could provide a more holistic view of it.

Concurrently, more computational tools geared towards SM pathway and NP discovery were being developed: CoExpNetViz, a computational tool to compare gene coexpression networks of different plant species through homology, was released at the beginning of 2016<sup>27</sup>. Their approach requires targeting genes of interest, “baits”, in the transcriptome of two plant species. The algorithm then identifies genes coexpressed with the bait genes, and then uses gene families from PLAZA<sup>28</sup> to group them according to cross-species orthology. The results are visualized as networks depicting the groups of orthologous genes that are all coexpressed with the bait genes across the two species. This tool was based on previous work that led to the elucidation of the biosynthetic pathways for  $\alpha$ -chaconine/solanine in potato by comparing its coexpression network with that of tomato<sup>29</sup>. This research in particular was also aided by the close phylogenetic relatedness of tomato and potato, and the fact that the pathway genes were physically clustered, demonstrating that comparative transcriptomics, phylogenomics and BGC identification can also be seamlessly integrated through computational solutions.

As more genomes continued to become available, it was undeniable that plant specialized metabolism research was on the verge of a computational revolution.

## 1.3. A new era in natural products discovery

The increasing availability of plant genomes and transcriptomes before the start of this PhD project presented a challenge and an opportunity within the field: new approaches needed to be developed to effectively and efficiently mine the large sets of data generated by high throughput -omics.

### 1.3.1. Biosynthetic Gene Clusters

As discussed in 1.2., before the start of this PhD, around 2016, just over twenty BGCs had been identified in plants through a variety of methods<sup>21</sup>. The first computational study identifying the potential of finding chromosomally colocated genes as a strategy to discover novel SM pathways in plants had been published a couple of years earlier; in 2014 Chae *et al.* examined the *A. thaliana*, *Sorghum bicolor*, soybean, and rice genomes and identified groups of genes that were chromosomally clustered; they found that between 22% and 31% of the plants' metabolic genes were located in these clusters, more than expected by chance<sup>30</sup>. Moreover, they queried the *A. thaliana* gene clusters within a large-scale microarray dataset and concluded that gene clusters with genes involved in SM were more likely to be coexpressed than the clusters without SM genes<sup>30</sup>. This, and the increasing number of characterized BGCs in plant genomes introduced a unique opportunity for the NP discovery and SM research fields: rapid computational discovery of SM BGCs took the neighboring bacteria and fungi kingdoms into a "gene cluster revolution"<sup>31</sup>, and the revolution could cross the border.

Concurrently to our development of plantiSMASH to tackle this opportunity (**see chapter 2**), other tools were being developed to mine plant genomes and predict BGCs that could be associated with SM pathways. Schlöpfer *et al.* developed PlantClusterFinder, an algorithm to detect metabolic gene clusters in plants<sup>32</sup>. Their method consisted of identifying SM genes in a genome, and then using a sliding window to search for chromosomal regions with clustered SM genes. With this strategy, they identified clusters in 21 plant species, and over 1,700 of the clusters (15%) had SM genes. Interestingly, they found that SM pathways were more than twice as likely to contain genes that are physically clustered than non-SM pathways. Lastly, of the thirteen characterized pathways known in the species they used in their study, Schlöpfer *et al.* identified twelve with the help of PlantClusterFinder. Shortly after, Töpfer *et al.* released the PhytoClust Tool for the metabolic gene cluster discovery in plant genomes<sup>33</sup>, and used it to mine 31 plant genomes, identifying thousands of clusters. Despite using a distinct computational approach and having studied different species, their findings were in line with those of Schlöpfer *et al.*, and both studies in turn effectively supported the previous work of Chae *et al.*<sup>30</sup>: the relation between plant specialized metabolism and gene clusters is ripe for exploitation and can easily guide plant NP discovery.

Studying plant BGCs has been shown to be beneficial not only to NP discovery, but also to better understand the evolution of SM pathways in plants. Along with identifying the gene cluster involved in the biosynthesis of the triterpenoid avenacin in oat, Qi *et al.* also found the cluster was located in a region of the genome not conserved in other cereals, suggesting this BGC had not been assembled in a

common ancestor<sup>20</sup>. Similar conclusions were reached with the characterization of the BGC responsible for the biosynthesis of the triterpenoid thalianol in *A. thaliana*: this cluster is unlikely to have a common origin with avenacin, indicating both triterpenoid clusters were assembled recently, and independently<sup>34</sup>. The characterization of the BGC for marneral, another triterpenoid produced by *A. thaliana*, provided another piece of the puzzle: Field *et al.* found that while the marneral and thalianol clusters likely shared a common ancestor, the marneral cluster was not found in *Arabidopsis lyrata* despite its close phylogenetic relatedness to *A. thaliana*<sup>35</sup>; this suggested the marneral cluster had been recently lost in *A. lyrata*. Taken together, these findings indicate the assembly and disassembly of plant BGCs is very dynamic; thus, understanding the processes responsible for BGC evolution can help us unlock the secrets behind the diverse range of metabolites that plants can produce.

Altogether, the research of BGCs in plants has the capacity to deeply contribute to many aspects of plant biology research, particularly regarding specialized metabolism. The development and improvement of algorithms aimed at identifying and predicting BGCs is thus crucial for the growth of the field.

### 1.3.2. Coexpression Networks

Identifying coexpressed genes has been a very successful approach to mine transcriptome datasets. Analyzing coexpression allows us to identify gene functions and characterize SM pathways. The level of coexpression among any two genes can be measured using a variety of metrics, such as the Pearson correlation coefficient (PCC) or Spearman's rank. This strategy has resulted in many discoveries within the field<sup>23,24,36,37</sup>, but it demands *a priori* knowledge of the function of some genes to propagate their annotations to those whose function is unknown. Acquiring such prior knowledge manually, however, may not be straightforward in plant genomes, many of which have over 40,000 genes<sup>38</sup>, with a large number of them being functionally uncharacterized.

One approach to overcome this obstacle is using coexpression networks; they are one of the multiple types of gene networks that can be generated from gene expression data<sup>39,40</sup>, and represent genes as nodes connected by edges/links if the genes show coexpression correlation above a predetermined threshold. In *A. thaliana*, coexpression networks have been shown to capture the functional categorization of genes by grouping them in clusters or modules of tightly coexpressed genes, which have similar function or regulation<sup>41</sup>. Moreover, evidence suggest that the function and members of some gene modules are conserved through speciation events<sup>42</sup>, allowing the propagation of annotations between species with more confidence. When handling large datasets, however, coexpression networks can become too complex to readily interpret<sup>22,39,43</sup>, in part due to unobserved factors that affect gene expression and may cause correlation among genes<sup>44</sup>, which in turn may lead to an increase in type I errors when interpreting the significance of correlated genes in a network or module. In summary, better gene prioritization methods need to be developed to improve the confidence of predictions from the analysis of coexpression networks.

To address this, many strategies are being developed. For example, a more robust statistical backdrop has been introduced<sup>45</sup>, and many network module detection algorithms have been developed or adapted for transcriptomic research.

Some of these methods have been determined to be particularly good at identifying sets of genes associated with a specific function<sup>46</sup>, or were even specifically designed to predict plant SM pathways<sup>47</sup>. Another approach to remedy the emergence of type I errors is to compare the coexpression networks of different species. As mentioned in 1.2., this strategy has already led to the identification of NPs in plants<sup>29</sup>, and more recently it has been proposed as a method to study the evolution of SM pathways in plants<sup>48</sup>. While the previously discussed CoExpNetViz<sup>27</sup> was designed to compare coexpression networks with the purpose of SM pathway discovery, it still requires *a priori* knowledge in the form of bait genes, making it less useful for genomes with lower-quality annotations, or those that are phylogenomically distant from well-annotated genomes (which makes it difficult to propagate annotations through homology). Other algorithms, though, have successfully analyzed the transcriptome of more evolutionarily distant species: for example, Netotea *et al.* developed ComPIEx, which they used to compare the coexpression network of *A. thaliana*, *Populus trichocarpa* and *Oryza sativa*, revealing common network motifs enriched with the same GO terms in all three species<sup>49</sup>. CoMPIEx, however, only uses these three genomes (and more recently, CoMPIEx 2 compares *A. thaliana*, *P. trichocarpa* and *Picea abies*<sup>50</sup>); this is in contrast to CoExpNetViz, which allows comparative analyses of transcriptomes input by the user, and severely limits its usage.

While the power of comparatively analyzing gene coexpression to guide plant SM pathway discovery is clear, no comprehensive computational strategy has been developed that can simultaneously tackle all the challenges it presents. A unique opportunity stands here, as comparative analyses can provide biological insights in multiple species. Thus, the development of computational tools that can provide high-confidence SM pathway predictions through comparative transcriptomics has potential to significantly extend our knowledge of plant biology.

### 1.3.3. Multi-omics Integration

Another way to improve the mining of -omics datasets is through integrated multi-omics analyses. Above, we briefly discussed how sequential transcriptomic and metabolomic analyses led to the discovery of several SM pathways and their associated metabolites<sup>23,24,26</sup>: the observations made in gene expression profiles allowed the authors of these studies to prepare metabolomic experiments to validate their hypotheses. This sequence of analyses differs from an integrated multi-omics analysis; the latter is performed as a single simultaneous comprehensive analysis. This strategy is particularly advantageous in complex biological systems, where a multi-omics analysis can provide a holistic view by combining advantages of different omics technologies. For example, in clinical research, many multi-omics integration tools are used for cancer prognosis prediction, which helps physicians decide treatments and improve patient survival<sup>51</sup>. Similarly, this strategy has a lot of potential to guide plant SM pathway discovery<sup>52</sup>.

The popularity and success of transcriptomic and metabolomic analyses for SM pathway discovery present them as prime candidates for multi-omic integration, and multiple methods have been developed to integrate their datasets. For example, the experiments can be designed to send one set of replicates to RNA-Seq, and another set to mass spectrometry (MS) analysis; alternatively, each sample can be split,

sending one half to each -omics analysis. Later, integrating the datasets can be done in different manners too. For example, Urbanczyk-Wochniak *et al.* performed a correlation analysis between transcripts and metabolite profiles in potato tubers and identified more than double the number of significantly correlated pairs than would be expected by chance<sup>53</sup>. They also mapped their transcriptome and metabolome onto known reactions, as well as the enzymes, substrates and products involved; their results indicated that the correlation between substrate and enzyme was higher than that of product and enzyme<sup>53</sup>. Similar results were later found in an integrated analysis of the transcriptome and metabolome of *Catharanthus roseus*: genes and metabolites involved in terpenoid indole alkaloid biosynthesis showed a high degree of correlation<sup>54</sup>.

These and other strategies for transcriptomic and metabolomic integration have been reviewed in detail before<sup>55</sup>, with each presenting unique advantages and limitations; however, a common thread among them is the high noise levels that often hide metabolite-gene associations. Indeed, gene and metabolites known to be closely related in pathways often do not show correlation. For example, ter Kuile and Westerhoff found that metabolic fluxes in parasitic protists did not correlate proportionally with the concentration of the corresponding enzymes<sup>56</sup>. Similarly, Moxley *et al.* showed that mRNA expression of enzyme-coding genes did not correlate well with metabolic flux changes in yeast<sup>57</sup>. While these two studies evaluated the integration of metabolomic and transcriptomic data in other organisms, plants have a vastly more complex metabolic system than yeast and protists, making it very likely that straightforward simple correlation of gene expression and metabolite abundance will not lead to high confidence predictions. It is therefore clear that, similar to what is the case for coexpression network analysis, multi-omics integration requires more advanced gene/metabolite association prioritization methods to improve their performance and power in plant SM research and NP discovery.

## 1.4. Contributions of this thesis

This PhD thesis aims at tackling some of the unique opportunities that computational genomics has brought to the plant SM pathway discovery field. In **1.2** and **1.3.**, we discussed many of these opportunities and obstacles, but the focus during my PhD research was directed at multi-omics solutions; namely, the development of computational methods to analyze data from multiple -omics sources for the purpose of SM pathway discovery in plants.

In **chapter 2**, we introduce plantiSMASH, a computational tool for the automated identification, annotation and expression analysis of plant biosynthetic gene clusters. This project aims at facilitating the study of plant BGCs, which, as discussed in **1.3.1.**, can guide NP discovery and SM pathway evolution research. To do this, we base our pipeline on antiSMASH<sup>58</sup>, following a similar workflow, user inputs and GUI, but customizing it specifically for the particularities of plant genomes and adding a coexpression analysis module. We use plantiSMASH to mine publicly-available chromosome-level plant genome assemblies and identify hundreds of candidate BGCs, including all experimentally characterized BGCs in the genomes we mine.

In **chapter 3**, we continue the research of plant BGCs, this time with the purpose

of understanding the manner in which they assemble during evolution and reconstructing the evolutionary relationships between closely related BGCs and BGC-like genomic structures. For this, we focus on a family of triterpene clusters in Brassicaceae, including the thalianol<sup>34</sup> and marneral<sup>35</sup> clusters discussed in 1.3.1.. Centering our analysis on the scaffold-generating enzyme responsible for the triterpene scaffold of both pathways, an oxidosqualene cyclase (OSC), we query the genome of thirteen Brassicaceae species to identify the genes flanking each OSC, a region we name their “genomic neighborhood” (GN). We then compare the evolution of the SM genes in the GNs with the known species phylogeny of Brassicaceae, which leads us to uncover the most likely historical scenarios for the assembly and metabolic diversification of the Brassicaceae triterpene BGCs.

In **chapter 4**, we explore using comparative analysis of transcriptomic networks to guide SM pathway discovery. For this we develop CADE-HEroN, a computational workflow designed to tackle limitations the analysis of coexpression networks discussed in 1.3.2.. CADE-HEroN allows users to input their own transcriptomes, genome assemblies and homolog relationships to identify conserved gene expression patterns among different species. Our algorithm allows users to either input bait genes and target specific metabolic processes of interest, or to generate a comparative time-series analysis that facilitates untargeted analysis. We use our workflow to study the phosphate starvation response in *A. thaliana*, tomato and rice, and identify a number of genes that have a conserved behavior across the three species and are likely to be coregulated. Some of the genes our analysis highlights are already known to be involved in phosphate starvation, proving this workflow can recover functionally related genes, while others were not yet functionally annotated, providing a source of candidate genes that may be prioritized for experimental characterization.

In **chapter 5**, we present MEANtools, an integrative multi-omics approach for metabolic pathway prediction. Our objective is to tackle some of the limitations of transcriptomic and metabolomic integration discussed in 1.3.3., and increase the confidence in SM pathway predictions generated from the correlation of gene expression and metabolite abundance datasets. We integrate the datasets not only through correlation, but through an enzymatic reaction database that links the relationships between known metabolites, reactions and enzymes. In this manner, MEANtools can predict potential metabolic pathways. Predicted pathways, and the genes, metabolites and reactions involved, are presented to the user in a graphic manner for easy interpretation. In addition, the user can use the algorithm in a variety of modes according to their needs and the data they have available. Lastly, we use MEANtools to generate pathway predictions based on a paired transcriptomic-metabolomic dataset in tomato that was previously used for the characterization of the faltarindiol pathway<sup>59</sup>; this shows that our algorithm was able to correctly predict multiple steps within this pathway. With this, we present a computational tool that can, in a rapid and automated manner, help scientists generate testable hypotheses about biosynthetic pathways, and the genes, reactions and metabolites involved in them.

Finally, in **chapter 6** we discuss the main findings, conclusions and perspectives gathered along the development of all research projects and the writing of this thesis.



## References

1. Lietava, J. Medicinal plants in a Middle Paleolithic grave Shanidar IV? *J. Ethnopharmacol.* **35**, 263–266 (1992).
2. Huffman, M. A. Animal self-medication and ethno-medicine: exploration and exploitation of the medicinal properties of plants. *Proc. Nutr. Soc.* **62**, 371–381 (2003).
3. Willis, K. J. *State of the world's plants report-2017*. (Royal Botanic Gardens, 2017). doi:978-1-84246-628-5
4. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl. Acad. Sci.* **115**, 6506–6511 (2018).
5. Weng, J. K. The evolutionary paths towards complexity: A metabolic perspective. *New Phytol.* **201**, 1141–1149 (2014).
6. Kessler, A. & Kalske, A. Plant Secondary Metabolite Diversity and Species Interactions. *Annu. Rev. Ecol. Evol. Syst.* **49**, 115–138 (2018).
7. Fang, C., Fernie, A. R. & Luo, J. Exploring the Diversity of Plant Metabolism. *Trends Plant Sci.* **24**, 83–98 (2019).
8. McChesney, J. D., Venkataraman, S. K. & Henri, J. T. Plant natural products: Back to the future or into extinction? *Phytochemistry* **68**, 2015–2022 (2007).
9. Kabera, J. N., Semana, E., Mussa, A. R. & He, X. Plant Secondary Metabolites: Biosynthesis, Classification, Function and Pharmacological Properties. *J. Pharm. Pharmacol.* **2**, 377–392 (2014).
10. Krishna, S., Bustamante, L., Haynes, R. K. & Staines, H. M. Artemisinins: their growing importance in medicine. *Trends Pharmacol. Sci.* **29**, 520–7 (2008).
11. Cornish, K. Alternative Natural Rubber Crops: Why Should We Care? *Technol. Innov.* **18**, 244–255 (2017).
12. British American Tobacco - The global market. Available at: [https://www.bat.com/group/sites/UK\\_\\_9D9KCY.nsf/vwPagesWebLive/DO9DCKFM](https://www.bat.com/group/sites/UK__9D9KCY.nsf/vwPagesWebLive/DO9DCKFM). (Accessed: 26th July 2020)
13. Hartmann, T. From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry* **68**, 2831–2846 (2007).
14. Petrovska, B. Historical review of medicinal plants' usage. *Pharmacogn. Rev.* **6**, 1 (2012).
15. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* **43**, 155–176 (2016).
16. Bentley, S. D. *et al.* Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147 (2002).
17. Medema, M. H. *et al.* antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, 339–346 (2011).
18. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, (2019).

19. Frey, M. *et al.* Analysis of a chemical plant defense mechanism in grasses. *Science* (80-. ). **277**, 696–699 (1997).
20. Qi, X. *et al.* A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 8233–8 (2004).
21. Nützmann, H.-W., Huang, A. & Osbourn, A. Plant metabolic clusters - from genetics to genomics. *New Phytol.* **211**, 771–89 (2016).
22. Medema, M. H. & Osbourn, A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Nat. Prod. Rep.* **33**, 951–62 (2016).
23. Lau, W. & Sattely, E. S. Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. *Science* **349**, 1224–8 (2015).
24. Rajniak, J., Barco, B., Clay, N. K. & Sattely, E. S. A new cyanogenic metabolite in *Arabidopsis* required for inducible pathogen defence. *Nature* **525**, 376–379 (2015).
25. Omranian, N. *et al.* Differential metabolic and coexpression networks of plant metabolism. *Trends Plant Sci.* **20**, 266–8 (2015).
26. Chen, X. & Facchini, P. J. Short-chain dehydrogenase/reductase catalyzing the final step of noscapine biosynthesis is localized to laticifers in opium poppy. *Plant J.* **77**, 173–184 (2014).
27. Tzfadia, O. *et al.* CoExpNetViz: Comparative Co-Expression Networks Construction and Visualization Tool. *Front. Plant Sci.* **6**, 1194 (2016).
28. Proost, S. *et al.* PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res.* **43**, D974–D981 (2015).
29. Itkin, M. *et al.* Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* **341**, 175–9 (2013).
30. Chae, L., Kim, T., Nilo-Poyanco, R. & Rhee, S. Y. Genomic signatures of specialized metabolism in plants. *Science* **344**, 510–3 (2014).
31. Jensen, P. R. Natural Products and the Gene Cluster Revolution. *Trends in Microbiology* **24**, 968–977 (2016).
32. Schlöpfer, P. *et al.* Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. *Plant Physiol.* **173**, 2041–2059 (2017).
33. Töpfer, N., Fuchs, L. & Aharoni, A. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Res.* **45**, 7049–7063 (2017).
34. Field, B. & Osbourn, A. E. Metabolic Diversification--Independent Assembly of Operon-Like Gene Clusters in Different Plants. *Science* (80-. ). **320**, 543–547 (2008).
35. Field, B. *et al.* Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 16116–21 (2011).
36. Horan, K. *et al.* Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol.* **147**, 41–57 (2008).
37. Sohrabi, R. *et al.* In Planta Variation of Volatile Biosynthesis: An Alternative Biosynthetic Route to the Formation of the Pathogen-Induced Volatile

- Homoterpene DMNT via Triterpene Degradation in Arabidopsis Roots. *Plant Cell* **27**, 874–890 (2015).
38. Sterck, L., Rombauts, S., Vandepoele, K., Rouze, P. & Van de Peer, Y. How many genes are there in plants (... and why are they there)? *Curr. Opin. Plant Biol.* **10**, 199–203 (2007).
  39. Hoang, T., Hong, S., Company, H. M. & Shin, D. Extending Biological Pathways by Utilizing Conditional Mutual Information Extracted from RNA-SEQ Gene Expression Data. **27**, (2010).
  40. Li, Y., Pearl, S. A. & Jackson, S. A. Gene Networks in Plant Biology: Approaches in Reconstruction and Analysis. *Trends Plant Sci.* **20**, 664–675 (2015).
  41. Mao, L., Van Hemert, J. L., Dash, S. & Dickerson, J. A. Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics* **10**, 346 (2009).
  42. Ficklin, S. P. & Feltus, F. A. Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice. *Plant Physiol.* **156**, 1244–1256 (2011).
  43. Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M. & Ligterink, W. Learning from Co-expression Networks: Possibilities and Challenges. *Front. Plant Sci.* **7**, 444 (2016).
  44. Furlotte, N. A., Kang, H. M., Ye, C. & Eskin, E. Mixed-model coexpression: Calculating gene coexpression while accounting for expression heterogeneity. *Bioinformatics* **27**, i288–i294 (2011).
  45. Gobbi, A. & Jurman, G. A null model for pearson coexpression networks. *PLoS One* **10**, (2015).
  46. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* **9**, 1090 (2018).
  47. Wisecaver, J. H. *et al.* A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* **29**, 944–959 (2017).
  48. Ruprecht, C. *et al.* Beyond Genomics: Studying Evolution with Gene Coexpression Networks. *Trends Plant Sci.* **0**, 64–69 (2017).
  49. Netotea, S., Sundell, D., Street, N. R. & Hvidsten, T. R. ComPIEx: Conservation and divergence of co-expression networks in *A. thaliana*, *Populus* and *O. sativa*. *BMC Genomics* **15**, 1–17 (2014).
  50. ComPIEx. Available at: <http://complex.plantgenie.org/>. (Accessed: 26th July 2020)
  51. Huang, S., Chaudhary, K. & Garmire, L. X. More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics* **8**, (2017).
  52. Rai, A., Saito, K. & Yamazaki, M. Integrated omics analysis of specialized metabolism in medicinal plants. *Plant J.* **90**, 764–787 (2017).
  53. Urbanczyk-Wochniak, E. *et al.* Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.* **4**, 989–993 (2003).
  54. Rischer, H. *et al.* Gene-to-metabolite networks for terpenoid indole alkaloid

- biosynthesis in *Catharanthus roseus* cells. *Proc. Natl. Acad. Sci.* **103**, 5614–5619 (2006).
55. Cavill, R., Jennen, D., Kleinjans, J. & Briedé, J. J. Transcriptomic and metabolomic data integration. *Brief. Bioinform.* bbv090- (2015). doi:10.1093/bib/bbv090
  56. ter Kuile, B. H. & Westerhoff, H. V. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* **500**, 169–171 (2001).
  57. Moxley, J. F. *et al.* Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 6477–82 (2009).
  58. Weber, T. *et al.* antiSMASH 3.0--a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43**, 1–7 (2015).
  59. Jeon, J. E. *et al.* A Pathogen-Responsive Gene Cluster for Highly Modified Fatty Acids in Tomato. *Cell* **180**, 176-187.e19 (2020).

# Chapter 2

## **plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters**

*Satria A. Kautsar<sup>1,2,†</sup>, Hernando G. Suarez Duran<sup>1,†</sup>, Kai Blin<sup>3</sup>, Anne Osbourn<sup>4</sup> and Marnix H. Medema<sup>1</sup>*

<sup>1</sup> *Bioinformatics Group, Wageningen University, 6708 PB Wageningen, The Netherlands*

<sup>2</sup> *Teknik Informatika, Universitas Lampung, Jln. Sumantri Brojonegoro No. 01, Lampung 35141, Indonesia*

<sup>3</sup> *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark*

<sup>4</sup> *Department of Metabolic Biology, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK*

*† These authors contributed equally to the paper as first authors.*

*Published in Nucleic Acids Research 45(W1):55-63, 2017*

## 2.1. Abstract

Plant specialized metabolites are chemically highly diverse, play key roles in host–microbe interactions, have important nutritional value in crops and are frequently applied as medicines. It has recently become clear that plant biosynthetic pathway-encoding genes are sometimes densely clustered in specific genomic loci: biosynthetic gene clusters (BGCs). Here, we introduce plantiSMASH, a versatile online analysis platform that automates the identification of candidate plant BGCs. Moreover, it allows integration of transcriptomic data to prioritize candidate BGCs based on the coexpression patterns of predicted biosynthetic enzyme-coding genes, and facilitates comparative genomic analysis to study the evolutionary conservation of each cluster. Applied on 48 high-quality plant genomes, plantiSMASH identifies a rich diversity of candidate plant BGCs. These results will guide further experimental exploration of the nature and dynamics of gene clustering in plant metabolism. Moreover, spurred by the continuing decrease in costs of plant genome sequencing, they will allow genome mining technologies to be applied to plant natural product discovery. The plantiSMASH web server, precalculated results and source code are freely available from <http://plantismash.secondarymetabolites.org>.

## 2.2. Introduction

Across Planet Earth, bacteria, fungi and plants produce an immense diversity of specialized metabolites, each with their own specific ecological roles in the manifold interorganismal interactions in which they engage. This diverse specialized metabolism is a rich source of natural products that are used widely in medicine, agriculture and manufacturing. In bacteria and fungi, where genes for most specialized metabolic pathways are physically clustered in so-called biosynthetic gene clusters (BGCs), the rapid accumulation of genome sequences has revolutionized the process of natural product discovery: indeed, genome mining has now become a dominant method for the discovery of novel molecules (1–4). In this genome mining process, BGCs are computationally identified in genome sequences and then linked to molecules through functional analysis (e.g. using metabolomic data, chemical structure predictions, mutant libraries and/or heterologous expression). Many sequence based aspects of this genome mining procedure are facilitated by the antiSMASH framework, which was launched in 2010 (5) and has seen continuous development since then (6,7). The genome mining procedure has two main purposes: (i) finding biosynthetic genes for important known compounds to allow heterologous production through fermentation in industrial strains, and (ii) identifying novel natural product chemistry guided by biosynthetic gene cluster diversity. Altogether, this development has appropriately been termed the ‘gene cluster revolution’ (1).

In recent years, it has become clear that not only microbial, but also plant biosynthetic pathways are frequently chromosomally clustered: after the initial discoveries of the cyclic hydroxamic acid 2,4-dihydroxy-1,4-benzoxazin-3-one (DIBOA) and avenacin gene clusters (8,9), around thirty plant BGCs have been discovered (10,11). Together, they encode the production of a wide range of different compounds, including cyclic hydroxamic acids, di- and triterpenes, steroidal and benzyloquinoline alkaloids, cyanogenic glucosides and polyketides. In the genome of the model plant species *Arabidopsis thaliana* alone, four BGCs have been linked

to specific metabolites and recent analyses based on epigenomic profiling indicate the presence of various additional uncharacterized ones (12).

Various technological developments in eukaryote genome sequencing (13) are finally making complete plant genome sequencing feasible at larger scales: high-quality plant genome sequences for almost 100 species are already publicly available, and more or less complete genomes can be sequenced for as little as 10–50k US dollars each. Hence, genome mining may become an important methodology in the study of plant natural products as well, and a realistic opportunity thus presents itself for the plant natural product research community to have a ‘gene cluster revolution’ of its own. Naturally, a key technology required to realize this is a computational framework specifically designed for the identification and analysis of plant BGCs. Importantly, tools available for bacterial and fungal genome mining do not suffice for plants (14), as (i) plant biosynthetic pathways involve unique enzyme families not found in bacteria and fungi; (ii) not all plant biosynthetic pathways are clustered (e.g. anthocyanins (15)), so identification of a biosynthetic gene does not equal identification of a BGC; (iii) intergenic distances in plant genomes are larger and much more variable (16–19); (iv) plant genomes contain clustered groups of genes (e.g. tandem arrays) whose products do not constitute a pathway; (v) several plant pathways are split across more than one BGC (20,21).

Here, we introduce antiSMASH for plants (or ‘plantiSMASH’ in short), which has been designed to tackle each of these challenges. Through a comprehensive library of profile Hidden Markov Models (pHMMs) for enzyme families known to be involved in plant biosynthetic pathways, combined with CD-HIT clustering of predicted protein sequences belonging to the same family, it allows the efficient identification of genomic loci encoding multiple different (sub)families of specialized metabolic enzymes. Moreover, comparative genomic analysis as well as analysis of gene expression patterns within these candidate BGCs allow assessment of each locus for its likelihood to encode genes working together in one pathway. Finally, coexpression analysis between candidate BGCs and with other genes across the genome allows identification of biosynthetic pathways that are encoded on multiple loci. To exploit this new framework, we offer an initial analysis of BGC diversity across the plant kingdom, which showcases the presence of many complex biosynthetic loci in diverse species.

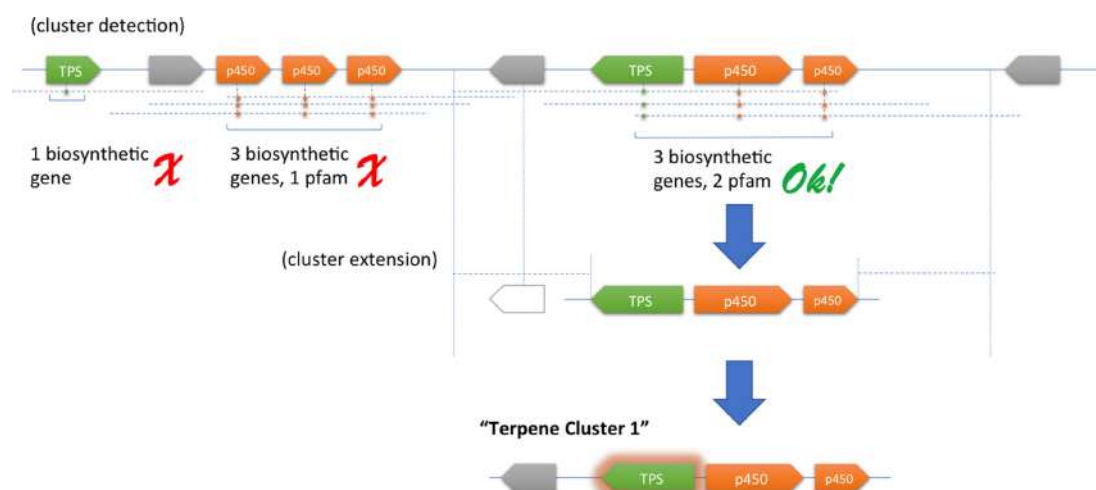
## 2.3. Methods and Implementation

### 2.3.1. A procedure for the identification of candidate plant biosynthetic gene clusters

The microbial version of antiSMASH (5) predicts BGCs by using HMMer (22) to identify specific (combinations of) signature protein domains that belong to scaffold generating enzymes specific for a class of biosynthetic pathways. Subsequently, hit genes are used as anchors from which gene clusters are extended upstream and downstream by a specified extension distance.

Although very effective for detecting biosynthetic clusters in bacteria and fungi,

this procedure is unfit to detect biosynthetic gene clusters in plants, for the reasons described above. To address this, a novel detection strategy was chosen (Figure 1): instead of identifying BGCs through the identification of core scaffold-generating genes alone, plantiSMASH identifies them by looking for all genes predicted to encode biosynthetic enzymes, including those required for tailoring of the scaffold.



**Figure 1.** General strategy followed by plantiSMASH for the identification of plant BGCs. First, plantiSMASH identifies biosynthetic genes (having a hit on one of the 62 pHMMs) that are located in close proximity to each other. Subsequently, it will look for the co-occurrence of at least three biosynthetic enzyme-coding genes, comprising at least two different enzyme types. (Based on the results of the CD-HIT clustering of encoded protein sequences, closely related duplicate genes will only be counted once). Afterward, identified clusters are extended to incorporate any flanking genes. Finally, each cluster is classified based on the presence of core enzymes (see Supplementary Table S1). In this example, the detected cluster is assigned to the ‘Terpene’ class due to the presence of a terpene synthase-encoding gene.

To determine what constitutes a high-potential candidate BGC, we make use of the recently proposed definition for plant BGCs as ‘genomic loci encoding genes for a minimum of three different types of biosynthetic reactions (i.e. genes encoding functionally different (sub)classes of enzymes)’ (14). (Albeit arbitrary, this definition correctly describes all known plant BGCs at the moment and is open to improvement as more are discovered.) Accordingly, with default settings plantiSMASH defines clusters as loci where at least three different enzyme subclasses belonging to at least two different enzyme classes are co-located on the same locus. Enzyme classes are identified using pHMMs specific for each class (Supplementary Table S1); to count the number of subclasses of each enzyme class at a certain locus, the CD-HIT algorithm (23) is employed for sequence-based clustering to identify groups of sequences within an enzyme class with (by default) >50% mutual amino acid sequence identity. This successfully distinguishes potentially real BGCs from tandem repeat regions that are also frequently found in genomes (Supplementary Table S2).

In order to identify all classes of biosynthetic enzymes known to be involved in plant specialized metabolic pathways, we performed a comprehensive literature search of previously characterized plant biosynthetic pathways, which resulted in a list of 62 protein domains that have been associated with specialized metabolic pathways in plants (see Supplementary Table S1). Fifty-seven of these protein domains are represented by pHMMs from the Pfam database (24), and custom pHMMs were generated for five enzyme families not (fully) covered by Pfam



domains. We consciously refrained from attempting to construct custom pHMMs for all enzyme families known to be involved in plant biosynthetic pathways, as the limited amount of training data available would lead to an overly strict prediction system that would no longer be able to detect biosynthetic novelty; instead, we assume that the broad enzyme families covered by Pfam domains are likely to be biosynthetically involved if multiple enzymes from these different families are encoded together in the same locus. As in the microbial version of antiSMASH, the presence of genes predicted to encode signature enzymes (defined as enzymes that determine the chemical class of the end compound, such as terpene synthases) in a candidate BGC are used to assign a cluster to a biosynthetic class (see Supplementary Table S3 for cluster rules). However, compared to the microbial version, the biosynthetic classes in 'plantiSMASH' are more of an approximation, since not all signature enzyme families used can be unequivocally used to predict the compound type; e.g. while strictosidine synthase (25) and norcoclaurine synthase (26) are well-characterized members of the Bet v1 enzyme family, it is not clear what proportion of this family have similar Pictet-Spenglerase(-like) catalytic activities.

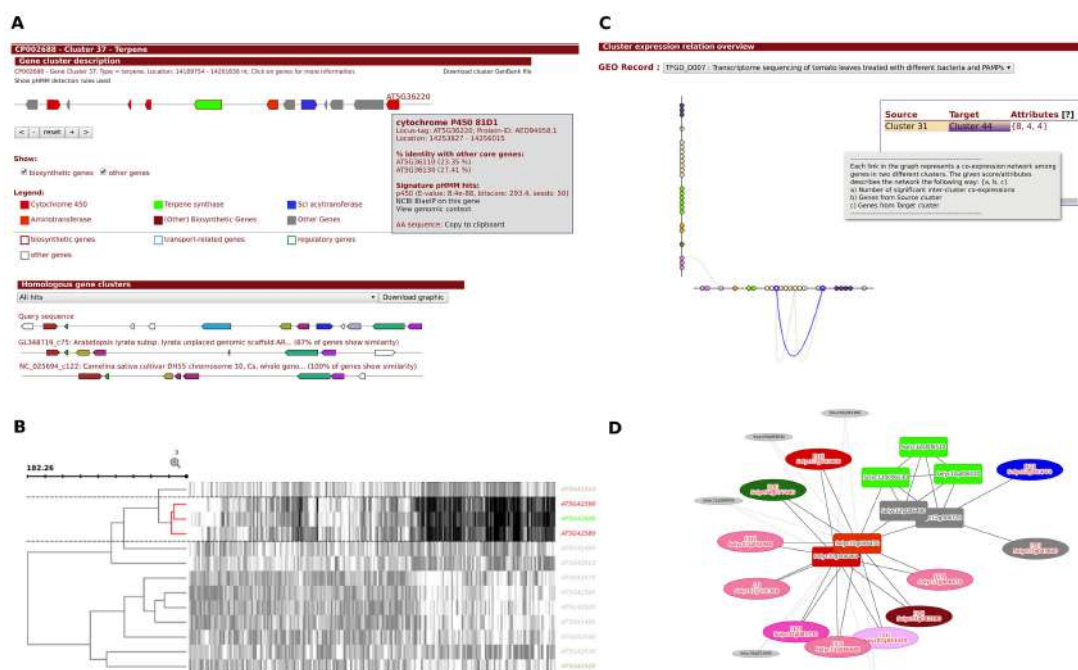
Another particular challenge for BGC detection in plant genomes is the large variation in gene density that occurs not only between but also within plant genomes (16–19). Replacing the static kilobase distance cut-off of microbial antiSMASH by a fixed cut-off based on the maximum number of genes that lie between each pHMM hit also does not provide a solution, as BGCs would then be allowed to cross large repeat regions or even centromeres. Therefore, we chose an alternative, more dynamic, cut-off that is a linear function of local gene density (defined as the gene density of the ten genes nearest to a pHMM hit), and applies a multiplier to calculate the cut-off in kb that is optimal for that specific genomic region (see Supplementary Table S2 and Figures S1 and S2 for results illustrating calibration of the defaults).

### 2.3.2. Flexible and user-friendly input and output

To obtain reliable BGC predictions, a high-quality annotation of gene features in a genome is essential. While we do make available the option to run GlimmerHMM (27) on plant genome sequences, performing *de novo* gene finding on a raw FASTA file is not desirable, given the relatively low accuracy of such a procedure. Because, additionally, the GenBank and EMBL input formats previously accepted for antiSMASH are not available for many plant genomes, we now allow users to supply input also in FASTA+GFF3 format, currently the most widely used format for describing plant genome annotations. For this, we implemented a new module based on Biopython's GFF parsing package ([http://biopython.org/wiki/GFF\\_Parsing](http://biopython.org/wiki/GFF_Parsing)) capable of combining the CDS features from the input sequence, if any, with those of a file compliant to the Generic Feature Format Version 3 as defined by The Sequence Ontology in 2003 (<https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md>). To properly match GFF3 CDS features to their correct sequence, the module demands record names (chromosome/scaffold/contigs) to be identical in both inputs; the only exception being if both inputs only contain one record, in which case the requirement is instead that no feature has coordinates outside the sequence range. This new module allows plantiSMASH to be used with genomes that are only annotated with GFF3 files, such as many of those present in the Joint Genome Institute's Phytozome database

(28).

Based on the biosynthetic gene cluster predictions, a rich and interactive HTML output is generated (Figure 2), which is largely reminiscent of the output of microbial antiSMASH jobs (5). Additionally, genes in the visualization page for each candidate BGC are colored based on the class of enzymes encoded, and a legend is provided that details the color scheme. On mouse click, panels for each gene provide information on the pHMMs that have hits against it, as well as on the amino acid identity to homologous genes within the same locus as calculated by CD-HIT.



**Figure 2.** Outputs generated by the plantiSMASH pipeline. The figure illustrates several visualized outputs generated by plantiSMASH, as they appear for various biosynthetic gene clusters of known natural products. **(A)** Visual overview generated for each gene cluster; in this case, the tirucalladienol cluster from *Arabidopsis thaliana* (47) is shown. Gene annotations and pHMM hit details appear on mouse click. Also, ClusterBlast output showing alignment of homologous genomic loci across other genomes of related species is provided. **(B)** Example of a gene expression heat map, showing coexpression among the core genes of the marnerial BGC from *A. Thaliana* (48) (and not with the flanking genes). **(C)** Hive plot on the overview page, which highlights pairs of candidate BGCs which show many coexpression correlations between their genes; in this example view, the coexpression links between the two loci encoding  $\alpha$ -tomatine biosynthesis in *Solanum lycopersicum* (20) are highlighted (clusters 31 and 44). **(D)** Example ego network that summarizes coexpression correlations between members of the  $\alpha$ -tomatine gene (cluster 44), as well as with genes in other gene clusters (including the other  $\alpha$ -tomatine biosynthetic locus, cluster 31) and with genes elsewhere on the genome.

### 2.3.3. Coexpression analysis identifies pathways within and between gene clusters

As plant scientists are just beginning to understand the phenomenon of metabolic gene clustering in plant genomes, it is currently unknown which proportion of genomic loci that encode multiple contiguous biosynthetic enzyme-encoding genes are *bona fide* BGCs in the sense that their constituent genes are involved in one specific pathway. One powerful strategy to predict whether genes are involved in the

same pathway is the use of coexpression analysis, in which their expression patterns are compared across a wide range of samples. This strategy has proven very effective in the *de novo* identification of gene sets involved in biosynthetic pathways, even if they are not physically clustered on the chromosome (29).

To allow detailed investigation of whether genes in a cluster show coexpression, we added a dedicated analysis module: CoExpress. This module reads transcriptomic datasets, either in SOFT format (from the NCBI Gene Expression Omnibus) or in comma-separated (CSV) format, and generates powerful visualizations of these data for each candidate BGC. Because combining many datasets into one coexpression analysis may blot out coexpression signals that are very specific to certain biological or chemical treatments (which often highly specifically incite expression of plant specialized metabolic pathways), we designed the module in such a way that it visualizes one transcriptomic dataset at a time. This has the added value that the user can browse through multiple datasets and can individually assess specific samples that are linked to a treatment of interest.

The visualizations of within-cluster coexpression patterns are 2-fold: first, a hierarchically clustered heatmap visualization, plotted using a modified version of the InChlib (<http://www.openscreen.cz/software/inchlib/home>) JavaScript library, offers a direct view of patterns in and relationships between the supplied normalized gene expression values. The dendrogram is generated using a coexpression distance metric with a complete-linkage hierarchical clustering method. In this metric, the Pearson Correlation Coefficient (PCC) is transformed directly into a distance value scaled from 0 to 200 (0 for PCC = 1, or positively correlated, and 200 for PCC = -1 or negatively correlated). In order to make correlations maximally visible, the color scheme is normalized per gene (row) by default; however, the user can also select for the color scheme to be normalized by sample (column). Second, a gene cluster-specific coexpression network (30) (with a default distance based cutoff of <50, dynamically adjustable) summarizes the correlations and helps to identify specific groups of genes in the locus that are highly coexpressed: these occur as connected components with high numbers of edges.

Coexpression analysis is not just useful for analysis of functional connections within a candidate BGC, but also allows prediction of functional links with other genomic loci. It is now well-understood that several plant BGCs do not act alone, but rather in concert with another BGC or with individual enzyme-coding genes elsewhere on the genome (11). Therefore, plantiSMASH leverages coexpression data to offer two analyses that identify these trans-genomic interactions: first, the BGC-specific coexpression network can be extended to display a first-order ego network that incorporates genes elsewhere on the genome that either (i) are members of another candidate BGC and show high gene expression correlation (>0.9 PCC) with at least one gene in the BGC, or (ii) contain a 'biosynthetic' domain (defined as being one of the domains in Supplementary Table S1) and show high gene expression correlation with at least two genes in the BGC, at least one of which being a biosynthetic gene itself. Second, interactions between candidate BGCs are summarized in a hive plot, in which pairs of clusters are connected by an edge if the genes of both clusters create at least one subnetwork that satisfies the following criteria: (i) all nodes belong to the same Louvain community (31), as determined by analyzing the full coexpression network of all candidate clusters' genes; (ii) all nodes have a transitivity greater than zero; (iii) the subnetwork contains at least two genes from each cluster; (iv) the subnetwork contains at least one gene per cluster that has

a biosynthetic domain; and (v) The subnetwork contains at least three genes with a biosynthetic domain. This highlights arrangements of pairs of clusters that may be linked functionally via coexpression, and is reminiscent of the characterized -tomatine biosynthetic pathway in *Solanum lycopersicum*, which is encoded in two separate clusters that are highly coexpressed (20). All in all, the coexpression analysis of candidate BGCs allows effective prioritization for, e.g. heterologous expression studies. Yet, it should still be kept in mind that loci that do not show high coexpression might still encode genes that are jointly involved in a biosynthetic pathway, e.g. if the transcriptomic samples available do not include any treatments that induce the expression of the pathway, or if expression of the pathway is sequestered either spatially across tissues or in terms of timing.

### **2.3.4. Comparative genomic analysis shows conservation and diversification**

Comparing a candidate BGC with homologous genomic loci in other plant genomes can give important information on its evolutionary conservation or diversification. Whereas strong conservation of clusteredness across larger periods of evolutionary time may point to a selective advantage of clustering for these genes, diversification of BGCs by cooption of other enzyme-coding genes may give clues to finding novel variants of natural products that have been generated through directional pathway evolution. In order to facilitate such comparative analysis on a case-by-case basis, we constructed a plant-specific version of the antiSMASH ClusterBlast module. To do so, we ran plantiSMASH on a collection of all publicly available plant genomes, obtained from NCBI's GenBank, JGI's Phytozome and Kazusa (32). In order to avoid cases where loci homologous to detected candidate BGCs would not be included in the database by not satisfying the identification criteria, the thresholds for this search were lowered to find all genomic loci with two or more different enzymes, where the CD-HIT cut-off was also set to a generously inclusive level of 0.9. A total of 7978 genomic loci were thus included in the plant ClusterBlast database. As in the microbial version of antiSMASH, the translated protein sequence of each predicted gene in a candidate BGC is searched against this database using the DIAMOND algorithm (33) and genomic loci are sorted based on the number of hits, conserved synteny and cumulative bit score. To also facilitate direct comparison with known plant BGCs, all plant BGCs with known products for which the sequence was available were added to the MIBiG repository (34), which allows users to find similarities between newly identified and known clusters with the KnownClusterBlast module of antiSMASH.

### **2.3.5. Precomputed results allow fast access to comprehensive plantiSMASH results**

In order to allow users to directly access plantiSMASH results for publicly available plant genomes, runs for 47 high quality plant genomes were precomputed and made available online at <http://plantismash.secondarymetabolites.org/precalc>. Importantly, publicly available gene expression datasets with sufficient numbers of samples to be suitable for coexpression analysis were loaded into these results. In total, 73 transcriptomic datasets were included for five species: *A. thaliana*, *S. lycopersicum*, *Oryza sativa*, *Zea mays* and *Glycine max* (Supplementary Tables S4–

7). As an indication for web server users: the computations took about 24 min per genome on average, on a 2 GHz CPU, depending on the size of the genome and pre-selected additional analyses including the co-expression analysis (see further details in Supplementary Table S4).

Sequences that are not publicly available (as well as available sequences with custom transcriptomic datasets) can be analyzed directly using the plantiSMASH web server at <http://plantismash.secondarymetabolites.org>. In this way, plantiSMASH results for all kinds of genomes and transcriptomes are optimally available to users.

## 2.4. Results and Discussion

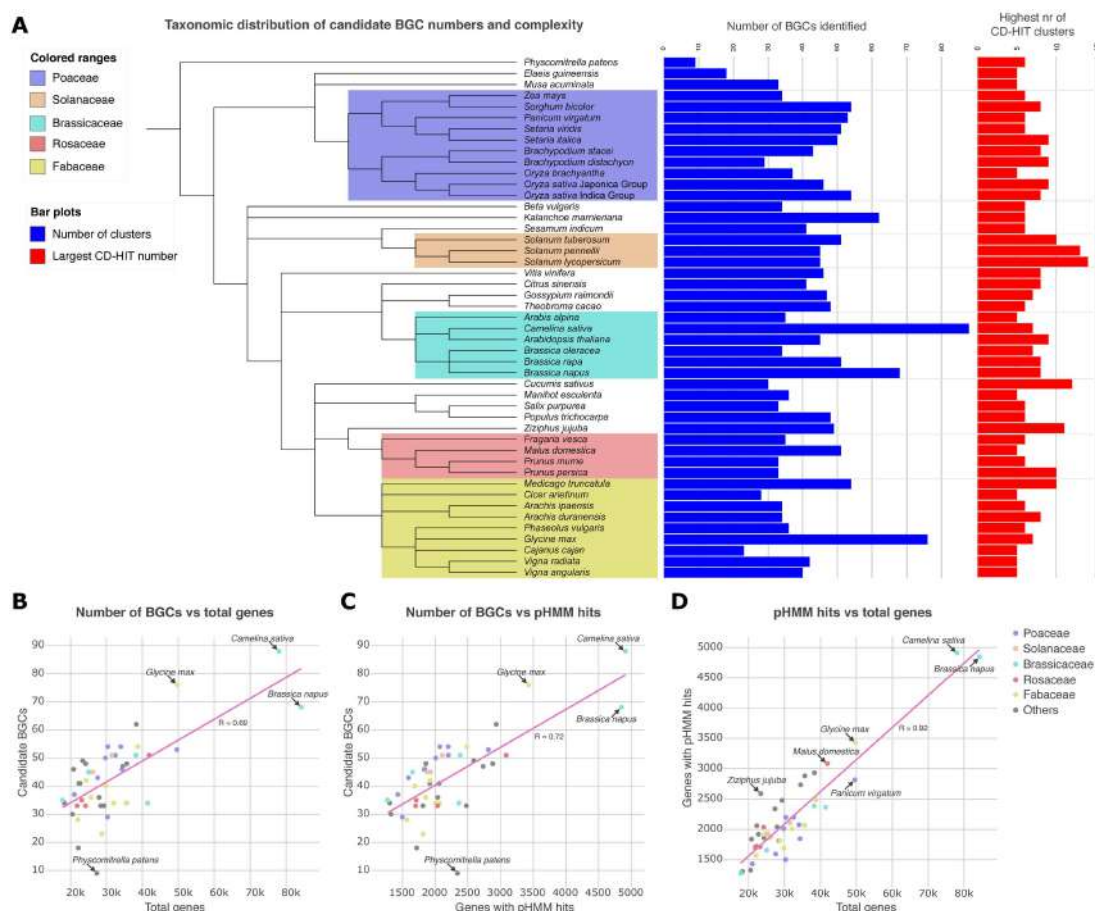
### 2.4.1. PlantiSMASH successfully detects all experimentally characterized plant biosynthetic gene clusters

Even though only a relatively small set of plant BGCs has been discovered, these ~30 BGCs still present the best objective test case for the BGC detection algorithm. Importantly, they range from complex BGCs with many different enzyme-coding genes, such as the noscapine and cucurbitacin BGCs (21,35), to relatively simple ones that only encode a couple of enzymes, such as the dhurrin and linamarin/lotaustralin BGCs (36). Of this set, only 19 BGCs have annotated sequence information publicly available. When plantiSMASH was run on a multi-GenBank file containing accurately annotated versions of these 19 known BGCs, all clusters were successfully detected with default settings. When run on different genome annotation versions available from GenBank or Phytozome, BGCs of low complexity (i.e. with a small number of enzyme-coding genes) were occasionally missed when key genes were missing from the structural annotations or when many false positive gene assignments were present in the region of interest (affecting the dynamic gene density-based cut-off of plantiSMASH): for example, the linamarin BGC from *Lotus japonicus* was not detected in assembly/annotation version 3.0, while it was detected in the older version 2.5. This highlights the importance of using high-quality genome annotations supported by transcriptomic data when using plantiSMASH to search for BGCs of interest. Alternatively, the stand-alone version of plantiSMASH provides additional cut-off methods (e.g. raw distance-based or gene-count-based) that can be attempted as well to mitigate such issues.

### 2.4.2. Plant genomes contain large numbers of complex biosynthetic gene clusters

When run on the 47 plant genomes for which chromosomelevel assemblies are currently available on either NCBI or Phytozome, plantiSMASH found a wide variety of candidate BGC numbers across plant taxonomy (Figure 3). In general, the numbers of candidate BGCs were relatively even between monocots and dicots (while very low in the only moss genome included), while the largest numbers of BGCs were found in dicot genomes. These outliers all corresponded to recent

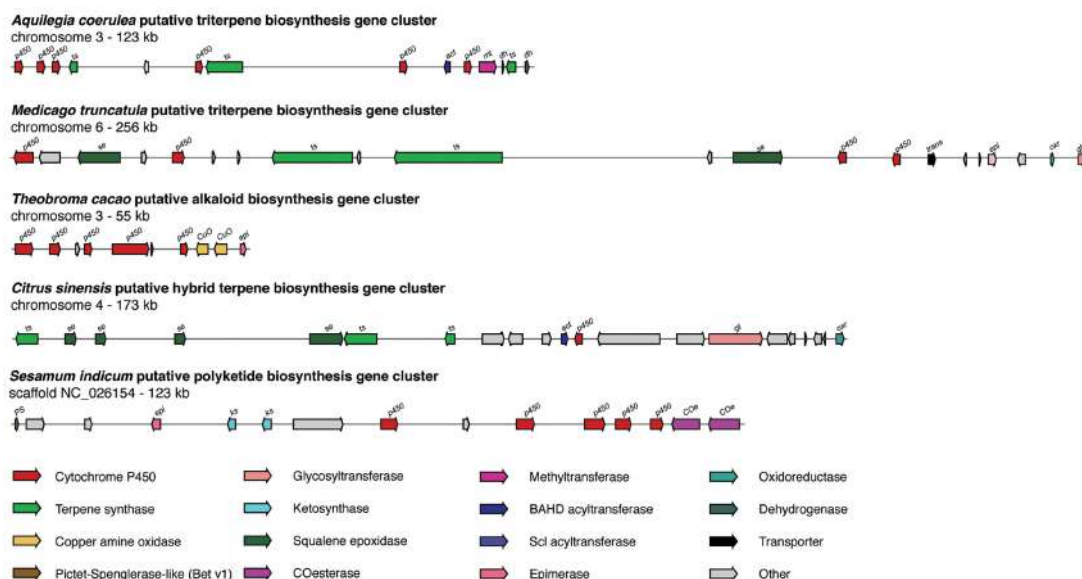
(partial) genome amplification events, such as in the case of *Camelina sativa* (37) with 88 candidate BGCs, *Brassica napus* (38) with 68 candidate BGCs and *G. max* (39) with 76 candidate BGCs.



**Figure 3.** Numbers of candidate BGCs identified across the Plant Kingdom. **(A)** PlantSMASH BGC predictions plotted onto a phylogenetic tree of plant species for which chromosome-level genome assemblies are available. The blue bars indicate the number of candidate BGCs per genome, the red bars indicate the most complex candidate BGC identified in each species (in terms of the number of unique enzymes encoded, as defined by CD-HIT groups). **(B)** Number of candidate BGCs plotted versus the total number of genes; as expected, more BGCs are found in larger genomes. Outliers represent genomes that have recently undergone whole-genome duplication, and the moss *Physcomitrella patens*, in the genome of which only a very low number of candidate BGCs is found. **(C)** Number of candidate BGCs plotted versus the number of genes with pHMM hits to biosynthetic domains. **(D)** Number of genes with biosynthetic domains plotted against the total number of genes; a linear correspondence is largely observed.

In many plant genomes, candidate BGCs of high complexity were identified, with as many as seven or eight different enzymatic classes encoded in the same tight genomic region. These constitutions are clearly non-random and make it promising to study candidate BGCs even in the absence of coexpression data. Dozens of such complex BGCs were found, which cover all known as well as putative pathway classes; examples are provided in Figure 4.





**Figure 4.** Example candidate BGCs identified by plantiSMASH. Five example candidate BGCs are shown, which cover a diverse range of enzymatic classes. Dozens of candidate BGCs of comparable complexity can be found across the precomputed plantiSMASH results that are available online.

### 2.4.3. Coexpression patterns can guide BGC prioritization

We subjected the candidate BGCs identified in the genome of *A. thaliana* to a more detailed statistical analysis using within-cluster coexpression in a merged transcriptomic dataset. For this, we compiled two sets of gene expression datasets, one containing transcriptomic experiments of biological treatments (defense; Supplementary Table S5) and one containing experiments of hormone treatments and non-biological stress inductions (Supplementary Table S6). Together, these datasets comprise transcriptomic measurements of 1047 samples. The Mann–Whitney U one-sided test was selected to test which of the *A. thaliana* BGCs have a statistically greater within-cluster coexpression distribution than the genome’s background coexpression distribution. Given a BGC consisting of  $x$  genes, the background distribution for the statistical test of this cluster contains all PCCs between pairs of genes that are  $x-1$ ,  $x-2$ , ..., 0 genes away from each other across the entire genome (except predicted BGCs). Only genes observed in all transcriptomic experiments were allowed in the test, and only PCCs between genes that each have a Median Absolute Deviation  $>0$  are added to the distributions. Lastly, the CD-HIT algorithm was run on the entire *A. thaliana* proteome at 0.5 identity cutoff (same as plantiSMASH’s default) to cluster all similar enzymes. The same statistical tests were repeated afterward, but this time discarding PCCs between genes that code for enzymes within the same CD-HIT cluster, ensuring both distributions only include coexpression of genes that produce enzymes of different classes, which more accurately resembles the type of interactions desired in a *bona fide* BGC. The results of these analyses (Supplementary Table S8 and Figure S3) show that at a significance level of 0.05, 11 predicted BGCs showed statistically higher within cluster coexpression than their respective background distribution even when discarding coexpression between genes in the same CD-HIT cluster. This list includes the four known *A. thaliana* BGCs, encoding the biosynthetic pathways for arabidiol/baruol ( $P = 2.92e-40$ ), thalianol ( $P = 1.94e-17$ ), marneral ( $P = 7.03e-10$ ) and tirucalla-7,24-dien3-ol ( $P = 1.10e-4$ ), which corroborates that coexpression is a

valid criterion to prioritize functional BGCs.

There are several explanations for the fact that strong coexpression is observed for some candidate BGCs but not others. A first explanation is that their coordinated expression is induced by conditions not included in these transcriptomic experiments; in other words, absence of evidence of coexpression is not evidence of absence of coexpression. A second explanation is that a number of candidate BGCs probably do not encode entire consistently coexpressed biosynthetic pathways by themselves; evidence for this comes from an analysis of characterized enzyme-coding genes inside these candidate BGCs (Supplementary Table S9); e.g. *AT1G24100* and *AT5G57220*, which occur in two different candidate BGCs, are known to each be involved in a different branch of glucosinolate biosynthesis (40,41), a complex multifurcated pathway that shows only partial and fragmented genomic clustering. Contrary to what might be expected, however, there was no strong correlation ( $R = 0.004$ , and  $P = 0.64$  when fitting linear regression) of coexpression with cluster size, which suggests that the default plantiSMASH BGC prediction cut-offs are not set too inclusively.

All in all, coexpression analysis provides a powerful tool to prioritize the candidate BGCs detected by plantiSMASH that are most likely to encode functional pathways.

## 2.5. Conclusions

The highly automated discovery of candidate BGCs by plantiSMASH and the powerful visualizations of coexpression data that allow their prioritization present a key technological step in the route toward high-throughput genome mining of plant natural products. As plant genome sequencing and assembly technologies continue to improve at a rapid pace, it is likely that high-quality plant genomes for thousands of species will soon be available; hence, ‘clustered’ biosynthetic pathways present low-hanging fruits for the discovery of novel molecules. Empowered by synthetic biology tools and powerful heterologous expression systems in yeast and tobacco (42–46), this will likely make it possible to scale up plant natural product discovery tremendously.

Continued development of the antiSMASH/plantiSMASH framework in the future is needed to further accelerate this process: e.g. the development of (machine-learning) algorithms that predict substrate specificities of key enzymes like terpene synthases, and the systematic construction of pHMMs for automated subclassification of complex enzyme families such as cytochrome P450s and glycosyltransferases, will allow more powerful predictions of the natural product structural diversity encoded in diverse BGCs. Additionally, detailed evolutionary genomic analysis of the phenomenon of gene clustering, including BGC birth, death and change processes, will further our understanding of how BGCs facilitate natural product diversification during evolution. As more plant BGCs are experimentally characterized, the algorithms will co-evolve with the knowledge gained, and more detailed class-specific cluster detection rules could be designed; moreover, it will become clearer what does and what does not constitute a *bona fide* BGC. Finally, when scientists further unravel the complexities of tissue-specific and differentially timed gene expression of plant biosynthetic pathways, we will learn more on how best to leverage coexpression data for biosynthetic pathway prediction.

Thus, a more comprehensive understanding of the remarkable successes of



evolution to generate an immense diversity of powerful bioactive molecules will hopefully make it possible for biological engineers to mimic nature's strategies and deliver many useful new molecules for use in agricultural, cosmetic, dietary and clinical applications.

## 2.6. Supplementary Information

Supplementary figures and tables are available to download from: <https://doi.org/10.1093/nar/gkx305>

## References

1. Jensen, P.R. (2016) Natural products and the gene cluster revolution. *Trends Microbiol.*, **24**, 968–977.
2. Medema, M.H. and Fischbach, M.A. (2015) Computational approaches to natural product discovery. *Nat. Chem. Biol.*, **11**, 639–648.
3. Rutledge, P.J. and Challis, G.L. (2015) Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat. Rev. Microbiol.*, **13**, 509–523.
4. Ziemert, N., Alanjary, M. and Weber, T. (2016) The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.*, **33**, 988–1005
5. Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E. and Breitling, R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, **39**, W339–W346.
6. Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E. and Weber, T. (2013) antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.*, **41**, W204–W212.
7. Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., Lee, S.Y., Fischbach, M.A., Muller, R., Wohlleben, W. et al. (2015) antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.*, **43**, W237–W243.
8. Frey, M., Chomet, P., Glawischnig, E., Stettner, C., Grun, S., Winklmair, A., Eisenreich, W., Bacher, A., Meeley, R.B., Briggs, S.P. et al. (1997) Analysis of a chemical plant defense mechanism in grasses. *Science*, **277**, 696–699.
9. Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R. and Osbourn, A. (2004) A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 8233–8238.
10. Nutzmans, H.-W. and Osbourn, A. (2014) Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.*, **26**, 91–99.
11. Nutzmans, H.W., Huang, A. and Osbourn, A. (2016) Plant metabolic gene clusters—from genetics to genomics. *New Phytol.*, **211**, 771–789.

12. Yu,N., Nutzmam,H.-W., MacDonald,J.T., Moore,B., Field,B., Berriri,S., Trick,M., Rosser,S.J., Kumar,S.V., Freemont,P.S. *et al.* (2016) Delineation of metabolic gene clusters in plant genomes by chromatin signatures. *Nucleic Acids Res.*, **44**, 2255–2265.
13. VanBuren,R., Bryant,D., Edger,P.P., Tang,H., Burgess,D., Challabathula,D., Spittle,K., Hall,R., Gu,J., Lyons,E. *et al.* (2015) Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature*, **527**, 508–511.
14. Medema,M.H. and Osbourn,A. (2016) Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Nat. Prod. Rep.*, **33**, 951–962.
15. Shi,M.-Z. and Xie,D.-Y. (2014) Biosynthesis and metabolic engineering of anthocyanins in *Arabidopsis thaliana*. *Recent Pat. Biotechnol.*, **8**, 47–60.
16. Ibarra-Laclette,E., Lyons,E., Hernandez-Guzman,G., Perez-Torres,C.A., Carretero-Paulet,L., Chang,T.-H., Lan,T., Welch,A.J., Juarez,M.J.A., Simpson,J. *et al.* (2013) Architecture and evolution of a minute plant genome. *Nature*, **498**, 94–98.
17. Keller,B. and Feuillet,C. (2000) Colinearity and gene density in grass genomes. *Trends Plant Sci.*, **5**, 246–251.
18. Kellogg,E.A. and Bennetzen,J.L. (2004) The evolution of nuclear genome structure in seed plants. *Am. J. Bot.*, **91**, 1709–1725.
19. Sandhu,D. and Gill,K.S. (2002) Gene-containing regions of wheat and the other grass genomes. *Plant Physiol.*, **128**, 803–811.
20. Itkin,M., Heinig,U., Tzfadia,O., Bhide,A.J., Shinde,B., Cardenas,P.D., Bocobza,S.E., Unger,T., Malitsky,S., Finkers,R. *et al.* (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, **341**, 175–179.
21. Shang,Y., Ma,Y., Zhou,Y., Zhang,H., Duan,L., Chen,H., Zeng,J., Zhou,Q., Wang,S., Gu,W. *et al.* (2014) Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science*, **346**, 1084–1088.
22. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
23. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
24. Finn,R.D., Bateman,A., Clements,J., Coghill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
25. Wu,F., Zhu,H., Sun,L., Rajendran,C., Wang,M., Ren,X., Panjekar,S., Cherkasov,A., Zou,H. and Stockigt,J. (2012) Scaffold tailoring by a newly detected Pictet-Spenglerase activity of strictosidine synthase: from the common tryptoline skeleton to the rare piperazino-indole framework. *J. Am. Chem. Soc.*, **134**, 1498–1500.
26. Lee,E.-J. and Facchini,P. (2010) Norcoclaurine synthase is a member of the pathogenesis-related 10/Bet v1 protein family. *Plant Cell*, **22**, 3489–3503
27. Majoros,W.H., Pertea,M. and Salzberg,S.L. (2004) TigrScan and GlimmerHMM:

- two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.
28. Goodstein,D.M., Shu,S., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U., Putnam,N. et al. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
  29. Rajniak,J., Barco,B., Clay,N.K. and Sattely,E.S. (2015) A new cyanogenic metabolite in *Arabidopsis* required for inducible pathogen defence. *Nature*, **525**, 376–379.
  30. Serin,E.A.R., Nijveen,H., Hilhorst,H.W.M. and Ligterink,W. (2016) Learning from co-expression networks: possibilities and challenges. *Front. Plant Sci.*, **7**, 444.
  31. Blondel,V., Guillaume,J., Lambiotte,R. and Lefebvre,E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.*, **10**, P10008.
  32. Sato,S., Nakamura,Y., Kaneko,T., Asamizu,E., Kato,T., Nakao,M., Sasamoto,S., Watanabe,A., Ono,A., Kawashima,K. et al. (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res.*, **15**, 227–239.
  33. Buchfink,B., Xie,C. and Huson,D.H. (2014) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
  34. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., de Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C. et al. (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
  35. Winzer,T., Gazda,V., He,Z., Kaminski,F., Kern,M., Larson,T.R., Li,Y., Meade,F., Teodor,R., Vaistij,F.E. et al. (2012) A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science*, **336**, 1704–1708.
  36. Takos,A.M., Knudsen,C., Lai,D., Kannangara,R., Mikkelsen,L., Motawia,M.S., Olsen,C.E., Sato,S., Tabata,S., Jørgensen,K. et al. (2011) Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in *Lotus japonicus* and suggests the repeated evolution of this chemical defence pathway. *Plant J.*, **68**, 273–286.
  37. Kagale,S., Koh,C., Nixon,J., Bollina,V., Clarke,W.E., Tuteja,R., Spillane,C., Robinson,S.J., Links,M.G., Clarke,C. et al. (2014) The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. *Nat. Commun.*, **5**, 3706.
  38. Chalhoub,B., Denoeud,F., Liu,S., Parkin,I.A.P., Tang,H., Wang,X., Chiquet,J., Belcram,H., Tong,C., Samans,B. et al. (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science*, **345**, 950–953.
  39. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
  40. Grubb,C.D., Zipp,B.J., Ludwig-Müller,J., Masuno,M.N., Molinski,T.F. and Abel,S. (2004) *Arabidopsis* glucosyltransferase UGT74B1 functions in glucosinolate biosynthesis and auxin homeostasis. *Plant J.*, **40**, 893–908.
  41. Pfalz,M., Vogel,H. and Kroymann,J. (2009) The gene controlling the indole glucosinolate modifier1 quantitative trait locus alters indole glucosinolate structures and aphid resistance in *Arabidopsis*. *Plant Cell*, **21**, 985–999.

42. Casini,A., Storch,M., Baldwin,G.S. and Ellis,T. (2015) Bricks and blueprints: methods and standards for DNA assembly. *Nat. Rev. Mol. Cell Biol.*, **16**, 568–576.
43. Liu,W., Yuan,J.S. and Stewart,C.N. (2013) Advanced genetic tools for plant biotechnology. *Nat. Rev. Genet.*, **14**, 781–793.
44. Patron,N.J. (2014) DNA assembly for plant biology: techniques and tools. *Curr. Opin. Plant Biol.*, **19**, 14–19.
45. Patron,N.J., Orzaez,D., Marillonnet,S., Warzecha,H., Matthewman,C., Youles,M., Raitskin,O., Leveau,A., Farre,G., Rogers,C. et al. (2015) Standards for plant synthetic biology: a common syntax for exchange of DNA parts. *New Phytol.*, **208**, 13–19.
46. Thimmappa,R., Geisler,K., Louveau,T., O'Maille,P. and Osbourn,A. (2014) Triterpene biosynthesis in plants. *Annu. Rev. Plant Biol.*, **65**, 225–257.
47. Boutanaev,A.M., Moses,T., Zi,J., Nelson,D.R., Mugford,S.T., Peters,R.J. and Osbourn,A. (2014) Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E81–E88.
48. Field,B. and Osbourn,A.E. (2008) Metabolic diversification–independent assembly of operon-like gene clusters in different plants. *Science*, **320**, 543–547.

# Chapter 3

## **Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae**

Zhenhua Liu<sup>1\*</sup>, Hernando G. Suarez Duran<sup>2\*</sup>, Yosapol Harnvanichvech<sup>2</sup>, Michael J. Stephenson<sup>1</sup>, M. Eric Schranz<sup>3</sup>, David Nelson<sup>4</sup>, Marnix H. Medema<sup>2</sup> and Anne Osbourn<sup>1</sup>

<sup>1</sup> Department of Metabolic Biology, John Innes Centre, Norwich Research Park, Colney Lane, Norwich, NR4 7UH, UK

<sup>2</sup> Bioinformatics Group, Wageningen University, Droevendaalsesteeg 1 6708PB, Wageningen, the Netherlands

<sup>3</sup> Biosystematics Group, Wageningen University, Droevendaalsesteeg 1 6708PB, Wageningen, the Netherlands

<sup>4</sup> Department of Microbiology, Immunology and Biochemistry, University of Tennessee, 858 Madison Avenue, Suite G01, Memphis, TN 38163, USA

*\* These authors contributed equally to the paper as first authors.*

*Published in New Phytologist 227:1109-1123*

### 3.1. Introduction

Plants are chemical engineers *par excellence*, and are collectively estimated to make over a million different specialized metabolites (Afendi *et al.*, 2012). These compounds have important ecological functions, providing protection against attack by pests and pathogens, inhibiting the growth of competing plants, shaping the plant microbiome, and serving as attractants for seed dispersal agents and pollinators (Weng *et al.*, 2012a; Huang *et al.*, 2019). Plant natural products also are a rich source of bioactives for medicinal, agricultural and industrial applications (Böttger *et al.*, 2018). Despite their tremendous chemical diversity, the mechanisms underpinning the evolution of new metabolic pathways are poorly understood. Although individual enzymes are known to be recruited primarily through gene duplication, often involving sub- or neo-functionalization, little is known about how pathways consisting of multiple biosynthetic steps originate (Weng, 2014). It has recently become apparent that the genes for the biosynthesis of various natural products, such as the plant defence compound avenacin A-1 from oat (antimicrobial triterpene) (Qi *et al.*, 2004), noscapine from opium poppy (anticancer alkaloid) (Winzer *et al.*, 2012; Guo *et al.*, 2018), and dhurrin from sorghum (an insect repellent cyanogenic glycoside) (Takos *et al.*, 2011) are clustered in plant genomes. These plant biosynthetic gene clusters (BGCs) consist of at least three different classes of enzyme-encoding genes (Medema & Osbourn, 2016) and bring unique perspectives towards metabolic innovation and diversification. First, plant BGCs have not originated from microbes via horizontal gene transfer. Instead, they appear to have arisen from plant genes by as yet unknown mechanisms, presumably honed by rounds of natural selection (Field & Osbourn, 2008). Those BGCs that make the same/very similar types of compounds are usually restricted to a narrow taxonomic window of closely related lineages (Nutzmann *et al.*, 2016), suggesting that these BGCs must have assembled relatively recently in evolutionary time. Second, plant BGCs usually contain a gene encoding an enzyme that makes the natural product scaffold, along with a combination of genes encoding other types of enzymes that modify this scaffold (tailoring enzymes) (Medema & Osbourn, 2016). The ‘rules’ that govern the evolutionary interplay between genes encoding different scaffold-generating enzymes and tailoring enzymes are not understood. Therefore, understanding the evolution and diversification of plant BGCs is expected to offer new insights into how plants have acquired the ability to synthesize such a remarkable diversity and complexity of specialized metabolites.

The terpenoids are the major class of plant natural products, comprising ~40% of the plant natural products discovered thus far (Chassagne *et al.*, 2019); of these, the triterpenes are the largest and most structurally complex (> 20 000 reported so far) (Christianson, 2017). We have previously shown that the genes for the biosynthesis of structurally diverse triterpenes are organized in BGCs in the genome of the model plant *Arabidopsis thaliana* (Field & Osbourn, 2008). *Arabidopsis thaliana* is a member of the mustard family, Brassicaceae (Cruciferae), which includes economically important crop plants such as turnip, cabbage and oilseed rape. When we started this project, 13 high-quality sequenced Brassicaceae genomes covering lineages I-II (Beilstein *et al.*, 2010) and early diverging species were available (Supporting Information Table S1). The evolutionary relationships among most of

these genomes are well-defined by phylogenomic analysis (Beilstein *et al.*, 2010). Furthermore, several explicit examples of gene duplication-promoted metabolic innovation have been reported in the Brassicaceae (Weng *et al.*, 2012b; Edger *et al.*, 2015; Liu *et al.*, 2016), demonstrating that Brassicaceae genomes are excellent working materials for exploring the genomic basis underpinning metabolic diversification. Herein we take advantage of the extensive resource of available high-quality genome sequences to systematically investigate the genomic mechanisms underlying triterpene diversification in the wider Brassicaceae.

## 3.2. Materials and Methods

### 3.2.1. Genome mining

Genomes were retrieved from NCBI (<https://www.ncbi.nlm.nih.gov/>), CoGE (<https://www.genomeevolution.org/coge/>) and Phytozome v.12.0 (<https://phytozome.jgi.doe.gov/>) in GenBank format. To thoroughly identify oxidosqualene cyclase (OSC) loci, both Hmmer3 (Finn *et al.*, 2011) and Blastp (Altschul *et al.*, 1990) were used to identify OSC homologues. The Hmmer profiles (pHMMs) were downloaded from the PFAM library (Finn *et al.*, 2016). PF13243 (targeting OSC N-terminal) and PF13249 (targeting OSC C-terminal) were used to search for OSC homologues with hmmsearch. The cut\_tc (trusted cut-off) option was used. For Blastp, protein identity  $\geq 40$  and bit score  $\geq 100$  were used as cut-offs, and AtCAS1 (At2g07050) was used as query sequence. Blastp identified the same OSCs present in the Hmmer analysis, but Hmmer showed slightly more candidates. The Hmmer output was aligned with outgroup protein PGGT1B\_sp (geranylgeranyltransferase type I) by using Muscle (Edgar, 2004), generating a multiple sequence alignment (MSA). The MSA then was trimmed manually to keep only the conserved domains and used to build a phylogenetic tree with FastTree 2.1 using standard parameters (Price *et al.*, 2010). Support for tree nodes was assessed using fast global bootstrap iterations (Tamatakis *et al.*, 2008; Price *et al.*, 2010). Bootstrap values (1000 iterations)  $> 0.7$  are shown in Fig. S1. Proteins grouped with outgroup PGGT1B\_sp were discarded (nine proteins). To fully annotate the tree, we propagated subfamily annotations present in *A. thaliana*, resulting in three distinct groups: clade I, clade II and the sterol clade (Field and Osbourn, 2008). Another phylogeny of OSCs was reconstructed using the RAxML method (Kozlov *et al.*, 2019), which generated a tree with 134 of 164 partitions identical to the tree generated by FastTree, differing mainly in some deep ancestral splits that are difficult to resolve (Fig. S2). Pfam domains PF00067 and PF02458 were used to identify members of the cytochrome P450 (CYP) gene/protein family and the acyltransferase (ACT) gene/protein family, respectively. The identified CYP and ACT genes were pooled separately, and we used the hmalign – trim option to trim nonconserved domains. Hmalign-based MSAs were taken as input to infer phylogenetic trees with FastTree. Well-annotated *A. thaliana* CYPs (Nelson, 2009) and ACTs (Tuominen *et al.*, 2011) were used as markers to annotate protein families

and subfamilies in the phylogenetic tree (Figs S3, S4). The CYPs that were functionally characterized in this study were formally assigned to subfamilies and named as CYP708A9, CYP708A10, CYP705A38, CYP708A11 and CYP705A37v2 according to procedures of the Cytochrome P450 homepage (Nelson, 2009).

### 3.2.2. Genomic Neighbourhoods (GN) association analysis

We developed a simple tool to identify and isolate the GNs of a given list of target genes based on a user-specified number of flanking genes. In this study, we defined 'GNs' as the OSC flanking regions extending five genes either side of an OSC gene. This resulted in most GNs consisting of 11 genes with two exceptions: OSCs located in a scaffold with < 11 genes result in smaller GNs, and OSCs in close proximity to each other result in overlapping GNs, which are merged into one large GN containing more than 11 genes. After identifying all OSC GNs across the Brassicaceae genomes, Pfam protein domain content was predicted using Hmmer.

Because of the differences in domain content among the GNs of the three OSC clades (Field & Osbourn, 2008), we explored the enrichment of protein domains in the OSC GNs separately for each OSC clade by comparing them to the rest of the genomes with a hypergeometric test (one-tailed exact Fisher's test) and the Bonferroni correction for multiple comparisons. The close phylogenetic relationships between the species in our study potentially presents a problem for standard statistical procedures, because various studies have shown that the common evolutionary history of related species results in an abundance of type I errors (Martins & Garland 1991; Martins *et al.*, 2002), in part due to the nonindependence of the samples. To address the possibility of an inflation of degrees of freedom within the enrichment test, we used the OSC phylogenetic tree that we had generated previously and grouped together all monophyletic groups of leaves that share the trait targeted by the test. We used the conservative assumption that each clade of OSCs for which all GNs contain a particular protein domain represents a group of vertically inherited orthologues. Based on this, we then selected only the leaf with the highest target protein domain count for the enrichment test, instead of each leaf individually, effectively increasing the independence of the samples and resulting in a reduction of successful counts ( $k$  in the hypergeometric test). Because of the high abundance of CYP and Transferase domains in the OSC GNs, we also tested the enrichment of CYP and Transferase subfamilies separately.

Multiple methods that incorporate phylogenetic information into statistical procedures have been developed, with no method being objectively better than all others in all cases. All of them have the disadvantage, compared to the Fisher's exact test, that they do not take into account the frequency at which a given domain occurs in the rest of the genome. To further reduce the problems that arise from phylogenetic relatedness, in our study, we selected two of these phylogenetic methods to complement our initial exploration of enriched domains: phylogenetic logistic regression (PLR), and phylogenetic generalized linear mixed models (PGLMM) (Ives & Garland, 2010), both of which have been shown to be robust and adequate for studying the evolution of binary traits (Garamszegi, 2014). The main difference between the models is how the phylogenetic component is applied to the regression models: in PLR models, the dependent variable (the trait of interest)



evolves to 0 or 1 through a lineage according to a switching rate, whereas PGLMM includes an additional hidden trait that evolves through the phylogenetic tree and determines the probability of the dependent variable evolving to 0 or 1 at the tip. When using regression methods such as PLR and PGLMM, the strength of association between the dependent and independent variables is not symmetrical, making their selection an important step for correct interpretation. We selected the absence and presence of OSC subfamilies in the GNs as the independent variables (i.e. the GN-centric OSC belonging either to clade I, clade II or the sterol clade), and the absence and presence of each of the other protein domains as dependent variables; this approach allows us to interpret significant associations in the regression models that indicate whether the absence/presence of a given OSC subfamily (as opposed to others) in a GN is a strong predictor of the absence/presence of specific protein domains in their GNs.

We performed the phylogenetic regressions in R/ape v.5.0 using the 'binaryPGLMM' function found in the package (Paradis *et al.*, 2004) for PGLMM, and R/phyloilm v.2.6 (Ho & Ané 2014) for PLR. Because we were interpreting the results of the two regression models in conjunction with other statistical tests and phylogenomic analysis, we set the significance threshold at  $P < 0.1$  and did not apply multiple testing correction.

### 3.2.3. All-vs-all comparison of OSC GNs

In order to assess the similarity amongst the identified OSC GNs, we first measured the average amino acid identity between protein domains that appear in compared GNs. In the case of GNs with multiple copies of any domain, the sequence identity of all possible domain pairs is identified and the Hungarian algorithm (Kuhn, 1955) is used to select the pairs resulting in the highest average. When a GN has extra copies of a protein domain, the highest identity scores are selected, and the additional domains are removed until both GNs have the same number of copies, after which the Hungarian algorithm is used. We next used two other tools from BiG-Scape (Navarro-Muñoz *et al.*, 2018) (<https://git.wageningenur.nl/medema-group/BiG-SCAPE>) to compare GNs. BiG-Scape compares loci with multiple genes based on three criteria, two of which are of interest to us: domain composition similarity, measured here with the Jaccard index of identified protein domains, and domain sequence similarity (DSS), measured by averaging the sequence similarity of shared domains, with a score penalization when a GN has extra copies of a protein domain that are not present in the other. To ensure a fair comparison, only OSC GNs with  $\geq 11$  genes were selected as input for the tool.

### 3.2.4. Mapping GNs to their whole-genome duplication (WGD)-derived syntenic blocks

The WGD-derived syntenic regions in *A. thaliana* are defined by 'anchor genes' which retain sequence similarity with its paralogue in its sister region, and the full list of anchor genes and syntenic regions is publicly available (Freeling *et al.*, 2007). We

identified the anchor genes upstream and downstream of each OSC GN and used them to define their corresponding sister regions. To compare each GN with its sister region, we generated a MSA with Muscle (Edgar, 2004) and a sequence similarity matrix with R/seqinr v.3.4 (Charif & Lobry, 2007).

### 3.2.5. Ancestral states reconstruction

We pruned and isolated clade II OSCs from the phylogenetic tree and then removed the 11 leaves corresponding to GNs with < 11 genes. Given that GNs with multiple OSCs appear as multiple leaves in the phylogenetic tree, only one leaf was selected as being representative of the neighborhood based on maximizing the DSS (domain sequence similarity) index with the surrounding leaves in the tree; the nonrepresentative OSC leaves were removed.

In order to explore the evolutionary history of the selected GNs, we assigned binary traits to each leaf according to the presence (=1) or absence (=0) of the most abundant tailoring enzymes within the clade II OSC GNs (CYP702A, CYP705A, CYP708A and ACT IIIa) and reconstructed their ancestral states through the tree via maximum parsimony by using the Mesquite software (Maddison & Maddison, 2008). We selected the maximum parsimony criterion because it does not require the assumption of an underlying model of evolution for the assembly of plant biosynthetic gene clusters (BGCs), a process that is still poorly understood.

We used the CYP and ACT phylogenetic trees to validate the ancestral state reconstruction of CYP705A and ACT IIIa in the OSC tree. For this, we pruned both trees to remove all genes in a scaffold with < 11 genes (32 CYP705As, 59 ACT IIIas) and isolated the CYP705A and ACT IIIa subtrees. Furthermore, as additional validation, we also reconstructed the ancestral states on an OSC phylogeny generated by RAxML. Although the resulting tree had some differences in the deep ancestral splits (Fig. S2), ancestral state reconstruction on this tree would lead to the same conclusions regarding multiple parallel origins of complex triterpene biosynthetic loci.

### 3.2.6. Evolutionary tests

The OSC protein-coding DNA sequences were aligned with TranslatorX (Abascal *et al.*, 2010). Genes with sequence length < 2200 nucleotides and poorly aligned genes were both filtered. The final alignment and the input tree used for the various evolutionary tests can be found on Zenodo (doi: 10.5281/zenodo.3531676). This alignment was subjected to evolutionary tests in the HyPhy package (Pond *et al.*, 2005). BUSTED analysis was used to infer whether a gene has experienced positive selection at at least one site on at least one branch (Murrell *et al.*, 2015). MEME analysis was used to detect individual sites evolving under positive selection in a proportion of branches (Murrell *et al.*, 2012). For both analyses, 'universal genetic code' was selected and  $P < 0.05$  was set for significance.

### 3.2.7. Plant material and growth conditions

*Capsella rubella* (Monte Gargano) and *Brassica rapa* (r-o-18) seeds were obtained from Lars Østergaard and *Arabidopsis lyrata* seeds (accession VLP6) from

Levi Yant (John Innes Centre). Plants were grown from these seeds in a controlled growth chamber at 22°C under long-day light conditions (16 h : 8 h, light : dark). *Nicotiana benthamiana* plants were grown in a glasshouse, under the same long-day light conditions.

### 3.2.8. RNA isolation and RT-qPCR analysis

Total RNA for *A. thaliana*, *A. lyrata* and *C. rubella* were isolated from post-flowering plants using TRIzol reagent (ThermoFisher, Carlsbad, CA, USA). Roots were dug out from soil and then washed thoroughly with water. Leaves were collected from rosettes and stems were collected from the basal second internodes. Plant material was frozen in liquid N<sub>2</sub> immediately after harvesting. Reverse transcription reactions were performed using  $\leq 2$   $\mu$ g of total RNA, random primers and a reverse transcription kit (Agilent, Santa Clara, CA, USA). The cDNAs were used as templates for quantitative (q)PCR analysis which was carried out using a CFX Real-time PCR system (Bio-Rad). PCR reactions (15  $\mu$ l) consisted of: 7.5  $\mu$ l SYBR Green I master mix solution (Roche), 1  $\mu$ l gene specific primers, 5 $\times$  diluted cDNAs and water. The amplification protocol involved denaturation at 95°C for 2 min, followed by 39 cycles of 95°C for 10 s and 62°C for 20 s. Amplicon dissociation curves (i.e. melting curves) were recorded after cycle 39 by heating from 65 to 95°C in 0.5°C increments. The specificity of the amplification products was verified by melting curve analysis. ACTIN or PP2A was used as a reference. Normalized gene expression levels were calculated using the  $2^{-\Delta\Delta C_t}$  method (Livak & Schmittgen, 2001). The qPCR experiments were conducted using three independent biological replicates, each consisting of three technical replicates. Primers are listed in Table S2.

### 3.2.9. Isolation of mutants of *C. rubella*

#### Isolation of TILLING (Targeting Induced Local Lesions IN Genomes) mutants

*Capsella rubella* TILLING mutants were ordered from RevGenUK (<https://jicbio.nbi.ac.uk/revgen.html>) at the John Innes Centre. Pre-screening was carried out with DNA pools from M2 progeny by sequencing of PCR products amplified using specific primer pairs (Table S2). Ten M3 seeds from each positive line were sown. Thirteen independent lines bearing a mutation in CYP708A9 and ten with a mutation in CYP705A38 gene were obtained. No CYP708A10 TILLING mutant lines were identified (Fig. S5a,b). Mutations were confirmed by sequencing. Only line Mcr705-8 showed accumulation of tirucalla-8,24-dien-3 $\beta$ ,23-diol (Ti2) in comparison to wild-type (WT). This line was then crossed with the WT line. The F2 population derived from these crosses was used for metabolite analysis and morphological phenotyping analysis.

#### Constructs for CRISPR-Cas9 vector and mutant isolation

Golden Gate (GG) assembly of a CRISPR-Cas9 vector for genome editing using

the seed FAST-RFP as screen marker (Shimada *et al.*, 2010) was carried out as described recently (Castel *et al.*, 2019). Two gene-specific guide RNA (gRNA)-targeting sequences before a protospacer adjacent motif (PAM, NGG) site were selected (Fig. S5c). Briefly, the gRNA targeting sequences were synthesized within forward primers. Together with a universal reverse primer, a c. 200-bp fragment including CRISPR-Cas9 targeting fragment and gRNA backbone was amplified by PCR. The level 1 GG reaction contained: 20 ng pICSL90002 plasmid, 2.5 ng gRNA fused fragment, 61 ng pICH47751 (for gRNA1) or pICH47761 plasmid (for gRNA2), 1 µl Bsa I HF New England Biolabs (NEB, Ipswich, MA, USA), 1.5 µl 10× Bovine serum albumin (BSA), 1.5 µl T4 ligase buffer (10×), 1 µl T4 ligase and water (final volume 15 µl). The GG assembly protocol was as follows: (37°C for 4 min, 16°C for 3 min) × 25 cycles, then 65°C for 10 min. The gRNA targeting sequence region after GG assembly was verified by sequencing. The level 2 GG reaction contained: 160 ng pICSL4723 plasmid, 83 ng pICSL11015 plasmid, 135 ng BCJJ358 plasmid, 59 ng pICH47751 assembled gRNA1 level 1 product, 59 ng pICH47761 assembled gRNA2 level 1 product, 41 ng pICH41780, 0.5 µl Bpi I, 1.5 µl 10× BSA, 1.5 µl T4 ligase buffer (10×), 0.5 µl T4 ligase and water (final volume 15 µl). The GG assembly protocol was set up as: (37°C for 3 min, 16°C for 4 min) × 25 cycles, 65°C for 10 min. HindIII digestion was used to verify the level 2 GG assembled product. *Agrobacterium tumefaciens* strain LBA4404 carrying the CRISPR-Cas9 construct was used for transformation of *C. rubella* plants. The floral dipping method used for *A. thaliana* (Clough & Bent, 1998) was adopted. The concentration of Silwet-70 was reduced to 0.15%. Thirty flowering plants were used for floral dipping. T0 fluorescent seeds were screened with a Leica M205FA stereo microscope with a 530 nm-red fluorescent protein (RFP) LED light source. Twelve fluorescent seeds were obtained. Two T1 lines that had undergone genome editing were identified. One of these (T1-#20) had a 100-bp deletion within the CYP708A10 gene in a T1 leaf sample and was therefore used for next-generation screening. Twenty-four nonfluorescent (to avoid CRISPR/Cas9 construct on genome) T1-#20 seeds were sown to generate T2 lines. Twelve independent T2 lines showed genome editing on the CYP708A10 gene and were then sown to generate T3 lines. Two independent homozygous lines in the T3 generation were used for metabolic analyses.

### 3.2.10. Transient expression in *N. benthamiana*

#### Construct generation

The cDNAs for the *A. lyrata* AL8G20190, AL8G20160, AL8G20150 and AL8G20140 genes were amplified from root cDNA, and those of the *C. rubella* genes Carubv10016727m, Carubv10017128m, Carubv10017289m, Carubv10017243m and Carubv10017044m were amplified from cDNA from flower buds. The cDNAs were cloned into the GATEWAY entry vector pDONR207 (Invitrogen) and the constructs were verified by sequencing. The corresponding genes also were amplified from genomic DNA and cloned and sequenced. AL8G20140 and Carubv10017243m were misannotated in the Phytozome v.12 database. The revised sequences and single nucleotide variants are listed in Table S3. The *B. rapa* genes could not be cloned from available cDNA libraries (root, leaves, stems, flowers and siliques). Genomic DNA was therefore used as template

to clone the Brara.I04560, Brara.I04561, Brara.I04562 and Brara.I04563 genes. The sequences of these genes (if different from those represented in the Phytozome v.12 database) are listed in Table S3. The cloned genes were then inserted into the pEAQ-Dest-1 expression vector (Sainsbury *et al.*, 2009). Constructs were verified by sequencing and introduced into *A. tumefaciens* strain LBA4404.

### **Agro-infiltration of *N. benthamiana* leaves**

The strains harbouring expression constructs were freshly grown on Lysogeny Broth (LB) plates with antibiotic selection (50 µg ml<sup>-1</sup> kanamycin, 50 µg ml<sup>-1</sup> rifamycin, 100 µg ml<sup>-1</sup> streptomycin). For small-scale analysis, multiple clones were picked and inoculated into 10 ml LB broth. Cultures were then incubated in a shaker (28°C and 220 rpm) for about 16 h until the OD600 reached c. 2.0. For large-scale analysis, the 10 ml culture was further inoculated to make 1 L culture. LBA4404 cells were pelleted by centrifuging at 4500 g for 20 min and supernatants were discarded. The pellets were then resuspended in freshly made MMA buffer (10 mM MgCl<sub>2</sub>, 10 mM MES/KOH pH5.6, 150 µM acetosyringone) and diluted to OD600 0.2. For combinatorial assay, strains harbouring different constructs were mixed and infiltrated by a syringe without a needle for small-scale analysis. For triterpene purification and NMR analysis, around 75 to 100 plants were infiltrated batch-wise by a previously described customized vacuum infiltration system (Reed *et al.*, 2017). Leaves were harvested 6 d post-infiltration and lyophilized.

## **3.2.11. Metabolite extraction and analysis**

### **Metabolite extraction**

Three dry *N. benthamiana* leaf disks (9 cm diameter) were ground and saponified by mixture of ethanol/H<sub>2</sub>O/KOH pellets in 9 : 1 : 1 (v/v/w) (1 ml) at 70°C for 1 h. The ethanol was removed by evaporation (1 h, 70°C), and the samples were extracted with 1 ml ethyl acetate/H<sub>2</sub>O in 1 : 1 (v/v). The suspensions were centrifuged at 16 000 g for 1 min and the supernatants collected. The supernatants were dried under N<sub>2</sub> and resuspended in 50 µl derivatizing reagent, 1-(Trimethylsilyl)imidazole-Pyridine mixture (Sigma-Aldrich). The samples were incubated at 70°C for 30 min before analysis by GC-MS analysis. For *C. rubella* metabolites extraction, c. 30 mg of fresh tissues (leaves or flowers) were ground. Samples were extracted and prepared for GC-MS as described above. Ground samples used for LC-MS analysis were extracted with 1 ml ethyl acetate (shaken overnight). The samples were then extracted with 500 µl of water. The supernatants were dried under N<sub>2</sub> and resuspended in 100 µl methanol. The samples were all cleaned via 0.22-µm nylon filter tube (Spin-x centrifuge tube; Costar, Cole-Pamer, St Neots, UK) prior to LC-MS analysis.

### **GC-MS and LC-MS-IT-TOF analysis**

GC was performed on an Agilent 7890B fitted with a Zebron ZB5-HT Inferno capillary column (Phenomenex, Torrance, CA, USA). A 1-µl aliquot of each sample

was injected by a splitless pulse method (2.07 bar pulse pressure) with a GC inlet temperature of 250°C. The oven temperature program began from 170°C (held for 2 min) to 290°C (held for 4 min) at a speed of 6°C min<sup>-1</sup> and switched to 340°C (held 1 min) at a rate of 20°C min<sup>-1</sup>. Helium was used as carrier gas and the flow rate was set as 1 ml min<sup>-1</sup>.

LC-MS analysis was carried out on a Prominence/Nexera UHPLC system attached to an Ion-trap time-of-flight (ToF) mass spectrometer (Shimadzu, Kyoto, Japan). Separation was performed on a 100 × 2.1 mm 2.6 µm Kinetex EVO reverse phase column (Phenomenemex), using the following gradient of methanol (solvent B) vs 0.1% formic acid in water (solvent A): 0 min, 70% B; 10 min, 95% B; 11 min, 95% B; 11.1 min, 70% B; 14.5 min, 70% B. The flow rate was 0.5 ml min<sup>-1</sup> and the run temperature was set at 40°C. Detection was collected by positive electrospray MS. Spectra were collected from m/z 200–2000 with automatic sensitivity control set to a target of 70% optimal base peak intensity. Spray chamber conditions were: 250°C for curved desorption line, 300°C for heat block, 1.3 l min<sup>-1</sup> for nebulizer gas, and drying gas is 'on'. The instrument was calibrated immediately before analysis, using sodium trifluoroacetate cluster ions according to the manufacturer's instructions.

### 3.2.12. Triterpene purification and NMR analysis

Freeze-dried *N. benthamiana* leaves were thoroughly extracted using a SpeedExtractor (E-914). The extraction method was set as: solvent: ethyl acetate; pressure: 130 bar; three cycles, Hold time Cycle 1: 0 min; Hold time Cycle 2: 5 min; Hold time Cycle 3: 5 min. The extracts were dried by rotatory evaporation. The crude extracts were dissolved in ethanol and saponified by strong basic anion exchange resin amberSEP 900 hydroxide beads (Sigma-Aldrich) for about 30 min (the solvent turned to yellow) (Stephenson *et al.*, 2018). The solvent was filtered by a mixture layers of Celite 545 (Sigma-Aldrich) and fat-free quartz sand (Buchi 037689). The eluted solvent was dried, loaded to an IsoleraOne system silica gel column (BioTage® SNAP Vitra K-Sil 100g, Uppsala, Sweden) and separated with a gradient of 0–100% ethyl acetate. Fractions (164 ml) were collected in 200 ml Duran bottles and aliquots (5 µl) monitored on thin layer chromatography (TLC) plate (70644-50EA, Sigma-Aldrich). Triterpenes were visualized by spraying of fresh-made vanillin-sulfuric acid reagent (1 g vanillin dissolved in 100 ml of 50% sulfuric acid.) Depending on the purity, additional separation of selected fractions was carried out using a smaller volume column (BioTage® SNAP Vitra KP-Sil 10g) with 5% ethyl acetate in dichloromethane (DCM) as eluent solvent. Fractions from each step were assessed both by TLC and GC-MS analysis. Around 2–5 mg purified compounds were collected and subjected to NMR analysis.

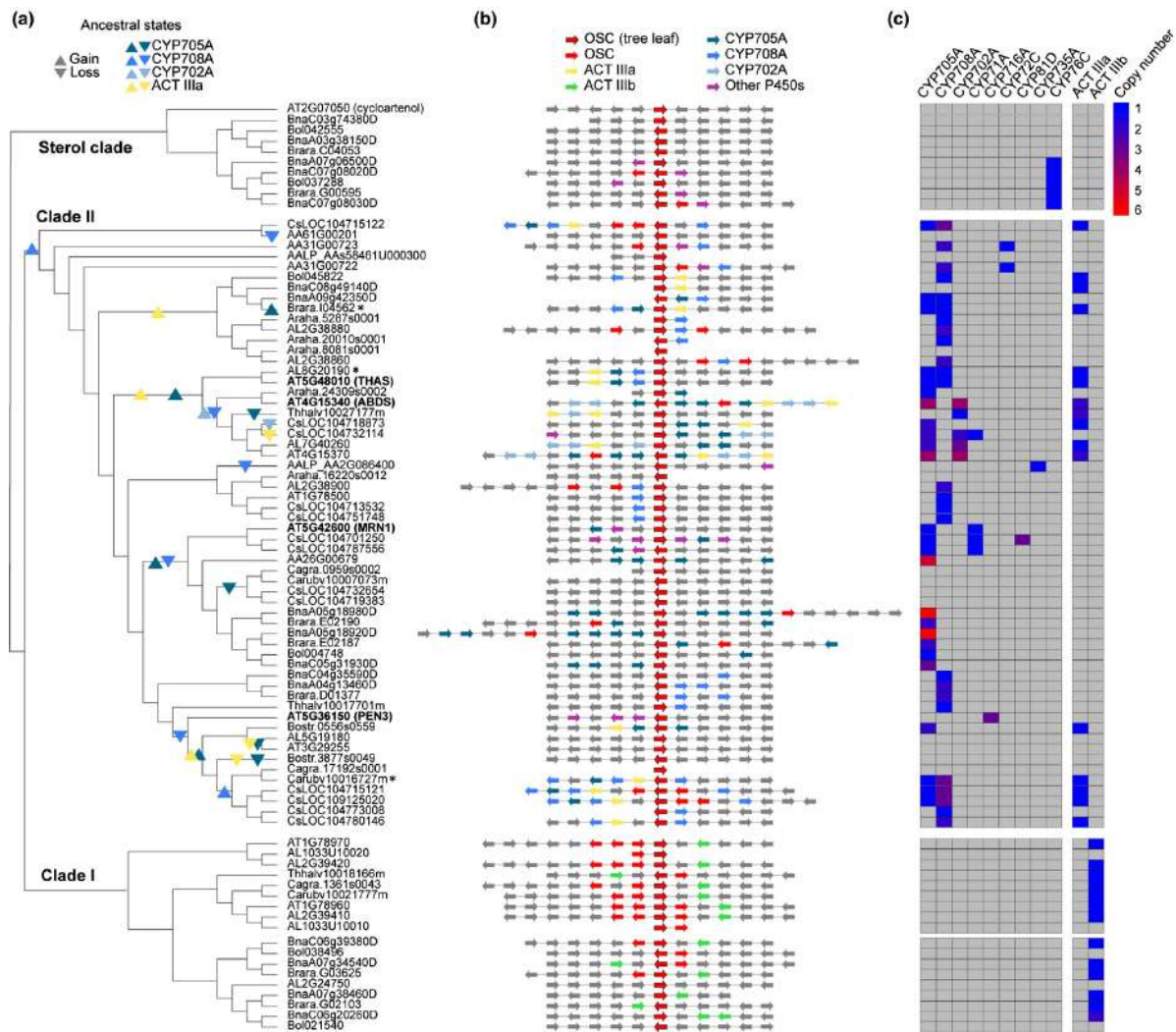
Purified compounds were analysed by NMR spectroscopy. DEPT-135, DEPT-edited-HSQC, COSY, and HMBC experiments were used to fully assign <sup>1</sup>H and <sup>13</sup>C spectra, or spectra were compared to the literature if reported previously. NMR spectra were recorded in Fourier transform mode at a nominal frequency of 400 MHz for <sup>1</sup>H NMR, and 100 MHz for <sup>13</sup>C NMR, using the specified deuterated solvent. Chemical shifts were recorded in ppm and referenced to the residual solvent peak or

to an internal tetramethylsilane (TMS) standard. Multiplicities are described as: s, singlet; d, doublet; dd, doublet of doublets; dt, doublet of triplets; t, triplet; q, quartet; m, multiplet; br, broad; appt, apparent; coupling constants are reported in Hz as observed and not corrected for second order effects (Table S4).

## 3.3. Results

### 3.3.1. Investigation of the genome neighborhoods around triterpene synthase genes

The first committed step in triterpene biosynthesis is catalyzed by triterpene synthases – also known as OSCs. Our systematic analysis of 13 sequenced Brassicaceae genomes identified a total of 163 predicted OSC genes (Table S1). These OSCs grouped into three major clades: the sterol clade, containing OSCs implicated in primary sterol biosynthesis, and two other clades (I and II), a topology consistent with our earlier investigations of OSCs in *A. thaliana* (Field & Osbourn, 2008; Field *et al.*, 2011). The OSCs located in the previously reported *A. thaliana* BGCs all belong to clade II (Figs 1a, S1a).



**Fig. 1.** Associations between clade II oxidosqualene cyclases (OSCs) and cytochrome P450 (CYP) and acyltransferase (ACT) subfamilies in the Brassicaceae, in rapidly evolving genomic regions. (a) Maximum-likelihood tree of clade II OSC protein sequences, including representative sterol and clade I OSCs (see Supporting Information Fig. S1 for the full 163 Brassicaceae OSCs). Characterized OSCs for *Arabidopsis thaliana* biosynthetic gene clusters (BGCs) are indicated in bold; \*, OSCs characterized in this study. The ancestral states of CYPs and ACTs in the clade II OSC gene neighbourhoods (GNs) were reconstructed with maximum parsimony and inferred changes in state (gene gains and losses) are shown. (b) OSC GNs. The genes encoding the OSC in each tree leaf in a are positioned in the middle. Arrows denote the strand directionality of genes. CYP and ACT subfamilies are denoted by colours (see key). (c) Heat map showing the CYP and ACT domains in the OSC GNs. The colour scale bar shows copy number values.

We then examined the immediate GNs around all 163 OSC genes, extending five genes on either side (Fig. 1b; Table S5; all GNs are publicly available in GenBank format at doi: 10.5281/zenodo.3531676) and tested which Pfam domains were over-represented in the OSC GNs relative to genome-wide distribution. To reduce potential phylogenetic bias due to genes likely derived from common ancestors, domains that appeared consistently in the GNs of monophyletic branches of OSCs were binned and only the leaf with the largest number of domain appearances was counted ('conservative' hypergeometric test,  $P < 0.01$ ; see Methods). The results indicate that the sterol, clade I and clade II OSC genes have distinct GN associations (Tables 1, S6). In general, the Pfam domains associated



with sterol OSC genes do not have any anticipated roles in specialized metabolism. Although significant associations between clade I OSC genes and ACT Pfam domains were detected (Tables 1, S6), there currently is no evidence that clade I OSCs and ACTs form functional BGCs. The clade II OSC genes, however, were significantly associated with both CYP and ACT genes, both of which encode potential triterpene scaffold-modifying enzymes (Tables 1, S6). Such associations were further supported by phylogenetic regression analyses (Table S6; Notes S1) and 'nonconservative' hypergeometric analysis (Table S7; Notes S1). Altogether, our analyses suggest that the evolutionary histories of clade I, clade II and sterol OSC GNs are distinct (Fig. 1a,b; Table 1), and importantly, that clade II OSC genes associated with CYP and ACT genes may reside in potential BGCs (Medema & Osbourn, 2016).

OSC	Associated Pfam domain	<i>P</i> < 0.01 (Fisher, bfr)
Sterol	NodS	3.68E-08
Sterol	Prefoldin_2	0.00000323
Sterol	MTS*	0.000208274
Sterol	rRNA_proc.arch	0.001700795
Sterol	DSHCT	0.004179688
Clade I	ACT_IIIb*	2.05E-17
Clade I	Fer4_7	5.86E-16
Clade I	Fer4_9	2.53E-15
Clade I	Fer4_4	6E-14
Clade I	Fer4	2.14E-12
Clade I	Fer4_10	3.56E-12
Clade I	MOSC	0.00000399
Clade I	MOSC_N	0.00000435
Clade I	Per1	0.0000147
Clade I	EPL1	0.00049452
Clade I	Methyltransf_32*	0.001632776
Clade I	Transferase*	0.001870304
Clade I	Noc2	0.003733658
Clade I	Aminotran_5*	0.006422319
Clade I	WAK_assoc	0.006540572
Clade II	CYP705A*	1.54E-30
Clade II	p450*	9.35E-24
Clade II	CYP708A*	1.88E-19
Clade II	ACT_IIIa*	1.71E-10
Clade II	CYP702A*	2.43E-10
Clade II	CYP716A*	0.000143077
Clade II	polyprenyl_synt*	0.002802102
Clade II	Transferase*	0.005441417

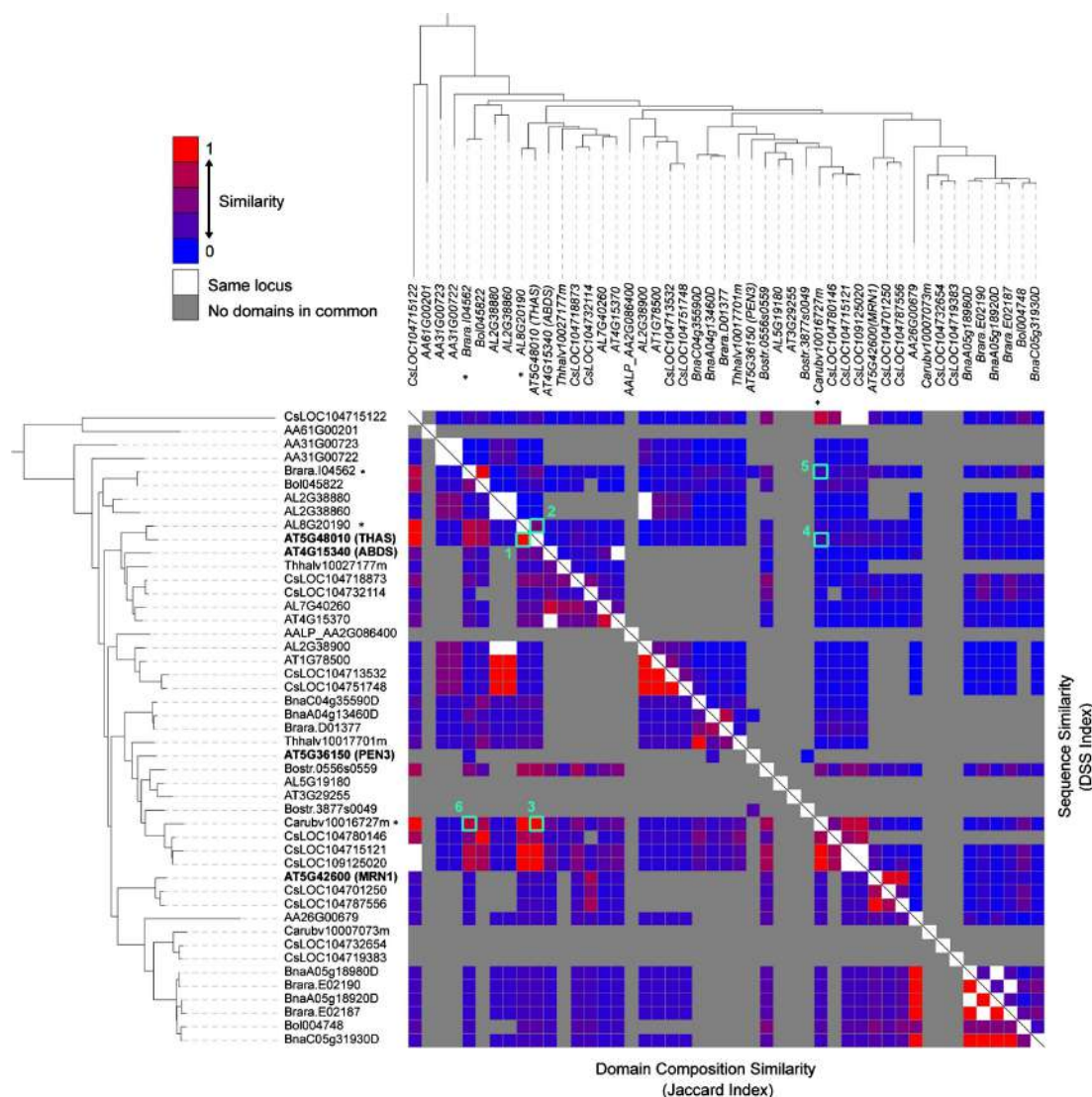
**Table 1.** Domains significantly associated with the Brassicaceae oxidosqualene cyclases (OSCs) from the three phylogenetic clades. Association outcome ( $P < 0.01$ ) of conservative Fisher's exact test (with Bonferroni correction) is listed. The full list can be found in Supporting Information Table S6. Enzymatic domains are indicated by \*

In order to further elucidate the relationship between clade II OSCs, CYPs and ACTs (in total, 58, 76 and 15 in the clade II OSC GNs, respectively), we identified all CYP and ACT genes in the 13 Brassicaceae and four outgroup genomes and annotated all of the CYP and ACT subfamilies by propagating the annotations of characterized genes from *A. thaliana*. Our analysis revealed that all ACT genes present in clade II OSC GNs belong to the ACT IIIa subfamily, whereas those in clade I OSC GNs belong to a distinct ACT subfamily, ACT IIIb (Fig. S3). The CYP

genes associated with clade II OSC genes belong to a total of eight families, of which 86% are in the CYP708, CYP702 and CYP705 families (Figs 1c, S4).

### 3.3.2. All-vs-all comparison of clade II OSC genome neighborhoods

Because of the abundance of genes encoding potential triterpene scaffold-modifying enzymes in clade II OSC GNs, 72% (36 of 50) of which fulfil our definition of BGCs (Medema & Osbourn, 2016), we focused on this clade to explore their evolutionary relationships. After pruning the clade to ensure that all GNs had at least five genes flanking each OSC, we performed an all-vs-all comparison of these GNs using three distinct methods: to measure architectural similarity (enzyme family content), we used the Jaccard index of Pfam domains, and as proxies for specific enzymatic function similarity or divergence, we measured the average amino acid identity of shared protein domains and the DSS index, which further considers the differing number of appearances of each domain in each GN (see Methods) (Fig. 2; Table S8). To avoid bias due to flanking genes not involved in specialized metabolism, we took into account only those domains known to be involved in specialized metabolic pathways in plants (Kautsar et al., 2017). Intriguingly, on the one hand, the domain content of clade II OSC GNs is very dynamic (439 of 741 GN pairs having no shared domains), suggesting rapid gene turnover in the OSC flanking regions. On the other, around 24% (73 of 302) of the GN pairs that have domains in common have highly similar domain compositions (Jaccard index  $\geq 0.5$ ), consistent with nonrandom associations of CYPs and ACTs with the OSCs from this clade. Surprisingly, in 58% of the GN pairs, the average amino acid identity between shared domains (177 of 302) falls below 50%, and 84% of GN pairs (257 of 302) have a DSS index  $\leq 0.3$ . This suggests that OSC-centric neighbourhoods with similar enzyme family compositions are either undergoing rapid evolution or have evolved multiple times independently. In line with the latter, ancestral state reconstruction, based on maximum parsimony and supported by detailed analysis in the phylogeny of the tailoring enzymes (Notes S1), indicates multiple parallel gene gain and loss events of the major CYPs and ACTs in the OSC-centric neighbourhoods (Figs 1a, S6). Interestingly, three of the six clade II OSC GNs (including the previously characterized thalianol and marnerial BGCs) in *A. thaliana* are located in dynamic chromosomal regions that do not show synteny with regions originating from the ancient WGD in Brassicaceae (Freeling et al., 2007). This again indicates that the loci around clade II OSCs are highly dynamic. Thus, the evolutionary dynamics of the OSC genomic neighbourhoods indicates a general pattern of independent and parallel evolution of the enzyme family compositions of these loci.

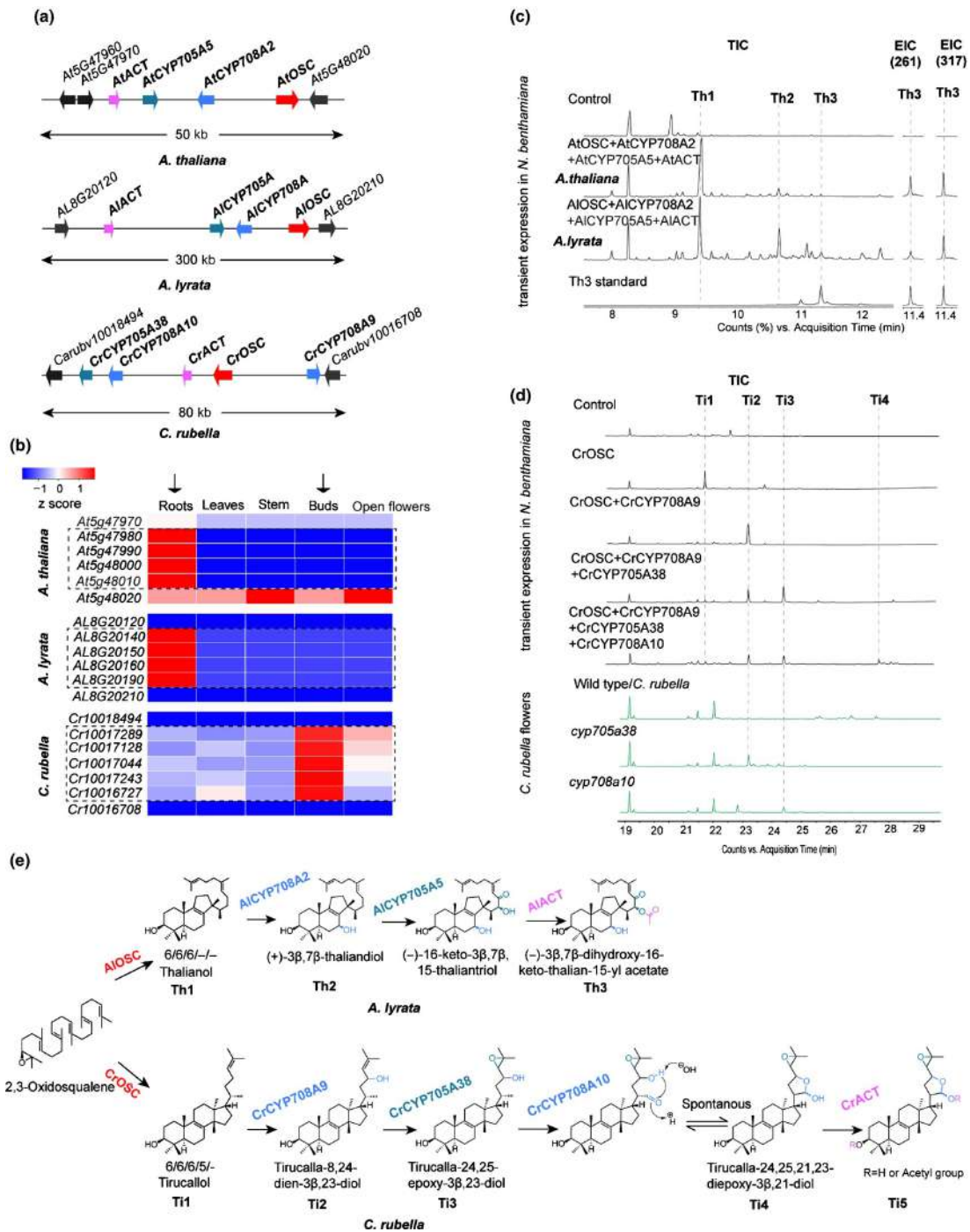


**Fig. 2.** Independently evolved triterpene biosynthetic gene clusters (BGCs) frequently converge towards similar enzyme family content, yet with low mutual sequence identity. Enzymatic domains from each genome neighbourhood (GN) are compared in an all-vs-all fashion. The Jaccard Index was used to measure the architectural similarity (enzyme family content) across the GNs, and the domain sequence similarity (DSS) index to quantify the similarity of the underlying protein sequences. The colour scale bar shows similarity score values. The clade II oxidosqualene cyclase (OSC) tree was used to shape the structure of the heat map. Characterized OSCs for *Arabidopsis thaliana* BGCs are indicated in bold. \*, OSCs characterized in this study. The diagonal line separates the Jaccard Index (left) and the DSS Index (right) comparisons. The numbers indicate: Jaccard (1) and DSS (2) index comparisons of the *A. thaliana* thalianol BGC GN with a putative BGC in *Arabidopsis lyrata*; Jaccard (3) and DSS (4) index comparisons of the *A. thaliana* thalianol BGC GN with a putative BGC in *Capsella rubella*; Jaccard (5) and DSS (6) index comparisons of the *C. rubella* BGC GN with a putative BGC in *Brassica rapa*.

### 3.3.3. Functional analysis of selected representative BGCs

In some cases, ancestral state reconstruction indicates that similar enzyme composition can be traced back to a recent common ancestry. For example, the CYP705A and ACT IIIa domains were present in some reconstructed ancestral OSC GNs (Fig. S6) at the base of certain monophyletic branches, including the one containing the previously characterized thalianol BGC from *A. thaliana* (Fig. 3a). The thalianol BGC consists of genes encoding an OSC, two CYPs (a CYP708A and a

CYP705A), and one ACT (Field & Osbourn, 2008; Huang et al., 2019). Our previous work and the current analysis indicate that a closely related (as yet uncharacterized) BGC is present in the sister species (Field et al., 2011) (Fig. 3a). The genes within these two BGCs (labelled 1 and 2, respectively in Fig. 2) share high nucleotide sequence identity and occur in the same genomic order. These clusters both have root-specific expression profiles (Fig. 3b) and are located in syntenic genomic blocks (Field et al., 2011) (Table S9). They are therefore likely to share a common evolutionary origin. To evaluate the function of the predicted *A. lyrata* BGC, we cloned the genes and expressed them in *N. benthamiana* using transient agro-infiltration (Sainsbury et al., 2009; Reed et al., 2017). These experiments confirmed that the *A. lyrata* OSC, CYP708A, CYP705A and ACT enzymes are functionally equivalent to their counterparts in *A. thaliana*, and when co-expressed together produce (-)-3 $\beta$ ,7 $\beta$ -dihydroxy-16-keto-thalian-15-yl acetate (Th3) (Huang et al., 2019) (Fig. 3c). The presence of thalianol pathway BGCs in *A. thaliana* and *A. lyrata* thus represents an example of conserved BGCs in closely related species.



**Fig 3.** Conservation and diversification of similar triterpene biosynthetic gene clusters (BGCs). (a) Schematic of three triterpene BGCs from *Arabidopsis thaliana*, *Arabidopsis lyrata* and *Capsella rubella*. (b) Expression profiles of the BGC and BGC-flanking genes for the three BGCs shown in (a). The numerical values for the blue-to-red gradient bar represent normalized expression levels relative to roots from reverse transcription quantitative (RT-q)PCR analysis ( $2^{-\Delta\Delta Ct}$ ). (c) Transient expression of *A. thaliana* thalianol BGC genes and the putative *A. lyrata* BGC genes in *Nicotiana benthamiana*. GC-MS total ion chromatograms (TICs) and extracted ion chromatograms (EICs) for two characteristic ions (261 and 317) of Th3 are shown. (d) TIC of extracts from: *N. benthamiana* leaves transiently co-expressing different combinations of the *C. rubella* BGC genes (upper panel); flower extracts from wild-type and mutant *C. rubella* lines (lower panel). The data are representative of at least two separate experiments. (e) The thalianol pathway in *A. lyrata* and the *C. rubella* tirucallol pathway. Enzymes are colour-coded according to the key in (a).

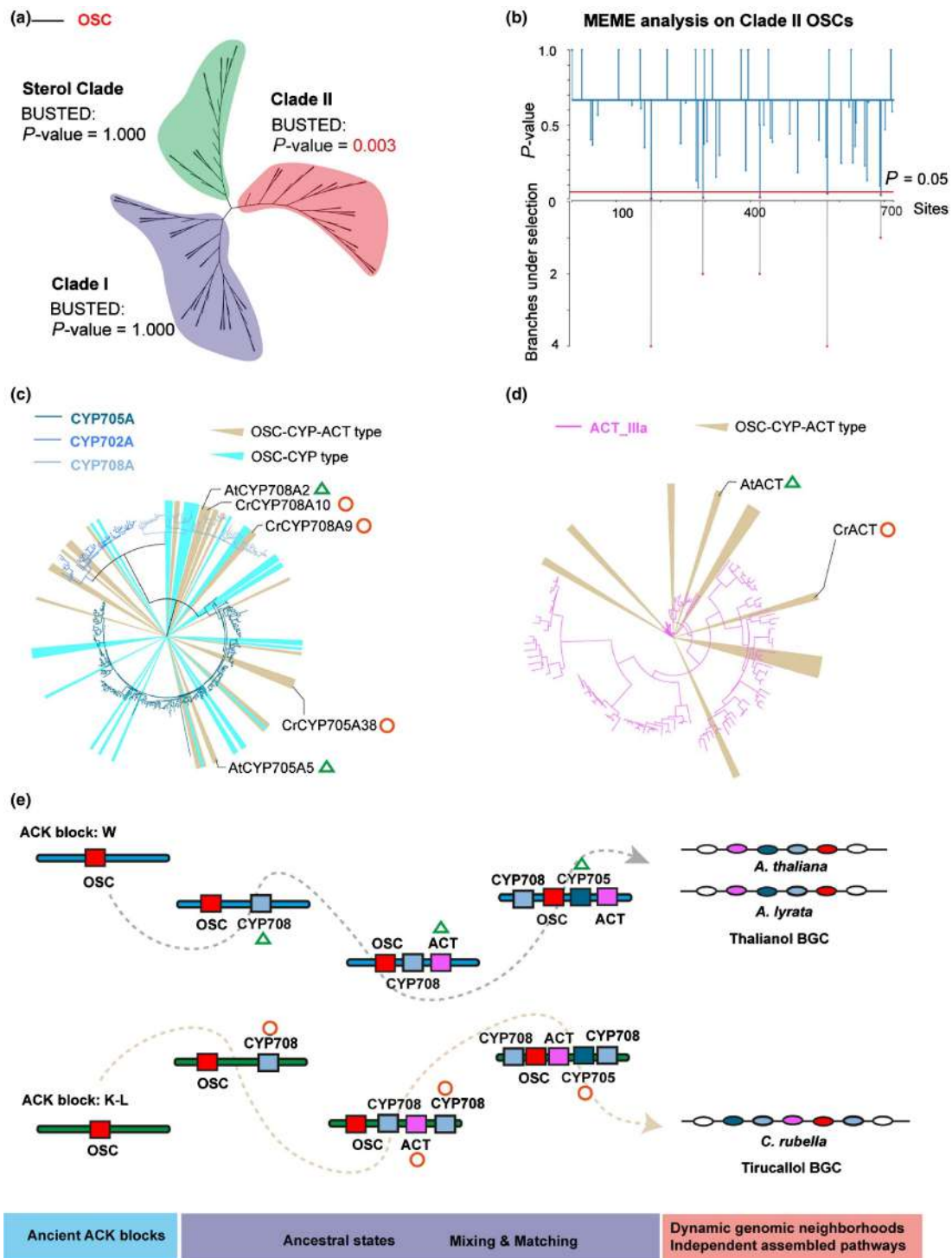
By contrast, many other pairs of putative BGCs identified in our large-scale analysis of Brassicaceae genomes appeared to have similar domain composition, but very limited sequence similarity (Fig. 2). Moreover, the ancestral state reconstruction suggested that CYPs and ACTs had been recruited to these GNs independently. An example of this is a predicted BGC in pink shepherd's purse (*C. rubella*). Just like the *A. thaliana* and *A. lyrata* thalianol clusters, this BGC contains domains for OSC, CYP708, CYP705 and ACT enzymes. However, the gene order is different, the sequence identity with the *A. thaliana* thalianol cluster is low (labelled 3 and 4, respectively in Fig. 2; average domain amino acid identity = 45% and DSS = 0.3) and there is an additional CYP708A gene (Fig. 3a). Unlike the thalianol BGC, the *C. rubella* gene cluster is expressed preferentially in the buds (Fig. 3b). The OSC from this cluster yielded a product (Ti1) when expressed in *N. benthamiana*, which we subsequently showed by NMR to be the triterpene tirucallol (Fig. 3d,e; Table S4A). Through combinatorial expression we then showed that the enzymes encoded by the neighbouring CYP and ACT genes were able to successively modify tirucallol. Specifically, co-expression of the OSC with CYP708A9 yielded (Ti2); further inclusion of CYP705A38 gave conversion of (Ti2) to (Ti3); compound (Ti3) was further modified by CYP708A10 to give (Ti4) (Fig. 3d). The ACT was able to further modify (Ti4) to give (Ti5) with very low conversion (Fig. S7). The structures of compounds Ti1, Ti2, Ti3 and Ti4 were determined by NMR (Table S4B–D). That of Ti5 was inferred by LC-MS (Fig. S7). The pathway is shown in Fig. 3e. Metabolite analysis of *C. rubella* TILLING/CRISPR-Cas9 mutants for CYP705A38 and CYP708A10 (Figs 3d, S5) provided further support for the pathway shown in Fig. 3e. Phylogenetic analysis indicates that CYPs and ACTs associated with the *A. thaliana* thalianol and *C. rubella* tirucallol BGCs are scattered across their respective enzyme family trees, rather than forming a subclade (Figs S3, S4). Furthermore, the *A. thaliana* thalianol BGC and *C. rubella* tirucallol BGC can be traced back to different ancestral crucifer karyotype blocks (Lysak et al., 2016) (W and K-L, respectively; Table S9). Collectively, our results indicate that although the *C. rubella* BGC has superficial similarities with the *A. thaliana* and *A. lyrata* thalianol pathway BGCs in terms of domain composition, it is functionally distinct and has evolved independently.

We also investigated the function of a candidate BGC from *Brassica rapa*, which belongs to Brassicaceae lineage II. This group separated from common ancestors of Brassicaceae lineage I species (which include *A. thaliana* and *C. rubella*) around 20 Myr ago (Hohmann et al., 2015). This candidate BGC contains genes predicted to encode an OSC, two CYPs (a CYP708A and a CYP705A), and an ACT. Expression of these genes in *N. benthamiana* revealed that the OSC produces the triterpene euphol, and that the associated CYP705A and ACT are able to metabolize this triterpene (Fig. S8). Activity was not, however, detected for the CYP708A. Interestingly, euphol is an epimer of tirucallol. Given the complex genomic history of the Brassicaceae, common ancestry cannot be fully excluded for these loci. Yet, the fact that the tirucallol and euphol OSCs are located in different ACK blocks (L and I, respectively; Table S9), the paralogy of these two OSCs in the OSC phylogeny, and the observation that the ACTs within these loci are not monophyletic, indicate that parallel events are likely to have taken place in the evolutionary history of these two loci (labelled 5 and 6 in Fig. 2), in this case with likely a parallel metabolic outcome.

We next applied Branch-Site Unrestricted Statistical Test for Episodic

Diversification (BUSTED) analysis across the three OSC clades, and we found evidence for gene-wide positive selection on clade II (Fig. 4a). In line with this, Mixed Effects Model of Evolution (MEME) analysis identified five sites in the clade II OSC alignment under positive selection (Fig. 4b). The episodic selection, however, is restricted to a limited number of branches, indicating clade II OSCs are evolving independently. Consistently, tailoring genes (CYP and ACT) contributing to clade II OSC/CYP or clade II OSC/CYP/ACT associations are scattered on their respective phylogenetic trees (Fig. 4c,d). Together with the ancestral state reconstruction, we propose that dynamic ancestral OSC GNs independently shuffle a core palette of decorating domains, forming divergent BGCs throughout the radiation of Brassicaceae species (Fig. 4e).





**Fig 4.** Independent evolution of clade II oxidosqualene cyclase (OSC) related biosynthetic gene clusters. (a) BUSTED (Branch-Site Unrestricted Statistical Test for Episodic Diversification) analysis on the three clades of Brassicaceae OSCs. Evidence for gene-wide positive selection was found for clade II OSC. (b) Mixed Effects Model of Evolution (MEME) analysis on clade II OSCs. Five sites from limited branches were found to be under positive selections ( $P < 0.05$ ). (c, d) Clade II OSC/CYP and clade II OSC/CYP/ACT associations are marked on cytochrome P450 (CYP) (c) and acyltransferase (ACT) (d) phylogenetic trees. The genes of *Arabidopsis thaliana* thalianol BGC and *Capsella rubella* tirucallol BGC are indicated. (e) Proposed model for the evolution of the *A. thaliana* and *A. lyrata* thalianol BGCs and the *C. rubella* tirucallol BGC. The plausible ancestral states of cluster assembly were drawn based on ancestral states reconstructions. The recruited domains can be traced on the phylogeny in (c) and (d).

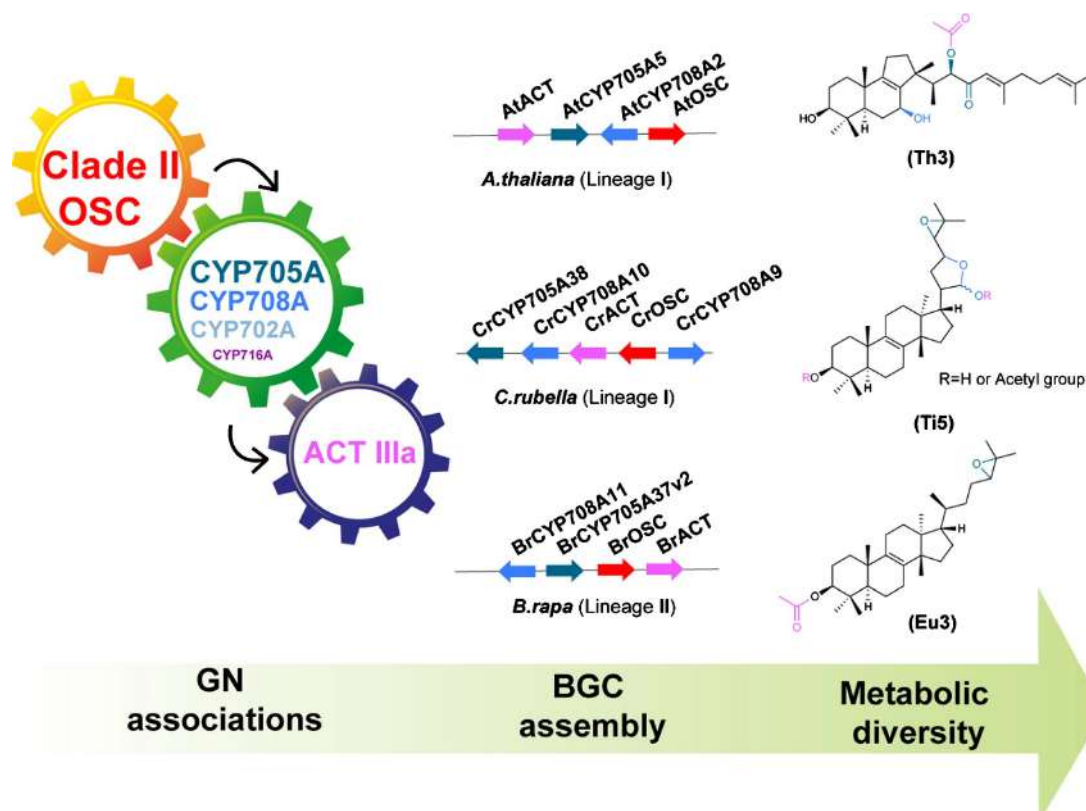


### 3.4. Discussion

How individual enzymes have evolved to achieve metabolic diversity is relatively well-understood (Benderoth *et al.*, 2006; Weng *et al.*, 2012a; Hofberger *et al.*, 2013; Hamberger & Bak, 2013; Moghe & Last, 2015; Barco & Clay, 2019), whereas the mechanisms of evolution of multi-step pathways are more elusive. In comparison to nonclustered pathways, biosynthetic gene clusters (BGCs) provide unique material with which to systematically study the evolutionary processes underpinning the birth of plant metabolic pathways. Here, our systematic genomic neighbourhood (GN) analysis of oxidosqualene cyclases (OSCs) across multiple Brassicaceae genomes has revealed that clade II OSC genes are predisposed to clustering with genes for potential triterpene scaffold-modifying enzymes. The number of genomes that we used in this study is small in relation to the >3000 species in the Brassicaceae. Ancestral state reconstruction with a larger dataset – once other genome sequences become available – may therefore produce a different outcome. However, our dataset is representative of the two major Brassicaceae lineages (I and II) and also includes the early diverging species *Aethionema arabicum*. The overall picture delineating parallel recruitment events in these dynamic genomic neighborhoods is highly supported and is corroborated by the evidence from cytochrome P450 (CYP) and acyltransferase (ACT) phylogenies as well as ancestral karyotype reconstruction.

Compared to previous analyses of plant BGCs (Kautsar *et al.*, 2017; Töpfer *et al.*, 2017; Schläpfer *et al.*, 2017), our phylogenomic analysis of OSCs across multiple Brassicaceae genomes provides, for the first time, a comprehensive picture of their genomic evolution across all relevant loci, whether they constitute gene clusters or not. Our phylogenetic analyses of the key gene families, along with their contextualization in whole-genome duplication (WGD)-derived subgenomes (Fig. S9), provide clear evidence for recurrent independent assembly of BGCs containing OSC, CYP and ACT genes across the Brassicaceae, leading to divergent or parallel metabolic outcomes (Fig. 4e).

Because of the importance of triterpenes in mediating interactions with herbivores and microbiota (Hostettmann & Marston, 1995; Papadopoulou *et al.*, 1999; Nielsen *et al.*, 2010; Sohrabi *et al.*, 2015; Zhou *et al.*, 2016; Huang *et al.*, 2019), we speculate that, similar to the case of resistance gene clusters in plant genomes (Michelmore & Meyers, 1998), triterpene GNs undergo rapid and dynamic evolution, forming ‘evolutionary playgrounds’ that enable rapid adaption to ever-changing environmental stresses (Field *et al.*, 2011). Triterpenoid scaffold diversification is achieved using a specific palette of CYPs (primarily CYP705A, CYP708A, CYP702A) and ACTs (ACT IIIa) (Fig. 5). The CYP716A family, which has been suggested as a major contributor to the diversification of triterpenoids in eudicot plants (Miettinen *et al.*, 2017), appeared only rarely among the CYPs in Brassicaceae in clade II OSC GNs. The molecular mechanisms by which BGCs form is not yet known, although miniature inverted-repeat transposable elements (MITEs) have been implicated in cluster assembly and/or regulation (Boutanaev & Osbourn, 2018).



**Fig. 5.** Drivers of triterpene diversification in the Brassicaceae. The biosynthesis of diverse triterpenes in the Brassicaceae is achieved by ‘mixing, matching and diverging’ a core palette of clade II oxidosqualene cyclases (OSCs), cytochrome P450s (CYPs) belonging primarily to the CYP705A, CYP708A, and CYP702A subfamilies, and acyltransferases (ACTs) belonging to the ACT IIIa subfamily. Examples of characterized biosynthetic gene clusters (BGCs) and their products are shown. Arrows denote the strand directionality of genes.

Our investigations reveal a novel genomic basis for metabolic diversification in plants through mixing, matching and diverging natural combinations of enzyme families. They further open up opportunities to mimic and expand on plant metabolic diversity by using synthetic biology approaches to engineer diverse bioactive molecules through combinatorial biosynthesis (Fig. 5), for which efficient heterologous expression platforms are now available in yeast (Scheler *et al.*, 2016; Arendt *et al.*, 2017; Bathe *et al.*, 2019) as well as tobacco (Reed *et al.*, 2017). Thus, our increased understanding of pathway formation paves the path to further explore and exploit the biological activities of triterpenes and other plant natural products toward applications in medicine and agriculture.

### 3.5. Supplementary Information

Supplementary figures and tables are available to download from: <https://doi.org/10.1111/nph.16338>

## References

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids*

*Research* **38**: 7– 13.

- Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin M, Darusman LK *et al.* 2012. KNApSACK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant and Cell Physiology* **53**: 1– 12.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403– 410.
- Arendt P, Miettinen K, Pollier J, De Rycke R, Callewaert N, Goossens A. 2017. An endoplasmic reticulum-engineered yeast platform for overproduction of triterpenoids. *Metabolic Engineering* **40**: 165– 175.
- Barco B, Clay NK. 2019. Evolution of glucosinolate diversity via whole-genome duplications, gene rearrangements, and substrate promiscuity. *Annual Review of Plant Biology* **70**: 585– 604.
- Bathe U, Frolov A, Porzel A, Tissier A. 2019. CYP76 oxidation network of abietane diterpenes in Lamiaceae reconstituted in yeast. *Journal of Agricultural and Food Chemistry* **49**: 13437– 13450.
- Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* **107**: 18724– 18728.
- Benderoth M, Textor S, Windsor AJ, Mitchell-Olds T, Gershenzon J, Kroymann J. 2006. Positive selection driving diversification in plant secondary metabolism. *Proceedings of the National Academy of Sciences, USA* **103**: 9118– 9123.
- Böttger A, Vothknecht U, Bolle C, Wolf A. 2018. Plant secondary metabolites and their general function in plants. In: Böttger A, Vothknecht U, Bolle C, Wolf A, eds. *Lessons on caffeine, cannabis & co*, Cham, Switzerland: Springer International, 3– 17.
- Boutanaev AM, Osbourn AE. 2018. Multigenome analysis implicates miniature inverted-repeat transposable elements (MITEs) in metabolic diversification in eudicots. *Proceedings of the National Academy of Sciences, USA* **28**: E6650– E6658.
- Castel B, Tomlinson L, Locci F, Yang Y, Jones JDG. 2019. Optimization of T-DNA architecture for Cas9-mediated mutagenesis in *Arabidopsis*. *PLoS ONE* **14**: e0204778.
- Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: U Bastolla, M Porto, HE Roman, M Vendruscolo, eds. *Structural approaches to sequence evolution: molecules, networks, populations*. Berlin/Heidelberg, Germany: Springer, 207– 232.
- Chassagne F, Cabanac G, Hubert G, David B, Marti G. 2019. The landscape of natural product diversity and their pharmacological relevance from a focus on the Dictionary of Natural Products. *Phytochemistry Reviews* **18**: 601– 622.
- Christianson DW. 2017. Structural and chemical biology of terpenoid cyclases.

*Chemical Reviews* **117**: 11570– 11648.

- Clough SJ, Bent AF. 1998. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *The Plant Journal* **16**: 735–743.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792– 1797.
- Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M *et al.* 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences, USA* **112**: 8362– 8366.
- Field B, Fiston-Lavier A-S, Kemen A, Geisler K, Quesneville H, Osbourn AE. 2011. Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proceedings of the National Academy of Sciences, USA* **108**: 16116–16121.
- Field B, Osbourn AE. 2008. Metabolic diversification—-independent assembly of operon-like gene clusters in different plants. *Science* **320**: 543– 547.
- Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**: W29– W37.
- Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A *et al.* 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44**: D279– D285.
- Freeling M, Rapaka L, Lyons E, Pedersen B, Thomas BC. 2007. G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in *Arabidopsis*. *Plant Cell* **19**: 1441– 1457.
- Garamszegi LZ, ed. 2014. *Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice*. Heidelberg, Germany: Springer.
- Guo L, Winzer T, Yang X, Li Y, Ning Z, He Z, Teodor R, Lu Y, Bowser TA, Graham IA *et al.* 2018. The opium poppy genome and morphinan production. *Science* **362**: 343– 347.
- Hamberger B, Bak S. 2013. Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philosophical transactions of the Royal Society of London. Series B: Biological Sciences* **368**: 20120426.
- Ho L, Ané C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology* **63**: 397– 408.
- Hofberger JA, Lyons E, Edger PP, Chris Pires J, Eric Schranz M. 2013. Whole genome and tandem duplicate retention facilitated glucosinolate pathway diversification in the mustard family. *Genome Biology and Evolution* **5**: 2155–2173.
- Hohmann N, Wolf EM, Lysak MA, Koch MA. 2015. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *The Plant Cell* **27**, 2770– 2784.

- Hostettmann K, Marston A. 1995. *Saponins*. Cambridge, UK: Cambridge University Press.
- Huang AC, Jiang T, Liu Y-X, Bai Y-C, Reed J, Qu B, Goossens A, Nützmann H-W, Bai Y, Osbourn A. 2019. A specialized metabolic network selectively modulates Arabidopsis root microbiota. *Science* **364**: eaau6389.
- Ives AR, Garland T. 2010. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology* **59**: 9– 26.
- Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH. 2017. PlantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research* **45**: W55– W63.
- Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**: 4453– 4455.
- Kuhn HW. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **2**: 83– 97.
- Liu Z, Tavares R, Forsythe ES, André F, Lugan R, Jonasson G, Boutet-Mercey S, Tohge T, Beilstein MA, Werck-Reichhart D *et al.* 2016. Evolutionary interplay between sister cytochrome P450 genes shapes plasticity in plant metabolism. *Nature Communications* **7**: 13026.
- Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2- $\Delta\Delta$ CT method. *Methods* **25**: 402– 408.
- Lysak MA, Mandáková T, Schranz ME. 2016. Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Current Opinion in Plant Biology* **30**: 108– 115.
- Martins EP, Diniz-Filho JAF, Housworth EA. 2002. Adaptive constraints and the phylogenetic comparative method: a computer simulation test. *Evolution* **56**: 1– 13.
- Martins EP, Garland T. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution* **45**: 534.
- Maddison WP, Maddison DR. 2008. Mesquite: A modular system for evolutionary analysis. *Evolution* **62**: 1103– 1118.
- Medema MH, Osbourn A. 2016. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Natural Product Reports* **33**: 951– 962.
- Michelmore RW, Meyers BC. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Research* **8**: 1113– 1130.
- Miettinen K, Pollier J, Buyst D, Arendt P, Csuk R, Sommerwerk S, Moses T, Mertens J, Sonawane PD, Pauwels L *et al.* 2017. The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis. *Nature Communications* **8**: 14153.
- Moghe GD, Last RL. 2015. Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiology* **169**:

1512– 1523.

- Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM *et al.* 2015. Gene-wide identification of episodic selection. *Molecular Biology and Evolution* **32**: 1365– 1371.
- Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics* **8**: e1002764.
- Navarro-Muñoz J, Selem-Mojica N, Mullooney M, Kautsar S, Tryon J, Parkinson E, De Los Santos E, Yeong M, Cruz-Morales P, Abubucker S *et al.* 2018. A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. *bioRxiv*: 445270.
- Nelson DR. 2009. The cytochrome p450 homepage. *Human Genomics* **4**: 59– 65.
- Nielsen JK, Nagao T, Okabe H, Shinoda T. 2010. Resistance in the plant, *Barbarea vulgaris*, and counter-adaptations in flea beetles mediated by saponins. *Journal of Chemical Ecology* **36**: 277– 285.
- Nutzmann HW, Huang A, Osbourn A. 2016. Plant metabolic clusters – from genetics to genomics. *New Phytologist* **211**: 771– 789.
- Papadopoulou K, Melton RE, Leggett M, Daniels MJ, Osbourn AE. 1999. Compromised disease resistance in saponin-deficient plants. *Proceedings of the National Academy of Sciences, USA* **96**: 12923– 12928.
- Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289– 290.
- Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**: 676– 679.
- Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**: e9490.
- Qi X, Bakht S, Leggett M, Maxwell C, Melton R, Osbourn A. 2004. A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proceedings of the National Academy of Sciences, USA* **101**: 8233– 8238.
- Reed J, Stephenson MJ, Miettinen K, Brouwer B, Leveau A, Brett P, Goss RJM, Goossens A, O'Connell MA, Osbourn A. 2017. A translational synthetic biology platform for rapid access to gram-scale quantities of novel drug-like molecules. *Metabolic Engineering* **42**: 185– 193.
- Sainsbury F, Thuenemann EC, Lomonossoff GP. 2009. PEAQ: Versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnology Journal* **7**: 682– 693.
- Scheler U, Brandt W, Porzel A, Rothe K, Manzano D, Božić D, Papaefthimiou D, Balcke GU, Henning A, Lohse S *et al.* 2016. Elucidation of the biosynthesis of carnosic acid and its reconstitution in yeast. *Nature Communications* **7**: 12942.
- Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T *et al.* 2017. Genome-wide prediction of metabolic

- enzymes, pathways, and gene clusters in plants. *Plant Physiology* **173**: 2041–2059.
- Shimada TL, Shimada T, Hara-nishimura I. 2010. A rapid and non-destructive screenable marker, FAST, for identifying transformed seeds of *Arabidopsis thaliana*. *The Plant Journal* **61**: 519– 528.
- Sohrabi R, Huh J-H, Badiéyan S, Rakotondraibe LH, Kliebenstein DJ, Sobrado P, Tholl D. 2015. In planta variation of volatile biosynthesis: an alternative biosynthetic route to the formation of the pathogen-induced volatile homoterpene DMNT via triterpene degradation in *Arabidopsis* roots. *Plant Cell* **27**: 874– 890.
- Stephenson MJ, Reed J, Brouwer B, Osbourn A. 2018. Transient expression in *Nicotiana benthamiana* leaves for triterpene production at a preparative scale. *JoVE* **138**: e58169.
- Takos AM, Knudsen C, Lai D, Kannangara R, Mikkelsen L, Motawia MS, Olsen CE, Sato S, Tabata S, Jørgensen K *et al.* 2011. Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in *Lotus japonicus* and suggests the repeated evolution of this chemical defence pathway. *The Plant Journal* **68**: 273– 286.
- Tamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology* **57**: 758– 771.
- Töpfer N, Fuchs LM, Aharoni A. 2017. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Research* **45**: 7049– 7063.
- Tuominen LK, Johnson VE, Tsai C-J. 2011. Differential phylogenetic expansions in BAHD acyltransferases across five angiosperm taxa and evidence of divergent expression among *Populus* paralogues. *BMC Genomics* **12**: 236.
- Weng JK. 2014. The evolutionary paths towards complexity: a metabolic perspective. *New Phytologist* **201**: 1141– 1149.
- Weng J-K, Li Y, Mo H, Chapple C. 2012b. Assembly of an evolutionarily new pathway for  $\alpha$ -pyrone biosynthesis in *Arabidopsis*. *Science* **337**: 960– 964.
- Weng J-K, Philippe RN, Noel JP. 2012a. The rise of chemodiversity in plants. *Science* **336**: 1667– 1670.
- Winzer T, Gazda V, He Z, Kaminski F, Kern M, Larson TR, Li Y, Meade F, Teodor R, Vaistij FE *et al.* 2012. A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science* **336**: 1704– 1708.
- Zhou Y, Ma Y, Zeng J, Duan L, Xue X, Wang H, Lin T, Liu Z, Zeng K, Zhong Y *et al.* 2016. Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nature Plants* **2**: 1– 8.

# Chapter 4

## **Comparative transcriptomics reveals a conserved core of the phosphate starvation response across monocots and eudicots**

*Hernando G. Suarez Duran<sup>1</sup>, Yanting Wang<sup>2</sup>, Yunmeng Zhang<sup>3</sup>, Harro Bouwmeester<sup>2</sup>, Marnix H. Medema<sup>1</sup>*

<sup>1</sup> *Bioinformatics Group, Wageningen University, 6708 PB Wageningen, The Netherlands*

<sup>2</sup> *Swammerdam Institute for Life Sciences, University of Amsterdam, 1000 BE Amsterdam, The Netherlands*

<sup>3</sup> *Plant Physiology Group, Wageningen University, 6708 PB Wageningen, The Netherlands*

*Manuscript in preparation.*



## 4.1. Abstract

Phosphorus (P) is an indispensable nutrient for most plant functions, and inorganic phosphate, from which  $P_i$  fertilizer is made, is a finite resource. Because of this, there is a large interest in understanding the mechanisms of  $P_i$  metabolism in plants, the adaption of plants to low-phosphate conditions and the regulatory mechanisms that underlie this adaptation. Due to the ubiquitous importance of P in plant metabolism, we hypothesised cellular systems related to  $P_i$  metabolism to be largely conserved across monocots and eudicots. Comparative transcriptomics is a useful framework for identifying such conserved responses: it allows de novo predictions in more than one species with higher confidence than when analysing transcriptomes independently, because spurious associations are unlikely to be observed repeatedly. To study the phosphate starvation response using comparative transcriptomics, we developed CADE-HERoN, a computational workflow that combines targeted and untargeted approaches to study the transcriptomes of multiple plants comparatively through coexpression and differential expression analyses across time-series data. We applied our approach to data obtained from phosphate starvation experiments in rice, tomato and Arabidopsis, which yielded several candidate genes for functional characterization that are hypothesised to be involved in the biosynthesis of strigolactones and related metabolites. Furthermore, integrating differential and coexpression analyses across species led to the identification of important metabolic pathways associated with  $P_i$  starvation responses, related to gibberellin, lipid and carotenoid biosynthesis. This association thus appears to be conserved throughout eudicots and monocots. Our analyses provide concrete guidance for genetic dissection of the phosphate starvation response across species, and may serve as a model for the utilization of comparative transcriptomics to characterize and visualize conserved metabolic processes.

## 4.2. Introduction

Phosphorus (P) is an indispensable nutrient for most plant functions and is mostly assimilated by plants from the soil as inorganic phosphate ( $P_i$ ). This process has low efficiency: only ~30% of  $P_i$  added to the soil as fertilizer is captured by plants, while the rest is unavailable due to fixation in the soil and microbial activity<sup>1</sup>. Because rock phosphate, from which  $P_i$  fertilizer is made, is a finite resource, there is a large interest in understanding the mechanisms of  $P_i$  metabolism in plants, and more importantly, the adaption of plants to low-phosphate conditions and the regulatory mechanisms that underlie this adaptation<sup>2</sup>. One such mechanism involves the biosynthesis of strigolactones, a class of plant hormones that not only regulate the plant response to low  $P_i$ <sup>3,4</sup>, but also increase  $P_i$  acquisition and reduce  $P_i$  utilization<sup>5</sup>. Recently, we used transcriptome analysis to better understand the phosphate starvation response and strigolactones biosynthesis in tomato (Wang et al., manuscript in preparation) and rice (Zhang et al., manuscript in preparation), which confirmed that the genes involved in strigolactones biosynthesis are upregulated under  $P_i$  starvation and showed that strigolactones have an integral role in the full P starvation transcriptional response of these two species. However, the strigolactones biosynthetic pathway and its regulation in response to low phosphate are still largely a black box. Due to the ubiquitous importance of P in plant processes, we expect processes surrounding  $P_i$  metabolism to be largely conserved

across monocots and eudicots. Indeed, the functions of certain  $P_i$  transporters and non-coding RNAs involved in  $P_i$  deficiency responses have already been shown to be conserved between *Arabidopsis thaliana* and *Oryza sativa*<sup>6</sup>. It is not known, however, how widespread this apparent conservation is, and which specific responses are conserved.

To deepen our understanding of differences and commonalities in  $P_i$  metabolism across the plant Kingdom, we can use comparative transcriptomics. This framework has the advantage of enabling one to better understand conserved metabolic processes, transfer knowledge from model to non-model species, and make *de novo* predictions in more than one species with higher confidence than when analysing transcriptomes independently, because false associations are unlikely to be observed repeatedly<sup>7,8</sup>. In recent studies, this methodology has been successfully applied for plant natural product discovery. The expression profiles of genes with known orthology relationships in closely related Cucurbitaceae were analysed and compared, leading to the characterization of genes involved in the biosynthesis of cucurbitacins and regulators of bitterness in melon, watermelon and cucumber<sup>9</sup>. In another study, the -solanin/-chaconine pathways in tomato and potato were characterized by analysing the expression and coexpression of genes involved in glycoalkaloid metabolism in solanaceous plants<sup>10</sup>. The latter analysis was aided by the CoExpNetViz tool, which allows users to analyse their own transcriptome, orthology relationships and specific target genes, and then displays networks showing which genes are coexpressed with the targets in more than one species<sup>11</sup>.

While these previous approaches are useful to research the conservation of a given subset of genes, such as those involved in the strigolactones biosynthetic pathway, they do not allow global comparisons of transcriptomic responses, such as for the phosphate starvation response. This requires an untargeted analysis that allows to concurrently study the expression of all genes across all timepoints, something especially critical when considering evolutionarily distant species, in which the rate of the metabolic response may differ. Some algorithms for untargeted comparative transcriptomics have been previously developed<sup>12,13</sup>, such as IsoRankN<sup>14</sup>, which was used to compare the coexpression networks of rice and maize, where common network motifs with the same GO term enrichment were identified in both networks, indicating functional conservation of orthologous genes<sup>15</sup>. However, more effective methods to analyse transcriptomic responses across time series data are desirable, as are tools that allow users to concurrently analyse and integrate targeted and untargeted approaches on different species for comparative transcriptome analysis.

To study the effects of phosphate starvation in the transcriptome of multiple plant species, we selected this comparative transcriptomic framework and studied the gene expression time series data sets of the phosphate starvation response in *A. thaliana*, *O. sativa* and *S. lycopersicum*. For this, we developed a computational workflow, CADE-HERoN (**C**omparative **A**nalysis of **D**ifferential **E**xpression, **H**omologs **E**xpression, **C**oexpression **N**etworks) that combines targeted and untargeted approaches to study the transcriptome of multiple plants comparatively by leveraging coexpression, differential expression, and user-set orthology relationships. We used this workflow to perform a targeted analysis of genes involved in strigolactones biosynthesis in all three species and examine common properties of their global coexpression networks in an unsupervised way to identify conserved responses related to gibberellin, xanthophylls and tocopherol biosynthesis and lipid metabolic

processes.

### 4.3. Material and Methods

**RNA-Seq datasets.** Gene expression datasets from Zhang *et al.* and Wang *et al.* (unpublished data) were used in combination with a dataset comprising similar experimental conditions in *A. thaliana* (AT), which were retrieved from NCBI's GEO, with identifier GSE74856<sup>16</sup>. The experimental setup from Zhang *et al.* is as follows: wildtype and D mutant rice plants were grown for a week on normal conditions, after which some were moved to a low phosphate environment. Root and shoot samples were taken from days 7, 8, 10, 14, 14.6 and 15 from the start of the experiment. The phosphate replenishment treatment was performed from day 14 onwards. The experimental setup from Wang *et al.* is as follows: wildtype and CCD8 mutant tomato plants were grown for a week on normal conditions. After this, some were subjected to phosphate starvation. Root samples were taken from days 9, 10, 11, 12 and 14 after the start of the experiment. Phosphate replenishment was performed on day 12. The experimental setup for GSE74856<sup>16</sup> is as follows: *Arabidopsis thaliana* ecotype Columbia plants were grown for 10 days. On this day, some plants were moved to a phosphate-deficient medium. Root samples were taken on days 10, 11 and 13 from the start of the experiment.

**RNA-Seq analysis.** Expression analysis was performed in R 3.6.1<sup>17</sup> with the edgeR package 3.26.8<sup>18</sup>. Genes with less than 1 count per million reads mapped (CPM) in at least two samples were removed from the analysis. Reads were normalized by library and exon size (reads per kilobase per million mapped reads, RPKM). Differential gene expression (DGE) test P-values were corrected with the Bonferroni method, and the significance threshold was set at  $P < 0.1$  and  $\log FC > 1$ .

**Coexpression analysis.** Coexpression was defined for each plant species and tissue type independently using Pearson's Correlation Coefficient (PCC). Gene-pairs with  $PCC < 0.7$  were removed from further analysis. The results of this analysis were used to create an independent transcriptomic network per plant species, with genes represented as nodes, and edges representing coexpressed gene pairs. The networks were then integrated with edges linking genes with known orthologous relationships, as retrieved from EnsemblPlants (release 39)<sup>19</sup> through the BioMart data mining tool<sup>20</sup>.

**Target genes.** The integrated transcriptomic network was used to extract subnetworks based on target genes known to be involved in strigolactones biosynthesis and phosphate starvation responses<sup>3</sup>: CCD7 (AT2G44990, Solyc01g090660, OS04G0550600), CCD8 (AT4G32810, Solyc08g066650, OS01G0746400), MAX1 (AT2G26170, Solyc08g062950, OS04G0550600), and D27 (AT1G03055, Solyc09g065750; we could not identify the D27 ortholog in the rice transcriptome).

**Orthologous Communities (OCs).** The OC of a target gene is a cross-species coexpression network defined by the group of genes that are coexpressed with the target in more than one species, as identified through orthologous relationships. OCs are generated by: (1) isolating the coexpression network neighbourhood of a target gene, (2) identifying the target's ortholog in another species and isolating the corresponding neighbourhood, (3) removing genes which have orthologs not coexpressed with the target gene.

### **Orthologous Communities of Differentially Expressed Genes (OC-DEGs).**

To identify common differential transcriptomic responses, results from our DGE analysis were used to generate networks which include experimental conditions as a new node type. In this network, edges do not have a weight and only link genes to experimental conditions in which they were found to be differentially expressed. By using the same methodology to identify OCs on this DGE network, OC-DEGs are identified.

**Untargeted analysis.** To identify genes that were coexpressed in a similar manner across *O. sativa* (OS) and *S. lycopersicum* (SL), a coexpression network was generated in which each node represents a “gene family” instead of a gene. These “gene families” are defined by the orthology relationship of the genes: most nodes represent one SL gene and its OS ortholog, while some nodes may represent e.g. two SL genes and one OS gene if the orthology relationship is one-to-many. In this network, edges link a gene family composed of genes that are all coexpressed with the genes of the same species in the gene family at the opposite end of the edge (**Suppl. Fig. 1**).

## **4.4. Results and Discussion**

### **4.4.1. Abbreviations**

AT = *Arabidopsis thaliana*

OS = *Oryza sativa*

SL = *Solanum lycopersicum*

OC = Orthologous Community

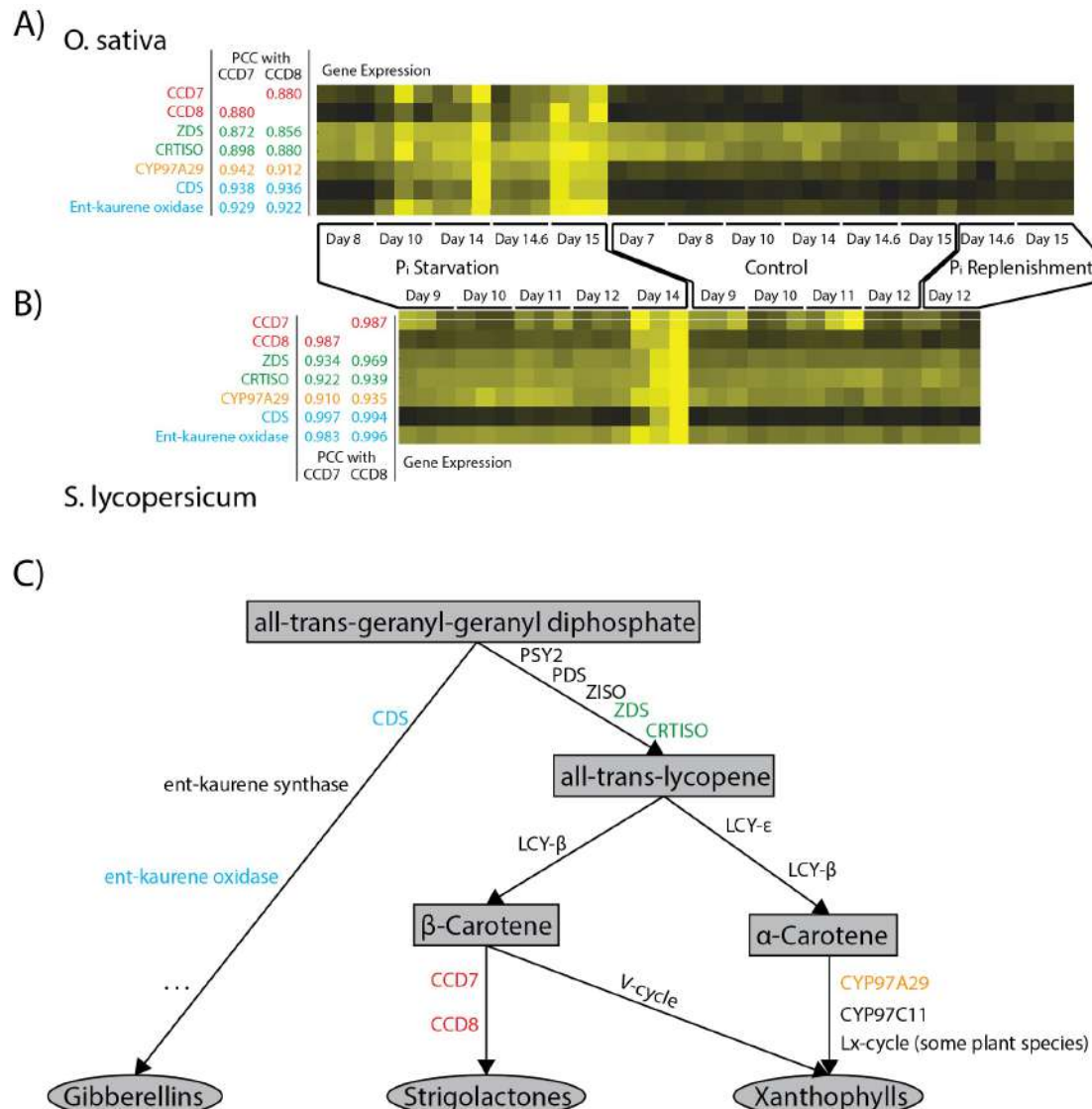
### **4.4.2. Conserved coexpression across rice and tomato indicates an ancient relationship between gibberellin, strigolactones and xanthophyllsbiosynthesis**

Strigolactones biosynthesis has been shown to be strongly affected by phosphate starvation<sup>3,4</sup> and our own research (Wang *et al.*, Zhang *et al.*, manuscripts in preparation) indicates strigolactones are also needed for the phosphate starvation response. In order to identify genes coexpressed with the core strigolactones biosynthetic genes (CCD7, CCD8, MAX1 and MAX2), we performed a targeted analysis to identify orthologous groups that are coexpressed with these bait genes in more than one species, so-called Orthologous Communities (OCs). We identified eight OCs around the four bait genes, which together describe the common transcriptomic response to phosphate starvation of the genes involved in strigolactones biosynthesis in three evolutionary distant species. A summary of this analysis can be seen in **Table 1** (the full list of genes in each OC is provided in **Suppl. Table 1**).

Bait alias	Bait genes	Plant species	Genes coexpressed with bait	Tissue	Community name
CCD8	Solyc08g066650.2;OS01G0746400	<i>S.lycopersicum</i> ;O.sativa	151	root	1o14
CCD7	Solyc01g090660.2;OS04G0550600	<i>S.lycopersicum</i> ;O.sativa	76	root	0o12
MAX1	Solyc08g062950.2;AT2G26170	<i>S.lycopersicum</i> ;A.thaliana	63	root	2o8
CCD8	Solyc08g066650.2;AT4G32810	<i>S.lycopersicum</i> ;A.thaliana	56	root	1o7
CCD8	AT4G32810;OS01G0746400	O.sativa;A.thaliana	43	root	7o14
MAX2	Solyc12g010900.1;AT2G42620	<i>S.lycopersicum</i> ;A.thaliana	13	root	3o9
CCD8	Solyc08g066650.2;AT4G32810;OS01G0746400	<i>S.lycopersicum</i> ;O.sativa;A.thaliana	10	root	1o7o14
MAX2	AT2G42620;OS06G0154200	O.sativa;A.thaliana	6	shoot	9o17

**Table 1.** Summary of the OC analysis in rice, tomato and *A. thaliana*, with the number of genes identified in each OC.

Four of the identified OCs among the three plant species were centred around CCD8, which includes the largest OC: 151 genes are coexpressed with CCD8 in OS and SL. The second largest OC, around CCD7 in OS and SL, only has 76 genes, all of which are part of the CCD8 OC as well. We performed a GO term enrichment analysis on these genes and found the CCD7/8 OC to be enriched in genes involved in lipid metabolism, terpenoid and gibberellin biosynthesis (**Suppl. Table 2** for the OS genes, and **Suppl. Table 3** for SL). The expression pattern of the genes annotated with these GO terms can be seen in **Fig. 1A)** for OS, and **Fig. 1B)** for SL. The genes annotated with terpenoid biosynthesis encode three enzymes within the xanthophylls biosynthetic pathway: a  $\zeta$ -carotene desaturase (ZDS) (Solyc01g097810/OS07G0204900), a prolycopene isomerase CRTISO (Solyc10g081650/OS11G0572700), and a carotenoid  $\beta$ -hydroxylase CYP97A29 (Solyc04g051190/OS02G0817900). The coexpression of ZDS and CRTISO with CCD7/8 is perhaps not surprising, as they both act a few enzymatic steps upstream of it: as seen in **Fig. 1C)**, ZDS acts directly upstream of CRTISO, and the reaction catalysed by CRTISO results in all-trans-lycopene, which if modified by the enzyme LCY- $\beta$ , results in  $\beta$ -carotene. This is the substrate CCD7 acts on to enter the strigolactones biosynthetic pathway. However, the coexpression of CYP97A29 is unexpected; this enzyme acts downstream of LCY- $\epsilon$ , which competes for all-trans-lycopene with LCY- $\beta$ , to produce zeinoxanthin<sup>21</sup>, the key precursor for xanthophylls pathways. A possible relationship between the strigolactones and xanthophylls pathway in a phosphate starvation context is intriguing:  $P_i$  deficiency has been shown to cause chlorophyll loss, which may lead to photo-oxidative stress if not counter-acted by non-photochemical quenching mechanisms such as the multiple xanthophylls cycles<sup>22</sup>. Interestingly, the taxonomically ubiquitous violaxanthin cycle (V-cycle) occurs parallel to the strigolactones pathway, downstream of  $\beta$ -carotene, whereas the taxonomically restricted lutein epoxide cycle (Lx-cycle) starts with lutein, two enzymatic steps downstream of  $\alpha$ -carotene, after the transformation of  $\alpha$ -carotene into zeinoxanthin by CYP97A29<sup>23</sup>. While the presence of an Lx-cycle has not been reported in any of the species analysed in this study<sup>24</sup>, we speculate that because the V-cycle and CCD7 would compete for the same substrate ( $\beta$ -carotene), an unknown mechanism could be promoting the Lx-cycle, or a similar undiscovered xanthophyll cycle in rice and tomato, to counter-act  $P_i$  deficiency-induced photo-oxidative stress.



**Fig. 1.** Expression pattern of CCD7/8 and the genes in **A)** rice and **B)** tomato coexpressed with these and annotated with gibberellin and carotenoid biosynthesis. The experimental conditions have been ordered for ease of comparison. **C)** The gibberellin, strigolactones and xanthophylls biosynthetic pathways.

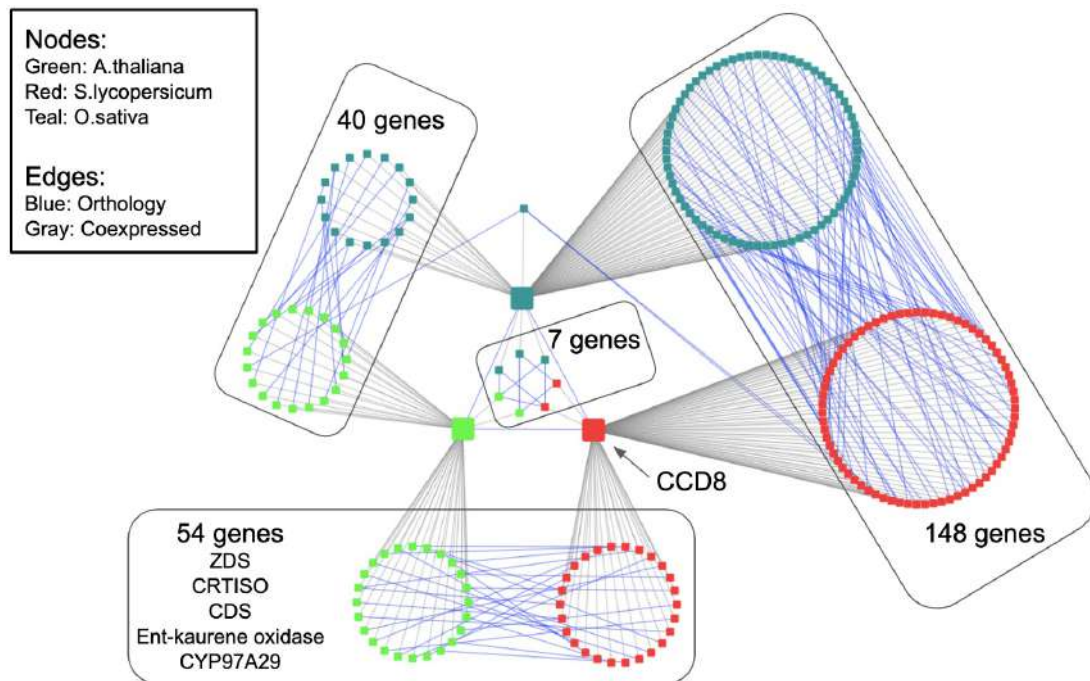
The genes annotated as involved in gibberellin biosynthesis in the CCD7/8 OC encode the enzymes ent-copalyl diphosphate synthase (CDS) (Solyc06g084240/OS02G0278700) and ent-kaurene oxidase (Solyc04g083160/OS06G0570100), two of the first three enzymes in the gibberellin biosynthesis pathway<sup>25</sup>. This pathway also has a close relationship with the carotenoid pathway by sharing the same precursor, all-*trans*-geranyl-geranyl diphosphate, on which CDS directly acts, as seen in **Fig. 1C**). Furthermore, just like strigolactones have a well-established involvement in P<sub>i</sub> deficiency responses in AT, OS and SL<sup>26,27</sup>, some studies have previously identified gibberellin as having a regulatory function in some P<sub>i</sub> deficiency responses in *A. thaliana*<sup>28</sup>.

A caveat in these analyses is that sequence-based orthology prediction does not guarantee functional equivalence at the biochemical level. However, similar expression patterns across similar experimental conditions of predicted orthologues strongly suggest a conserved function. Altogether, we speculate that identifying

xanthophylls and gibberellin biosynthesis genes closely coexpressed with CCD7/8 in tomato and rice suggests that they may have interlinked functions in the phosphate starvation response, that a similar regulation mechanism may exist across monocots and eudicots, and that much of the genetic machinery and functions behind these metabolic processes, such as the regulatory role of gibberellin, might have been conserved from before the divergence of monocots and eudicots.

#### **4.4.3. A conserved transcriptomic response across three plant species uncovers new candidate enzymes for strigolactones biosynthesis**

As our knowledge of the strigolactones biosynthesis pathway is still incomplete<sup>29</sup>, we next investigated whether cross-species coexpression analysis could generate new hypotheses regarding additional enzymes that might be involved in this pathway. The full CCD8 OC can be generated by identifying the CCD8 OC for each pair of species (43 genes in AT+OS, 151 genes in OS+SL, and 56 genes in AT+SL) in addition to the CCD8 OC among all three species, as seen in **Fig. 2**. We concatenated this list and removed duplicate genes and target genes from the list (CCD7, CCD8, MAX1 and MAX2); this results in a full CCD8 OC of 233 genes. Interestingly, some genes may appear in multiple of these OCs but not in the three-species OC, which is a consequence of incomplete orthology relationships in the input annotations (for example, the input data may show orthology between gene SL1 and OS1 and between SL1 and AT1, but not between AT1 and OS1, causing them to not appear in the three-species OC). The genes that appear in the OC of multiple bait genes are prime candidates for functional characterization and experimental validation, as these show conserved and strong coexpression with all known core strigolactones biosynthetic genes. Much is still unknown regarding these genes; for example, we queried the 233 genes from the CCD8 OC in UniProt<sup>30</sup> and retrieved annotations for 226 of them (**Suppl. Table 4**). Among the 91 SL genes present in this table, we found that 35 of them contain the words “uncharacterized protein”. Further research on these genes could lead to the annotation of enzymes, transporters or regulators associated with P<sub>i</sub> deficiency and/or strigolactones biosynthesis across three distinct plant species.



**Fig. 2.** The CCD8 OCs. Nodes represent genes, coloured by species: *A. thaliana* genes are green, *S. lycopersicum* genes are red, and *O. sativa* genes are teal. The big nodes are the target genes (CCD8). Gray edges link genes that are coexpressed ( $PCC > 0.7$ ), blue edges link orthologous genes.

Only two orthology groups are coexpressed with CCD8 in all three species, making up seven genes: two from AT, two from SL and three from OS. We analysed this group of genes in more detail. The two orthologous groups of genes in the AT-OS-SL CCD8 OC correspond to a gene family encoding U-box proteins (AT5G51270, OS06G0140800 and Solyc05g051610), and a gene family encoding cytochrome P450 enzymes that has been duplicated in OS (AT4G22690, Solyc08g079300, OS01G0700500 and OS08G0547900 [64.3% protein sequence identity]). These genes are specifically interesting, as the U-Box domain has been found to be involved in several abiotic stress responses in plants, including having a regulatory role in strigolactones and gibberellin signalling<sup>31</sup>, while cytochrome P450s have already been shown to be involved in the strigolactones biosynthetic pathway via MAX1, an enzyme that acts downstream of CCD7/8<sup>32,33</sup>.

#### 4.4.4. Communities of differentially expressed genes are conserved across eudicots and monocots

To study orthologous groups of genes that are specifically induced or inhibited by the phosphate starvation response, we then evaluated the common transcriptomic differential responses between OS and SL by identifying OC-DEGs. We did not include the AT dataset in this and the next part of the study because only the normalized gene expression (in RPKM) was publicly available, and DEG analysis requires the raw counts for accurate results. Due to the design of the experiments, only one sample is directly comparable among the two experiments: day 2 after the start of phosphate starvation (day 10 from the start of experiments). We identified an OC-DEG targeting this experimental condition consisting of 46 differentially expressed genes, corresponding to 20 gene families. Interestingly, the majority of



these genes consist of upregulated genes (**Suppl. Table 5**) with only two exceptions: OS04G0665600 and its ortholog Solyc10g080460.1, with a logFC of -1.21 and -1.28 respectively when compared to the control sample.

Because we did not expect the phosphate starvation response of these two distantly related plant species that were, moreover, grown under different conditions, to occur at the same rate, we repeated the OC-DEG search across all experimental condition comparison groups (phosphate starvation and replenishment vs control samples at the same timepoints) in both species, but only considered gene families with genes that are upregulated or downregulated in both species (**Tables 2-3**). While we previously showed that, at day 2 of phosphate starvation (day 10 of the experiment), 20 gene families were differentially expressed in OS and SL (19 upregulated and 1 downregulated), there is high overlap of upregulated genes families in both species between days 10-12 in SL and days 14-15 in OS: 37 genes in average throughout these comparisons, with the highest overlap occurring at day 14.6 of the OS starvation experiment, and day 11 for SL, where a total of 42 gene families are upregulated in both species. In contrast, very few genes are commonly upregulated in both species during the  $P_i$  replenishment process, where downregulation plays a more important and conserved role: 31 gene families in average are downregulated in both species in response to  $P_i$  replenishment. These results show that conserved transcriptomic responses during  $P_i$  deficiency across monocots and eudicots are not constrained to coexpression across time points, but also to differential gene expression at individual time points, and specifically strong induction of significant sets of genes. This is important because previous evidence shows that the function and members of coexpressed gene modules can be conserved through speciation events<sup>15</sup>, making comparative transcriptomics a powerful methodology to prioritize new candidate genes for functional characterization in conserved processes.

			O. sativa						
			P Starvation					P Replenishment	
			Day 8	Day 10	Day 14	Day 14.6	Day 15	Day 14.6	Day 15
S. lycopersicum	P Starvation	Day 9	0	11	14	15	14	1	2
		Day 10	0	19	33	37	36	2	7
		Day 11	0	20	35	41	37	1	8
		Day 12	0	20	33	37	35	1	7
	P Replenishment	Day 12	0	0	0	2	0	1	4

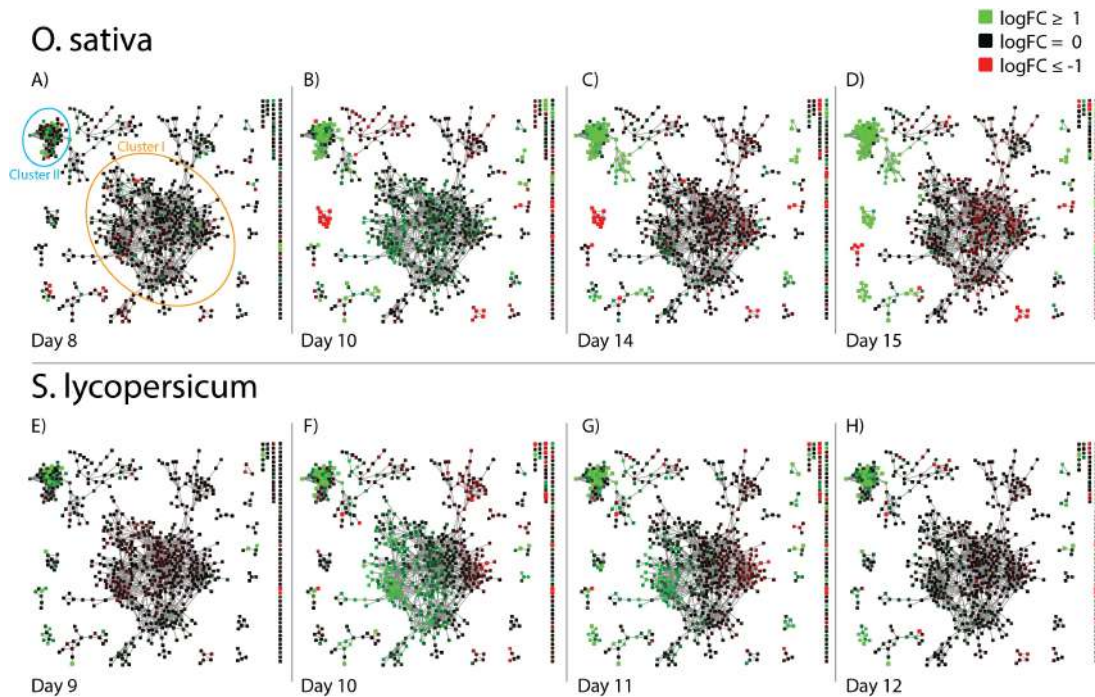
**Table 2.** Summary of the OC-DEG analysis in rice and tomato, showing the number of gene families that are upregulated in both species for any two comparison groups.

			O. sativa						
			P Starvation					P Replenishment	
			Day 8	Day 10	Day 14	Day 14.6	Day 15	Day 14.6	Day 15
S. lycopersicum	P Starvation	Day 9	0	0	0	0	0	0	0
		Day 10	0	1	2	5	4	1	1
		Day 11	0	2	6	8	6	2	2
		Day 12	0	0	1	1	0	0	0
	P Replenishment	Day 12	0	1	5	5	3	26	34

**Table 3.** Summary of the OC-DEG analysis in rice and tomato, showing the number of genes families that are downregulated in both species for any two comparison groups.

#### 4.4.5. An untargeted comparative time-series gene expression analysis reveals a conserved phosphate starvation response core gene set

While the results above show that a targeted approach can successfully generate interesting hypotheses about specific pathways, such an analysis is always biased by the specific target genes that are used. Hence, in order to compare the phosphate starvation response time series data sets in a more unbiased and comprehensive manner, we developed an untargeted mode for our analysis tools (see Methods) that reconstructs a global coexpression network of orthologous genes across species, and applied this on the OS and SL datasets. This analysis resulted in a complex coexpression network of 713 gene families and 2201 coexpression links among them, as seen in **Fig. 3**. This reveals that a total of 791 OS genes form a coexpression network under phosphate starvation with a similar topology as their 811 ortholog genes in SL.



**Fig. 3.** Timelapse view of the expression fold changes of gene families that are coexpressed during  $P_i$  starvation in *S. lycopersicum* (SL) and *O. sativa* (OS). Nodes represent gene families (e.g. one gene of each species if one-to-one orthology relationship). Edges link coexpressed gene

families. Nodes are coloured according to the highest absolute logFC value within that node, separately for each timepoint and each species (see color key). **A)** logFC values at day 8 in OS, **B)** day 10 in OS, **C)** day 14 in OS, and **D)** day 15 in OS. **E)** logFC values at day 9 in SL, **F)** day 10 in SL, **G)** day 11 in SL, and **H)** day 12 in SL.

To distinguish which groups of gene families are likely to be directly involved in  $P_i$  starvation or replenishment, we integrated this coexpression network analysis with DGE analysis and generated eight distinct graphical representations of the same network: one for either plant species in each of the four days of the  $P_i$  starvation experiment. For the OS graphs, we coloured each node according to the logFC value of the OS gene within the gene family, and in case the node represented two or more OS genes (in the case of one-to-many or many-to-many ortholog relationships), we selected as representative the gene with the highest absolute logFC value. We then repeated the procedure for the SL set of graphs. This effectively resulted in a “timelapse view” of the expression fold changes of gene families that are coexpressed during  $P_i$  starvation in both plants, as seen in **Fig. 3**. We can observe the expression changes in OS genes during  $P_i$  starvation on days 8, 10, 14 and 15 in **Fig. 3A-D)**, and the expression changes in SL genes during days 9, 10, 11 and 12 in **Fig. 3E-H)**.

In both species, the main cluster at the center of the network, highlighted as “cluster I” in the figure, mostly starts being upregulated in the second timepoint (day 10, **Fig. 3B,F)**, and in SL a number of genes remain upregulated on day 11 (**Fig. 3G)**, but the expression of most return to normal on day 12 (**Fig. 3H)** or day 14 in OS (**Fig. 3C)**. Of note, this cluster has upregulated genes involved in tocopherol biosynthesis, which, like xanthophylls, have been linked to protection against photo-oxidative stress and low phosphate availability<sup>34,35</sup>. The most striking commonality between the species is a cluster of gene families on the top left of the network, highlighted as “cluster II”, which show consistent upregulation by  $P_i$  starvation at all timepoints, with most gene families being upregulated already at the second timepoint (**Fig. 3B,F)**. Interestingly the upregulation in cluster II is more pronounced in OS, and by the last timepoint (**Fig. 3D)**, only 4 gene families in this cluster are not strongly upregulated in this species. There are 53 gene families in this cluster, representing 76 OS genes and their 54 orthologous genes in SL, as seen in **Suppl. Table 6**. A GO term analysis revealed cluster II is enriched in OS genes annotated with organophosphate catabolism, and lipid metabolic processes, in particular glycerol- and phospholipids: OS04G0394100, OS02G0514500 and OS08G0535700, paralogs that belong to the same orthology group with Solyc02g094400. Indeed, under  $P_i$  starvation plants undergo multiple lipid metabolic processes to replace phospholipids across different types of membranes at the cell level and conserve phosphate<sup>36–38</sup>. In the future, further analysis of yet uncharacterized gene families in this cluster may lead to additional insights regarding the phosphate starvation response, and annotation of functionally conserved genes across multiple plant species: 16 out of 54 SL genes in this cluster are annotated as uncharacterized in UniProt.

Altogether, this analysis suggests that the conserved core of the phosphate starvation response across monocots and eudicots involves coregulated lipid metabolic processes, and coregulated metabolic pathways that lead to the biosynthesis of gibberellin, strigolactones, xanthophylls and tocopherols, among others.

## 4.5. Conclusions

Focusing the analysis around the genes known to be involved in strigolactones biosynthesis yielded several candidate genes for functional characterization in regard to their involvement within the strigolactones pathway. Furthermore, the differential expression and untargeted analyses retrieved important metabolic pathways known to be associated with  $P_i$  starvation responses that thus appear to be conserved throughout eudicots and monocots. These associated responses constituted a small portion of the global common networks, and further research within the other network modules may yield multiple other conserved processes associated to  $P_i$  starvation in plants. In the future, our framework for comparative analysis of differential and coexpression networks may be used as a powerful methodology to characterize conserved pathways and transcriptomic responses in other biological systems and conditions as well.

## 4.6. Tool availability

We have made CADE-HErON, our computational workflow to identify OCs and OC-DEGs, fully available at <https://git.wageningenur.nl/medema-group/cade-heron/>

## 4.7. Supplementary Information

Supplementary figures and tables are available to download from: <https://doi.org/10.5281/zenodo.4056494>

## References

1. López-Arredondo, D. L., Leyva-González, M. A., González-Morales, S. I., López-Bucio, J. & Herrera-Estrella, L. Phosphate Nutrition: Improving Low-Phosphate Tolerance in Crops. *Annu. Rev. Plant Biol.* **65**, 95–123 (2014).
2. Veneklaas, E. J. *et al.* Opportunities for improving phosphorus-use efficiency in crop plants. *New Phytol.* **195**, 306–320 (2012).
3. Mayzlish-Gati, E. *et al.* Strigolactones Are Involved in Root Response to Low Phosphate Conditions in Arabidopsis. *Plant Physiol.* **160**, 1329–1341 (2012).
4. Ruyter-Spira, C. *et al.* Physiological Effects of the Synthetic Strigolactone Analog GR24 on Root System Architecture in Arabidopsis: Another Belowground Role for Strigolactones? *Plant Physiol.* **155**, 721–734 (2011).
5. Czarnecki, O., Yang, J., Weston, D. J., Tuskan, G. A. & Chen, J.-G. A dual role of strigolactones in phosphate acquisition and utilization in plants. *Int. J. Mol. Sci.* **14**, 7681–701 (2013).
6. Wang, F., Deng, M., Xu, J., Zhu, X. & Mao, C. Molecular mechanisms of phosphate transport and signaling in higher plants. *Semin. Cell Dev. Biol.* **74**, 114–122 (2018).

7. Movahedi, S., Van Bel, M., Heyndrickx, K. S. & Vandepoele, K. Comparative co-expression analysis in plant biology. *Plant, Cell Environ.* **35**, 1787–1798 (2012).
8. Janowski, M., Musialak-Lange, M., Hansen, B. O., Vaid, N. & Mutwil, M. Elucidating gene function and function evolution through comparison of co-expression networks of plants. *Front. Plant Sci.* **5**, 1–9 (2014).
9. Zhou, Y. *et al.* Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nat. Plants* **2**, 1–8 (2016).
10. Itkin, M. *et al.* Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* **341**, 175–9 (2013).
11. Tzfadia, O. *et al.* CoExpNetViz: Comparative Co-Expression Networks Construction and Visualization Tool. *Front. Plant Sci.* **6**, 1194 (2016).
12. Netotea, S., Sundell, D., Street, N. R. & Hvidsten, T. R. ComPIEx: Conservation and divergence of co-expression networks in *A. thaliana*, *Populus* and *O. sativa*. *BMC Genomics* **15**, 1–17 (2014).
13. Movahedi, S., Van de Peer, Y. & Vandepoele, K. Comparative Network Analysis Reveals That Tissue Specificity and Gene Function Are Important Factors Influencing the Mode of Expression Evolution in Arabidopsis and Rice. *Plant Physiol.* **156**, 1316–1330 (2011).
14. Liao, C. S., Lu, K., Baym, M., Singh, R. & Berger, B. IsoRankN: Spectral methods for global alignment of multiple protein networks. *Bioinformatics* **25**, 253–258 (2009).
15. Ficklin, S. P. & Feltus, F. A. Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice. *Plant Physiol.* **156**, 1244–1256 (2011).
16. Liu, T.-Y. *et al.* Identification of plant vacuolar transporters mediating phosphate storage. *Nat. Commun.* **7**, 11095 (2016).
17. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing* (2017).
18. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
19. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
20. Smedley, D. *et al.* BioMart - Biological queries made easy. *BMC Genomics* **10**, 1–12 (2009).
21. Isaacson, T., Ronen, G., Zamir, D. & Hirschberg, J. Cloning of tangerine from Tomato Reveals a Carotenoid Isomerase Essential for the Production of  $\beta$ -Carotene and Xanthophylls in Plants. *Plant Cell* **14**, 333–342 (2002).
22. Hernández, I. & Munné-Bosch, S. Linking phosphorus availability with photo-oxidative stress in plants. *J. Exp. Bot.* **66**, 2889–2900 (2015).
23. Stigliani, A. L., Giorio, G. & D'Ambrosio, C. Characterization of P450 Carotenoid  $\beta$ - and  $\epsilon$ -Hydroxylases of Tomato and Transcriptional Regulation of Xanthophyll Biosynthesis in Root, Leaf, Petal and Fruit. *Plant Cell Physiol.* **52**, 851–865 (2011).

24. García-Plazaola, J. I., Matsubara, S. & Osmond, C. B. The lutein epoxide cycle in higher plants: its relationships to other xanthophyll cycles and possible functions. *Funct. Plant Biol.* **34**, 759 (2007).
25. Hedden, P. & Thomas, S. G. Gibberellin biosynthesis and its regulation. *Biochem. J.* **444**, 11–25 (2012).
26. Gomez-Roldan, V. *et al.* Strigolactone inhibition of shoot branching. *Nature* **455**, 189–194 (2008).
27. Umehara, M. *et al.* Inhibition of shoot branching by new terpenoid plant hormones. *Nature* **455**, 195–200 (2008).
28. Jiang, C., Gao, X., Liao, L., Harberd, N. P. & Fu, X. Phosphate Starvation Root Architecture and Anthocyanin Accumulation Responses Are Modulated by the Gibberellin-DELLA Signaling Pathway in Arabidopsis. *Plant Physiol.* **145**, 1460–1470 (2007).
29. Wang, Y. & Bouwmeester, H. J. Structural diversity in the strigolactones. *J. Exp. Bot.* **69**, 2219–2230 (2018).
30. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
31. Shu, K. & Yang, W. E3 Ubiquitin Ligases: Ubiquitous Actors in Plant Development and Abiotic Stress Responses. *Plant Cell Physiol.* **58**, 1461–1476 (2017).
32. Zhang, Y. *et al.* The tomato MAX1 homolog, SIMAX1, is involved in the biosynthesis of tomato strigolactones from carlactone. *New Phytol.* **219**, 297–309 (2018).
33. Abe, S. *et al.* Carlactone is converted to carlactonoic acid by MAX1 in Arabidopsis and its methyl ester can directly interact with AtD14 in vitro. *Proc. Natl. Acad. Sci.* **111**, 18084–18089 (2014).
34. Hernández, I. & Munné-Bosch, S. Linking phosphorus availability with photo-oxidative stress in plants. *J. Exp. Bot.* **66**, 2889–2900 (2015).
35. Demmig-Adams, B. *et al.* Emerging trade-offs - impact of photoprotectants (PsbS, xanthophylls, and vitamin E) on oxylipins as regulators of development and defense. *New Phytol.* **197**, 720–729 (2013).
36. Hartel, H., Dormann, P. & Benning, C. DGD1-independent biosynthesis of extraplastidic galactolipids after phosphate deprivation in Arabidopsis. *Proc. Natl. Acad. Sci.* **97**, 10649–10654 (2000).
37. Andersson, M. X., Stridh, M. H., Larsson, K. E., Liljenberg, C. & Sandelius, A. S. Phosphate-deficient oat replaces a major portion of the plasma membrane phospholipids with the galactolipid digalactosyldiacylglycerol. *FEBS Lett.* **537**, 128–132 (2003).
38. Jouhet, J. *et al.* Phosphate deprivation induces transfer of DGDG galactolipid from chloroplast to mitochondria. *J. Cell Biol.* **167**, 863–874 (2004).

# Chapter 5

## MEANtools: An Integrative Multi-omics Approach for Metabolic Pathway Prediction

*Hernando G. Suarez Duran<sup>1</sup>, Olga Zafra Delgado<sup>1</sup>, Justin J. J. van der Hooff<sup>1</sup>, Elizabeth S. Sattely<sup>2</sup>, Marnix H. Medema<sup>1</sup>*

<sup>1</sup> *Bioinformatics Group, Wageningen University, 6708 PB Wageningen, The Netherlands*

<sup>2</sup> *Department of Chemical Engineering, 94305 Stanford University, Stanford, CA, USA*

*Manuscript in preparation.*

## 5.1. Abstract

Plants harbour a highly diverse and complex specialized metabolism, with key functions in pollination, stress response, microbiome modulation and defence against herbivores and pathogens. Contemporary omics-based approaches to identify the underlying enzymatic pathways have so far largely been dominated by genomic and transcriptomic methods that use known metabolites or enzymes as a starting point. Multi-omics integration has the potential to solve this limitation by allowing the prediction of metabolic pathway to be made through simultaneous observations in multiple -omic sources that are used to generate a single hypothesis. However, no systematic, unsupervised multi-omics method has been developed that integrates transcriptomic, metabolomic and genomic data for untargeted specialized metabolic pathway discovery. Here, we present MEANtools, a multi-omics analysis workflow that integrates genomic, transcriptomic and metabolomic data with enzymatic reaction databases to predict metabolic pathways, by identifying mass differences between metabolites that are co-abundant with transcripts whose enzymatic products are capable of catalyzing reactions that can explain these. MEANtools has a flexible and user-friendly input and output, and allows users to tune the enzymatic reactions that are retrieved from the database to generate metabolic pathway predictions; for example, by retrieving only experimentally validated reactions, or by selecting a taxa of origin. We illustrate MEANtools' usage with a paired transcriptomic-metabolomic experiment and show that it is able to predict multiple steps in the recently characterized falcarindiol biosynthetic pathway in tomato. This demonstrates its potential to generate testable hypotheses on metabolites, enzymes and reactions in metabolic pathways.

## 5.2. Introduction

Plants produce a wide array of natural products (NPs) with abundant chemical diversity and complexity<sup>1</sup>. Specialized metabolites have a wide range of uses in a variety of industries, ranging from pharmacological, to flavors and fragrances<sup>2</sup>. This has fueled an interest in new methodologies to predict and identify specialized metabolites and the metabolic pathways different species use to produce them<sup>3</sup>. Indeed, the field studying specialized metabolism has come a long way since the discovery of penicillin in fungi ~90 years ago: from mostly phenotypic screening methods in the 1940s, to the bioinformatics- and genomics-based methods developed in recent years to study both microorganisms and plants<sup>4,5</sup>. In recent years, the high level of success and many advances in bioinformatics and -omics analysis have resulted in an ever-increasing number of high-quality genome assemblies, as well as transcriptome, metabolome, and enzymatic reaction datasets. Moreover, advances in synthetic biology allow the results of *in silico* analyses to be more easily validated *in vivo*, increasing the rate at which novel NPs and their producing enzymes can be characterized. For instance, transient expression of plant enzymes in tobacco allows enzyme and metabolic pathway characterization within a matter of weeks<sup>6</sup>. Taken together, the plant functional genomics field has entered a new era<sup>7</sup>.

This computational era of plant NP functional genomics has so far largely been defined by genomic and transcriptomic strategies<sup>5</sup>. Examples of the former include computational tools such as plantiSMASH<sup>8</sup> and PhytoClust<sup>9</sup>, which identify



chromosomally co-localized clusters of genes that may encode specialized metabolic pathways. Transcriptomic strategies rely on the guilt-by-association principle<sup>10</sup> to identify enzyme-coding genes with concerted expression that are therefore likely to function within the same metabolic pathway; however, these methods require prior knowledge about expected enzyme functions to link the enzyme-coding genes to specific metabolites: for example, by targeting known metabolites or enzymes, their structural or functional annotations can be used to generate hypotheses about the metabolic pathways they are associated with<sup>11,12</sup>. More recently, tools like CoExpNetViz<sup>13</sup> and CADE-HEroN (Suarez et. al, manuscript in preparation) leverage sequence homology for this purpose by using transcriptomes from multiple species, however many plant enzymatic functions are known to be species-specific<sup>14</sup>, and therefore unlikely to be predicted by these methods.

A promising solution to this limitation may be found in the integration of data from multiple –omic sources. Plant specialized metabolism is a complex interaction of multiple biological systems working synergistically, and multi-omics techniques offer a holistic view of the process. They allow researchers to acquire a more accurate understanding of metabolism and reduce type I and type II errors in the predicted associations among genes, metabolites and enzymatic reactions by integrating observations from various data sources to predict metabolic pathways<sup>15</sup>. Multi-omics integration strategies can be broadly separated into four categories: conceptual, statistical, model and pathway-based. Each strategy presents distinct challenges and all have been reviewed in detail before, with multiple examples of successful usage<sup>15,16</sup>. Despite this no systematic, unsupervised multi-omics method has been developed that integrates transcriptomic, metabolomic and genomic data for untargeted specialized metabolic pathway discovery.

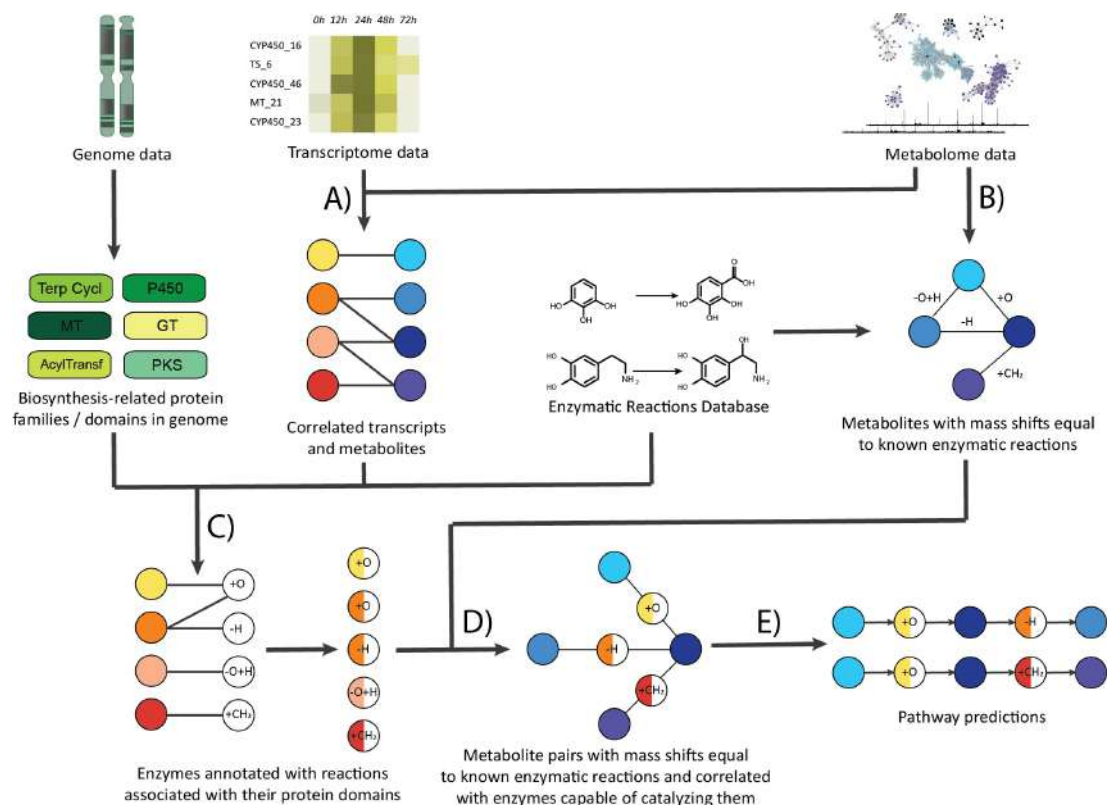
In this paper, we present MEANtools, a multi-omics analysis workflow that combines correlation-based and pathway-based integration techniques along with the guilt-by-association principle to predict metabolic pathways, and the enzyme-coding genes and metabolites associated with each prediction, presenting researchers with hypotheses that can be prioritized for experimental validation. (**Fig. 1**). MEANtools predicts metabolic pathways by correlating the expression of enzyme-coding genes with co-abundant metabolites in paired transcriptomic-metabolomic experiments. Although this methodology has aided the characterization of diverse metabolic processes in plants<sup>17,18</sup> by reducing the dimensionality of the problem and thus generating a small set of testable hypotheses<sup>19</sup>, it is known to result in a high number of false positive metabolite-transcript associations<sup>20</sup>. Therefore, we leverage RetroRules<sup>21</sup>, a retrosynthesis-oriented database of enzymatic reactions annotated with known and predicted protein domains and enzymes linked to each reaction, to assess whether observed chemical changes of metabolites (identified as mass shifts) can be explained by positing subsequent catalytic steps that are known to be catalyzed by protein families encoded in coexpressed gene modules. MEANtools generates a metabolic network based on enzymatic reaction databases, a workflow used by many computational tools for retrosynthetic metabolic pathway design or reconstruction<sup>22</sup>; however, MEANtools uses metabolome, transcriptome and metabolome data as well in this process, allowing users to explore the biosynthetic potential of any molecular structure and generate concrete hypotheses about possible pathways leading up to (or from) a given metabolite, which can be tested in the laboratory. Results are displayed in a variety of formats for users to interact with, describing predicted metabolic pathways

along with the metabolites, enzymes and reactions that are potentially involved in them. Altogether, MEANtools serves as strong basis for the development of methodologies to explore ways in which paired genomic, transcriptomic and metabolomic data can be used to analyze biosynthetic diversity.

## 5.3. Methods and Implementation

### 5.3.1. MEANtools workflow

MEANtools generates metabolic pathway predictions by integrating metabolomic, transcriptomic and genomic data with enzymatic reaction databases, as seen in **Fig. 1**. For this purpose, the user can annotate a (sub)set of mass signatures (mass-to-charge ratios of measured ions) in the metabolomic dataset with metabolite structures, and/or MEANtools can assign structure predictions by identifying adducts in the metabolome and querying NPDB (Natural Product Database), a comprehensive database of natural product structures which has been compiled by combining a range of public databases (*Stokman et al.*, in preparation; <http://...>). Possible links between metabolites and transcripts are first identified by calculating correlations across samples, and the correlated pairs are then queried in RetroRules to identify transcript-metabolite pairs in which the transcript encodes an enzyme with a protein domain capable of catalyzing an enzymatic reaction that has the correlated metabolite as possible substrate or product. MEANtools then maps the products of each successful reaction to other mass signatures in the metabolome, or to unmeasured “ghost” mass signatures (explained further down below). This procedure is then repeated iteratively according to the user’s specifications to predict further reactions from the new products and generate a reaction network. Finally, MEANtools analyzes the resulting network to provide the user with possible metabolic pathway predictions that can then be browsed in Cytoscape<sup>23</sup>, or displayed as figures indicating each metabolite, reaction, and possible enzymes that may explain the transformations required.



**Fig. 1.** MEANtools predicts metabolic pathways by integrating transcriptomic, metabolomic and genomic data. **A)** First, correlations are computed between expression levels of transcripts and abundances of metabolites. **B)** Mass signatures in the metabolome are queried against RetroRules to identify pairs with mass differences associated with known enzymatic reactions. **C)** The protein families/domains encoded by the genes in the correlated transcript-metabolite pairs are used to query RetroRules and identify which enzymatic reactions may be associated with each transcript. **D)** MEANtools then integrates the results of previous steps to identify cases in which metabolite pairs are correlated to a transcript that encodes an enzyme capable of catalyzing a reaction that explains their mutual mass difference. **E)** Finally, MEANtools maps the product of these reactions to other mass signatures in the metabolome, and repeats the procedure to generate pathway predictions.

### 5.3.2. Correlation-based integration generates testable associations

In stage A, MEANtools processes transcriptome and metabolome data from paired datasets. The transcriptome must be input in a CSV table of normalized gene expression data (RPKM/FPKM), and the metabolome as a CSV table of metabolite abundance, with mass signatures as rows and samples as columns. Both inputs are converted into DataFrames with the Pandas Python package (v0.24.2)<sup>24</sup>, leveraging its merge function for each data integration and annotation step. MEANtools uses SciPy (v1.2.1)<sup>25</sup> to calculate Spearman's rank correlations and identify pairs of transcripts and mass signatures with positively or negatively correlated expression (0.7) and abundance throughout the input, according to user-set correlation thresholds. Additionally, MEANtools also provides the user the option to use Pearson's correlation coefficient instead, and/or select a custom correlation cutoff to increase the number of resulting associations at the expense of reliability.

Global reconstruction of co-expression modules in gene expression data has been shown to be a powerful method to identify groups of genes involved in the same metabolic pathway when querying for modules with genes that encode

biosynthetic enzymes<sup>26</sup>. Because of this, selecting as input for MEANtools only genes present in this type of coexpression modules has the potential to result in increased confidence of the predictions as they will have a higher chance of being associated with the same metabolic pathway than otherwise. To facilitate this process, MEANtools includes our own implementation of the approach to identify gene coexpression clusters described by Wisecaver *et al.*<sup>26</sup> as an optional pre-processing step. Furthermore, MEANtools allows users to visualize the expression of each cluster in heatmaps with genes sorted in three categories according to the protein domains they encode, following the same categorization used by plantiSMASH<sup>8</sup>: scaffold-generating enzymes, tailoring enzymes, and the remaining genes.

### **5.3.3. Mass-shifts associated to reactions serve as templates for pathway prediction**

In stage B, MEANtools scans the metabolome to identify mass signatures that could represent metabolites in same metabolic pathway. For this purpose, MEANtools queries all enzymatic reactions in RetroRules and cross-references them with MetaNetX<sup>27</sup>, a repository of metabolic networks that MEANtools uses to identify the mass differences between the main products and substrates of all reactions. MEANtools then annotates all reactions with an associated mass shift. This only needs to be performed once, upon retrieving or updating the RetroRules database.

MEANtools uses the generated associations between reactions and mass shifts to scan the input metabolome and identify pairs of mass signatures with a difference in mass-charge ratio that can be explained by a known reaction, annotating one mass signature as possible substrate and the other as possible product. Moreover, because a mass shift may be explained by more than one reaction, and many reactions are reversible, any pair of mass signatures can be annotated with multiple reactions in both directions. Furthermore, because it cannot be assumed that all metabolites in a metabolic pathway will be present at sufficient levels to be detected in the metabolome, MEANtools generates “ghost mass signatures” which serve as an intermediate unmeasured metabolite between any two metabolites with measured mass signatures, a concept that was recently applied by Network to generate metabolic networks based on MS/MS spectra<sup>28</sup>. Based on the combination of reaction-mass-shift associations and bridging ghost mass signatures, MEANtools generates a reaction network composed of mass signatures linked by annotated reactions that serves as the basis for the prediction of metabolic pathways.

### **5.3.4. Curated Pfam-reaction annotations constrain metabolite-enzyme associations**

Stage C occurs concurrently: MEANtools integrates the metabolomic-transcriptomic correlation data from the first stage as a constraint to identify reactions in the generated reaction network that can be explained by one of the enzymes encoded by genes that are correlated to the mass signatures linked by the reaction. To this end, we adapted the RetroRules database, which is populated with ~17,000 reactions annotated with enzymes that are predicted to be associated with




them<sup>21</sup>. The majority of these annotated enzyme-reaction associations, however, are the result of propagating the annotation of characterized reactions to other reactions with the same or similar enzyme commission (EC) number, and are therefore likely to contain too many errors to function as an effective constraint for our applications. In order to increase confidence in the enzymatic annotations we cross-referenced each reaction in RetroRules to the manually curated reaction databases Rhea<sup>29</sup> and KEGG<sup>30</sup>, identified reaction-enzyme associations supported by experimental evidence and then propagated these annotations through KEGG Orthology groups. This resulted in a curated set of 5,168 reactions and 2,706,626 respective high-confidence reaction-enzyme associations (**Suppl. Table 1**). To increase the number of enzymatic transformations available for metabolic pathway predictions, we generated a second, less strict, set of associations. For this, we used ECDomainMiner<sup>31</sup>, a tool designed to infer reaction associations between EC numbers and Pfam<sup>32</sup> domains for automatic annotation. Using the first set of curated rules as a starting point, we identified a set of associations thus inferred with at least 0.95 confidence. We then selected all other reactions from RetroRules with an EC number and Pfam annotations identical to this high-confidence set, and included them in a second set of associations, resulting in 7,070 reactions and 3,135,908 reaction-enzyme associations. We generated a third set of 10,650 reactions and 3,203,412 reaction-enzyme associations by identifying RetroRules reactions annotated with fewer than 7 Pfams, an arbitrary cut off that serves as a tradeoff between enzymatic annotation specificity and the number of preserved reactions. MEANtools includes these three different reaction-enzyme associations datasets as settings (strict, medium and loose, respectively) to allow the user to constrain the predictions for specific purposes and find the right balance between sensitivity and specificity, considering the tradeoff between enzymatic annotation confidence and diversity of the resulting set of enzymatic reactions.


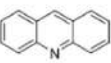

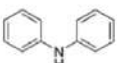
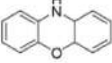
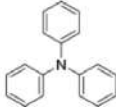
By cross-checking these reaction-enzyme association data sets with sets of correlated enzyme-coding genes and metabolites, MEANtools effectively filters the set of possible mass shift-reaction associations based on -omics evidence. Thus, MEANtools generates a reaction network where each node is a mass signature within the metabolome, or an unmeasured ghost mass signature. In this network, nodes are linked by directed edges representing enzymatic reactions that can be catalyzed by at least one of the enzyme families encoded by the genes correlated to one of the two mass signatures the reaction links.

### 5.3.5. MEANtools predicts metabolic pathways supported by multi-omics input and enzymatic reaction databases

In stage D, MEANtools uses the reaction network to generate pathway predictions. To this end, it first predicts possible metabolites and their corresponding molecular structures for each mass signature by identifying possible adducts and querying NPDB (*Stokman et al.*, in preparation), or a user-defined metabolite database that can be supplied in CSV format. MEANtools then uses the rdkit Python package (v 2019.03.2.0)<sup>33</sup> to generate *in silico* molecules resulting from each reaction associated substrate(s). Because of the large number of reactions in RetroRules, generating all product molecules for the metabolite structures predicted in a metabolome by querying all reactions can become time-consuming and computationally taxing. Every successful reaction will result in new metabolites that

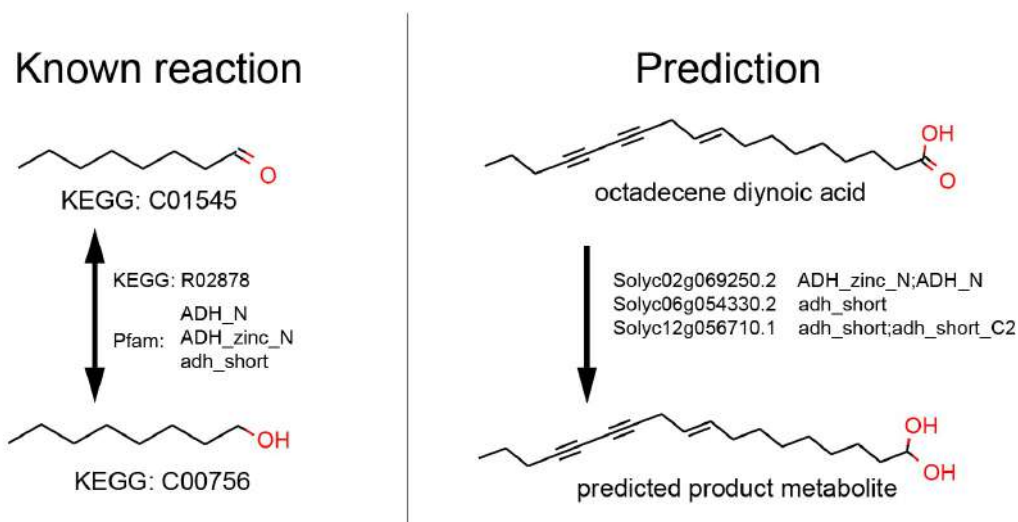
are ready to be queried again to predict a potential pathway, leading to iterations that are increasingly larger. To speed up this process, MEANtools guides the generation of *in silico* molecules by -omics evidence, namely the reaction-substrate pairs identified in the previous stage. To further expedite each iteration, MEANtools queries reactions according to key substructures present in the substrate molecule using a divide-and-conquer strategy as seen in **Fig. 2**. For each metabolite structure (**Fig. 2 A**), MEANtools starts by querying the presence of specific atoms in the structure, like N or C (**Fig. 2 B**). Upon success, in the next step MEANtools queries reactions that only involve simple substructures, like N=N and C=C, and if both atoms are present, reactions centered on the substructures C=N and C-N are also queried (**Fig. 2 C**). In the next round, MEANtools queries more complex substructures according to which substructures have already been identified, for example, only metabolites with the substructure C=N are queried for reactions centered on the C=N-C substructure (**Fig. 2 D**), and this process ends for each metabolite when a round does not result in any successful queries. This process is expedited further by only querying reactions that would result in metabolites with a mass signature that can be mapped in the metabolome, or as a ghost mass signature. MEANtools precomputes this map of substructures once, upon retrieving or updating the RetroRules database. Altogether, this strategy lets MEANtools utilize computing time efficiently when generating *in silico* molecules.

 Substructure in metabolite  
 Substructure not in metabolite  
 Substructure untested

A)		Input metabolites					
							
B)	Reactant substructures						
	N	✓	✓	✓	✓	✓	✓
	C	✓	✓	✓	✓	✓	✓
C)	C-C	✓	✓	✓	✓	✓	✓
	C=C	✓	✓	✓	✓	✓	✓
	N=N	✗	✗	✗	✗	✗	✗
	N-C	✓	✓	✓	✓	✓	✓
	N=C	✗	✓	✓	✗	✗	✗
D)	C-N-C	✗	✗	✗	✓	✓	✓
	C=N-C	○	✓	✓	○	○	○
	N-C-N	✗	✗	✗	✗	✗	✗
	N=C-N	○	✗	✓	○	○	○

**Fig. 2.** MEANtools identifies possible reactions for a molecular structure according to a divide-and-conquer strategy. For each metabolite, MEANtools first queries the presence of key atoms, and then continues to query, in rounds, increasingly complex reactant substructures according to which substructures have already been identified. For example, **A)** a set of metabolites is initially queried for the presence of **B)** nitrogen and carbon atoms. **C)** Metabolites that pass these criteria are then be queried for more complex substructures like C-N or C=C. **D)** In the following round, MEANtools queries substructures with more complexity according to which substructures have already been identified: in this manner, only metabolites with the N=C substructure are queried for the N=C-N substructure.

As a result of this stage, MEANtools generates series of ensuing reactions with predicted products for all mass signature-pairs, associated correlated enzyme-coding genes, and references to the characterized reactions and enzymes that served as rules to predict this reaction. An example of this can be seen in **Fig. 3**, where a characterized dehydrogenation serves as template to predict a dehydrogenation of octadecene diynoic acid. This last stage can be iterated multiple times (as desired by the user), to generate pathway predictions that extend beyond one enzymatic reaction away from the initial query molecule.



**Fig. 3.** MEANtools uses reaction rules from known enzymatic reactions to identify predict structures in metabolic pathways. For example, the dehydrogenation of 1-Octanol into 1-Octanal (KEGG reaction: R02878) is used as template to predict the dehydrogenation of octadecene diynoic acid.

### 5.3.6. Easy-to-browse untargeted and targeted metabolic pathway predictions

In stage E, MEANtools analyses the resulting reaction network generated in previous stages to predict candidate metabolic pathways of interest to the user. For this purpose, MEANtools uses the NetworkX Python package (v2.4)<sup>34</sup> to generate one subnetwork for each of the initial metabolites input by the user. MEANtools converts each subnetwork into a directed acyclic graph (DAG) by identifying all cycles within the network that represent predicted reversible reactions, only retaining links that are able to move the reaction forward and away from the initial metabolite. In the case of cycles between metabolites at the same reaction distance from the initial metabolite, the edge with the weakest enzyme-metabolite correlation is removed. This method effectively generates various DAGs rooted at the initial metabolites, for which candidate metabolic pathways can be predicted by identifying the longest reaction path in each subnetwork that starts from the initial metabolite. The method is repeated to generate a DAG with each initial metabolite at the end of the reaction, resulting in two pathway predictions for each input structure. MEANtools then outputs the full reaction network and all DAGs as CSV tables that are easy to import and browse in Cytoscape<sup>23</sup>. Lastly, pathway predictions are output as SVG image files, detailing the metabolites, reactions and genes involved, and their respective correlations.

To aid users in exploring the predictions, MEANtools provides an option to generate SVG files for each of the molecular structures predicted in previous stages, which allows users to identify and prioritize structures or reactions of interest. Upon identification, MEANtools can generate DAGs and pathway predictions rooted at any molecule selected by the user. Finally, the user can also select specific metabolites and, if available, MEANtools will generate a pathway prediction containing all metabolites specified by the user.



### 5.3.7. Flexible and user-friendly input and output

Although MEANtools is focused on the analysis of paired transcriptomic-metabolomic datasets, it can also be used with only metabolome data, or to analyse the biosynthetic potential of a specific metabolite structure. In total, MEANtools provides five tool modes according to which data the user inputs, as seen in **Table 1**. Tool modes 1 and 2 were described above and generate pathway predictions supported by the metabolome and transcriptome, with their only difference being the origin of the metabolite structure prediction: database prediction or targeted user input. Tool modes 3 and 4 generate pathway predictions only using the metabolome, and thus will not use enzyme expression-metabolite abundance correlation evidence as a restriction. However, by adding the genome annotation as input, MEANtools will add to each pathway prediction all genes that encode enzymes capable of each reaction step and restrict the output to Pfam domains encoded in the genome sequence. Finally, tool mode 5 allows the user to generate the unrestricted network of biosynthetic potential starting from a specific metabolite structure according to known enzymatic reactions.

Tool mode	User input:		
	Metabolite structures	Metabolome	Transcriptome
1	X	X	X
2		X	X
3	X	X	
4		X	
5	X		

**Table 1.** MEANtools can be used in five tool modes, according to the data used as input. Tool modes 1 and 2 generate pathway predictions supported by the metabolome and transcriptome. Tool modes 3 and 4 generate pathway predictions only using only the metabolome. Tool mode 5 allows the user to generate the biosynthetic potential starting of specific metabolite structures according to known enzymatic reactions.

## 5.4. Results and Discussion

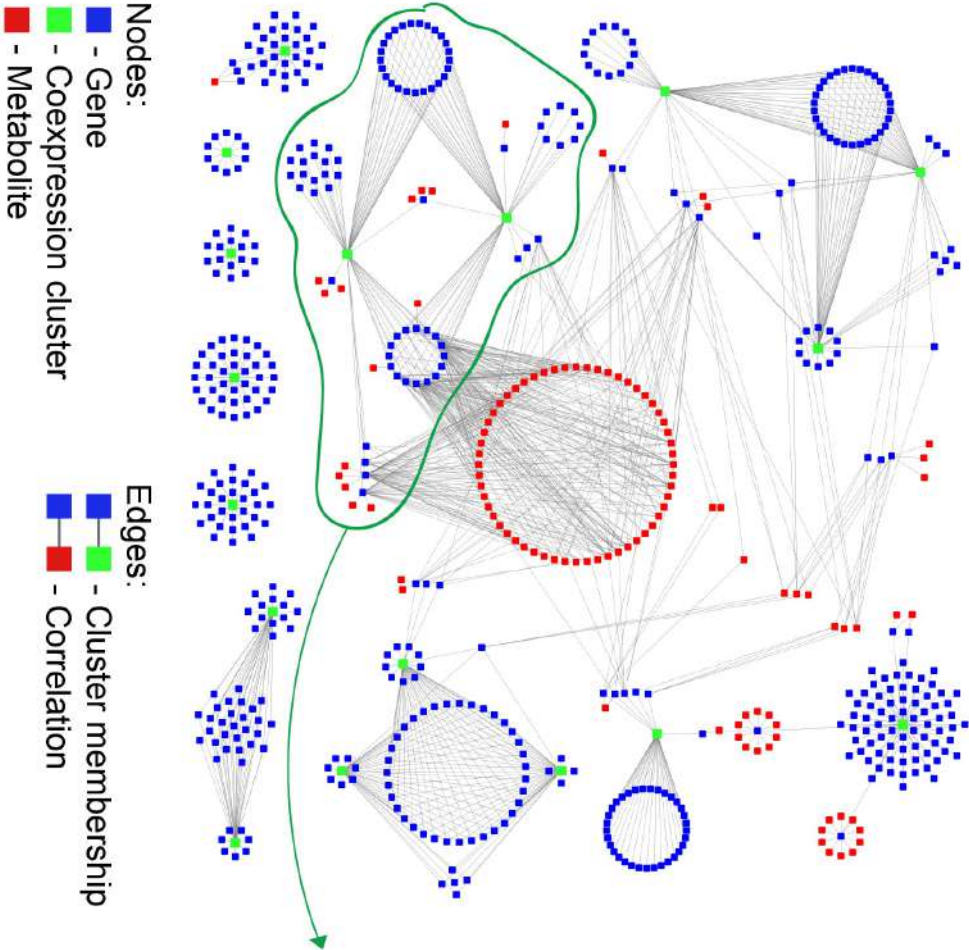
### 5.4.1. Identification of multi-omics correlation clusters help prioritize datapoints in the transcriptome and metabolome

To demonstrate MEANtools as a metabolic pathway predictions workflow, we used the paired transcriptomic-metabolomic dataset generated and described by Jeon *et al.*<sup>35</sup>. This dataset is the result of seven fungal and bacterial elicitations applied to tomato leaves, sampled over two days in triplicates, resulting in 87 samples for which gene expression data were collected and for which abundances of 11,266 mass signatures were measured.

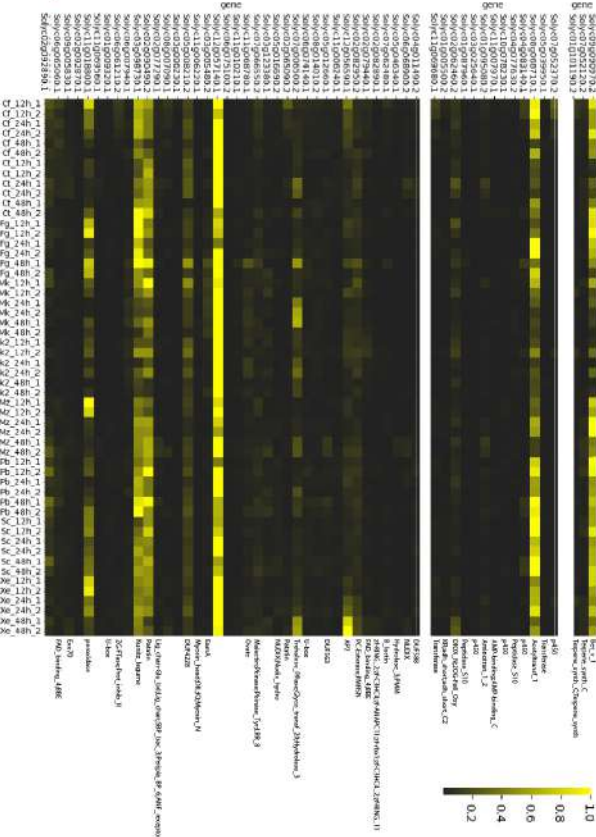
Because of the high number of mass signatures, using MEANtools in an

unsupervised manner (tool modes 2 and 4) can be time-consuming and computationally taxing, as it will produce numerous structure predictions. To reduce the number of datapoints, we used MEANtools to identify clusters of coexpressed genes that encode enzymes with known biosynthetic potential, along with the mass signatures correlated to them (see Methods). The results of this analysis module are visualized in **Fig. 4 A)** with an automatically generated heatmap describing the expression changes in the genes within the clusters identified, and **Fig. 4 B)**, which details the correlation of the coexpressed genes with mass signatures. By browsing these heatmaps and this network, users can easily prioritize genes and mass signatures of interest. An example of a gene coexpression of this process can be seen in **Fig. 4**: we identified a cluster (circled in **Fig. 4A)**) with enzyme members with potentially interesting biosynthetic potential (detailed in **Fig. 4B)**), such as terpene synthases and protein domains known to modify terpene scaffolds, such as acyltransferases and P450s<sup>36</sup>. As seen in **Fig 4. A)**, many of these genes (blue nodes) correlate with many mass signatures (red nodes), making it a group of 75 genes and 70 mass signatures that could be used as a basis for unsupervised pathway prediction using any of the tool modes of MEANtools, in order to guide further experimental analysis and characterization.

A) Transcript-metabolite correlation network



B) Gene expression heatmap

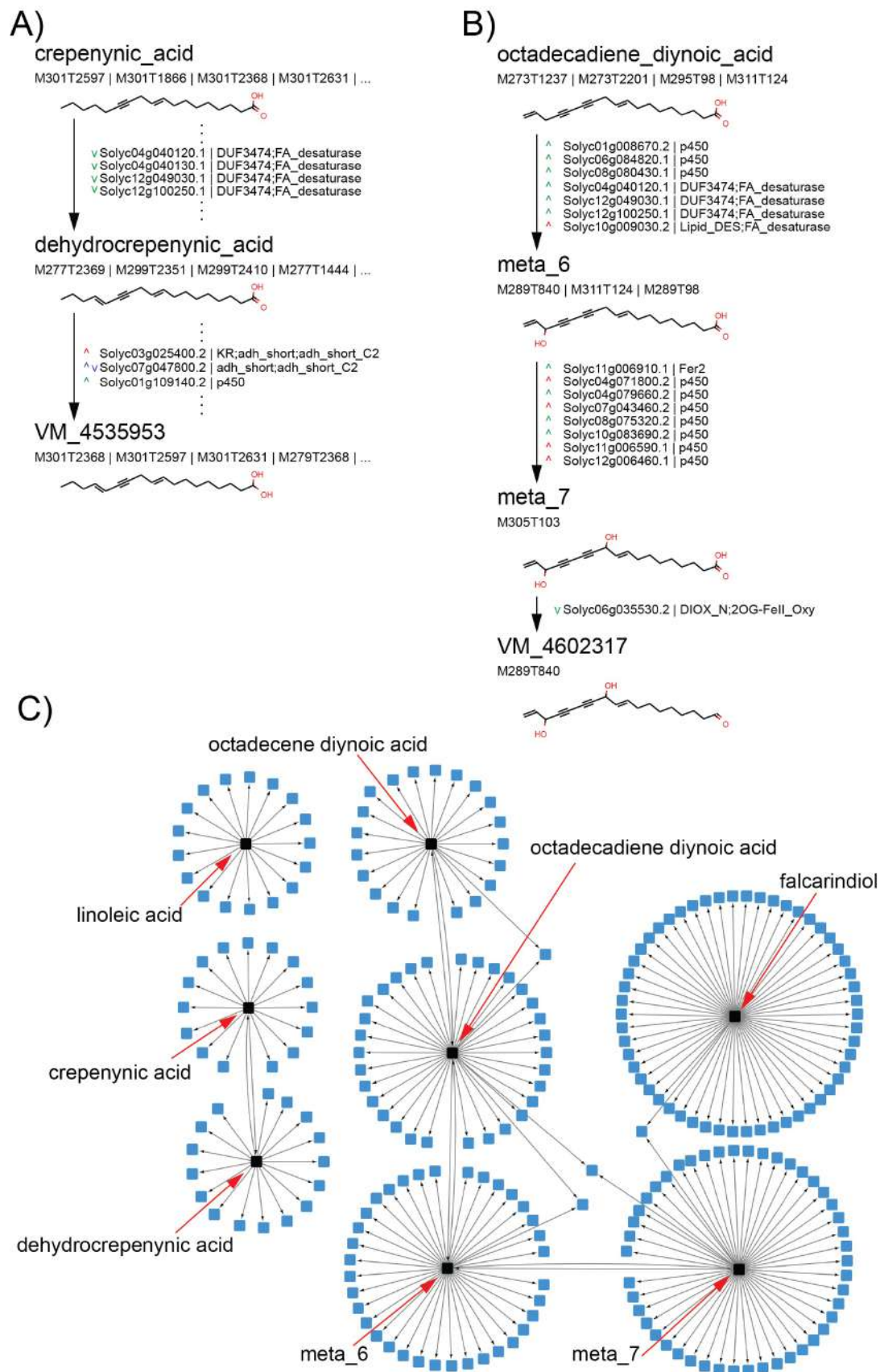


**Fig. 4.** MEANtools identifies gene expression clusters with genes that encode biosynthetic protein domains. **A)** Transcript-metabolite correlation network of genes within a biosynthetic coexpression cluster. Nodes represent genes (blue), metabolites (red) and clusters (green). **B)** MEANtools generates a gene expression heatmap for each identified biosynthetic coexpression cluster. Expression is normalized by samples (columns) and genes are sorted in three categories according to encoded protein domains: scaffold-generating enzymes at the top, followed by tailoring enzymes, and lastly the remaining genes.

### 5.4.2. MEANtools predicts parts of the proposed falcarindiol biosynthesis pathway

Jeon *et al.*'s research led to a proposed metabolic pathway for the biosynthesis of falcarindiol, the candidate enzyme classes for each reaction step, and the identification of some genes (potentially) involved in the pathway. To validate MEANtools' predictions, we identified all mass signatures that can be predicted for the eight structures in the falcarindiol pathway. Because Jeon *et al.* focused mainly on M+Na and M+H adducts, we selected only mass signature predictions corresponding to these adducts with a tolerance of 60ppm. This resulted in a total of 45 mass signatures mapped to 8 structures (**Suppl. Table 2**), which were used as the metabolite structure input for a MEANtools analysis in mode 1.

By using only experimentally validated enzyme-reaction associations ("strict" settings), and a minimum absolute Spearman correlation coefficient of 0.5, MEANtools predicts the second step of the falcarindiol biosynthesis pathway proposed by Jeon *et al.* (crepenynic acid -> dehydrocrepenynic acid), seen in **Fig. 5 A)**. Among the genes that MEANtools predicts may be involved in this reaction is Solyc12g100250, which Jeon *et al.* identified as a major desaturase in the falcarindiol pathway that was linked to this reaction using transient expression. MEANtools also predicts steps five and six of the pathway proposed by Jeon *et al.* (octadecadiene diynoic acid -> metabolite\_6 -> metabolite\_7), seen in **Fig. 5 B)**, and provides candidate genes encoding enzymes with a protein domain that has been characterized as able to perform each reaction. Lastly, by using only the structure predictions, metabolome data and genome annotations (tool mode 3), MEANtools anticipates the second step of the pathway (crepenynic acid -> dehydrocrepenynic acid) and steps four, five and six (octadecene diynoic acid -> octadecadiene diynoic acid -> metabolite\_6 -> metabolite\_7), pictured in **Fig. 5 C)** as a metabolite network. Interestingly, in this tool mode, MEANtools anticipates a step of the pathway it does not with strict settings. To further explore the predictive power of MEANtools, we repeated the analysis with "medium" and "loose" settings, and a distribution of the correlation coefficients of all predicted gene-metabolite associations for the falcarindiol pathway at each setting can be seen in **Suppl. Fig. 1**. Here, we can see MEANtools predicts 7 genes with ADH\_N and NAD\_binding\_10 Pfam domains that are, with lower confidence than the other predictions, predicted to potentially perform the dehydrogenation of octadecene diynoic acid. A table with all of the predictions can be found in **Suppl. Table 3**.



**Fig. 5.** MEANtools predicts parts of the falcarindiol pathway as proposed by Jeon *et al.*<sup>35</sup>, and the genes responsible for each enzymatic step. A “^” or “v” sign next to each gene indicates whether it is correlated to the abundance of the substrate or the product of each step: green indicates positive correlation, red indicates negative correlation, and blue indicates the gene is correlated to multiple

mass signatures, with some positively, and some negatively. **A)** MEANtools predicts 55 possible genes for the second step of faltarindiol biosynthesis (crepenynic acid -> dehydrocrepenynic acid), including 4 genes with the FA\_desaturase Pfam domain, and Solyc12g100250, the gene *Jeon et al.* identified as the desaturase that catalyses this step in the pathway. MEANtools also anticipates the possible production of another metabolite outside the faltarindiol pathway (VM\_4535953) and provides a set of candidate genes that could be involved. **B)** MEANtools predicts steps five and six of faltarindiol biosynthesis (octadecadiene diynoic acid -> metabolite\_6 -> metabolite\_7). Among the genes predicted for the fifth step are six genes with p450 or FA\_desaturase protein domains. While the cytochrome p450 enzymes are the most likely candidates for this hydroxylation reaction, the fact that MEANtools also indicates FA\_desaturase enzymes as candidates for the reaction is explained by the evidence that AlkB1 (not pictured), a known *Pseudomonas* FA\_desaturase, is capable of hydroxylation<sup>37</sup>. MEANtools also anticipates the possible production of another metabolite outside the faltarindiol pathway (here designated VM\_4602317) and identifies a redox enzyme that could potentially be involved. **C)** An overview of the biosynthetic potential of the metabolites in the faltarindiol biosynthetic pathway generated using MEANtools based on genome information only, without filtering them with expression correlation data. Each node represents a unique metabolite, and arrows link metabolites that can be transformed into one another through a known enzymatic reaction. Black nodes indicate the metabolites in the faltarindiol biosynthetic pathway. MEANtools anticipates the second step of the pathway (crepenynic acid -> dehydrocrepenynic acid) and steps four, five and six (octadecene diynoic acid -> octadecadiene diynoic acid -> metabolite\_6 -> metabolite\_7).

## 5.5. Conclusions

MEANtools can generate testable hypotheses on metabolic pathways rapidly, with little to no prior knowledge, by integrating multi-omics data. This method effectively automates the identification of key Pfam domains required for a specific reaction and allows users to tune the reaction-Pfam domain associations according to their level of confidence, or based on the taxa of origin. For this purpose, MEANtools queries RetroRules, a retrosynthesis-oriented enzymatic reactions database, showing that tools and methods within the retrosynthetic biology and synthetic pathway design fields have considerable application potential for metabolic pathway prediction and potentially NP discovery. While we present MEANtools as a metabolic pathway prediction approach, its different modes allow for diverse usage, such as for retrosynthetic pathway design by targeting specific metabolite structures without metabolome or transcriptome data (tool mode 5). To disentangle reaction networks into useful predictions, MEANtools converts them into DAGs, which allow the quick identification of linear metabolic pathways, which are presented to the user along with the metabolites, enzymes and reactions involved in them.

## 5.6. Future perspectives

Because of MEANtools' flexible and modular design, there is room for improvement in many of its individual processing steps. Annotating mass signatures with predicted structures can be improved by using MS/MS data to increase accuracy and allow validation, in a similar way as done by Network<sup>28</sup>. Converting predicted reaction networks into DAGs was the only method we used to study and present unsupervised predictions, but more complex manipulations of the network may allow for predictions better tailored for the user, such as prioritizing specific reactions or molecular substructures. Further curating the reaction-Pfam domain associations, or allowing the user better control over them, could improve the method as well: some enzyme domains may be linked to large numbers of reactions, likely to lead to false positives when the objective is to predict pathways, but could be

useful when exploring the biosynthetic potential of a structure when designing a synthetic pathway. Altogether, we present a novel computational method to predict metabolic pathways and is guided by multi-omic evidence, allowing researchers to quickly generate testable and easy-to-browse hypotheses. Furthermore, we anticipate that our work provides the basis for future work to expand the numbers of ways in which paired genomic, transcriptomic and metabolomic data can be used to analyze biosynthetic diversity.

## 5.7. Tool availability

MEANtools is fully available at <https://git.wageningenur.nl/medema-group/meantools/>

## 5.8. Supplementary Information

Supplementary figures and tables are available to download from: <https://doi.org/10.5281/zenodo.4056591>

## References

1. Fang, C., Fernie, A. R. & Luo, J. Exploring the Diversity of Plant Metabolism. *Trends Plant Sci.* **24**, 83–98 (2019).
2. Kabera, J. N., Semana, E., Mussa, A. R. & He, X. Plant Secondary Metabolites: Biosynthesis, Classification, Function and Pharmacological Properties. *J. Pharm. Pharmacol.* **2**, 377–392 (2014).
3. Owen, C., Patron, N. J., Huang, A. & Osbourn, A. Harnessing plant metabolic diversity. *Curr. Opin. Chem. Biol.* **40**, 24–30 (2017).
4. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* **43**, 155–176 (2016).
5. Medema, M. H. & Osbourn, A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Nat. Prod. Rep.* **33**, 951–62 (2016).
6. Reed, J. & Osbourn, A. Engineering terpenoid production through transient expression in *Nicotiana benthamiana*. *Plant Cell Rep.* **37**, 1431–1441 (2018).
7. Rai, A., Yamazaki, M. & Saito, K. A new era in plant functional genomics. *Curr. Opin. Syst. Biol.* **15**, 58–67 (2019).
8. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63 (2017).
9. Töpfer, N., Fuchs, L. & Aharoni, A. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Res.* **45**, 7049–7063 (2017).
10. Tohge, T. & Fernie, A. R. Co-expression and co-responses: within and beyond transcription. *Front. Plant Sci.* **3**, 248 (2012).



11. Klein, A. P. & Sattely, E. S. Two cytochromes P450 catalyze S-heterocyclizations in cabbage phytoalexin biosynthesis. *Nat. Chem. Biol.* **11**, 837–839 (2015).
12. Klein, A. P. & Sattely, E. S. Biosynthesis of cabbage phytoalexins from indole glucosinolate. *Proc. Natl. Acad. Sci.* **114**, 1910–1915 (2017).
13. Tzfadia, O. *et al.* CoExpNetViz: Comparative Co-Expression Networks Construction and Visualization Tool. *Front. Plant Sci.* **6**, 1194 (2016).
14. Ghosh, S. Triterpene structural diversification by plant cytochrome P450 enzymes. *Front. Plant Sci.* **8**, 1886 (2017).
15. Rai, A., Saito, K. & Yamazaki, M. Integrated omics analysis of specialized metabolism in medicinal plants. *Plant J.* **90**, 764–787 (2017).
16. Cavill, R., Jennen, D., Kleinjans, J. & Briedé, J. J. Transcriptomic and metabolomic data integration. *Brief. Bioinform.* bbv090- (2015). doi:10.1093/bib/bbv090
17. Wedow, J. M. *et al.* Metabolite and transcript profiling of Guinea grass (*Panicum maximum* Jacq) response to elevated [CO<sub>2</sub>] and temperature. *Metabolomics* **15**, 51 (2019).
18. Perez de Souza, L. *et al.* Multi-tissue integration of transcriptomic and specialized metabolite profiling provides tools for assessing the common bean (*Phaseolus vulgaris*) metabolome. *Plant J.* **97**, 1132–1153 (2019).
19. Urbanczyk-Wochniak, E. *et al.* Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.* **4**, 989–993 (2003).
20. Redestig, H. & Costa, I. G. Detection and interpretation of metabolite–transcript coresponses using combined profiling data. *Bioinformatics* **27**, i357–i365 (2011).
21. Duigou, T., Du Lac, M., Carbonell, P. & Faulon, J. L. Retrorules: A database of reaction rules for engineering biology. *Nucleic Acids Res.* **47**, D1229–D1235 (2019).
22. Wang, L., Dash, S., Ng, C. Y. & Maranas, C. D. A review of computational tools for design and reconstruction of metabolic pathways. *Synth. Syst. Biotechnol.* **2**, 243–252 (2017).
23. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2005).
24. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* **1697900**, 51–56 (2010).
25. Virtanen, P. *et al.* SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python. (2019).
26. Wisecaver, J. H. *et al.* A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* **29**, 944–959 (2017).
27. Moretti, S. *et al.* MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* **44**, D523–D526 (2016).
28. Beauxis, Y. & Genta-Jouve, G. Network: a web server for natural products anticipation. *Bioinformatics* 1–2 (2018). doi:10.1093/bioinformatics/bty864



29. Alcántara, R. *et al.* Rhea—a manually curated resource of biochemical reactions. *Nucleic Acids Res.* **40**, D754–D760 (2012).
30. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
31. Alborzi, S. Z., Devignes, M.-D. & Ritchie, D. W. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. *BMC Bioinformatics* **18**, 107 (2017).
32. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
33. Landrum, G. RDKit: Open-source cheminformatics. (2006). Available at: <http://www.rdkit.org>.
34. Hagberg, A. A., -Los, Ianlgov, Schult, D. A. & Swart, P. J. *Exploring Network Structure, Dynamics, and Function using NetworkX*. (2008).
35. Jeon, J. E. *et al.* A Pathogen-Responsive Gene Cluster for Highly Modified Fatty Acids in Tomato. *Cell* **180**, 176–187.e19 (2020).
36. Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P. & Osbourn, A. Triterpene Biosynthesis in Plants. *Annu. Rev. Plant Biol.* **65**, 225–257 (2014).
37. Smits, T. H. M., Witholt, B. & van Beilen, J. B. Functional characterization of genes involved in alkane oxidation by *Pseudomonas aeruginosa*. *Antonie Van Leeuwenhoek* **84**, 193–200 (2003).

# **Chapter 6**

## **General Discussion**

## 6.1. General remarks

As discussed in the introduction to this thesis, before the start of this PhD, by early 2016, plant specialized metabolism research was on the verge of a computational revolution. This brought multiple opportunities and challenges to the field, and multiple contributions to maximize the power of novel computational methods to discover, characterize or predict specialized metabolic pathways and their associated natural products in plants were reviewed. The common thread among the strategies discussed is the aim to increase confidence in associations and predictions, by extracting more knowledge or information from -omics data sources. During my PhD research, I developed several strategies towards achieving this aim.

In summary, we can classify all strategies under two main categories: single-omic and multi-omics strategies. Besides their eponymous characteristics (the usage of one or more -omics), their main difference is that of direction: single-omic methods reveal insight by looking inwards into the dataset to reveal new information, while multi-omics methods do so by looking outwards, and comparing what is learned with what can be learned from other data sources.

**Chapter 2** provides a good example of an inward-looking strategy with plantiSMASH, mainly focusing on the analysis of genomics data. The tool, however, does provide a form of multi-omics analysis through a coexpression analysis on the input transcriptome, as it pertains to the biosynthetic gene clusters (BGC) predicted in the input genome; and in the background, through the library of plant SM protein domains that we curated based on many known SM pathways in literature, that in turn undoubtedly used many other -omics sources during characterization. Nevertheless, this “external knowledge” is used differently than in multi-omics analyses: it is used to annotate the genome and generate a set of assumptions based on which a prediction can be proposed. In the case of plantiSMASH, annotations of protein domains and SM biosynthesis functions constitute the assumptions made, and the predicted BGCs are the proposition. In multi-omics analyses, the external knowledge from another -omic source is a main part of the prediction itself, such as the association between mass signatures abundances and transcripts expression made by MEANtools discussed in **chapter 5**.

plantiSMASH itself, and its framework, has the potential to be used as a tool within larger workflows. For example, we previously mentioned the characterization of avenacin<sup>1</sup>, thalianol<sup>2</sup>, marneral<sup>3</sup> and  $\alpha$ -chaconine/solanine<sup>4</sup> BGCs, and how comparative genomics provided insight about their evolutionary trajectories: the authors compared the specific BGC with the matching genomic region in other species, which allowed them to gain evolutionary insights into the BGC itself. We exploited this strategy in **chapter 3** with some differences. Rather than setting a specific BGC as our object of study like in previous research, we focused our comparative analyses around a more general type of genomic elements: genomic neighborhoods (GN) containing genes encoding oxidosqualene cyclases (OSC), cytochrome P450 (CYP) and acyltransferases (ACT), constrained only by the Brassicaceae genomes available at the time. This allowed us to generate insights into all GNs at the same time, rather than about one or two specific BGC as had been done in previous research.

As we discussed before, the need for better prioritization is a common limitation

across the plant SM research field. We tackled this limitation in **chapter 4**, where we developed CADE-HEroN, a workflow to prioritize groups of genes based on a comparative transcriptomic analysis. We continued to explore gene prioritization solutions in **chapter 5** with the integration of transcriptomic and metabolomic analyses using an enzymatic reaction library. Both projects also benefitted from modules aimed at facilitating untargeted analysis: CADE-HEroN through a comparative time-series gene expression analysis, and MEANtools through the predicted reaction networks that can be browsed in multiple formats.

The idea of a future integration of these two tools is tantalizing: predicted metabolic pathways involving genes with expression patterns conserved across species would be great candidates for experimental validation. Moreover, all of the computational tools and pipelines we developed throughout this thesis could potentially be integrated within one single workflow; in this way, we could maximally use the guilt-by-association principle and provide SM pathway predictions based on observations that come from the genomes, transcriptomes, metabolomes and phylogenetic relationships of multiple species. While this would be a seemingly large endeavor to undertake for the discovery of one or two pathways, a generic computational tool could eventually be developed for a comprehensive automated analysis across many plant species at the same time.

## 6.2. Where plants meet microorganisms

We have previously discussed how plant SM research and NP discovery has learned from the neighboring bacteria and fungi kingdoms: BGC discovery and characterization was the dominant guide in NP discovery in microorganisms before it gained popularity within the plant kingdom. In **chapter 2** we introduced plantiSMASH, a computational tool for BGC discovery in plants, based on antiSMASH which has successfully aided in the characterization of thousands of BGCs in microorganisms to date<sup>5,6</sup>. Throughout this thesis, we have discussed how using BGCs and BGC-like structures has resulted in several NP discoveries and contributed immensely to understanding plant specialized metabolism. Whether the so-called “gene cluster revolution”<sup>7</sup> crosses the border to neighboring kingdoms or not, it is safe to conclude that studying BGCs will be a mainstay of plant SM research and NP discovery going forward.

This, however, begs one logical question, what else can plant researchers learn from those working on microorganisms?

The BiG-SCAPE/CORASON framework<sup>8</sup> is a good example: after identifying BGCs through antiSMASH, BiG-SCAPE generates a similarity network among predicted BGCs and those in a database of BGCs. This similarity network is used to group BGCs into gene cluster families (GCFs), which are then queried by CORASON with a multi-locus phylogenetic analysis to reveal their evolutionary relationships. The authors demonstrated the usefulness of this strategy to guide NP discovery by mining 3,080 actinobacterial genomes, which led to the discovery of seven new detoxins<sup>8</sup>.

A tool like plantiSMASH could in the future be integrated with a phylogenetic analysis similar to the one we developed in **chapter 3** and with the BiG-SCAPE/CORASON framework that has demonstrably successful results with

microbial BGCs. This would result in a larger comprehensive multi-omics pipeline. For example, a user could input a few closely related genomes of their choice, along with a tree describing their phylogenetic relatedness; the tool would predict BGCs and GNs in all genomes, and then perform a multiple sequence analysis of the scaffold-generating enzymes identified in all clusters and an automated analysis with the BiG-SCAPE/CORASON framework. Our previous analysis suggests that, unlike in microorganisms, the subsequent GCFs would tend to group BGCs and GNs that may have evolved independently, which requires the inclusion of phylogenetic analyses per gene family, customized for plants. The framework we developed in **chapter 3** would be key for this: the new computational tool would need to differentiate well between protein subfamilies, generate ancestral state reconstructions of the GCFs and integrate the phylogenetic tree of the genomes into the analysis. The user would then receive information on the predicted BGCs and GNs, accompanied by their predicted evolutionary relationships. This would in turn allow the user to make better decisions regarding experimental validation.

Another advance in microbial NP discovery that has yet to be adapted to plants is the capacity to predict chemical structures directly from genome sequences. The substrate specificity of NRPS adenylation domains and PKS acyltransferase domains can be analyzed by a range of tools and algorithms, such as antiSMASH, to predict the chemical structures of the metabolites associated with a set of metabolic genes<sup>9</sup>. No similar strategy has been developed yet for plant genomes, and some obstacles as to why are readily apparent: even when focusing on a specific family of metabolites, like triterpenes, some tailoring enzymes like cytochrome P450s have a diverse range of catalytic activities and regioselectivities, some of them being species-specific<sup>10</sup>. While more research focused on characterizing and categorizing the functions of plant enzymes will be key to develop a similar predictive strategy in plants, a good start would lie in computationally mining the vast amount of existing literature on the subject. Some of the work we presented in **chapter 5** provides a foundation for this: MEANtools mines the RetroRules database to associate enzymatic domains to specific molecular substructures. MEANtools cannot provide a metabolic structure prediction *de novo* (and uses the transcriptome and metabolome in addition to the genome), but very often the precursors of SM pathways are metabolites from primary metabolism<sup>11</sup>; therefore, targeting a library of known metabolites in the primary metabolism of the species being queried with MEANtools could provide good set of predicted NPs. Moreover, the continuing development of databases annotating and associating metabolic enzymes, and reactions, like RetroRules<sup>12</sup>, and tools using these databases, will have a great positive impact. Altogether, chemical structure prediction based on plant SM pathway discovery is not entirely out of reach.

### 6.3. Pathway prediction and retrosynthetic analysis: two sides of the same coin

In **chapter 5** we introduced RetroRules, a database of enzymatic reactions for metabolic engineering and retrosynthesis workflows<sup>12</sup>. This database was primarily designed to aid in retrosynthetic analysis, a kind of analysis that is more closely associated with the synthetic route design and drug discovery fields than plant SM research. This strategy aims at identifying the reaction mechanisms towards a target

molecule, which in essence is the same as metabolic pathway discovery, but with a different set of constraints due to their *in vivo* or *in vitro* nature. The main difference, however, lies in the direction in which the information is processed: retrosynthetic analysis is a top-down approach where pathway prediction is bottom-up.

We can picture a reaction pathway as a pyramid. At the top of the pyramid rests the final metabolite of the reaction: a NP of interest or a synthetic drug for which we intend to discover the reaction mechanisms towards their production. The bricks in the pyramid constitute the myriad of mechanisms that permit the synthesis of the final metabolite, including the main reactions in the pathway, the production of necessary cofactors and catalysts, and the conditions leading to them.

Retrosynthetic analysis explores this pyramid from the top down: it aims at identifying the reaction mechanism of a target molecule by converting the chemical structure into simpler structures, and then further simplifying those and repeating the process until molecules that can be made by known reaction mechanisms are reached. Researchers start from the top of the pyramid (the final metabolite of the pathway) and lay bricks below it until they design an architecture that they are comfortable with. This is distinct from most strategies in SM and NP discovery, which more generally follow a bottom-up approach: individual analyses and observations are integrated piece by piece until a prediction about the system can be confidently reached, such as generating insights about the SM system or the prediction of a pathway; by laying bricks one a time, they may be laid on top of each other too, in the same manner as new analyses rely on the knowledge from previous ones to reach an adequate interpretation. This strategy is well exemplified in the previously discussed studies that led to the characterization of the biosynthetic pathways for podophyllotoxin in mayapple<sup>13</sup>, 4-hydroxyindole-3-carbonyl nitrile (4-OH-ICN) in *Arabidopsis thaliana*<sup>14</sup> and noscapine in opium poppy<sup>15</sup> in which observations from transcriptomic experiments were used to appropriately design metabolomic experiments that the authors could use to “map the pyramid” of the metabolic pathways.

In reality, researchers often have a final or intermediate metabolite, class of metabolites, or a genomic feature of interest in hand when designing and interpreting their -omics experiments and analyses. This makes the *de facto* strategies more akin to a middle-out approach: exploring the pyramid with a mix of bottom-up and top-down approaches.

This is an advantageous aspect in SM research: knowledge of at least the metabolite class of a NP can give important information regarding its structural characteristics, which in turn provides information about certain enzymes required for their biosynthesis, a “guide” to explore the pyramid. A common example of this is querying for scaffold-generating enzymes when targeting a specific class of metabolites, but some information about tailoring enzymes involved in a pathway can also be deduced: Jeon *et al.* used characteristics of the fatty acid backbone of faltarindiol to guide the identification of the fatty acid desaturases involved in the pathway<sup>16</sup>. Another genomic feature that can serve as the pyramid’s guide are BGCs: the characterization of the  $\alpha$ -chaconine/solanine BGC in potato was guided by the characteristics of the  $\alpha$ -tomatine BGC in tomato<sup>4</sup>.

In **chapter 5** we presented MEANtools, which leveraged the RetroRules database for retrosynthetic analysis to generate testable hypotheses about biosynthetic pathways and facilitating plant SM research and NP discovery. We

developed a cheminformatic framework based on the generic reaction databases that RetroRules consolidates to predict the reactions involved in plant metabolic pathways. The integration of this database is based on the assumption that knowledge and tools from the synthetic route design/drug discovery fields can be transferred successfully to the field of plant biology (and vice versa), which has the potential of benefitting both. MEANtools shows us how this integration can benefit plant biology by using reaction rules derived from characterized enzymes across the tree of life to predict possible metabolic pathways in plants, and as researchers discover and characterize new metabolic pathways, retrosynthetic databases like RetroRules can grow, in turn benefitting the synthetic route design/drug discovery fields.

Another example of a set of retrosynthetic analysis tools with potential to guide plant metabolic pathway discovery are molecular similarity algorithms. These algorithms are used to query novel molecules in large small-molecule databases to identify similar bioactive molecules with properties that may be shared by the molecule of interest. This method is called virtual screening, and has been used successfully for a long time in drug discovery<sup>17</sup>. There exists an overlap between this method and metabolic pathway discovery. The metabolites within a metabolic pathway often constitute a group of metabolites with high degree of molecular similarity; a method able to identify similar molecules could potentially identify metabolites within the same pathway. This idea has been explored in the past: KCF-S is a method to describe and compare molecules, which has been designed with the specific purpose of identifying pairs of metabolites that can possibly be converted into each other through an enzymatic reaction<sup>18</sup>. The authors demonstrated this strategy by querying a database of molecules, and their algorithm grouped molecules that were less structurally diverse than clusters generated by other tools, with many metabolite pairs being one enzymatic reaction away from each other<sup>18</sup>; recently, they released their method as a Python package<sup>19</sup>. Above, we discussed how *in silico* prediction of *de novo* metabolic structures awaits in the future of plant SM research, and molecular similarity algorithms could be key to link *de novo* structure predictions with known precursors from a metabolite database. Integrating these methods with a cheminformatic framework capable of predicting reaction pathways and associated substrates based on molecular substructures (such as the one we developed for MEANtools in **chapter 5**) will effectively create a fully top-down pathway discovery strategy within the plant SM and NP discovery fields.

## 6.4. A cornucopia of data, methods, and tools: Sophie's choice?

Chief among the contributions of computational genomics to plants NP discovery and SM research is the myriad of distinct tools and methods that researchers can take advantage of for any -omics dataset. In this thesis we discussed a few of these methods to guide metabolic pathway discovery: genome sequences can be queried for BGCs<sup>20,21</sup> and/or compared with other sequences and BGCs<sup>2-4</sup>; transcriptomes can be used to identify coexpressed genes<sup>13,14,22,23</sup>, or to generate coexpression networks and identify coexpressed modules<sup>24-26</sup>, all of which can be compared with other transcriptomes<sup>4,27,28</sup>; also, metabolomes can be analyzed along with the previously mentioned workflows either sequentially<sup>13-15</sup> or in an integrated manner<sup>29-</sup>

<sup>31</sup> to generate and validate hypotheses.

Moreover, computational genomics has facilitated the usage of previously acquired knowledge: the continuously growing databases of genomes, transcriptomes, metabolomes, characterized pathways, enzymatic reactions and BGCs that can be used to reinforce observations in -omics analyses performed in newer data. While new “wet lab” -omics experiments are always needed to acquire new knowledge, uncovering how it is related to previously acquired knowledge in literature is a powerful strategy to produce discoveries with higher confidence than otherwise. This is evidenced in the main chapters of this thesis: a connecting thread of this research has been the development of methods that integrate publicly available resources for the analyses.

When using any particular -omics dataset to elucidate a metabolic pathway or to identify a natural product, it is undeniably cheaper to run many computational analyses, or compare the dataset to literature, than it is to run more “wet lab” experiments. This generates interest in running as many computational analyses as possible, but in a world where time and other extra-scientific limitations are at play, the large number of tools, methods, datasets and databases in literature can be a problem for researchers: it can be taxing to select the appropriate set of analyses and tools for a specific research project. While many authors will say their tool, method or workflow is the best available for what they do, as supported by a variety of qualifications or metrics, the truth may lie in the opposite end: instead of selecting the *best* set of analyses and tools for the job, a researcher could default to running *all* analyses that are possible and that can contribute to elucidating a targeted objective.

First, this would demand a unified repository of tools and methods with proper categorization and documentation. Something similar exists in Galaxy, a web-based scientific biomedical analysis platform for genomic, proteomic and metabolomic analyses, bringing together >5,500 tools<sup>32</sup>. A similar platform for NP discovery could host or package the computational tools mentioned in this thesis and more, along with a user-friendly GUI and documentation to allow easier access to bioinformatic tools to plant biologists. Developing this platform would certainly be a large undertaking, but extensive literature already covers the advantages, disadvantages, and best way of using many of these methods and tools<sup>26,31,33,34</sup>. Furthermore, the majority of plant NP discovery computational tools have similar or interchangeable file formats; for example, the majority of transcriptomic analyses discussed in this thesis start with counts or F/RPKM values as input, and all methods we discussed involving protein prediction do so through the PFAM library<sup>35</sup> or EC numbers, two annotation types (sequence/structure and function) that were recently shown to be highly associated<sup>36</sup>.

It may be possible that some analyses and tools will not contribute towards the objective a researcher has set, or the data they have available. For example, a BGC prediction analysis may not have much to offer when using a low-quality genome assembly with short contigs, or when targeting an un-clustered pathway, and differential expression analyses are not well suited to identify meaningful patterns in pairs of transcriptome samples that have the majority of genes differentially expressed. In other words, a simple database or repository of tools would not entirely solve the problem of *how* to select the appropriate set of analyses and tools, although it would make it easier. To properly solve this problem, a big characteristic



of any prospective unified repository of plants NP discovery tools would require a good degree of automation; the database could effectively make this selection process happen *in silico* instead.

Many automated tools discussed, and developed, through this thesis demonstrate a part of how such platform could be automated: plantiSMASH and MEANtools integrate different automated analyses according to the input (and their formats) and allow advanced users to further customize parameters of specific steps or how the output is formatted. A unified repository of tools could work similarly, starting with a preliminary format and quality control analysis of the inputs to determine the best set of tools according to predetermined rulesets or recipes. Moreover, as more users use the database, they could save and make public their recipes/rulesets for other researchers to use. For the less advanced users, this could be presented to the user as suggestions or pre-checked defaults. Altogether, this could result in a platform in which biologists input their raw multi-omics data, select a set of analyses and their specified parameters and/or targets, click run, and receive within a day a large number of high-confidence predictions ready for human interpretation.

An example of a platform that aims to provide such a unified repository framework can be found in KBase, a knowledgebase that consolidates a number of tools to analyze -omics data and predict biological functions<sup>37</sup>. This knowledgebase not only solves the problem of discoverability, accessibility and utilization of many tools, but also integrates data from genome, metabolite and reaction databases, facilitating their selection process too. Moreover, KBase bolsters a user-friendly GUI, making the process more straightforward for biologists with less expertise in bioinformatic analyses. This could be a good foundation whence a plant NP discovery equivalent could be constructed; although KBase focuses mainly on microbial genome analysis, it does include plant genomes in its knowledgebase too. Furthermore, many of the tools developed through this PhD, and discussed as possible follow-ups within this thesis, provide a good initial set of tools to be added and integrated into recipes and rulesets uniquely aimed at plant genomes.

The development of this platform would be a large undertaking, but it would bring unmeasurable benefits to the field: it would make the plethora of existing bioinformatic tools, methods and databases more easily discoverable and accessible to researchers, especially if framed within a user-friendly interface. As we have discussed above, the computational aspect of the plants NP discovery and SM research fields is still rather young, making this a good moment to start the development of this kind of unified platform. As the impact of computational genomics in the field continues to grow, the number of tools, methods and databases will multiply: not only will this make the development of such platform harder, but by then a one-stop-platform for all NP discovery needs might become a necessity rather than a boon.

## 6.5. Closing remarks

It is undeniable that computational genomics has provided a cornucopia of opportunities for SM plant research in the shape of computational tools and methods, each of them with their own set of challenges and limitations. More importantly, each new method and tool has brought new promises too, of the myriad of new ideas and integrated methods that could be implemented in the future; each

generation increasing the speed at which advancements are made. With this thesis I contributed to this process. The plant SM research field is richer today than it was four years ago, and it will likely be exponentially richer four years from now.

## References

1. Qi, X. *et al.* A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 8233–8 (2004).
2. Field, B. & Osbourn, A. E. Metabolic Diversification--Independent Assembly of Operon-Like Gene Clusters in Different Plants. *Science (80-. )*. **320**, 543–547 (2008).
3. Field, B. *et al.* Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 16116–21 (2011).
4. Itkin, M. *et al.* Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* **341**, 175–9 (2013).
5. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
6. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, (2019).
7. Jensen, P. R. Natural Products and the Gene Cluster Revolution. *Trends in Microbiology* **24**, 968–977 (2016).
8. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
9. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).
10. Ghosh, S. Triterpene structural diversification by plant cytochrome P450 enzymes. *Front. Plant Sci.* **8**, 1886 (2017).
11. Hartmann, T. From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry* **68**, 2831–2846 (2007).
12. Duigou, T., Du Lac, M., Carbonell, P. & Faulon, J. L. Retrorules: A database of reaction rules for engineering biology. *Nucleic Acids Res.* **47**, D1229–D1235 (2019).
13. Lau, W. & Sattely, E. S. Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. *Science* **349**, 1224–8 (2015).
14. Rajniak, J., Barco, B., Clay, N. K. & Sattely, E. S. A new cyanogenic metabolite in *Arabidopsis* required for inducible pathogen defence. *Nature* **525**, 376–379 (2015).
15. Chen, X. & Facchini, P. J. Short-chain dehydrogenase/reductase catalyzing the final step of noscapine biosynthesis is localized to laticifers in opium poppy. *Plant J.* **77**, 173–184 (2014).

16. Jeon, J. E. *et al.* A Pathogen-Responsive Gene Cluster for Highly Modified Fatty Acids in Tomato. *Cell* **180**, 176–187.e19 (2020).
17. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015).
18. Kotera, M. *et al.* KCF-S: KEGG Chemical Function and Substructure for improved interpretability and prediction in chemical bioinformatics. *BMC Syst. Biol.* **7 Suppl 6**, S2 (2013).
19. Sato, M., Suetake, H., Kotera, M. & Sato, M. KCF-Convoy: efficient Python package to convert KEGG Chemical Function and Substructure fingerprints. *bioRxiv* (2018). doi:10.1101/452383
20. Schlöpfer, P. *et al.* Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. *Plant Physiol.* **173**, 2041–2059 (2017).
21. Töpfer, N., Fuchs, L. & Aharoni, A. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Res.* **45**, 7049–7063 (2017).
22. Horan, K. *et al.* Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol.* **147**, 41–57 (2008).
23. Sohrabi, R. *et al.* In Planta Variation of Volatile Biosynthesis: An Alternative Biosynthetic Route to the Formation of the Pathogen-Induced Volatile Homoterpene DMNT via Triterpene Degradation in Arabidopsis Roots. *Plant Cell* **27**, 874–890 (2015).
24. Mao, L., Van Hemert, J. L., Dash, S. & Dickerson, J. A. Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics* **10**, 346 (2009).
25. Wisecaver, J. H. *et al.* A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* **29**, 944–959 (2017).
26. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* **9**, 1090 (2018).
27. Tzfadia, O. *et al.* CoExpNetViz: Comparative Co-Expression Networks Construction and Visualization Tool. *Front. Plant Sci.* **6**, 1194 (2016).
28. Netotea, S., Sundell, D., Street, N. R. & Hvidsten, T. R. ComPIEx: Conservation and divergence of co-expression networks in *A. thaliana*, *Populus* and *O. sativa*. *BMC Genomics* **15**, 1–17 (2014).
29. Urbanczyk-Wochniak, E. *et al.* Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.* **4**, 989–993 (2003).
30. Rischer, H. *et al.* Gene-to-metabolite networks for terpenoid indole alkaloid biosynthesis in *Catharanthus roseus* cells. *Proc. Natl. Acad. Sci.* **103**, 5614–5619 (2006).
31. Cavill, R., Jennen, D., Kleinjans, J. & Briedé, J. J. Transcriptomic and metabolomic data integration. *Brief. Bioinform.* bbv090- (2015). doi:10.1093/bib/bbv090
32. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).

33. Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M. & Ligterink, W. Learning from Co-expression Networks: Possibilities and Challenges. *Front. Plant Sci.* **7**, 444 (2016).
34. Li, Y., Pearl, S. A. & Jackson, S. A. Gene Networks in Plant Biology: Approaches in Reconstruction and Analysis. *Trends Plant Sci.* **20**, 664–675 (2015).
35. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
36. Alborzi, S. Z., Devignes, M.-D. & Ritchie, D. W. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. *BMC Bioinformatics* **18**, 107 (2017).
37. Arkin, A. P. *et al.* KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* **36**, 566–569 (2018).

## Summary

The field of plant natural product (NP) discovery has changed substantially since the isolation of morphine from opium poppy in 1806. The largest of these changes came in the last decade with the integration of computational genomics into the field, which resulted in the development of a plethora of computational methods that aid in the discovery of new plant NPs and the biosynthetic pathways, metabolites and enzymes associated with them. As more computational methods continue to be developed, and more plant genomes are sequenced, the NP discovery field is ripe with obstacles and opportunities uniquely suited to multi-omics solutions.

In **chapter 1**, we explore the history of the plant NP discovery field and highlight some of the strategies used throughout. We also discuss the state of the field and some of the obstacles and opportunities that computational genomics introduced, mainly regarding the identification of plant biosynthetic gene clusters (BGCs), analysis of coexpression networks and multi-omics integration.

In **chapter 2**, we present a computational tool for the automated identification, annotation and expression analysis of plant BGCs: plantiSMASH. Here, we show how BGC identification can guide plant NP discovery by mining all publicly available chromosome-level plant genome assemblies and recovering all BGCs experimentally characterized at the time, along with a wide range of putative novel ones. Furthermore, we develop a coexpression analysis module, which facilitates the integration of a transcriptomic analysis to any predicted BGC. In **chapter 3**, we leverage BGCs and BGC-like genomic structures to study the evolution of triterpene biosynthetic pathways in plants. Here, we queried 13 Brassicaceae plant genomes to identify all oxidosqualene cyclases (OSC), and the genomic regions flanking them. We use these regions to perform a series of phylogenetic analyses to compare the evolution of biosynthetic genes with that of the associated Brassicaceae species and uncover the most likely evolutionary events that led to the assembly and diversification of Brassicaceae triterpene BGCs. In **chapter 4**, we introduce CADE-HEroN: a workflow for comparative analysis of the coexpression networks of multiple species to guide the discovery of plant specialized metabolic (SM) pathways. We use this workflow to study the SM pathways associated with the phosphate starvation response in *Arabidopsis thaliana*, tomato and rice. This resulted in the identification of many genes of known and unknown function that have a conserved behavior under phosphate starvation across the three species. In **chapter 5**, we describe the development of an integrative multi-omics approach for plant SM pathway prediction: MEANtools. This computational tool integrates data from paired transcriptomic-metabolomic datasets to predict potential metabolic pathways, and the reactions, metabolites and genes involved in them. In this manner, MEANtools can help scientists generate testable hypotheses about biosynthetic pathways, which we showcase by using our pipeline with a recently published paired transcriptomic-metabolomic dataset.

We conclude this thesis in **chapter 6**, discussing how the plant NP discovery and SM research fields have benefitted from cross-pollination with adjacent fields, and how to better take advantage of this and the many other opportunities discussed throughout the thesis.

## Envoi

Firstly, thank you for reading these words; it is not always a delight to write, but it is always to be read.

The work of this thesis is only the result of a collaboration of many people, and many are cited throughout this text, but some are deserving of further acknowledgments. None of this could have been possible without *Marnix Medema* and *Dick de Ridder* who gave me the opportunity to take on this journey, for which I cannot thank you enough. Later as supervisors, your guidance was crucial and your support abundant; all successes I could claim from this research, are thanks to you both.

Thanks to everyone in the WUR Bioinformatics lab too, you are all too many to list, but you were all very helpful and enjoyable to be around.

I would like to thank too *Harro Bouwmeester* and *Eric Schranz*, with whom I had very enjoyable collaborative research experiences in Wageningen, and *Anne Osbourn* and *Elizabeth Sattely*, who provided me with very memorable research opportunities away from Wageningen.

Lastly, thanks to my friends and family, without whom sanity would have run out long ago.

Hunt dragons, slay beasts, sleep when you are dead.

-H

The research described in this thesis was financially supported by NWO.