



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/189024/>

Version: Accepted Version

---

**Article:**

Zhan, Chuansong, Shen, Shuangqian, Yang, Chenkun et al. (2022) Plant metabolic gene clusters in the multi-omics era. TRENDS IN PLANT SCIENCE. ISSN: 1360-1385

<https://doi.org/10.1016/j.tplants.2022.03.002>

---

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

[Click here to view linked References](#)

# 1 **Plant metabolic gene cluster in the multi-omics era**

2 Chuansong Zhan<sup>1,2\*</sup>, Shuangqian Shen<sup>1,3\*</sup>, Chenkun Yang<sup>3</sup>, Zhenhua Liu<sup>4</sup>, Alisdair R.

3 Fernie<sup>5,6</sup>, Ian A. Graham<sup>7</sup> and Jie Luo<sup>1,2,&</sup>

4 <sup>1</sup> College of Tropical Crops, Hainan University, Haikou 570228, China.

5 <sup>2</sup> Sanya Nanfan Research Institute of Hainan University, Hainan Yazhou Bay Seed  
6 Laboratory, Sanya 572025, China.

7 <sup>3</sup> National Key Laboratory of Crop Genetic Improvement and National Center of Plant Gene  
8 Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China.

9 <sup>4</sup> Joint Center for Single Cell Biology, School of Agriculture and Biology, Shanghai Jiao  
10 Tong University, Shanghai 200240, China.

11 <sup>5</sup> Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476  
12 Potsdam-Golm, Germany.

13 <sup>6</sup> Center of Plant Systems Biology and Biotechnology, 4000 Plovdiv, Bulgaria.

14 <sup>7</sup> Center for Novel Agricultural Products, Department of Biology, University of York, York,  
15 UK.

16 \* These authors contributed equally to this article

17 Correspondence: [jie.luo@hainanu.edu.cn](mailto:jie.luo@hainanu.edu.cn) (J. Luo)

18 **Keywords:** Metabolic gene cluster, natural variation, GWAS,

19

## 20 **Abstract**

21 Secondary metabolism in plants gives rise to a vast array of small molecule natural  
22 products. The discovery of operon-like gene clusters in plants has provided a new  
23 perspective on the evolution of specialized metabolism and the opportunity to rapidly  
24 advance the metabolic engineering of natural product production. Here we review  
25 historical aspects of the study of **plant metabolic gene clusters** as well as general  
26 strategies for identifying plant metabolic gene clusters in the **multi-omics** era. We also  
27 emphasize the exploration of their **natural variation** and evolution, as well as new  
28 strategies for the prospecting of plant metabolic gene clusters and deeper understanding  
29 as to how their structure influences their function.

## 30 **Plant secondary metabolism and metabolic gene clusters**

31 More than 200,000 primary and secondary metabolites have been identified in plants,  
32 with the majority categorized as secondary (or specialized) metabolites [1-4]. Generally,  
33 primary metabolites such as amino acids, sugar, and nucleic acids are essential for  
34 growth and development and are ubiquitously produced by most cell types of all plant  
35 species. Different classes of secondary metabolites including terpenoids,  
36 phenylpropanoids and alkaloids assist in survival across ecological niches where they  
37 provide protection against biotic and abiotic stress [5, 6] and assist in sexual  
38 reproduction and dispersal. These metabolites also provide humankind with a huge  
39 catalog of compounds with pharmacological and other industrial properties [5, 7-10].  
40 Synthesis of such an array of secondary metabolites has been underpinned by the  
41 evolution of gene families with hundreds of members encoding enzymes such as  
42 cytochrome P450 (P450) oxidases and methyl transferases that are responsible for  
43 building the structural complexity of secondary metabolites [11-13].

44 Understanding the biosynthetic pathways, regulatory mechanisms and transport  
45 processes responsible for production of secondary metabolites will be essential to fully  
46 explore the potential of this treasure trove of natural products for the benefit of human

47 society and the environment [14-18]. In contrast to the situation in prokaryotes, genes  
48 involved in plant **secondary metabolism** are generally randomly distributed across the  
49 plant genome, which typically means that genes encoding the enzymes of a biochemical  
50 pathway have to be discovered one step at a time. Even today with the advent of gene  
51 sequence information [19-26], discovery of the full complement of genes responsible  
52 for biochemical pathways underlying plant specialized metabolism remains a  
53 considerable challenge.

54 Plant metabolic gene clusters can be defined as being composed of two or more  
55 non-homologous and closely linked genes that encode enzymes from the same  
56 biosynthetic pathway [27]. Moreover, the genes within the clusters are usually  
57 coordinately regulated [28]. These features render plant metabolic gene clusters a  
58 valuable tool for the functional characterization of biosynthetic pathways that they are  
59 associated with [29]. Meanwhile, the development of multi-omics approaches  
60 (combining two or more of genomics, transcriptomics, metabolomics or epigenomics)  
61 offers new strategies and opportunities to discover natural product pathways. For  
62 example, a metabolite-based genome-wide association study (mGWAS) was performed  
63 and successfully identified a subspecies-specific diterpene (5,10-diketo-casbene) gene  
64 cluster and a hydroxycinnamoyl-tyramine gene cluster in rice (*Oryza sativa* L.) [30, 31].  
65 Recently, a pathogen-responsive gene cluster that is responsible for biosynthetic  
66 faltarindiol was identified by using a combination of metabolomics and RNA  
67 sequencing analysis in tomato (*Solanum lycopersicum*) [32]. Here we review recent  
68 advances in the field of plant metabolic gene cluster discovery. For this purpose, we  
69 provide a historical framework prior to discussing emerging strategies in the post-  
70 genomic era as well as emphasizing the exploration of natural variation in plant  
71 metabolic gene clusters using forward genetic approaches such as genome-wide  
72 association studies. Finally, we present a perspective for a more comprehensive  
73 understanding of plant metabolic gene clusters.

## 74 **Historical Aspects and General Strategies for Plant Metabolic Gene Cluster** 75 **Identification**

76 In 1960, the term operon was coined by *Francois Jacob* and *Jacques Monod* who  
77 discovered and characterized the lac operon in *Escherichia coli* (*E. coli*). The lac operon  
78 confers the ability to grow on lactose as sole carbon source (Figure 1). The discoveries  
79 of the operon structure provided *Francois Jacob*, *Andre Michel Lwoff*, and *Jacques*  
80 *Monod* the opportunity to receive the 1965 Nobel Prize for Physiology and Medicine.  
81 Originally, operons were thought to be a uniquely microbial phenomenon. Indeed,  
82 about 50% of the genes in prokaryotes are clumped together as gene cluster [33, 34].  
83 The first operon-like cluster in plants was identified in 1997 (Figure 1 and Table 1) [35].  
84 After that, more than 30 metabolic gene clusters from distant phylogenetic clades across  
85 the plant kingdom have been reported (Figure 1 and Table 1) [36-39]. For instance, the  
86 first diterpene gene cluster in *Oryza sativa* and triterpene gene cluster in arabidopsis  
87 (*Arabidopsis thaliana*) was identified in 2004 and 2008, respectively (Figure 1 and  
88 Table 1) [40, 41]. With the exception of a few metabolic gene clusters, most plant  
89 metabolic gene clusters were identified with a time lag following the publication of the  
90 plant genomes to which they belong (Figure 1). It appears that while metabolic gene  
91 clusters are generally found in the genomes of all plant species, they remain the  
92 exception rather than the rule in describing the organization of genes associated with  
93 metabolic pathways unlike the situation in microbes.

94 Generally, two main strategies have been used for gene identifying and  
95 characterizing: forward genetic (from phenotype to gene) strategies and the reverse  
96 genetics (from gene to phenotype) strategies (Figure 2). In the identification of plant  
97 gene cluster, forward genetic strategies are one of the powerful strategies that have been  
98 verified multiple times. Genome-wide association studies (GWAS), map-based cloning,  
99 and bulked-segregant analyses are three common forward genetic approaches used for  
100 causal gene(s) identification and characterization. Among these, the genome-wide  
101 association studies (GWAS) and map-based cloning demonstrated great utility in the

102 identification of metabolic gene clusters in plants (Figure 2A and 2B). For example,  
103 five clustered genes encoding enzymes required for the biosynthesis of cucurbitacins  
104 were successfully identified through genome-wide association analysis using variation  
105 maps of 115 different cucumber varieties [42] (Figure 2A). In tomato, a natural  
106 population consisting of 600 lines were studied by using systematic metabolome and  
107 genomic analysis strategies. A potential gene cluster on chromosome 10 containing a  
108 P450 oxidoreductase, an acyltransferase, an acetyl-CoA dehydrogenase and an UDP-  
109 glucosyltransferase, in addition to the previously identified gene cluster on  
110 chromosome 7 was uncovered [43, 44]. Further study showed that this locus was  
111 responsible for the natural variation of the toxic anti-nutritional factor  $\alpha$ -solanine in  
112 tomato [44]. Recently, Zhan and Shen *et al.* performed metabolic GWAS (mGWAS) in  
113 monocot rice populations and revealed three brand-new gene clusters: diterpenoid gene  
114 cluster on chromosome 7, *DGC7*; hydroxycinnamoyl tyramine gene cluster, a *HT* gene  
115 cluster; a hydroxycinnamoyl putrescine gene cluster and a *HP* gene cluster [30, 31, 45].  
116 They also demonstrated that all end-products synthesized by these three gene clusters  
117 can confer disease resistance in rice [30, 31, 38, 45, 46]. Comparison of the  
118 transcriptome of stems and capsules from opium poppy varieties HN1, HM1 and HT1  
119 (which producing high levels of noscapine, high morphine, and high thebaine,  
120 respectively) revealed 10 co-expressed genes specifically existed in HN1 [47]. By  
121 screening the HN1 Bacterial Artificial Chromosome (BAC) library and analyzing the  
122 F2 population, a 10 gene metabolic gene cluster specific to HN1 was found, which is  
123 responsible for the production of noscapine [47]. Through using seventeen putative  
124 mutants that were crossed into an inbred line of maize, *Bx1/Bx1*. The first plant  
125 metabolic gene cluster – the DIMBOA gene cluster was identified [35]. A few years  
126 later, Qi *et al.* used the recombinant inbred lines that derived from *A. strigose* C13815  
127 x *A. wiestii* C11994 to successfully map the gene, *AsbAS1*, which responsible for  
128 biosynthesize the *beta*-amyrin which is the skeleton of avenacins [48]. Further studies  
129 uncovered that this gene was part of an antimicrobial triterpenoid gene cluster at the  
130 locus, containing a total of 12 genes [49, 50]. In barley, *Cer-Cqu* was mapped to a

131 discrete location on chromosome arm 2HS using population mapping method.  
132 Combined with BAC library data and sequencing, three candidate genes (*Cer-C*, *Cer-*  
133 *Q* and *Cer-U*) were identified [51]. These three genes were distributed over a 101 Kb  
134 chromosomal interval and were highly co-expressed in leaf sheath tissue, confirming  
135 that they formed a metabolic gene cluster that catalyzed the biosynthesis of  $\beta$ -diketone.  
136 Taken together these studies clearly illustrate the power of forward genetics in the  
137 identification of plant metabolic gene clusters.

138 The advent of multiple omics technologies has offered us new strategies for the  
139 identification of the plant metabolic gene clusters (Figure 2C). The bio-informatic  
140 computational pipeline strategy is a special reverse strategy worthy to mention here.  
141 For example, algorithms such as PlantiSMASH and PhytoClust have been developed  
142 and applied to predict the secondary metabolic gene clusters in plants [52, 53]. Both  
143 tools adopt accurate Hidden Markov Model profiles (pHMMS) to judge different  
144 biosynthesis genes and predict candidate gene clusters in combination with genome  
145 locations. Generally, most of the retrieval rules of these computer algorithms are based  
146 on the typical combination of the “signature enzymes” and “tailoring enzymes” (Figure  
147 2C). For example, the terpene synthase (TPS) and cytochrome P450 enzyme (CYP450)  
148 are the main types of enzymes that are involved in terpenoid metabolic pathways. Of  
149 these, the terpene synthases are considered to act as the “signature enzymes”, whereas  
150 cytochrome P450 enzymes are “tailoring enzymes”. Based on these basic rules, Töpfer  
151 Nadine *et al.*, searched TPS/CYP450 combinations across multiple plant genomes and  
152 identified both known and novel terpene gene clusters [53]. It is known that this  
153 approach is able to render the identification of plant metabolic gene clusters more facile  
154 and the accuracy of such predictions will be increased through integration of genome  
155 and transcriptome data (Figure 2D and 2E) [52, 53]. Availability of user-friendly  
156 interfaces for such algorithms in combination with developments in the fields of next-  
157 generation sequencing, analytical chemistry, synthetic biology, and systems biology  
158 will considerably accelerate the speed of discoveries of gene clusters in diverse non-  
159 model plants. Specifically, the combination of different strategies, such as different

160 omics, will provide more clues to narrow the gap between phenotypic diversity and  
161 genetic variation. For example, through parallel mGWAS and a gene-based association  
162 analysis using metabolic, genetic, and phenotypic data, new candidate gene clusters for  
163 the natural variation in content of tyramine were identified [31].

## 164 **Cluster Constituents and Organization**

165

### 166 *Constituents of Metabolic Gene Cluster*

167 Plant metabolic gene clusters, reported so far, range from ~ 35 kb to several hundred  
168 kb in size and consist of three to 15 genes [49, 54]. In addition to the signature enzyme  
169 that initiate the metabolic pathway, the metabolic gene cluster usually contains various  
170 modifying enzymes, such as: cytochrome oxidases, glycosyltransferases,  
171 acyltransferases, methyltransferases, dioxygenases, carboxylesterases,  
172 dehydrogenase/reductases, transaminases, *etc.* The detailed characteristics and  
173 examples of plant metabolic gene clusters components mentioned above that have been  
174 summarized in previous reviews [29, 55]. In general, the richer the variety and number  
175 of modifying enzymes in a gene cluster, the larger the gene cluster needs to be.

176 Interestingly, recent studies have shown that in addition to genes encoding enzymes  
177 other genes, encoding transporters and cofactor synthases, can also be associated with  
178 plant metabolic clusters [31, 56]. Darbani Behrooz *et al.* have reported that the  
179 cyanogenic glucoside gene cluster consists of four different genes: *CYP79D3*,  
180 *CYP79D4*, *CYP736A2* and *UGT85K3*. Another study uncovered that the gene  
181 *SbMATE2*, which encodes a transporter that is required for the transport of non-  
182 endogenous cyanogenic glucosides is located within the same cluster in *Sorghum*  
183 *bicolor* [56]. A cofactor is a non-protein substance which is required for a protein to be  
184 catalytically active. In the hydroxycinnamoyl tyramine (HT) gene cluster, besides the  
185 biosynthetic genes (tyrosine decarboxylase, OsTyDC1; tyrosine decarboxylase,  
186 OsTyDC1; acyl transferases, OsTHT1 and OsTHT2), a pyridoxal 5-phosphate (PLP)

187 cofactor synthetase OsPDX3 is also embedded in these gene clusters [31]. PLP is a type  
188 of cofactor that is required for the catalysis of enzymes such as transaminases,  
189 isomerases, decarboxylases, racemases, aldolases, deaminases, and aminotransferases.  
190 *In vitro* enzyme analyses demonstrated that the HT gene cluster member OsPDX3 acted  
191 as a cofactor donor for the PLP-dependent tyrosine decarboxylase OsTyDC1,  
192 suggesting that the cofactor synthase was indirectly necessary for the production of the  
193 end products (Figure 3A). Such step-by-step characterization not only enriches our  
194 understanding of the scope of their function but also broadens our understanding of the  
195 enzyme catalog of gene cluster components. Indeed, the presence of these novel  
196 members suggest that bioinformatic tools will need to be refined in order to  
197 accommodate such members of plant metabolic gene clusters.

198

#### 199 *Organization of Metabolic Gene Cluster*

200 Various plant gene clusters have been described and most fall into the compact gene  
201 cluster type. Here we will also discuss the type of super metabolic gene clusters (Figure  
202 3).

203 The thalianol gene cluster in arabidopsis is the smallest plant compact gene cluster  
204 with only 35~38 Kb [41]. Other clusters belonging to this structural form are the  
205 faltarindiol gene cluster in tomato [32], the hydroxycinnamoyl-tyramine gene cluster  
206 [31], the 5,10-diketo-casbene gene cluster in rice and the dhurrin gene cluster (Figure  
207 3A) [30, 56, 57]. Interestingly, combined with the analysis of metabolite biosynthesis  
208 pathway, it was found that the distribution order of the compact gene cluster members  
209 was roughly collinear with the reaction steps, revealing a new pattern for plant  
210 metabolic gene cluster assembly [49]. For example, the genes within the noscapine gene  
211 cluster in opium poppy could be roughly divided into three reaction sequence modules.  
212 The early module contains *CYP82Y1*, *PSMT3*, *CYP719A21* and *PSMT1*; the middle  
213 module contains *CYP82X1*, *CYP82X2*, *PSAT1* and *PSMT2*; the late module contains  
214 *PSSDR1* and *PSCXE1*, which exactly corresponds to sequentially genome organization  
215 in poppy [47]. Moreover, the organization of the avenacin cluster components appears

216 to be broadly collinear with the order of the biosynthetic pathway on oat chromosome  
217 1. Specifically, the gene encoding the first step *bASI/Sad1* is located closest to the  
218 telomere and the late pathway genes including *CYP72A476*, *UGT91G16* and *TG1/Sad3*  
219 that are also required for avenacin biosynthesis are more distal to the telomere [49]. The  
220 authors proposed that placing *UGT91G16* and *TG1/ SAD3* genes farthest from the  
221 telomeres may be a gene arrangement strategy to mitigate the occurrence of toxin  
222 accumulation caused on telomere deletions. These examples show that collinearity of  
223 gene order and biosynthetic pathway reactions is quite common in compact pathways  
224 identified to date and may provide some insight into how gene clusters have evolved in  
225 response to natural selection.

226 One possibility is that compact gene clusters offer a selective advantage in their  
227 co-expression, co-inheritance or in the construction of metabolons. Metabolon is a  
228 complex formed by the non-covalently bound interactions of enzymes that promote  
229 substrate channeling between successive steps in metabolic pathways [58, 59]. This  
230 organization type may promote the efficient delivery of intermediates and prevents  
231 unnecessary metabolic crossovers to maintain metabolic flexibility. Until now, with  
232 neither the glycolytic metabolon nor the TCA cycle metabolon [60, 61] forming plant  
233 metabolic gene clusters, the dhurrin gene cluster is the only metabolic gene cluster that  
234 is able to form metabolons [62, 63]. *UGT85B1* interacts with *CYP79A1* and *CYP71E1*  
235 to form a channel complex that guides the rapid flow of metabolic intermediates to  
236 dhurrin biosynthesis [57]. Gene fusions that contain multiple domains can be  
237 considered as a tighter physical association of a metabolon. One such example is  
238 *STORR* [(*S*) - to (*R*)-reticuline] a fusion of a cytochrome P450 and oxidoreductase genes  
239 that resulted in the key gateway reaction essential for morphine biosynthesis in opium  
240 poppy [64]. Interestingly, *STORR* is a member of the 15 gene BIA cluster in opium  
241 poppy [65]. Collectively, this modular assembly implies that for some metabolites  
242 plants may have experienced selective pressures that has resulted in not only gene  
243 clustering but specific ordering of genes within a cluster. Whether or not this relates to  
244 metabolon function remains to be determined but the evidence to date suggests such

245 ordering is the exception rather than the rule.

246 Loose gene clusters are defined as closely adjacent core gene cluster components  
247 and distantly distributed metabolic pathway initiation enzymes, modification enzymes  
248 or regulators, indicative of a fragmented pathway (Figure 3B). For example, the  
249 majority of the genes that comprise the cucurbitenol gene cluster including the gene for  
250 the oxidosqualene cyclase, three different types of CYP genes and an acyltransferase  
251 gene are clustered on chromosome 6. However, the other four CYP genes that from the  
252 gene cluster that is also required for cucurbitacins biosynthesis are located on  
253 chromosome 3 and chromosome 1, respectively (Figure 3B). In addition, two  
254 transcription factors which can regulate the synthesis of cucurbitacin C are located on  
255 different chromosomes to the core gene cluster [66]. Similarly, both the  $\alpha$ -solanine  
256 biosynthetic gene cluster of tomato and the  $\alpha$ -solanine synthetic gene cluster of potato  
257 are typical loose gene clusters, whose components are mainly distributed on  
258 chromosome 7 and chromosome 12, with the major structural genes being collinear  
259 between tomato and potato (Figure 3B) [43]. This phenomenon implies that loosely  
260 arranged gene clusters among close- homology species may have experienced common  
261 evolutionary trajectories. In summary, the loose gene cluster is a broader definition of  
262 metabolic gene cluster that may reflect an intermediate form of dynamic clustering gene  
263 cluster components from related pathways. Whether these components will continue to  
264 operate remotely or, in future, form more tightly packed clusters remain to be seen.

265 Super gene clusters may be defined as different metabolic gene clusters coming  
266 together as hotspots in the genome. Recently, studies have shown that the gene cluster  
267 responsible for the synthesis of middle-chain acyl sugars in tomato is composed of  
268 tricyclic specific *Sl-AACS* (acyl-CoA synthase) and *Sl-AECH* (ethyl CoA hydase)  
269 genes, which are closely arranged on chromosome 7 (Figure 3C). Interestingly, the  
270 organization of these genes, along with the sterol alkaloid gene cluster, form such a  
271 "super metabolic gene cluster" in tomato [43, 67]. Tomato steroid alkaloids and  
272 acylsugars both play defensive roles in plants, but are structurally distinct and stored in  
273 different tissues. Co-localization of these gene clusters may confer a selective

274 advantage through an additive or synergistic effect of numerous defensive metabolites.  
275 Similarly, another recent study based on the obtained high-quality complete genome  
276 information, using plantiSMASH algorithm analysis and cluster density score revealed  
277 that the terminal 100 Mb region of chromosome 1 in *A. strigosa* genome is a gene  
278 cluster hotspot that contains a total of 19 putative gene clusters, of which 17 clusters  
279 include at least three co-expressed genes, where the avenacin gene cluster is located  
280 [49]. This poses the question why so many different gene clusters are grouped in the  
281 same locations, such as the sub-telomeric regions of eukaryotic genomes? Answering  
282 this question will require that we continue mining the metabolic gene clusters from a  
283 wider range of plant species and analyzing the effects of gene clusters in both  
284 evolutionary and ecological contexts.

#### 285 **Classes of Metabolites synthesized by gene clusters**

286 Clusters of non-homologous genes responsible for the biosynthesis of diverse classes  
287 of specialized metabolites have been reported in arabidopsis, rice (*Oryza sativa*) and a  
288 range of other plant species (Figure 4).

289

##### 290 *Oxylipins*

291 Falcarindiol (FAD, FaDOH, (3R,8S)-Falcarindiol), a cytotoxic and anti-inflammatory  
292 polyacetylenic oxylipin, present in many edible crops such as tomatoes, carrots and  
293 celery, exhibits antifungal, anti-bacterial, antimutagenic and anticancer activities, and  
294 it could be potentially used as a food additive (Figure 4) [32, 68-70]. The genes  
295 (ACET1a, *Solyc12g100250*; ACET1b, *Solyc12g100270*), which encode a desaturase  
296 and a decarboxylase respectively, have been proved to form a falcarindiol gene cluster  
297 (Figure 4) [32].

298

##### 299 *Terpenoids*

300 Terpenoids are the most structurally diverse group of plant metabolites and more than  
301 half the known metabolic gene clusters are associated with pathways for terpene

302 biosynthesis. A noteworthy example is the diterpene gene clusters in rice. Rice can  
303 produce large quantities of labdane-related (which includes the ubiquitous gibberellins)  
304 and casbene-type diterpenoids. The former includes momilactones A&B [71-74],  
305 phytocassanes A-E [75, 76], oryzalexins A-F [77-80], oryzalexin S [81] and the latter  
306 include 5,10-diketo-casbene [30]. Most of these specialized metabolites are produced  
307 by biosynthetic pathways encoded by metabolic gene clusters and additionally exhibit  
308 antimicrobial properties [30, 82]. A recent study revealed that the labdane-related  
309 diterpenoid in rice not only play important roles in rice disease resistance and act as  
310 important allelochemicals, but may also act as a regulatory switch that triggers stomatal  
311 closure [83-85]. Results suggest that CPS2 and/or CPS4 knockout lines exhibit  
312 significantly increased susceptibility to drought [85]. While casbene-type diterpenoids  
313 have only so far been reported in rice from among the Poaceae [86, 87], they are  
314 widespread in the Euphorbiaceae family of plants where they are recognized for their  
315 pharmacological activities [88-91]. These diterpenoids are produced by gene clusters  
316 that are evolutionary conserved across the Euphorbiaceae [92, 93]. Most interestingly,  
317 casbene synthesizing enzymes have evolved independently in the Poaceae and  
318 Euphorbiaceae but both have adopted a strategy of forming gene clusters for production  
319 of the same diterpenoid class of molecules [30].

320 Other terpenoid compounds associated with metabolic clusters include thalianol  
321 [41], arabidiol [94], tirucalladienol [95] and marneral [96] in *Arabidopsis thaliana*,  
322 avenacin [48] in *Avena strigose*, kauralexins [97] and zealexins [97] in maize,  
323 cucurbitacin C [42, 66] in cucumber, cucurbitacin B [66] in melon and cucurbitacin E  
324 [66] in watermelon, 20-hydroxy-betulinic acid in *Lotus japonicus* and monoterpenes in  
325 *Solanum lycopersicum* [9, 97, 98]. The triterpene gene clusters were reported to play  
326 important roles in modulating the *Arabidopsis thaliana* root microbiota [95]; disk  
327 assays for antifungal activity revealed that Avenacin A-1 is an antifungal triterpenoid  
328 [99]; Cucurbitacin C is associated with the distinctive taste of cucumber and confers  
329 bitterness on the entire plant [42]. Similarly, 20-hydroxy-betulinic acid may be involved  
330 in the process of nodulation [100], however, the functions of arabidopsis marneral,

331 maize kauralexin and zealexin and the *S. lycopersium* monoterpenes are at present less  
332 clear.

333

#### 334 *Phenylpropanoids*

335 Phenylpropanoids are large, structurally diverse, and widely distributed compounds  
336 [101] and to date only two metabolic gene clusters have been shown to be associated  
337 with these compounds [31, 45]. A combination of metabolite-based genome-wide  
338 association studies (mGWAS), biochemical validation and co-expression data  
339 identified gene clusters associated with biosynthesis of the aromatic  
340 hydroxycinnamoyl-tyramine [31] and aliphatic hydroxycinnamoyl-putrescine (Figure  
341 5) [45] phenolamines in rice. Further pathogen incubation assays with transgenic  
342 material demonstrated that both aromatic and aliphatic phenolamines contribute to  
343 enhanced disease resistance to *Magnaporthe oryzae* (*M. oryzae*). In addition, the  
344 aromatic hydroxycinnamoyl-tyramine also displayed broad-spectrum disease  
345 resistance to bacterial blight (Figure 5). Together, these results indicate that the  
346 phenomenon of gene clustering also extends to the biosynthesis of phenylpropanoid  
347 pathway derivatives [31, 45]. Similarly, in this respect is the recent extension of the  
348 flavonol-phenylacyltransferase (*FPT*) cluster in a recent study examining the evolution  
349 of high light responses suggest clustering is involved in some steps of phenylpropanoid  
350 biosynthesis [39].

351

#### 352 *Benzoxazinoids*

353 A further set of widely distributed compounds – the Benzoxazinoids (Bxs), are a class  
354 of specialized metabolites that were discovered in the 1950's in cereals [102].  
355 Benzoxazines have been shown to be involved in a range of biological processes, such  
356 as defense against pathogens and resistance to insects [103, 104]. 2,4-dihydroxy-7-  
357 methoxy-1,4-benzoxazin-3-one (DIMBOA) is the key defensive compound in maize  
358 (*Zea mays*) (Figure 4). As a representative Bxs, DIMBOA biosynthesis has been  
359 reported to be mediated by a metabolic gene cluster [35]. The complete biosynthetic

360 pathway involves nine enzymes (*Bx1* to *Bx9*) which act sequentially in the synthesis of  
361 DIMBOA-glucoside from indole-3-glycerol phosphate. In the beginning, the 2,4-  
362 dihydroxy-1,4-benzoxazin-3-one (DIBOA) gene cluster was defined as a group of five  
363 genes (*Bx 1-5*). However, further experiments revealed that there are four additional  
364 genes (*Bx6-Bx9*) that are required for biosynthesis of DIMBOA [105-107]. Interestingly,  
365 this cluster is split in other plants of the Poaceae [108, 109]. For example, the cluster  
366 genes are split in two parts in wheat. One part (*Bx3*, *Bx4* and *Bx5*) of them is located on  
367 the short arm of chromosome 5 (A-, B- and D-genome), another part (an additional *Bx3*  
368 copy) was detected on the long arm of chromosome 5B [107]. Similar to the metabolites  
369 biosynthesized by already characterized plant metabolic gene clusters, both DIBOA and  
370 BIMBOA can confer the pathogen resistance and also contribute to defense against  
371 herbivores [109].

372

### 373 *Alkaloids*

374 Alkaloids are a class of basic nitrogen containing natural products and in plants are best  
375 known for their pharmacological activities. In *Papaver somniferum* (opium poppy) a  
376 cluster of 10 genes encode enzymes for production of the antitussive and anticancer  
377 compound noscapine which is a member of the phthalideisoquinoline subclass of  
378 benzyloisoquinoline alkaloids (BIAs; 118). Assembly of the opium poppy genome led  
379 to the discovery that the noscapine gene cluster is part of a larger 15 gene cluster that  
380 also encodes five enzymes involved in the pathway leading to production of the  
381 morphinan class of BIAs which include the well-known analgesic painkillers codeine  
382 and morphine (Figure 4) [47, 65, 110, 111].

383 Steroidal glycoalkaloids (SGAs) in species of the *Solanaceae* can act as  
384 antinutritional alkaloids [112]. Comparative analysis between potato and tomato has  
385 revealed an array of ten genes encoding enzymes of SGAs biosynthesis [43]. Six of  
386 these genes are located in an adjacent region of chromosome 7, whereas two others are  
387 on chromosome 12 [43].

388

389 *Cyanogenic glucosides*

390 Over 2,600 plant species, including a number of cereals (i.e. barley, *Hordeum vulgare*;  
391 rye, *Secale cereal*; oat, *Avena sativa*; wheat, *Triticum aestivum*; sorghum, *Sorghum*  
392 *bicolor*; sugar cane, *Saccharum officinarum*; millet, *Setaria italica*; maize, *Zea mays*  
393 and rice, *Oryza sativa*), have been confirmed to contain cyanogenic glycosides (CGs)  
394 (Figure 4) [113]. Up to now, about 60 kinds of cyanogenic glycosides have been found  
395 [62]. Interestingly, three different kinds of amino acids (L-valine, L-isoleucine and L-  
396 tyrosine) are involved as precursors, and the genes that are responsible for their  
397 biosynthesis are also clustered [62]. The CYP79D3 gene in *Lotus japonicus* encodes a  
398 cytochrome P450 enzyme that is responsible for the first step in cyanogenic glucoside  
399 biosynthesis. Meanwhile, the other two genes (*CYP736A2* and *UGT85K3*) which are  
400 located around the *CYP79D3*, together with *CYP79D3* constitute the entire pathway for  
401 cyanogenic glucoside biosynthesis [62]. Interestingly, the gene of SbMATE2 in  
402 *Sorghum bicolor* that encode a transporter is also located in the cluster and is co-  
403 expressed with the other biosynthesis genes [56]. Evidence suggests that these CGs may  
404 play an important role in survival against pathogens or herbivores [114]. Earlier reports  
405 suggest that CGs can act as a kind of herbivore deterrents to protect the *Arabidopsis*  
406 *thaliana* and *Sorghum bicolor* [113, 115]. Another study found a relationship between  
407 Fusarium wilt resistance in flax and HCN release in roots [116].

408

409 *Other metabolites*

410 The  $\beta$ -diketones are polyketides that are also encoded by a gene cluster (Figure 4) [51].  
411 As the main components of leaf surface wax, the  $\beta$ -diketones protect against pathogens  
412 and pests [51]. Recently, a tomato gene cluster on chromosome 7 that is involved in  
413 acyl-sugar accumulation has been identified. This cluster co-localizes with the steroidal  
414 alkaloid gene cluster [65, 67, 117]. Interestingly, both acylsugars and alkaloids are  
415 active defensive compounds in plants.

## 416 **Regulation**

417 It has been shown that plant metabolic cluster genes are intended to be co-expressed or  
418 accordingly co-regulated. In general, the spatiotemporal expression patterns of gene  
419 clustering components are closely related to the accumulation patterns of metabolites.  
420 Acylsugars are mainly found in the glandular trichomes of *Solanaceae* [67]; the  
421 noscapine and pro-morphinan genes of the BIA gene cluster are coordinately regulated  
422 with noscapine and morphinan accumulation in the stems and capsules of opium poppy  
423 [47]; avenacin preferentially accumulated in oat root tips [48, 49]. These studies  
424 demonstrate that expression of genes in metabolic clusters is consistent with the tissue  
425 specific accumulation of the corresponding metabolites. While the discovery of  
426 metabolic gene clusters in plants has advanced our understanding of the related  
427 metabolic pathways, we are only beginning to understand the relevance of gene  
428 expression for gene cluster formation. In the following section, we summarize current  
429 knowledge of the regulatory mechanism for metabolic gene clusters starting from  
430 transcription factor to chromatin modifications.

431

### 432 *Transcriptional regulation*

433 Not surprisingly transcription factors play a role in regulation of genes that are clustered.  
434 Momilactone A&B, phytocassane A-E, oryzalide A-C and oryzalexin A-F are  
435 diterpenoids, and their biosynthesis is closely related to two classical diterpene gene  
436 clusters in rice. A basic leucine zipper (bZIP) family transcription factor OsTGAP1 was  
437 reported to be involved in regulating the synthesis of diterpenes in rice, it was found  
438 that this could cooperatively but indirectly regulate the transcript level of the  
439 diterpenoid gene cluster components [28]. Indeed, a couple of homologous basic helix-  
440 loop-helix transcription factors controls expression of the cucurbitacin clusters in  
441 cucumber, melon, and watermelon. Individual members of the group mediate diverse  
442 fruit-, leaf-, and root-specific cluster expression patterns [66].

443 Different from the idea of co-regulation, some transcription factors reported to

444 regulate certain components of gene cluster specifically. For example, GAME9, an AP2  
445 family transcription factor, regulates the transcription of  $\alpha$ -solanine genes cluster  
446 components GMAE4&7 by binds to another transcription factor, MYC2 in *Solanaceae*.  
447 Notably, recently reports demonstrate that *GLYCOALKALOID METABOLISM 9*  
448 (*GAME9*) is the transcription factor which regulates the biosynthesis of SGAs in potato  
449 and tomato [118]. Transformation analysis of tomato and potato showed that expression  
450 of genes associated with SGAs and the upstream mevalonate pathway are altered in  
451 GAME9 knockdown and overexpression plants [118]. Similarly, the bZIP transcription  
452 factor OsAPIP5, a negative regulator of cell death, directly binds the  
453 hydroxycinnamoyl-putrescine gene cluster component *OsPHT4* promoter, repressing  
454 its transcription [45]. Together, these cases suggest that the transcriptional regulation of  
455 plant metabolic gene clusters may operate under mechanisms that we have not yet fully  
456 explored. Further studies of transcription factor regulation cases are needed to gain  
457 deeper insight into such mechanisms.

458

#### 459 *The epigenetic regulation*

460 Chromatin modification plays an important role in the regulation of gene clusters in  
461 plants. For instance, the chromatin mark of histone H3 lysine 27 trimethylation  
462 (H3K27me3) is associated with repression of cluster expression. On the contrary, the  
463 histone variant H2A.Z marks are associated with activation of cluster expression. As  
464 reported, two clusters in *A. thaliana* are associated with chromatin decondensation.  
465 These clustered pathways (thalianol and marneral clusters) are characterized by  
466 chromatin signatures of trimethylation of histone H3 lysine 27 (H3K27me3) [119, 120].  
467 The expression levels of the thalianol and marneral cluster genes were altered in the  
468 CURLY LEAF (CLF) and PICKLE (PKL) mutants and these changes were restricted  
469 to the clusters and did not extend to the genes that directly flank the clusters [119, 120].  
470 Besides, another exciting finding concerning the chromatin regulation of these two gene  
471 clusters is that they have also been positively regulated by the SWR1 chromatin  
472 remodeling complex [119, 121, 122]. Further study revealed that ARP6 is indispensable

473 for the incorporation of H2A.Z into nucleosomes and its mutant can alter the expression  
474 of all four genes of the cluster [119]. Another interesting story of epigenetic regulation  
475 of plant gene clusters is a histone demethylase JMJ705 that can directly regulate genes  
476 from *DGC7* (a rice diterpenoid gene cluster) via methyl jasmonate-mediated epigenetic  
477 control [30]. Further research uncovered that this gene cluster is implicated in rice  
478 disease resistance [30, 46].

## 479 **Natural Variation and Evolution**

480 The genetic linkage of enzyme-coding genes in plant metabolic gene clusters confer to  
481 them some features of coinheritance [29]. However, recent research revealed that this  
482 phenomenon is only suit for the mature or fixed clusters [92]. Zhan *et al.* report the  
483 identification of one terpene synthase (*OsTPS28*) and two cytochrome P450 oxidases  
484 (*OsCYP71Z2* and *CYP71Z21*) form a metabolic gene cluster in rice [30]. The pan-  
485 genome data of *DGC7* demonstrated that the intact *DGC7* is highly enriched in the  
486 *japonica* varieties (102/109) compared to the *indica* varieties (13/313) (Figure 6).  
487 Moreover, the results of *Fst* and  $\pi$  studies further revealed that the *DGC7* was located  
488 in the sweep region. These results suggested that the *DGC7* was subject to selection  
489 during the domestication in *japonica* while not in *indica* or in the wild rice ancestor *O.*  
490 *rufipogon* [30]. Similarly, recent research uncovered that the natural variation of  
491 chromosomal inversion exists in the triterpene gene cluster in *Arabidopsis thaliana* [41,  
492 123]. This natural selection shuffles the distant genes into the thalianol cluster thereby  
493 rendering it compact.

494 Apart from the structural variation, single nucleotide polymorphisms (SNPs) and  
495 small indels are also an important part of natural variations and have been identified in  
496 several different plant metabolic gene clusters. For instance, Shen *et al.*, suggest that  
497 the coordinated transcription of *OsTyDC1* and *OsTHT1* are influenced by natural  
498 variation and this may be a reason for the combination of genes for favorable traits [31].  
499 Genomic co-linear analysis of wild and cultivated rice species shows that due to lack  
500 of the *OsTyDC1* homologs, the *Oryza punctata* (BB genome lineage), *Oryza*

501 *brachyantha* (FF genome lineage) have not formed the HT gene cluster. However, this  
502 cluster is conserved in the AA genome lineage. Unlike the HT gene cluster, the  
503 acylsugar gene cluster is missing or incomplete in most *Solanaceae* family species [67].  
504 Another example is the steroidal glycoalkaloid gene cluster in tomato. A natural  
505 variation of a *Solyc10g085230* introduces a premature stop codon to this gene and this  
506 variation can reduce the steroidal glycoalkaloid content during ripening [44, 65, 123-  
507 127].

## 508 **Concluding Remarks and Future Perspectives**

509 DIMBOA is the first reported plant metabolic gene cluster, identified 24 years ago [35].  
510 At that time no plant genomes had been published and scientists used a range of  
511 molecular biology cloning methods to identify genes associated with specific proteins.  
512 Although great achievements were made with these laborious approaches, they were  
513 low-throughput and focused on specific enzyme activities. The first reported complete  
514 sequence of a plant genome was that of *Arabidopsis thaliana* in 2000 [128]. This  
515 landmark event greatly accelerated the process of functional annotation of plant genes.  
516 The resulting gain in genome-level information sparked a rapid development period for  
517 research on plant metabolic gene clusters. During the last decade, the advent of next-  
518 generation sequencing and the development of multi-omics technologies greatly  
519 improved our ability to identify and dissect metabolic gene clusters in plants. Many  
520 aspects of the research of plant metabolic gene clusters have been considerably  
521 expanded in this period, providing insights on biosynthetic genes and regulatory genes  
522 [42], transcriptional regulation and epigenetic regulation [30, 119], and secondary  
523 metabolism and primary metabolism [32]. However, the current rate of discovery of  
524 plant metabolic gene clusters suggests that our catalog is far from complete.  
525 Furthermore, in addition to existing tools such as genomics, transcriptomics,  
526 metabolomics, epigenomics, proteomics, phenomics, next-generation sequencing and  
527 advanced bioinformatics an ever-increasing arsenal of tools is being used to crack the  
528 mysteries of plant metabolic gene clusters (Figure 7) [129-135]. Applying the advances

529 in artificial intelligence (AI) are also worth considering. One of the primary means in  
530 AI is deep learning which has been applied already in different fields related to plant  
531 science. For instance, to improve the accuracy of protein 3D structure prediction [136].  
532 In the foreseeable future, we believe that AI will also play an important role in the  
533 research of plant metabolic gene cluster.

534         Apart from the enzymes that are responsible for the synthesis of compounds, co-  
535 enzymes can also be an important part of plant metabolic gene clusters [31]. This hints  
536 to the possibility that other kinds of genes or proteins, for example, transcription factors,  
537 may also be located within gene clusters. In addition to the general structure of  
538 individual gene clusters, super gene clusters represent a very compelling area for future  
539 research. Given their characteristics, representing a combination of different metabolic  
540 traits, it is worth thinking about why these combinations were selected during plant  
541 genome evolution and which set of circumstances may have led to this. There are still  
542 many mysteries embedded in plant genomes. The integration of association analysis  
543 technology, including GWAS, rapid and efficient plant transformation [137-141], and  
544 epigenetic and synthetic biology technologies, should render the analysis [137, 142-  
545 146], discovery, and utilization of plant metabolic gene clusters more efficient as well  
546 as allowing us deeper understanding of the mechanisms underlying their structure,  
547 formation and utility. Of particular note in this respect is mGWAS which has proven a  
548 highly effective manner of identifying plant metabolic gene clusters. Indeed, many of  
549 the plant metabolic gene clusters reported in these studies were not present in the now-  
550 defunct plant metabolic gene cluster databases such as Planti-SMASH and PhytoClust.  
551 A second advantage of this approach is that it highlights only physiologically relevant  
552 gene clusters, i.e. those whose variance controls the genetic architecture of the  
553 accumulation of the pathway end-product, thereby ensuring the biological relevance of  
554 their assemblies. As such, expansion of the scope of mGWAS to encompass a broader  
555 range of plant species will likely prove instrumental in the identification and genetic  
556 dissection of plant metabolic gene clusters in the next decades (see also outstanding  
557 questions).

558

## 559 **Acknowledgments**

560 Research in our laboratories was supported by the Hainan Major Science and  
561 Technology Project (No. ZDKJ202002), the Hainan Academician Innovation Platform  
562 HD-YSZX-202003 and HD-YSZX-202004, the National Natural Science Foundation  
563 of China (no. 32100318), China Postdoctoral Science Foundation (2021TQ0093), the  
564 Hainan Yazhou Bay Seed Laboratory (B21Y10904), the Hainan University Startup  
565 Fund (KYQD(ZR)1866), and Hainan Provincial Natural Science Foundation of China  
566 (321QN184).

## 567 **References**

- 568 1. Sumner, L.W. et al. (2015) Modern plant metabolomics: advanced natural product gene discoveries,  
569 improved technologies, and future prospects. *Nat. Prod. Rep.* 32 (2), 212-229.
- 570 2. Fernie, A.R. and Tohge, T. (2017) The genetics of plant metabolism. *Annu. Rev. Genet.* 51 (1), 287-  
571 310.
- 572 3. Lacchini, E. and Goossens, A. (2020) Combinatorial control of plant specialized metabolism:  
573 mechanisms, functions, and consequences. *Annu. Rev. Cell Dev. Biol.* 36, 291-313.
- 574 4. Dixon, R.A. and Strack, D. (2003) Phytochemistry meets genome analysis, and beyond.  
575 *Phytochemistry* 62 (6), 815-6.
- 576 5. Venegas-Molina, J. et al. (2021) Why and how to dig into plant metabolite–protein interactions. *Trends*  
577 *Plant Sci.* 26, 472-483.
- 578 6. Boutanaev, A.M. et al. (2015) Investigation of terpene diversification across multiple sequenced plant  
579 genomes. *Proc. Natl. Acad. Sci. U. S. A.* 112 (1), E81-E88.
- 580 7. Jacobowitz, J.R. and Weng, J.-K. (2020) Exploring uncharted territories of plant specialized  
581 metabolism in the postgenomic era. *Annul. Rev. Plant Biol.* 71 (1), 631-658.
- 582 8. Gülck, T. and Møller, B.L. (2020) Phytocannabinoids: origins and biosynthesis. *Trends Plant Sci.* 25  
583 (10), 985-1004.
- 584 9. Schmelz, E.A. and Tumlinson, J.H. (2011) Identity, regulation, and activity of inducible diterpenoid  
585 phytoalexins in maize. *Proc. Natl. Acad. Sci. U. S. A.* 108 (13), 5455-60.
- 586 10. Luo, D. et al. (2016) Oxidation and cyclization of casbene in the biosynthesis of Euphorbia factors  
587 from mature seeds of Euphorbia lathyris L. *Proc. Natl. Acad. Sci. U. S. A.* 113 (34), E5082-E5089.
- 588 11. Erb, M. and Kliebenstein, D.J. (2020) Plant secondary metabolites as defenses, regulators, and  
589 primary metabolites: the blurred functional trichotomy. *Plant Physiol.* 184 (1), 39-52.
- 590 12. Chen, F. et al. (2011) The family of terpene synthases in plants: a mid-size family of genes for  
591 specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* 66 (1), 212-229.

- 592 13. Tian, X. et al. (2018) Characterization of gossypol biosynthetic pathway. Proc. Natl. Acad. Sci. U. S.  
593 A. 115 (23), E5410-E5418.
- 594 14. Yang, D. et al. (2014) Transcriptomics, proteomics, and metabolomics to reveal mechanisms  
595 underlying plant secondary metabolism. Eng. Life Sci. 14 (5), 456-466.
- 596 15. Fang, C. and Luo, J. (2019) Metabolic GWAS-based dissection of genetic bases underlying the  
597 diversity of plant metabolism. Plant J. 97 (1), 91-100.
- 598 16. Butelli, E. et al. (2008) Enrichment of tomato fruit with health-promoting anthocyanins by expression  
599 of select transcription factors. Nat. Biotechnol. 26 (11), 1301-1308.
- 600 17. Peng, M. et al. (2017) Differentially evolved glucosyltransferases determine natural variation of rice  
601 flavone accumulation and UV-tolerance. Nat. Commun. 8 (1), 1975.
- 602 18. Li, Y. et al. (2021) Benefiting others and self: Production of vitamins in plants. J. Integr. Plant Biol.  
603 63 (1), 210-227.
- 604 19. Naoumkina, M.A. et al. (2010) Genomic and coexpression analyses predict multiple genes involved  
605 in triterpene saponin biosynthesis in *Medicago truncatula*. Plant Cell 22 (3), 850-866.
- 606 20. Fernie, A.R. and Schauer, N. (2009) Metabolomics-assisted breeding: a viable option for crop  
607 improvement? Trends Genet. 25 (1), 39-48.
- 608 21. Wang, W. et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature  
609 557 (7703), 43-49.
- 610 22. Huang, X. et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature  
611 490, 497.
- 612 23. Qin, P. et al. (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden  
613 genomic variations. Cell 184, 3542-3558.
- 614 24. Su, W. et al. (2021) Polyploidy underlies co-option and diversification of biosynthetic triterpene  
615 pathways in the apple tribe. Proc. Natl. Acad. Sci. U. S. A. 118 (20), e2101767118.
- 616 25. Shi, T. et al. (2020) Metabolomics analysis and metabolite-agronomic trait associations using kernels  
617 of wheat (*Triticum aestivum*) recombinant inbred lines. Plant J. n/a (n/a).
- 618 26. Chen, J. et al. (2020) Metabolite-based genome-wide association study enables dissection of the  
619 flavonoid decoration pathway of wheat kernels. Plant Biotechnol J. n/a (n/a), n/a.
- 620 27. Boycheva, S. et al. (2014) The rise of operon-like gene clusters in plants. Trends Plant Sci. 19 (7),  
621 447-459.
- 622 28. Okada, A. et al. (2009) OsTGAP1, a bZIP transcription factor, coordinately regulates the inductive  
623 production of diterpenoid phytoalexins in rice. J. Biol. Chem. 284 (39), 26510-8.
- 624 29. Nützmann, H.-W. et al. (2016) Plant metabolic clusters – from genetics to genomics. New Phytol.  
625 211 (3), 771-789.
- 626 30. Zhan, C. et al. (2020) Selection of a subspecies-specific diterpene gene cluster implicated in rice  
627 disease resistance. Nat. Plants 6 (12), 1447-1454.
- 628 31. Shen, S. et al. (2021) An *Oryza*-specific hydroxycinnamoyl tyramine gene cluster contributes to  
629 enhanced disease resistance. Sci. Bull. 66 (23), 2369-2380.
- 630 32. Jeon, J.E. et al. (2020) A pathogen-responsive gene cluster for highly modified fatty acids in tomato.  
631 Cell 180 (1), 176-187.
- 632 33. Price, M.N. et al. (2006) The life-cycle of operons. PLoS Genet. 2 (6), e96.
- 633 34. Bratlie, M.S. et al. (2010) Relationship between operon preference and functional properties of  
634 persistent genes in bacterial genomes. BMC Genom. 11 (1), 71.
- 635 35. Frey, M. et al. (1997) Analysis of a chemical plant defense mechanism in grasses. Science 277 (5326),

636 696-699.

637 36. Yang, J. et al. (2017) Haplotype-resolved sweet potato genome traces back its hexaploidization  
638 history. *Nat. Plants* 3 (9), 696-703.

639 37. Matsuba, Y. et al. (2013) Evolution of a complex locus for terpene biosynthesis in *Solanum*. *Plant*  
640 *Cell* 25 (6), 2022-2036.

641 38. Polturak, G. and Osbourn, A. (2021) The emerging role of biosynthetic gene clusters in plant defense  
642 and plant interactions. *PLoS Pathog.* 17, e1009698.

643 39. Tohge, T. et al. (2016) Characterization of a recently evolved flavonol-phenylacyltransferase gene  
644 provides signatures of natural light selection in Brassicaceae. *Nat. Commun.* 7, 12399.

645 40. Wilderman, P.R. et al. (2004) Identification of syn-pimara-7,15-diene synthase reveals functional  
646 clustering of terpene synthases involved in rice phytoalexin/allelochemical biosynthesis. *Plant Physiol.*  
647 135 (4), 2098.

648 41. Field, B. and Osbourn, A.E. (2008) Metabolic diversification—-independent assembly of operon-like  
649 gene clusters in different plants. *Science* 320 (5875), 543-547.

650 42. Shang, Y. et al. (2014) Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science*  
651 346 (6213), 1084-1088.

652 43. Itkin, M. et al. (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by  
653 clustered genes. *Science* 341 (6142), 175-179.

654 44. Zhu, G. et al. (2018) Rewiring of the fruit metabolome in tomato breeding. *Cell* 172 (1), 249-261.e12.

655 45. Fang, H. et al. (2021) A monocot-specific hydroxycinnamoylputrescine gene cluster contributes to  
656 immunity and cell death in rice. *Sci. Bull.* n/a, n/a.

657 46. Liang, J. et al. (2021) Rice contains a biosynthetic gene cluster associated with production of the  
658 casbane-type diterpenoid phytoalexin ent-10-oxodepressin. *New Phytol.* n/a (n/a), n/a.

659 47. Winzer, T. et al. (2012) A papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid  
660 noscapine. *Science* 336 (6089), 1704-1708.

661 48. Qi, X. et al. (2004) A gene cluster for secondary metabolism in oat: Implications for the evolution of  
662 metabolic diversity in plants. *Proc. Natl. Acad. Sci. U. S. A.* 101 (21), 8233-8.

663 49. Li, Y. et al. (2021) Subtelomeric assembly of a multi-gene pathway for antimicrobial defense  
664 compounds in cereals. *Nat. Commun.* 12 (1), 2563.

665 50. Qi, X. et al. (2006) A different function for a member of an ancient and highly conserved cytochrome  
666 P450 family: from essential sterols to plant defense. *Proc. Natl. Acad. Sci. U. S. A.* 103 (49), 18848-  
667 18853.

668 51. Schneider, L.M. et al. (2016) The Cer-cqu gene cluster determines three key players in a  $\beta$ -diketone  
669 synthase polyketide pathway synthesizing aliphatics in epicuticular waxes. *J. Exp. Bot.* 67 (9), 2715-  
670 2730.

671 52. Kautsar, S.A. et al. (2017) plantiSMASH: automated identification, annotation and expression  
672 analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* 45 (W1), W55-W63.

673 53. Töpfer, N. et al. (2017) The PhytoClust tool for metabolic gene clusters discovery in plant genomes.  
674 *Nucleic Acids Res.* 45 (12), 7049-7063.

675 54. Hans-Wilhelm Nutzmann, H.W. et al. (2016) Plant metabolic clusters - from genetics to genomics.  
676 *New Phytol.* 211 (3), 771-789.

677 55. Nützmann, H.-W. et al. (2018) Metabolic gene clusters in eukaryotes. *Annu. Rev. Genet.* 52 (1), 159-  
678 183.

679 56. Darbani, B. et al. (2016) The biosynthetic gene cluster for the cyanogenic glucoside dhurrin in

680 Sorghum bicolor contains its co-expressed vacuolar MATE transporter. *Sci. Rep.* 6 (1), 37079.  
681 57. Laursen, T. et al. (2016) Characterization of a dynamic metabolon producing the defense compound  
682 dhurrin in sorghum. *Science* 354 (6314), 890-893.  
683 58. Zhang, Y. and Fernie, A.R. (2021) Metabolons, enzyme–enzyme assemblies that mediate substrate  
684 channeling, and their roles in plant metabolism. *Plant Commun.* 2 (1), 100081.  
685 59. Sweetlove, L.J. and Fernie, A.R. (2018) The role of dynamic enzyme assemblies and substrate  
686 channelling in metabolic regulation. *Nat. Commun.* 9 (1), 2136.  
687 60. Zhang, Y. et al. (2017) Protein-protein interactions and metabolite channelling in the plant  
688 tricarboxylic acid cycle. *Nat. Commun.* 8 (1), 15212.  
689 61. Zhang, Y. et al. (2020) A moonlighting role for enzymes of glycolysis in the co-localization of  
690 mitochondria and chloroplasts. *Nat. Commun.* 11 (1), 4509.  
691 62. Takos, A.M. et al. (2011) Genomic clustering of cyanogenic glucoside biosynthetic genes aids their  
692 identification in *Lotus japonicus* and suggests the repeated evolution of this chemical defence pathway.  
693 *Plant J.* 68 (2), 273-286.  
694 63. Takos, A. et al. (2010) Genetic screening identifies cyanogenesis-deficient mutants of *Lotus*  
695 *japonicus* and reveals enzymatic specificity in hydroxynitrile glucoside metabolism. *Plant Cell* 22 (5),  
696 1605-1619.  
697 64. Winzer, T. et al. (2015) Morphinan biosynthesis in opium poppy requires a P450-oxidoreductase  
698 fusion protein. *Science* 349 (6245), 309-312.  
699 65. Guo, L. et al. (2018) The opium poppy genome and morphinan production. *Science* 362 (6412), 343-  
700 347.  
701 66. Zhou, Y. et al. (2016) Convergence and divergence of bitterness biosynthesis and regulation in  
702 Cucurbitaceae. *Nat. Plants* 2 (12), 16183.  
703 67. Fan, P. et al. (2020) Evolution of a plant gene cluster in Solanaceae and emergence of metabolic  
704 diversity. *eLife* 9, e56717.  
705 68. Jin, H.R. et al. (2012) The antitumor natural compound falcarindiol promotes cancer cell death by  
706 inducing endoplasmic reticulum stress. *Cell Death Dis.* 3 (8), e376-e376.  
707 69. Miyazawa, M. et al. (1996) Antimutagenic activity of falcarindiol from *Peucedanum praeruptorum*.  
708 *J. Agric. Food Chem.* 44 (11), 3444-3448.  
709 70. Villegas, M. et al. (1988) Isolation of the antifungal compounds falcarindiol and sarisan from  
710 *Heteromorpha trifoliata*. *Planta Med.* 54 (1), 36-37.  
711 71. Kato, T. et al. (1973) Momilactones, growth inhibitors from rice, *oryza sativa* L. *Tetrahedron Lett.*  
712 14 (39), 3861-3864.  
713 72. Cartwright, D. et al. (1977) Chemical activation of host defence mechanisms as a basis for crop  
714 protection. *Nature* 267 (5611), 511-513.  
715 73. Cartwright, D.W. et al. (1981) Isolation and characterization of two phytoalexins from rice as  
716 momilactones A and B. *Phytochemistry* 20 (3), 535-537.  
717 74. Kitaoka, N. et al. (2020) Interdependent evolution of biosynthetic gene clusters for momilactone  
718 production in rice. *Plant Cell* 33, 290-305.  
719 75. Koga, J. et al. (1995) Phytocassanes A, B, C and D, novel diterpene phytoalexins from rice, *Oryza*  
720 *sativa* L. *Tetrahedron* 51 (29), 7907-7918.  
721 76. Koga, J. et al. (1997) Functional moiety for the antifungal activity of phytocassane E, a diterpene  
722 phytoalexin from rice. *Phytochemistry* 44 (2), 249-253.  
723 77. Akatsuka, T. et al. (1985) Novel phytoalexins (oryzalexins A, B and C) isolated from rice blast leaves

724 infected with *Pyricularia oryzae*. *J. Ceram. Soc. Jp.* 49 (6), 1689-1701.

725 78. Sekido, H. et al. (1986) Oryzalexin D (3, 7-dihydroxy-(+)-sandaracopimaradiene), a new phytoalexin  
726 isolated from blast-infected rice leaves. *J. Pestic. Sci* 11 (3), 369-372.

727 79. Kato, H. et al. (1993) Oryzalexin E, A diterpene phytoalexin from UV-irradiated rice leaves.  
728 *Phytochemistry* 33 (1), 79-81.

729 80. Kato, H. et al. (1994) Oryzalexin F, a diterpene phytoalexin from UV-irradiated rice leaves.  
730 *Phytochemistry* 36 (2), 299-301.

731 81. Kodama, O. et al. (1992) Oryzalexin S, a novel stemarane-type diterpene rice phytoalexin. *Biosci.*  
732 *Biotech. Biochem.* 56 (6), 1002-1003.

733 82. Vanetten, H.D. et al. (1994) Two classes of plant antibiotics: phytoalexins versus "phytoanticipins".  
734 *Plant Cell* 6 (9), 1191-1192.

735 83. Guo, L. et al. (2017) *Echinochloa crus-galli* genome analysis provides insight into its adaptation and  
736 invasiveness as a weed. *Nat. Commun.* 8 (1), 1031.

737 84. Lu, X. et al. (2018) Inferring roles in defense from metabolic allocation of rice diterpenoids. *Plant*  
738 *Cell* 30 (5), 1119-1131.

739 85. Zhang, J. et al. (2021) A (conditional) role for labdane-related diterpenoid natural products in rice  
740 stomatal closure. *New Phytol.* n/a (n/a), n/a.

741 86. Inoue, Y. et al. (2013) Identification of a novel casbane-type diterpene phytoalexin, ent-10-  
742 oxodepressin, from rice leaves. *Biosci. Biotech. Bioch.* 77 (4), 760-765.

743 87. Horie, K. et al. (2016) Ultraviolet-induced amides and casbene diterpenoids from rice leaves.  
744 *Phytochem. Lett.* 15, 57-62.

745 88. Panizza, B.J. et al. (2019) Phase I dose-escalation study to determine the safety, tolerability,  
746 preliminary efficacy and pharmacokinetics of an intratumoral injection of tigilanol tiglate (EBC-46).  
747 *Ebiomedicine* 50, 433-441.

748 89. Hezareh, M. (2005) Prostratin as a new therapeutic agent targeting HIV viral reservoirs. *Drug News*  
749 *Perspect.* 18 (8), 496-500.

750 90. Johnson, H.E. et al. (2008) Variability in content of the anti-AIDS drug candidate prostratin in samoan  
751 populations of *homalanthus nutans*. *J. Nat. Prod.* 71 (12), 2041-2044.

752 91. Lebwohl, M. et al. (2012) Ingenol mebutate gel for actinic keratosis. *N. Engl. J. Med.* 366 (11), 1010-  
753 1019.

754 92. King, A.J. et al. (2014) Production of bioactive diterpenoids in the Euphorbiaceae depends on  
755 evolutionarily conserved gene clusters. *Plant Cell* 26 (8), 3286-3298.

756 93. King, A.J. et al. (2016) A cytochrome P450-mediated intramolecular carbon-carbon ring closure in  
757 the biosynthesis of multidrug-resistance-reversing lathyrene diterpenoids. *Chembiochem* 17 (17), 1593-  
758 1597.

759 94. Castillo, D.A. et al. (2013) An effective strategy for exploring unknown metabolic pathways by  
760 genome mining. *J. Am. Chem. Soc.* 135 (15), 5885-5894.

761 95. Huang, A.C. et al. (2019) A specialized metabolic network selectively modulates *Arabidopsis* root  
762 microbiota. *Science* 364 (6440), eaau6389.

763 96. Field, B. et al. (2011) Formation of plant metabolic gene clusters within dynamic chromosomal  
764 regions. *Proc. Natl. Acad. Sci. U. S. A.* 108 (38), 16116-16121.

765 97. Ding, Y. et al. (2020) Genetic elucidation of interconnected antibiotic pathways mediating maize  
766 innate immunity. *Nat. Plants* 6, 1375-1388.

767 98. Huffaker, A. et al. (2011) Novel acidic sesquiterpenoids constitute a dominant class of pathogen-

768 induced phytoalexins in maize. *Plant Physiol.* 156 (4), 2082-2097.

769 99. Geisler, K. et al. (2013) Biochemical analysis of a multifunctional cytochrome P450 (CYP51) enzyme  
770 required for synthesis of antimicrobial triterpenes in plants. *Proc. Natl. Acad. Sci. U. S. A.* 110 (35),  
771 E3360-E3367.

772 100. Krokida, A. et al. (2013) A metabolic gene cluster in *Lotus japonicus* discloses novel enzyme  
773 functions and products in triterpene biosynthesis. *New Phytol.* 200 (3), 675-690.

774 101. N., B.R. and M., W.R. (1994) Secondary metabolites in plant defence mechanisms. *New Phytol.*  
775 127 (4), 617-633.

776 102. Virtanen, A.I. et al. (1955) 2(3)-benzoxazolinone, an anti-fusarium factor in rye seedlings. *Acta.*  
777 *Chem. Scand.* 9, 1543-1544.

778 103. Héctor et al. (1996) Antialgal and antifungal activity of natural hydroxamic acids and related  
779 compounds. *J. Agric. Food Chem.* 44 (6), 1569–1571.

780 104. Niemeyer, H.M. et al. (1982) Reaction of a cyclic hydroxamic acid from gramineae with thiols.  
781 *Phytochemistry* 21 (9), 2287-2289.

782 105. Jonczyk, R. et al. (2008) Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in  
783 maize: characterization of Bx6 and Bx7. *Plant Physiol.* 146 (3), 1053-1063.

784 106. von Rad, U. et al. (2001) Two glucosyltransferases are involved in detoxification of benzoxazinoids  
785 in maize. *Plant J.* 28 (6), 633-642.

786 107. Frey, M. et al. (2009) Benzoxazinoid biosynthesis, a model for evolution of secondary metabolic  
787 pathways in plants. *Phytochemistry* 70 (15), 1645-1651.

788 108. Niemeyer, H.M. (1988) Hydroxamic acids (4-hydroxy-1,4-benzoxazin-3-ones), defence chemicals  
789 in the gramineae. *Phytochemistry* 27 (11), 3349-3358.

790 109. Sue, M. et al. (2011) Dispersed benzoxazinone gene cluster: molecular characterization and  
791 chromosomal localization of glucosyltransferase and glucosidase genes in wheat and rye. *Plant Physiol.*  
792 157 (3), 985-997.

793 110. Li, Y. and Smolke, C.D. (2016) Engineering biosynthesis of the anticancer alkaloid noscapine in  
794 yeast. *Nat. Commun.* 7, 12137-12137.

795 111. Ye, K. et al. (1998) Opium alkaloid noscapine is an antitumor agent that arrests metaphase and  
796 induces apoptosis in dividing cells. *Proc. Natl. Acad. Sci. U. S. A.* 95 (4), 1601-6.

797 112. Roddick, J.G. et al. (2001) Membrane disruption and enzyme inhibition by naturally-occurring and  
798 modified chacotriose-containing *Solanum* steroidal glycoalkaloids. *Phytochemistry* 56 (6), 603-10.

799 113. Jones, D.A. (1998) Why are so many food plants cyanogenic? *Phytochemistry* 47 (2), 155-162.

800 114. Kakes, P. (1989) An analysis of the costs and benefits of the cyanogenic system in *Trifolium repens*  
801 *L. Theor. Appl. Genet.* 77 (1), 111-118.

802 115. Tattersall, D.B. et al. (2001) Resistance to an herbivore through engineered cyanogenic glucoside  
803 synthesis. *Science* 293 (5536), 1826-1828.

804 116. Biere, A. et al. (2004) Plant chemical defense against herbivores and pathogens: generalized defense  
805 or trade-offs? *Oecologia* 140 (3), 430-41.

806 117. Wiemann, P. et al. (2013) Prototype of an intertwined secondary-metabolite supercluster. *Proc. Natl.*  
807 *Acad. Sci. U. S. A.* 110 (42), 17065-17070.

808 118. Cárdenas, P.D. et al. (2016) GAME9 regulates the biosynthesis of steroidal alkaloids and upstream  
809 isoprenoids in the plant mevalonate pathway. *Nat. Commun.* 7 (1), 10654.

810 119. Yu, N. et al. (2016) Delineation of metabolic gene clusters in plant genomes by chromatin signatures.  
811 *Nucleic Acids Res.* 44 (5), 2255-2265.

812 120. Nützmann, H.-W. and Osbourn, A. (2015) Regulation of metabolic gene clusters in *Arabidopsis*  
813 *thaliana*. *New Phytol.* 205 (2), 503-510.

814 121. Nützmann, H.-W. and Osbourn, A. (2014) Gene clustering in plant specialized metabolism. *Curr.*  
815 *Opin. Biotechnol.* 26, 91-99.

816 122. Nützmann, H.-W. et al. (2020) Active and repressed biosynthetic gene clusters have spatially distinct  
817 chromosome states. *Proc. Natl. Acad. Sci. U. S. A.* 24, 13800-13809.

818 123. Liu, Z. et al., Formation and diversification of a paradigm biosynthetic gene cluster in plants, *Nat.*  
819 *Commun.*, 2020, p. 5354.

820 124. Liu, Z. et al. (2020) Drivers of metabolic diversification: how dynamic genomic neighbourhoods  
821 generate new biosynthetic pathways in the Brassicaceae. *New Phytol.* 227 (4), 1109-1123.

822 125. Peters, R.J. (2020) Doing the gene shuffle to close synteny: dynamic assembly of biosynthetic gene  
823 clusters. *New Phytol.* n/a, n/a.

824 126. Rai, A. et al. (2021) Chromosome-level genome assembly of *Ophiorrhiza pumila* reveals the  
825 evolution of camptothecin biosynthesis. *Nat. Commun.* 12 (1), 405.

826 127. Li, Q. et al. (2020) Gene clustering and copy number variation in alkaloid metabolic pathways of  
827 opium poppy. *Nat. Commun.* 11 (1), 1190.

828 128. The *Arabidopsis* Genome, I. (2000) Analysis of the genome sequence of the flowering plant  
829 *Arabidopsis thaliana*. *Nature* 408 (6814), 796-815.

830 129. Varshney, R.K. et al. (2021) Designing future crops: genomics-assisted breeding comes of age.  
831 *Trends Plant Sci.* 26, 631-649.

832 130. Luo, C. et al. (2020) Single-cell genomics and epigenomics: technologies and applications in plants.  
833 *Trends Plant Sci.* 25, 1030-1040.

834 131. Gaquerel, E. et al. (2014) Revealing insect herbivory-induced phenolamide metabolism: from single  
835 genes to metabolic network plasticity analysis. *Plant J.* 79 (4), 679-692.

836 132. Li, D. and Gaquerel, E. (2021) Next-generation mass spectrometry metabolomics revives the  
837 functional analysis of plant metabolic diversity. *Annu. Rev. Plant Biol.* 72, 867-891.

838 133. Yang, W. et al. (2020) Crop phenomics and high-throughput phenotyping: past decades, current  
839 challenges, and future perspectives. *Mol. Plant* 13 (2), 187-214.

840 134. Wu, X. et al. (2021) Using high-throughput multiple optical phenotyping to decipher the genetic  
841 architecture of maize drought tolerance. *Genome Biol.* 22 (1), 185.

842 135. Yamamuro, C. et al. (2016) Epigenetic modifications and plant hormone action. *Mol. Plant* 9 (1),  
843 57-70.

844 136. Tunyasuvunakool, K. et al. (2021) Highly accurate protein structure prediction for the human  
845 proteome. *Nature* 596 (7873), 590-596.

846 137. Forestier, E.C.F. et al. (2021) Developing a *Nicotiana benthamiana* transgenic platform for high-  
847 value diterpene production and candidate gene evaluation. *Plant Biotechnol. J.* n/a (n/a).

848 138. Reed, J. et al. (2017) A translational synthetic biology platform for rapid access to gram-scale  
849 quantities of novel drug-like molecules. *Metab. Eng.* 42, 185-193.

850 139. Shamloul, M. et al. (2014) Optimization and utilization of agrobacterium-mediated transient protein  
851 production in *Nicotiana*. *J. Vis. Exp.* 86 (86), e51204-e51204.

852 140. Sainsbury, F. and Lomonosoff, G.P. (2008) Extremely high-level and rapid transient protein  
853 production in plants without the Use of Viral Replication. *Plant Physiol.* 148 (3), 1212.

854 141. Zhao, X. et al. (2017) Pollen magnetofection for genetic modification with magnetic nanoparticles  
855 as gene carriers. *Nat. Plants* 3 (12), 956-964.

856 142. Strobbe, S. et al. (2021) Metabolic engineering of rice endosperm towards higher vitamin B1  
857 accumulation. *Plant Biotechnol. J.* n/a (n/a).

858 143. Zhu, Q. et al. (2020) Plant synthetic metabolic engineering for enhancing crop nutritional quality.  
859 *Plant Commun.* 1 (1), 100017.

860 144. Zhu, Q. et al. (2018) From golden rice to aSTARice: bioengineering astaxanthin biosynthesis in rice  
861 endosperm. *Mol. Plant* 11 (12), 1440-1448.

862 145. Ro, D.-K. et al. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered  
863 yeast. *Nature* 440, 940–943.

864 146. Paddon, C.J. et al. (2013) High-level semi-synthetic production of the potent antimalarial  
865 artemisinin. *Nature* 496 (7446), 528-532.

866 147. Matsuba, Y. et al. (2015) Biosynthesis of the Diterpenoid Lycosantalolol via Nerylneryl  
867 Diphosphate in *Solanum lycopersicum*. *PLoS One* 10 (3), e0119302.

868 148. Mao, L. et al. (2020) Genomic evidence for convergent evolution of gene clusters for momilactone  
869 biosynthesis in land plants. *Proc. Natl. Acad. Sci. U. S. A.* 117 (22), 12472-12480.

870 149. Shimura, K. et al. (2007) Identification of a biosynthetic gene cluster in rice for momilactones. *J.*  
871 *Biol. Chem.* 282 (47), 34013.

872 150. Swaminathan, S. et al. (2009) CYP76M7 is an ent-cassadiene C11 $\alpha$ -hydroxylase defining a second  
873 multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell* 21 (10), 3315-3325.

874 151. Mugford, S.T. et al. (2009) A serine carboxypeptidase-like acyltransferase is required for synthesis  
875 of antimicrobial compounds and disease resistance in oats. *Plant Cell* 21 (8), 2473-2484.

876 152. Mugford, S.T. et al. (2013) Modularity of plant metabolic gene clusters: a trio of linked genes that  
877 are collectively required for acylation of triterpenes in oat. *Plant cell* 25 (3), 1078-1092.

878 153. Sohrabi, R. et al. (2015) In planta variation of volatile biosynthesis: an alternative biosynthetic route  
879 to the formation of the pathogen-induced volatile homoterpene DMNT via triterpene degradation in  
880 *arabidopsis* roots. *Plant Cell* 27 (3), 874-890.

881 154. Frey, M. et al. (2003) A 2-oxoglutarate-dependent dioxygenase is integrated in DIMBOA-  
882 biosynthesis. *Phytochemistry* 62 (3), 371-376.

883 155. Carere, J. et al. (2018) BdACT2a encodes an agmatine coumaroyl transferase required for pathogen  
884 defence in *Brachypodium distachyon*. *Physiol. Mol. Plant P.* 104, 69-76.

885

886

887 **Glossary**

888 **Plant metabolic gene cluster:** a group of closely linked non-homologous genes  
889 encoding enzymes from a multi-step process such as the biosynthesis of a  
890 secondary/primary metabolite in plants.

891 **Super gene cluster:** a large (two or more) metabolic gene clusters with related  
892 functions colocalizing in a genomic region.

893 **Multi-omics:** the analysis that integrate more than one profiling technology – capturing,  
894 for instance, the genome, transcriptome, metabolome, proteome and epigenome –  
895 across a common set of the samples.

896 **Natural variation:** the genetic diversity of an individual organism under natural  
897 conditions.

898 **Specialized metabolism:** The metabolites which have various functions, including  
899 been used by humans as medicines, dyes, pigments, cosmetics, agrochemicals and so  
900 on.

901

902 **Figure Legends**

903 **Figure 1. Timeline of the plant gene clusters and its related genomes.** Left, the  
904 timeline of the discovered plant gene cluster; Right, the timeline of the reported plant  
905 genomes. DIMBOA, 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one; FPT, flavonol-  
906 phenylacyltransferase; Zx, zealexin.

907

908 **Figure 2. Strategies for plant gene cluster identification.** (A) Identification of the  
909 plant gene cluster through genome-wide association studies. (B) Identification of the  
910 plant gene cluster through quantitative trait loci (QTLs). (C) Identification of the plant  
911 gene cluster through special algorithms. (D) Identification of the plant gene cluster  
912 through genome mining. (E) Identification of the plant gene cluster through the  
913 combination analysis of omics data. The figures are modified from refs<sup>31,49,80,118</sup>.

914

915 **Figure 3. Types of cluster organization.** (A) Compact gene clusters. (B) Loose gene  
916 clusters. (C) Super gene cluster – combination of different metabolic gene clusters.

917

918 **Figure 4. Main categories of plant metabolic cluster products and their**  
919 **agronomic/medical functions.** Terpenoid: Avenacin A-1 in *Avena strigose*;  
920 Diterpenoid: Casbene diterpenoid: Casbene diterpenoid in *Ricinus communis*;  
921 Phenylpropanoids: Feruloyl-tyramine in *Oryza sativa*; Benzoxazinoids: DIMBOA-  
922 glucoside in *Zea mays*; Alkaloid: Noscapine in *Papaver somniferum*; Cyanogenic  
923 glycoside: dhurrin in *Sorghum bicolor*; Fatty acids:  $\beta$ -diketones in *Hordeum vulgare*;  
924 Falcarindiol in *Solanum lycopersicum*.

925

926 **Figure 5. Two distinct phenolamide gene clusters confer broad spectrum disease**  
927 **resistance in rice.** Aromatic phenolamide gene cluster: hydroxycinnamoyl tyramine  
928 (HT) gene cluster include a pyridoxamine 5'-phos-phate oxidase (OsPDX3) producing  
929 the cofactor pyridoxal 5'-phosphate (PLP), a PLP-dependent tyrosine decarboxylase  
930 (OsTyDC1), and two duplicated hydroxycinnamoyl transferases (OsTHT1 and  
931 OsTHT2) and this gene cluster conserved in *Oryza* AA genome lineage; Aliphatic  
932 phenolamide gene cluster: hydroxycinnamoyl putrescine (HP) gene cluster include a  
933 decarboxylase (OsODC) and two tandem-duplicated genes encoding putrescine  
934 hydroxycinnamoyl acyltransferases (OsPHT3 and OsPHT4) and this gene cluster  
935 conserved in monocots. Chr., chromosome.

936

937 **Figure 6. The evolution of *DGC7*.** The relative proportion of six types of gene modules.  
938 The intact *DGC7* is highly enriched in the *japonica* varieties (102/109) compared to the  
939 *indica* varieties (13/313), suggesting the selection of *DGC7* during domestication.

940

941 **Figure 7. The omics data can be used to crack the mysteries of plant gene clusters.**  
942 Various interaction networks exist both within each omics network and also between  
943 omics networks. The features can help to crack the mysteries of plant gene clusters.

944

945 Table 1. Clustered pathways for the biosynthesis of plant natural products

Major classes of compound	Class of compound	Secondary metabolite	Phyto group	Plant species	Expression pattern	Method of cluster discovery	Refs.
Terpenes	Monoterpenes	$\beta$ -Phellandrene	Eudicot	<i>Solanum lycopersicum</i>	Induced co-expression	Characterized biosynthetic genes to cluster	[37]
	Diterpene	Lycosantalanol	Eudicot	<i>Solanum lycopersicum</i>	Induced co-expression	Characterized biosynthetic genes to cluster	[147]
	Diterpene	Casbene diterpenoids	Eudicot	<i>Euphorbia peplus</i>	Root	Characterized biosynthetic genes to cluster	[92]
	Diterpene	Casbene diterpenoids	Eudicot	<i>Jatropha curcas</i>	Root	Genome mining; genetics	[92]
	Diterpene	Casbene diterpenoids	Eudicot	<i>Ricinus communis</i>	/	Cluster mining; characterized biosynthetic genes to cluster	[6, 92]
	Diterpene	Casbene diterpenoids	Monocots	<i>Oryza sativa</i>	root/leaf	mGWAS-based discovery	[30]
	Diterpene	Momilactones	Bryophyte	<i>Calohyphnum plumiforme</i>	Induced co-expression	Characterized biosynthetic genes to cluster; genomics	[148]
Terpenes	Diterpene	Momilactones	Monocots	<i>Echinochloa crus-galli</i>	Induced co-expression	Induced co-expression based discovery	[83]
	Diterpene	Momilactones	Monocots	<i>Oryza sativa</i>	Induced co-expression	Characterized biosynthetic genes to cluster	[40, 149]
	Diterpene	Phytocassanes /oryzalides	Monocots	<i>Oryza sativa</i>	Induced co-expression	Characterized biosynthetic genes to cluster	[150]
	Diterpene	Zealexin	Monocots	<i>Zea mays</i>	Induced co-expression	Characterized biosynthetic genes to cluster	[97]
	Triterpene	Avenacins	Monocots	<i>Avena strigosa</i>	Root	Forward screen mutants	[41, 48-50, 151, 152]
	Triterpene	Thalianol	Eudicot	<i>Arabidopsis thaliana</i>	Root	Induced co-expression based discovery	
	Triterpene	Marneral	Eudicot	<i>Arabidopsis thaliana</i>	Root	Cluster mining	[96]
	Triterpene	Tirucalla-7,24-dien-3b-ol	Eudicot	<i>Arabidopsis thaliana</i>	Root	Cluster mining	[6]
	Triterpene	Arabidiol	Eudicot	<i>Arabidopsis thaliana</i>	Induced co-expression	Cluster mining	[94] [153]
Terpenes	Triterpene	Cucurbitacins C	Eudicot	<i>Cucumis sativus</i>	Stem/leaf/fruit	Forward screen the Bi locus for bitterness; GWAS	[6, 42]
	Triterpene	Cucurbitacins B	Eudicot	<i>Cucumis melo L.</i>	Root/fruit	Comparative genomics	[66]
	Triterpene	Cucurbitacins E	Eudicot	<i>Citrullus lanatus L.</i>	Root/fruit	Comparative genomics	[66]
	Triterpene	Thalianol	Eudicot	<i>Arabidopsis lyrata</i>	Root	Comparative genomics	[95, 124]
	Triterpene	Tirucallol	Eudicot	<i>Capsella rubella</i>	Buds	Comparative genomics	[124]
	Triterpene	20-Hydroxy-betulinic acid	Eudicot	<i>Lotus japonicus</i>	Root/induced co-expression	Cluster mining	[100]
N-containing compounds	Cyanogenic glycoside	Linamarin/lotaustralin	Eudicot	<i>Lotus japonicus</i>	Not strictly co-expression	Isolation of cyanogenesis deficient mutants; genomics	[62]
	Cyanogenic glycoside	Linamarin/lotaustralin	Eudicot	<i>Manihot esculenta</i>	Not strictly co-expression	Comparative genomics	[62]
	Cyanogenic glycoside	Dhurrin	Monocots	<i>Sorghum bicolor</i>	Not strictly co-expression	Comparative genomics	[62]

Alkaloid	Benzylisoquinoline alkaloid	Noscapine	Eudicot	<i>Papaver somniferum</i>	Stem	Forward screen; Tissue-specific coexpression	[47]
	Steroidal alkaloid	$\alpha$ -Tomatine	Eudicot	<i>Solanum lycopersicum</i>	Fruit	Characterized biosynthetic genes to cluster	[43]
Alkaloid	Teroidal alkaloid	$\alpha$ -Chaconine $\alpha$ -Solanine	Eudicot	<i>Solanum tuberosum</i>	Tubers	Characterized biosynthetic genes to cluster; Comparative genomics	[43]
Benzenoids	Hydroxamic acid	2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA)	Monocots	<i>Zea mays</i>	Induced co-expression	Forward screen screen for <i>bx1</i> mutants	[35, 105, 106, 154]
	Hydroxamic acid	2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA)	Monocots	<i>Echinochloa crus-galli</i>	Induced co-expression	Cluster mining	[83]
Phenylpropanoids	Phenylpropanoid derivatives	Hydroxycinnamoyl-tyramine	Monocots	<i>Oryza sativa</i>	Induced co-expression	mGWAS-based discovery	[31]
	Phenylpropanoid derivatives	Hydroxycinnamoyl-putrescine	Monocots	<i>Oryza sativa</i>	Induced co-expression	mGWAS-based discovery	[45]
	Phenylpropanoid derivatives	Hydroxycinnamoyl-agmatine	Monocots	<i>Brachypodium distachyon</i>	Induced co-expression	Induced co-expression based discovery	[155]
Fatty acids	Polyketide	$\beta$ -Diketones	Monocots	<i>Hordeum vulgare</i>	Leaf sheath	Forward screen the Cerqu leaf wax locus	[51]
	Modified fatty acids	Falcarindiol	Eudicot	<i>Solanum lycopersicum</i>	Induced co-expression	Induced co-expression based discovery	[32]
	Sugar aliphatic esters	Medium chain acylsugar	Eudicot	<i>Solanum lycopersicum</i>	Trichome	Forward screen; Tissue-specific coexpression	[67]
	Sugar aliphatic esters	Medium chain acylsugar	Eudicot	<i>Solanum pennellii</i>	Trichome	Comparative genomics	[67]
	Sugar aliphatic esters	Medium chain acylsugar	Eudicot	<i>Solanum melongena</i>	Trichome	Comparative genomics	[67]

946

947

1 **Outstanding Questions**

2 How to dissect the regulatory mechanisms, natural variation, evolution, constituents  
3 and function of plant metabolic gene clusters more directly and efficiently with multi-  
4 omics strategies?

5

6 How to reveal and simulate the full life cycle (birth, life and death) of plant metabolic  
7 gene cluster?

8

9 Why do most of the class of compounds biosynthesized by the plant gene cluster belong  
10 to secondary (or specialized) metabolites rather than primary metabolites?

11

12 How can we rationally develop the synthetic biology strategies for the production of  
13 bioactive compounds biosynthesized by the plant gene cluster?

Figure 1

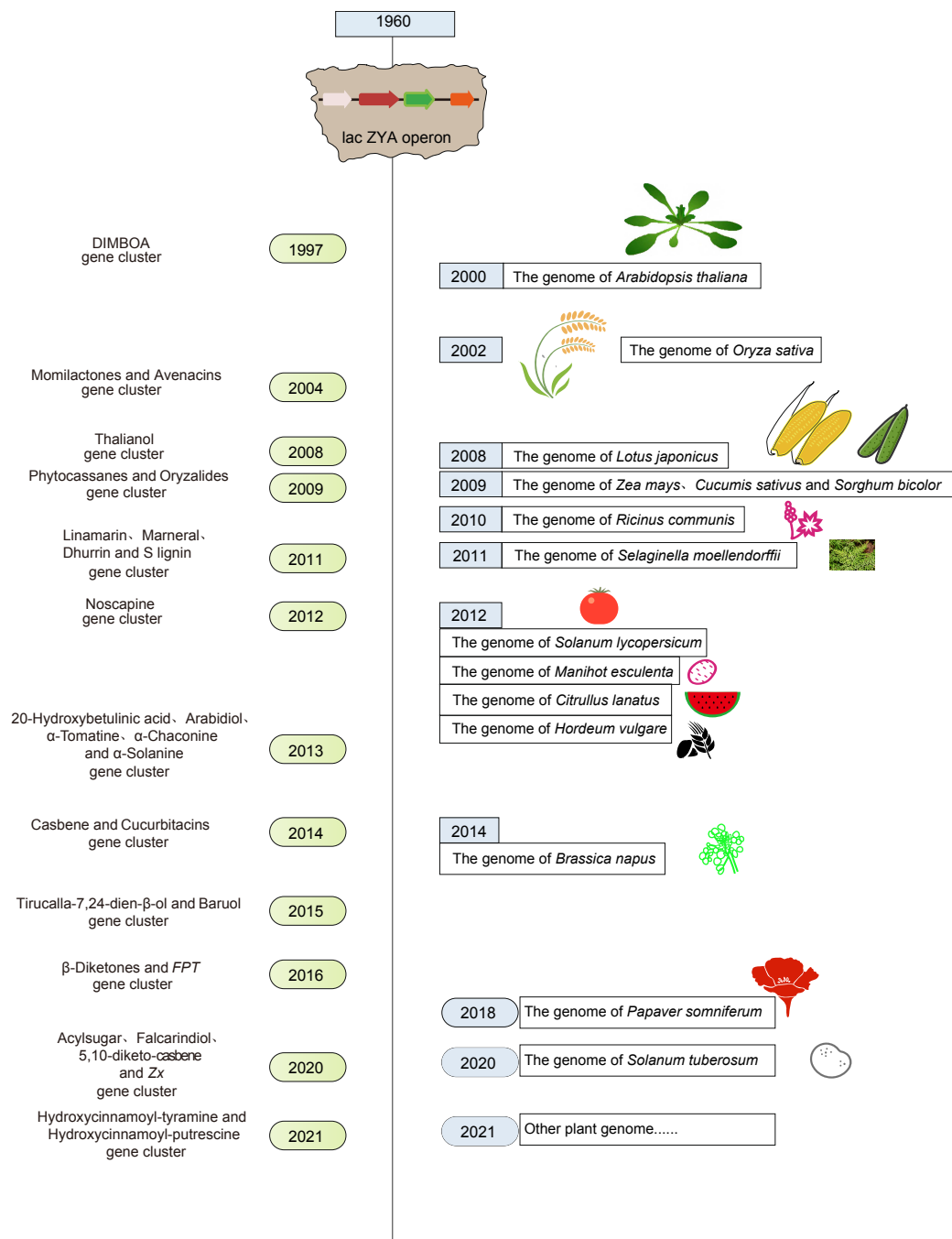


Figure 2

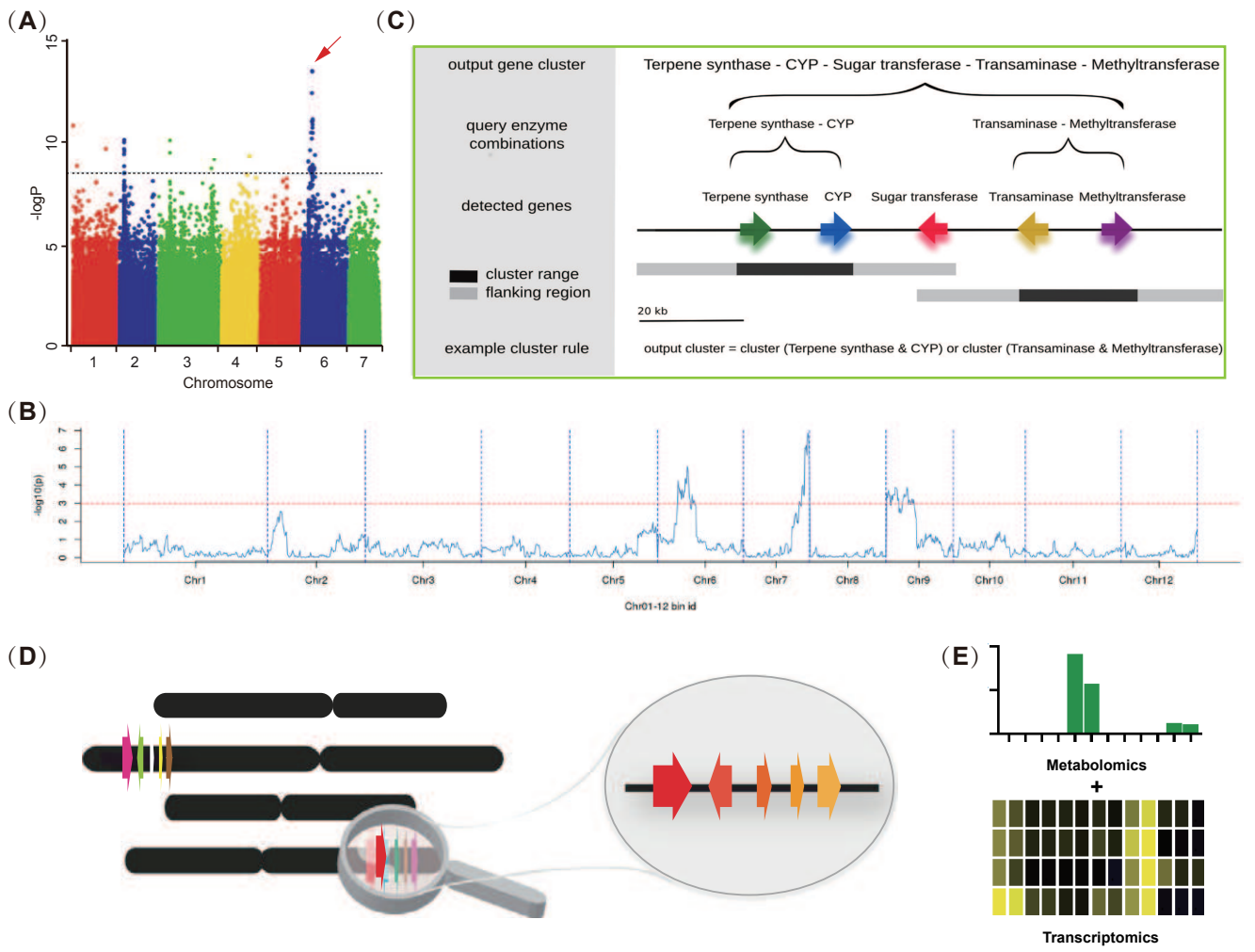


Figure 3

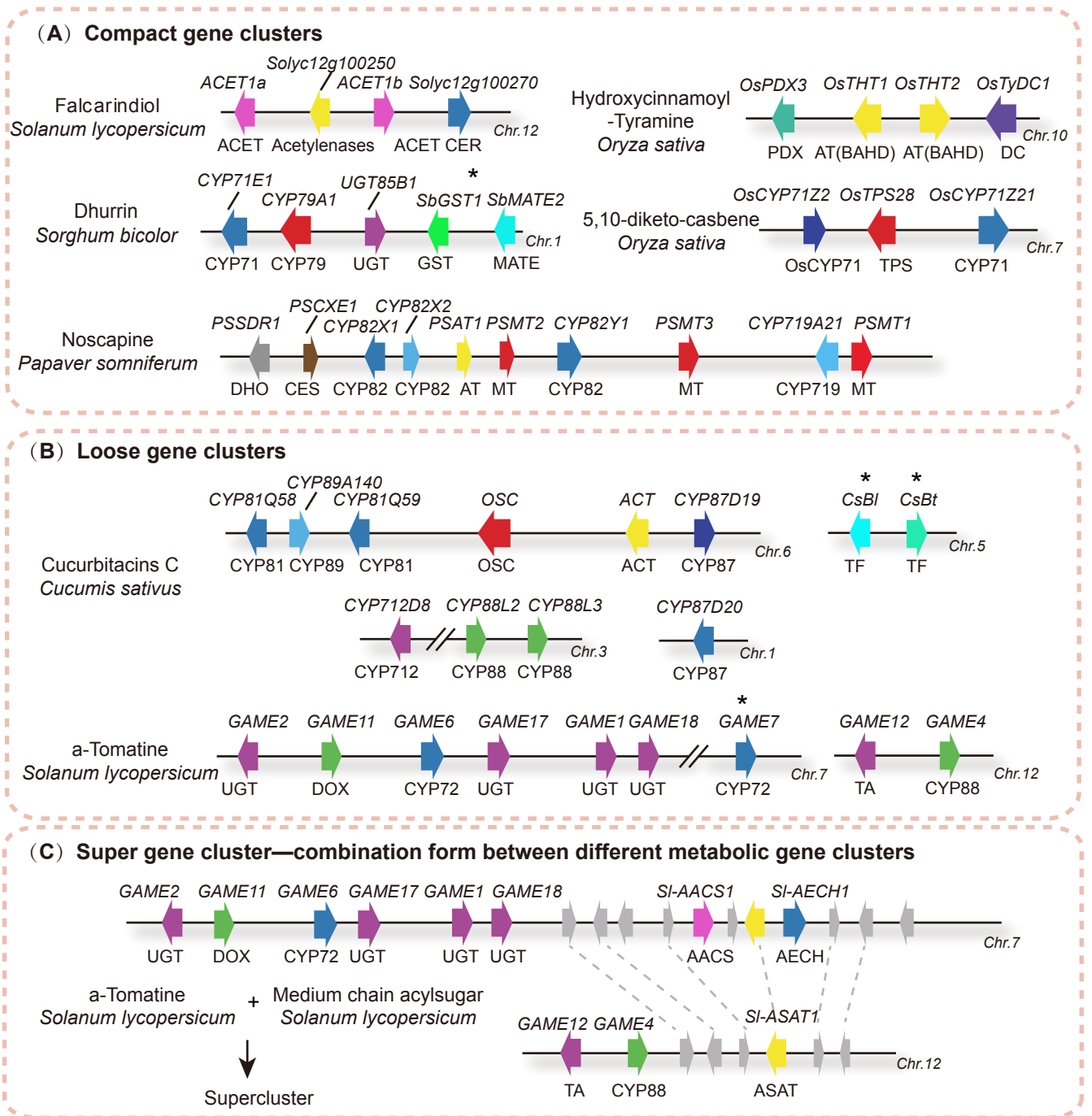


Figure 4

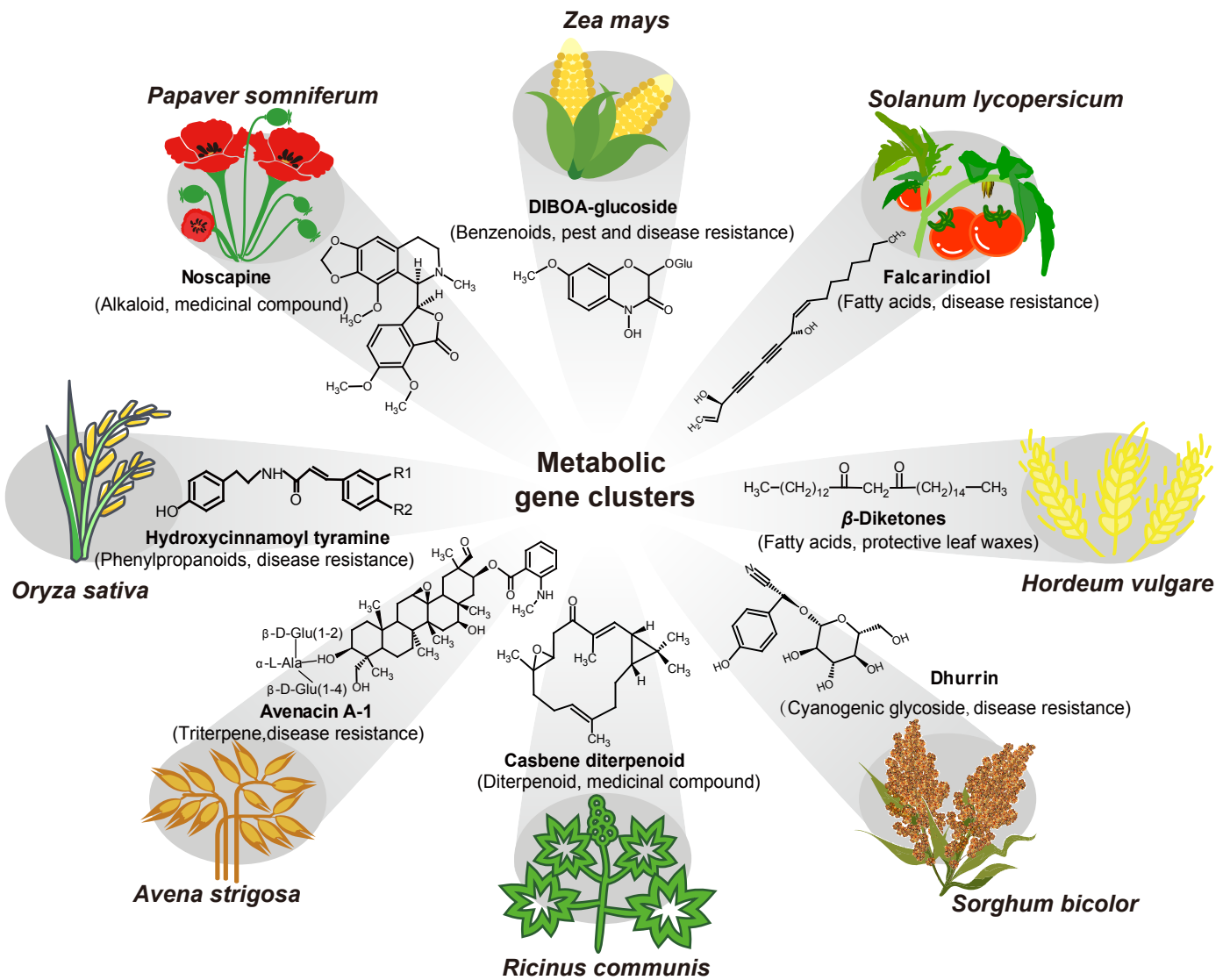


Figure 5

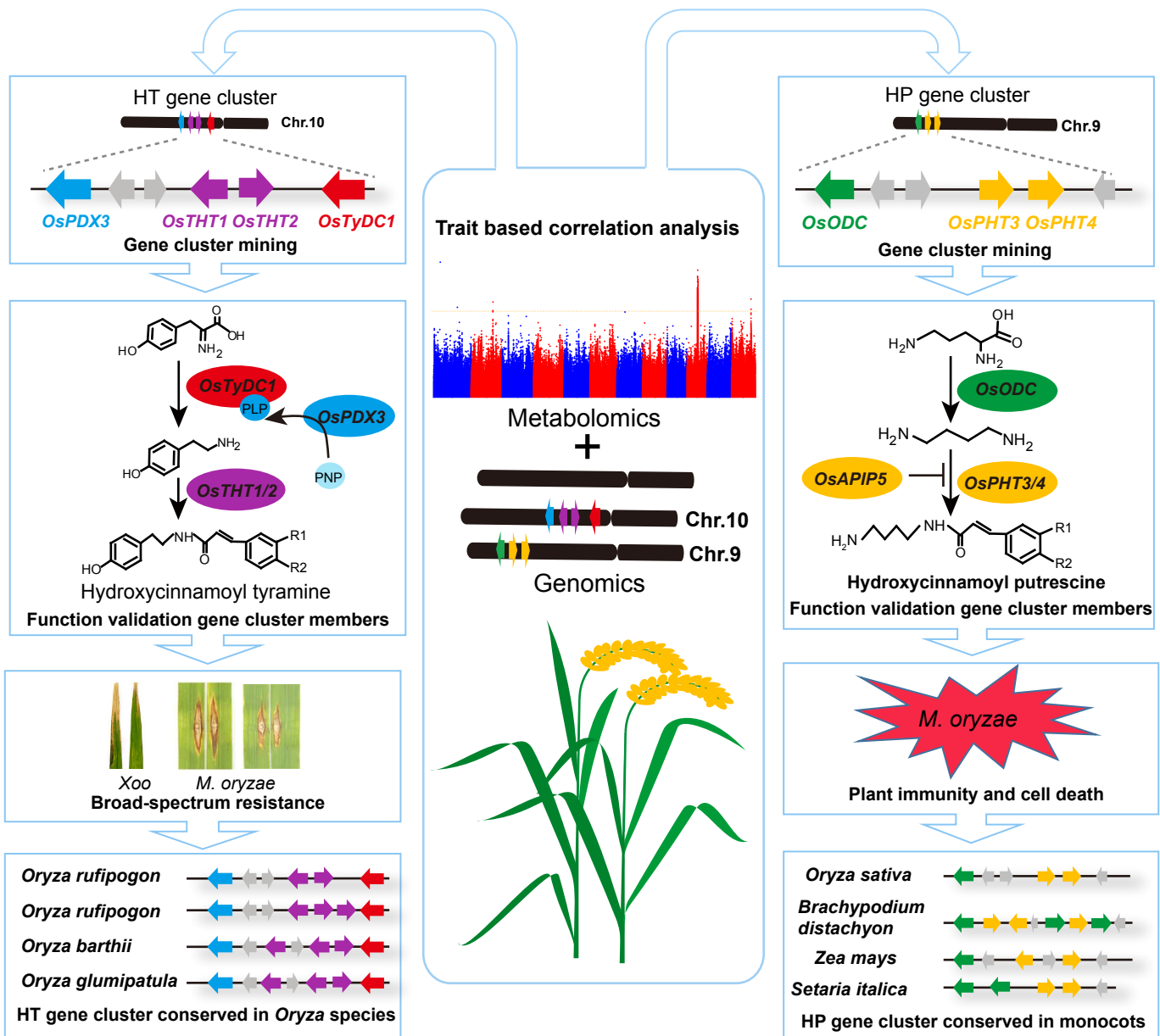


Figure 6

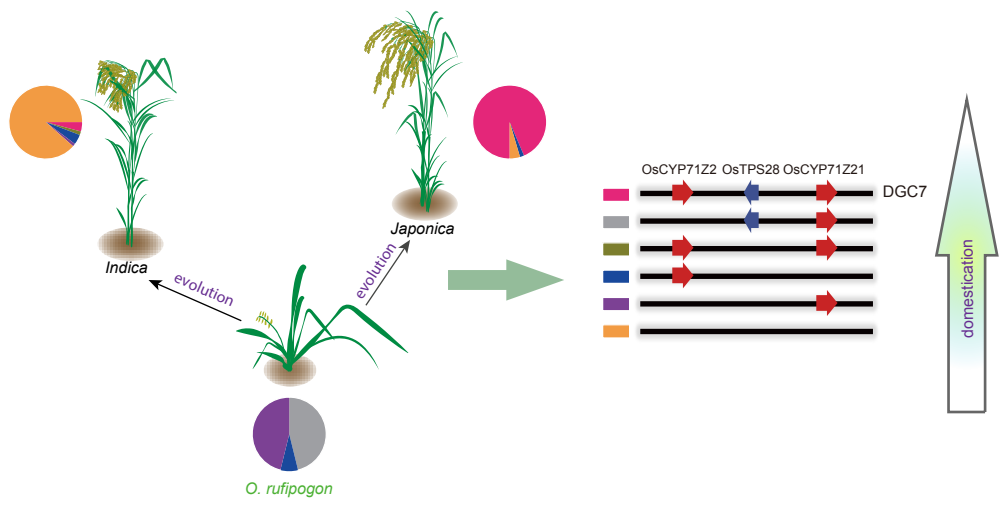


Figure 7

