Plant metabolic gene cluster in the multi-omics era

Chuansong Zhan[1,2]*, Shuangqian Shen[1,3]*, Chenkun Yang[3], Zhenhua Liu[4], Alisdair R. Fernie[5,6], Ian A. Graham[7] and Jie Luo[1,2,&]

[1] College of Tropical Crops, Hainan University, Haikou 570228, China.

[2] Sanya Nanfan Research Institute of Hainan University, Hainan Yazhou Bay Seed Laboratory, Sanya 572025, China.

[3] National Key Laboratory of Crop Genetic Improvement and National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China.

[4] Joint Center for Single Cell Biology, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China.

[5] Max-Planck-Institut fur Molekulare Pflanzenphysiologie, Am Muhlenberg 1, 14476 Potsdam-Golm, Germany.

[6] Center of Plant Systems Biology and Biotechnology, 4000 Plovdiv, Bulgaria.

[7] Center for Novel Agricultural Products, Department of Biology, University of York, York, UK.

* These authors contributed equally to this article

Correspondence: jie.luo@hainanu.edu.cn (J. Luo)

Keywords: Metabolic gene cluster, natural variation, GWAS,

Abstract

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

Plant secondary metabolism and metabolic gene clusters

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

Understanding the biosynthetic pathways, regulatory mechanisms and transport processes responsible for production of secondary metabolites will be essential to fully explore the potential of this treasure trove of natural products for the benefit of human society and the environment [14-18]. In contrast to the situation in prokaryotes, genes involved in plant **secondary metabolism** are generally randomly distributed across the plant genome, which typically means that genes encoding the enzymes of a biochemical pathway have to be discovered one step at a time. Even today with the advent of gene sequence information [19-26], discovery of the full complement of genes responsible for biochemical pathways underlying plant specialized metabolism remains a considerable challenge.

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

Historical Aspects and General Strategies for Plant Metabolic Gene Cluster Identification

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

Generally, two main strategies have been used for gene identifying and characterizing: forward genetic (from phenotype to gene) strategies and the reverse genetics (from gene to phenotype) strategies (Figure 2). In the identification of plant gene cluster, forward genetic strategies are one of the powerful strategies that have been verified multiple times. Genome-wide association studies (GWAS), map-based cloning, and bulked-segregant analyses are three common forward genetic approaches used for causal gene(s) identification and characterization. Among these, the genome-wide association studies (GWAS) and map-based cloning demonstrated great utility in the identification of metabolic gene clusters in plants (Figure 2A and 2B). For example, five clustered genes encoding enzymes required for the biosynthesis of cucurbitacins were successfully identified through genome-wide association analysis using variation maps of 115 different cucumber varieties [42] (Figure 2A). In tomato, a natural population consisting of 600 lines were studied by using systematic metabolome and genomic analysis strategies. A potential gene cluster on chromosome 10 containing a P450 oxidoreductase, an acyltransferase, an acetyl-CoA dehydrogenase and an UDP-glucosyltransferase, in addition to the previously identified gene cluster on chromosome 7 was uncovered [43, 44]. Further study showed that this locus was responsible for the natural variation of the toxic anti-nutritional factor *α*-solanine in tomato [44]. Recently, Zhan and Shen *et al.* performed metabolic GWAS (mGWAS) in monocot rice populations and revealed three brand-new gene clusters: diterpenoid gene cluster on chromosome 7, *DGC7*; hydroxycinnamoyl tyramine gene cluster, a *HT* gene cluster; a hydroxycinnamoyl putrescine gene cluster and a *HP* gene cluster [30, 31, 45]. They also demonstrated that all end-products synthesized by these three gene clusters can confer disease resistance in rice [30, 31, 38, 45, 46]. Comparison of the transcriptome of stems and capsules from opium poppy varieties HN1, HM1 and HT1 (which producing high levels of noscapine, high morphine, and high thebaine, respectively) revealed 10 co-expressed genes specifically existed in HN1 [47]. By screening the HN1 Bacterial Artificial Chromosome (BAC) library and analyzing the F2 population, a 10 gene metabolic gene cluster specific to HN1 was found, which is responsible for the production of noscapine [47]. Through using seventeen putative mutants that were crossed into an inbred line of maize, *Bx1*/*Bx1*. The first plant metabolic gene cluster – the DIMBOA gene cluster was identified [35]. A few years later, Qi *et al.* used the recombinant inbred lines that derived from *A. strigose* C13815 x *A. wiestii* C11994 to successfully map the gene, *AsbAS1*, which responsible for biosynthesize the *beta*-amyrin which is the skeleton of avenacins [48]. Further studies uncovered that this gene was part of an antimicrobial triterpenoid gene cluster at the locus, containing a total of 12 genes [49, 50]. In barley, *Cer-Cqu* was mapped to a discrete location on chromosome arm 2HS using population mapping method. Combined with BAC library data and sequencing, three candidate genes (*Cer-C*, *Cer-Q* and *Cer-U*) were identified [51]. These three genes were distributed over a 101 Kb chromosomal interval and were highly co-expressed in leaf sheath tissue, confirming that they formed a metabolic gene cluster that catalyzed the biosynthesis of β-diketone. Taken together these studies clearly illustrate the power of forward genetics in the identification of plant metabolic gene clusters.

The advent of multiple omics technologies has offered us new strategies for the identification of the plant metabolic gene clusters (Figure 2C). The bio-informatic computational pipeline strategy is a special reverse strategy worthy to mention here. For example, algorithms such as PlantiSMASH and PhytoClust have been developed and applied to predict the secondary metabolic gene clusters in plants [52, 53]. Both tools adopt accurate Hidden Markov Model profiles (pHMMS) to judge different biosynthesis genes and predict candidate gene clusters in combination with genome locations. Generally, most of the retrieval rules of these computer algorithms are based on the typical combination of the “signature enzymes” and “tailoring enzymes” (Figure 2C). For example, the terpene synthase (TPS) and cytochrome P450 enzyme (CYP450) are the main types of enzymes that are involved in terpenoid metabolic pathways. Of these, the terpene synthases are considered to act as the “signature enzymes”, whereas cytochrome P450 enzymes are “tailoring enzymes”. Based on these basic rules, Töpfer Nadine *et al.*, searched TPS/CYP450 combinations across multiple plant genomes and identified both known and novel terpene gene clusters [53]. It is known that this approach is able to render the identification of plant metabolic gene clusters more facile and the accuracy of such predictions will be increased through integration of genome and transcriptome data (Figure 2D and 2E) [52, 53]. Availability of user-friendly interfaces for such algorithms in combination with developments in the fields of next-generation sequencing, analytical chemistry, synthetic biology, and systems biology will considerably accelerate the speed of discoveries of gene clusters in diverse non-model plants. Specifically, the combination of different strategies, such as different omics, will provide more clues to narrow the gap between phenotypic diversity and genetic variation. For example, through parallel mGWAS and a gene-based association analysis using metabolic, genetic, and phenotypic data, new candidate gene clusters for the natural variation in content of tyramine were identified [31].

Cluster Constituents and Organization

Constituents of Metabolic Gene Cluster

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

Interestingly, recent studies have shown that in addition to genes encoding enzymes other genes, encoding transporters and cofactor synthases, can also be associated with plant metabolic clusters [31, 56]. Darbani Behrooz *et al.* have reported that the cyanogenic glucoside gene cluster consists of four different genes: *CYP79D3*, *CYP79D4*, *CYP736A2* and *UGT85K3*. Another study uncovered that the gene *SbMATE2*, which encodes a transporter that is required for the transport of non-endogenous cyanogenic glucosides is located within the same cluster in *Sorghum bicolor* [56]. A cofactor is a non-protein substance which is required for a protein to be catalytically active. In the hydroxycinnamoyl tyramine (HT) gene cluster, besides the biosynthetic genes (tyrosine decarboxylase, OsTyDC1; tyrosine decarboxylase, OsTyDC1; acyl transferases, OsTHT1 and OsTHT2), a pyridoxal 5-phosphate (PLP) cofactor synthetase OsPDX3 is also embedded in these gene clusters [31]. PLP is a type of cofactor that is required for the catalysis of enzymes such as transaminases, isomerases, decarboxylases, racemases, aldolases, deaminases, and aminotransferases. *In vitro* enzyme analyses demonstrated that the HT gene cluster member OsPDX3 acted as a cofactor donor for the PLP-dependent tyrosine decarboxylase OsTyDC1, suggesting that the cofactor synthase was indirectly necessary for the production of the end products (Figure 3A). Such step-by-step characterization not only enriches our understanding of the scope of their function but also broadens our understanding of the enzyme catalog of gene cluster components. Indeed, the presence of these novel members suggest that bioinformatic tools will need to be refined in order to accommodate such members of plant metabolic gene clusters.

Organization of Metabolic Gene Cluster

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

The thalianol gene cluster in arabidopsis is the smallest plant compact gene cluster with only 35~38 Kb [41]. Other clusters belonging to this structural form are the falcarindiol gene cluster in tomato [32], the hydroxycinnamoyl-tyramine gene cluster [31], the 5,10-diketo-casbene gene cluster in rice and the dhurrin gene cluster (Figure 3A) [30, 56, 57]. Interestingly, combined with the analysis of metabolite biosynthesis pathway, it was found that the distribution order of the compact gene cluster members was roughly collinear with the reaction steps, revealing a new pattern for plant metabolic gene cluster assembly [49]. For example, the genes within the noscapine gene cluster in opium poppy could be roughly divided into three reaction sequence modules. The early module contains *CYP82Y1*, *PSMT3*, *CYP719A21* and *PSMT1*; the middle module contains *CYP82X1*, *CYP82X2*, *PSAT1* and *PSMT2*; the late module contains *PSSDR1* and *PSCXE1*, which exactly corresponds to sequentially genome organization in poppy [47]. Moreover, the organization of the avenacin cluster components appears to be broadly collinear with the order of the biosynthetic pathway on oat chromosome 1. Specifically, the gene encoding the first step *bAS1/Sad1* is located closest to the telomere and the late pathway genes including *CYP72A476*, *UGT91G16* and *TG1/Sad3* that are also required for avenacin biosynthesis are more distal to the telomere [49]. The authors proposed that placing *UGT91G16* and *TG1/ SAD3* genes farthest from the telomeres may be a gene arrangement strategy to mitigate the occurrence of toxin accumulation caused on telomere deletions. These examples show that collinearity of gene order and biosynthetic pathway reactions is quite common in compact pathways identified to date and may provide some insight into how gene clusters have evolved in response to natural selection.

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

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

Super gene clusters may be defined as different metabolic gene clusters coming together as hotspots in the genome. Recently, studies have shown that the gene cluster responsible for the synthesis of middle-chain acyl sugars in tomato is composed of tricyclic specific *Sl-AACS* (acyl-CoA synthase) and *Sl-AECH* (ethyl CoA hydrase) genes, which are closely arranged on chromosome 7 (Figure 3C). Interestingly, the organization of these genes, along with the sterol alkaloid gene cluster, form such a "super metabolic gene cluster" in tomato [43, 67]. Tomato steroid alkaloids and acylsugars both play defensive roles in plants, but are structurally distinct and stored in different tissues. Co-localization of these gene clusters may confer a selective advantage through an additive or synergistic effect of numerous defensive metabolites. Similarly, another recent study based on the obtained high-quality complete genome information, using plantiSMASH algorithm analysis and cluster density score revealed that the terminal 100 Mb region of chromosome 1 in *A. strigosa* genome is a gene cluster hotspot that contains a total of 19 putative gene clusters, of which 17 clusters include at least three co-expressed genes, where the avenacin gene cluster is located [49]. This poses the question why so many different gene clusters are grouped in the same locations, such as the sub-telomeric regions of eukaryotic genomes? Answering this question will require that we continue mining the metabolic gene clusters from a wider range of plant species and analyzing the effects of gene clusters in both evolutionary and ecological contexts.

Classes of Metabolites synthesized by gene clusters

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

Oxylipins

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

Terpenoids

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

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

Phenylpropanoids

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

Benzoxazinoids

A further set of widely distributed compounds – the Benzoxazinoids (Bxs), are a class of specialized metabolites that were discovered in the 1950’s in cereals [102]. Benzoxazines have been shown to be involved in a range of biological processes, such as defense against pathogens and resistance to insects [103, 104]. 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA) is the key defensive compound in maize (*Zea mays*) (Figure 4). As a representative Bxs, DIMBOA biosynthesis has been reported to be mediated by a metabolic gene cluster [35]. The complete biosynthetic pathway involves nine enzymes (*Bx1* to *Bx9*) which act sequentially in the synthesis of DIMBOA-glucoside from indole-3-glycerol phosphate. In the beginning, the 2,4-dihydroxy-1,4-benzoxazin-3-one (DIBOA) gene cluster was defined as a group of five genes (*Bx 1-5*). However, further experiments revealed that there are four additional genes (*Bx6-Bx9*) that are required for biosynthesis of DIMBOA [105-107]. Interestingly, this cluster is split in other plants of the Poaceae [108, 109]. For example, the cluster genes are split in two parts in wheat. One part (*Bx3*, *Bx4* and *Bx5*) of them is located on the short arm of chromosome 5 (A-, B- and D-genome), another part (an additional *Bx3* copy) was detected on the long arm of chromosome 5B [107]. Similar to the metabolites biosynthesized by already characterized plant metabolic gene clusters, both DIBOA and BIMBOA can confer the pathogen resistance and also contribute to defense against herbivores [109].

Alkaloids

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

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

Cyanogenic glucosides

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

Other metabolites

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

Regulation

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

Transcriptional regulation

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

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

The epigenetic regulation

Chromatin modification plays an important role in the regulation of gene clusters in plants. For instance, the chromatin mark of histone H3 lysine 27 trimethylation (H3K27me3) is associated with repression of cluster expression. On the contrary, the histone variant H2A.Z marks are associated with activation of cluster expression. As reported, two clusters in *A. thaliana* are associated with chromatin decondensation. These clustered pathways (thalianol and marneral clusters) are characterized by chromatin signatures of trimethylation of histone H3 lysine 27 (H3K27me3) [119, 120]. The expression levels of the thalianol and marneral cluster genes were altered in the CURLY LEAF (CLF) and PICKLE (PKL) mutants and these changes were restricted to the clusters and did not extend to the genes that directly flank the clusters [119, 120]. Besides, another exciting finding concerning the chromatin regulation of these two gene clusters is that they have also been positively regulated by the SWR1 chromatin remodeling complex [119, 121, 122]. Further study revealed that ARP6 is indispensable for the incorporation of H2A.Z into nucleosomes and its mutant can alter the expression of all four genes of the cluster [119]. Another interesting story of epigenetic regulation of plant gene clusters is a histone demethylase JMJ705 that can directly regulate genes from *DGC7* (a rice diterpenoid gene cluster) via methyl jasmonate-mediated epigenetic control [30]. Further research uncovered that this gene cluster is implicated in rice disease resistance [30, 46].

Natural Variation and Evolution

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

Apart from the structural variation, single nucleotide polymorphisms (SNPs) and small indels are also an important part of natural variations and have been identified in several different plant metabolic gene clusters. For instance, *Shen et al.*, suggest that the coordinated transcription of OsTyDC1 and OsTHT1 are influenced by natural variation and this may be a reason for the combination of genes for favorable traits [31]. Genomic co-linear analysis of wild and cultivated rice species shows that due to lack of the OsTyDC1 homologs, the *Oryza punctata* (BB genome lineage), *Oryza* *brachyantha* (FF genome lineage) have not formed the HT gene cluster. However, this cluster is conserved in the AA genome lineage. Unlike the HT gene cluster, the acylsugar gene cluster is missing or incomplete in most *Solanaceae* family species [67]. Another example is the steroidal glycoalkaloid gene cluster in tomato. A natural variation of a *Solyc10g085230* introduces a premature stop codon to this gene and this variation can reduce the steroidal glycoalkaloid content during ripening [44, 65, 123-127].

Concluding Remarks and Future Perspectives

DIMBOA is the first reported plant metabolic gene cluster, identified 24 years ago [35]. At that time no plant genomes had been published and scientists used a range of molecular biology cloning methods to identify genes associated with specific proteins. Although great achievements were made with these laborious approaches, they were low-throughput and focused on specific enzyme activities. The first reported complete sequence of a plant genome was that of *Arabidopsis thaliana* in 2000 [128]. This landmark event greatly accelerated the process of functional annotation of plant genes. The resulting gain in genome-level information sparked a rapid development period for research on plant metabolic gene clusters. During the last decade, the advent of next-generation sequencing and the development of multi-omics technologies greatly improved our ability to identify and dissect metabolic gene clusters in plants. Many aspects of the research of plant metabolic gene clusters have been considerably expanded in this period, providing insights on biosynthetic genes and regulatory genes [42], transcriptional regulation and epigenetic regulation [30, 119], and secondary metabolism and primary metabolism [32]. However, the current rate of discovery of plant metabolic gene clusters suggests that our catalog is far from complete. Furthermore, in addition to existing tools such as genomics, transcriptomics, metabolomics, epigenomics, proteomics, phenomics, next-generation sequencing and advanced bioinformatics an ever-increasing arsenal of tools is being used to crack the mysteries of plant metabolic gene clusters (Figure 7) [129-135]. Applying the advances in artificial intelligence (AI) are also worth considering. One of the primary means in AI is deep learning which has been applied already in different fields related to plant science. For instance, to improve the accuracy of protein 3D structure prediction [136]. In the foreseeable future, we believe that AI will also play an important role in the research of plant metabolic gene cluster.

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

Acknowledgments

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

References

1. Sumner, L.W. et al. (2015) Modern plant metabolomics: advanced natural product gene discoveries, improved technologies, and future prospects. Nat. Prod. Rep. 32 (2), 212-229.

2. Fernie, A.R. and Tohge, T. (2017) The genetics of plant metabolism. Annu. Rev. Genet. 51 (1), 287-310.

3. Lacchini, E. and Goossens, A. (2020) Combinatorial control of plant specialized metabolism: mechanisms, functions, and consequences. Annu. Rev. Cell Dev. Biol. 36, 291-313.

4. Dixon, R.A. and Strack, D. (2003) Phytochemistry meets genome analysis, and beyond. Phytochemistry 62 (6), 815-6.

5. Venegas-Molina, J. et al. (2021) Why and how to dig into plant metabolite–protein interactions. Trends Plant Sci. 26, 472-483.

6. Boutanaev, A.M. et al. (2015) Investigation of terpene diversification across multiple sequenced plant genomes. Proc. Natl. Acad. Sci. U. S. A. 112 (1), E81-E88.

7. Jacobowitz, J.R. and Weng, J.-K. (2020) Exploring uncharted territories of plant specialized metabolism in the postgenomic era. Annul. Rev. Plant Biol. 71 (1), 631-658.

8. Gülck, T. and Møller, B.L. (2020) Phytocannabinoids: origins and biosynthesis. Trends Plant Sci. 25 (10), 985-1004.

9. Schmelz, E.A. and Tumlinson, J.H. (2011) Identity, regulation, and activity of inducible diterpenoid phytoalexins in maize. Proc. Natl. Acad. Sci. U. S. A. 108 (13), 5455-60.

10. Luo, D. et al. (2016) Oxidation and cyclization of casbene in the biosynthesis of Euphorbia factors from mature seeds of Euphorbia lathyris L. Proc. Natl. Acad. Sci. U. S. A. 113 (34), E5082-E5089.

11. Erb, M. and Kliebenstein, D.J. (2020) Plant secondary metabolites as defenses, regulators, and primary metabolites: the blurred functional trichotomy. Plant Physiol. 184 (1), 39-52.

12. Chen, F. et al. (2011) The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant J. 66 (1), 212-229.

13. Tian, X. et al. (2018) Characterization of gossypol biosynthetic pathway. Proc. Natl. Acad. Sci. U. S. A. 115 (23), E5410-E5418.

14. Yang, D. et al. (2014) Transcriptomics, proteomics, and metabolomics to reveal mechanisms underlying plant secondary metabolism. Eng. Life Sci. 14 (5), 456-466.

15. Fang, C. and Luo, J. (2019) Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. Plant J. 97 (1), 91-100.

16. Butelli, E. et al. (2008) Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors. Nat. Biotechnol. 26 (11), 1301-1308.

17. Peng, M. et al. (2017) Differentially evolved glucosyltransferases determine natural variation of rice flavone accumulation and UV-tolerance. Nat. Commun. 8 (1), 1975.

18. Li, Y. et al. (2021) Benefiting others and self: Production of vitamins in plants. J. Integr. Plant Biol. 63 (1), 210-227.

19. Naoumkina, M.A. et al. (2010) Genomic and coexpression analyses predict multiple genes involved in triterpene saponin biosynthesis in Medicago truncatula. Plant Cell 22 (3), 850-866.

20. Fernie, A.R. and Schauer, N. (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet. 25 (1), 39-48.

21. Wang, W. et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557 (7703), 43-49.

22. Huang, X. et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature 490, 497.

23. Qin, P. et al. (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell 184, 3542-3558.

24. Su, W. et al. (2021) Polyploidy underlies co-option and diversification of biosynthetic triterpene pathways in the apple tribe. Proc. Natl. Acad. Sci. U. S. A. 118 (20), e2101767118.

25. Shi, T. et al. (2020) Metabolomics analysis and metabolite-agronomic trait associations using kernels of wheat (Triticum aestivum) recombinant inbred lines. Plant J. n/a (n/a).

26. Chen, J. et al. (2020) Metabolite-based genome-wide association study enables dissection of the flavonoid decoration pathway of wheat kernels. Plant Biotechnol J. n/a (n/a), n/a.

27. Boycheva, S. et al. (2014) The rise of operon-like gene clusters in plants. Trends Plant Sci. 19 (7), 447-459.

28. Okada, A. et al. (2009) OsTGAP1, a bZIP transcription factor, coordinately regulates the inductive production of diterpenoid phytoalexins in rice. J. Biol. Chem. 284 (39), 26510-8.

29. Nützmann, H.-W. et al. (2016) Plant metabolic clusters – from genetics to genomics. New Phytol. 211 (3), 771-789.

30. Zhan, C. et al. (2020) Selection of a subspecies-specific diterpene gene cluster implicated in rice disease resistance. Nat. Plants 6 (12), 1447-1454.

31. Shen, S. et al. (2021) An Oryza-specific hydroxycinnamoyl tyramine gene cluster contributes to enhanced disease resistance. Sci. Bull. 66 (23), 2369-2380.

32. Jeon, J.E. et al. (2020) A pathogen-responsive gene cluster for highly modified fatty acids in tomato. Cell 180 (1), 176-187.

33. Price, M.N. et al. (2006) The life-cycle of operons. PLoS Genet. 2 (6), e96.

34. Bratlie, M.S. et al. (2010) Relationship between operon preference and functional properties of persistent genes in bacterial genomes. BMC Genom. 11 (1), 71.

35. Frey, M. et al. (1997) Analysis of a chemical plant defense mechanism in grasses. Science 277 (5326),

696-699.

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

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

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

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

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

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

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

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

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

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

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

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

48. Qi, X. et al. (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 (21), 8233-8.

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

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

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

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

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

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

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

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

Sorghum bicolor contains its co-expressed vacuolar MATE transporter. Sci. Rep. 6 (1), 37079.

57. Laursen, T. et al. (2016) Characterization of a dynamic metabolon producing the defense compound dhurrin in sorghum. Science 354 (6314), 890-893.

58. Zhang, Y. and Fernie, A.R. (2021) Metabolons, enzyme–enzyme assemblies that mediate substrate channeling, and their roles in plant metabolism. Plant Commun. 2 (1), 100081.

59. Sweetlove, L.J. and Fernie, A.R. (2018) The role of dynamic enzyme assemblies and substrate channelling in metabolic regulation. Nat. Commun. 9 (1), 2136.

60. Zhang, Y. et al. (2017) Protein-protein interactions and metabolite channelling in the plant tricarboxylic acid cycle. Nat. Commun. 8 (1), 15212.

61. Zhang, Y. et al. (2020) A moonlighting role for enzymes of glycolysis in the co-localization of mitochondria and chloroplasts. Nat. Commun. 11 (1), 4509.

62. Takos, A.M. 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 (2), 273-286.

63. Takos, A. et al. (2010) Genetic screening identifies cyanogenesis-deficient mutants of Lotus japonicus and reveals enzymatic specificity in hydroxynitrile glucoside metabolism. Plant Cell 22 (5), 1605-1619.

64. Winzer, T. et al. (2015) Morphinan biosynthesis in opium poppy requires a P450-oxidoreductase fusion protein. Science 349 (6245), 309-312.

65. Guo, L. et al. (2018) The opium poppy genome and morphinan production. Science 362 (6412), 343-347.

66. Zhou, Y. et al. (2016) Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. Nat. Plants 2 (12), 16183.

67. Fan, P. et al. (2020) Evolution of a plant gene cluster in Solanaceae and emergence of metabolic diversity. eLife 9, e56717.

68. Jin, H.R. et al. (2012) The antitumor natural compound falcarindiol promotes cancer cell death by inducing endoplasmic reticulum stress. Cell Death Dis. 3 (8), e376-e376.

69. Miyazawa, M. et al. (1996) Antimutagenic activity of falcarindiol from Peucedanum praeruptorum. J. Agric. Food Chem. 44 (11), 3444-3448.

70. Villegas, M. et al. (1988) Isolation of the antifungal compounds falcarindiol and sarisan from Heteromorpha trifoliata. Planta Med. 54 (1), 36-37.

71. Kato, T. et al. (1973) Momilactones, growth inhibitors from rice, oryza sativa L. Tetrahedron Lett. 14 (39), 3861-3864.

72. Cartwright, D. et al. (1977) Chemical activation of host defence mechanisms as a basis for crop protection. Nature 267 (5611), 511-513.

73. Cartwright, D.W. et al. (1981) Isolation and characterization of two phytoalexins from rice as momilactones A and B. Phytochemistry 20 (3), 535-537.

74. Kitaoka, N. et al. (2020) Interdependent evolution of biosynthetic gene clusters for momilactone production in rice. Plant Cell 33, 290-305.

75. Koga, J. et al. (1995) Phytocassanes A, B, C and D, novel diterpene phytoalexins from rice, Oryza sativa L. Tetrahedron 51 (29), 7907-7918.

76. Koga, J. et al. (1997) Functional moiety for the antifungal activity of phytocassane E, a diterpene phytoalexin from rice. Phytochemistry 44 (2), 249-253.

77. Akatsuka, T. et al. (1985) Novel phytoalexins (oryzalexins A, B and C) isolated from rice blast leaves

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

150. Swaminathan, S. et al. (2009) CYP76M7 is an ent-cassadiene C11α-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. Plant Cell 21 (10), 3315-3325.

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

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

153. Sohrabi, R. et al. (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 (3), 874-890.

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

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

Glossary

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

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

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

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

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

Figure Legends

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

Figure 2. Strategies for plant gene cluster identification. (A) Identification of the plant gene cluster through genome-wide association studies. (B) Identification of the plant gene cluster through quantitative trait loci (QTLs). (C) Identification of the plant gene cluster through special algorithms. (D) Identification of the plant gene cluster through genome mining. (E) Identification of the plant gene cluster through the combination analysis of omics data. The figures are modified from refs[31,49,80,118].

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

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

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

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

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

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

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

Outstanding Questions

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

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

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

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

Figure 1

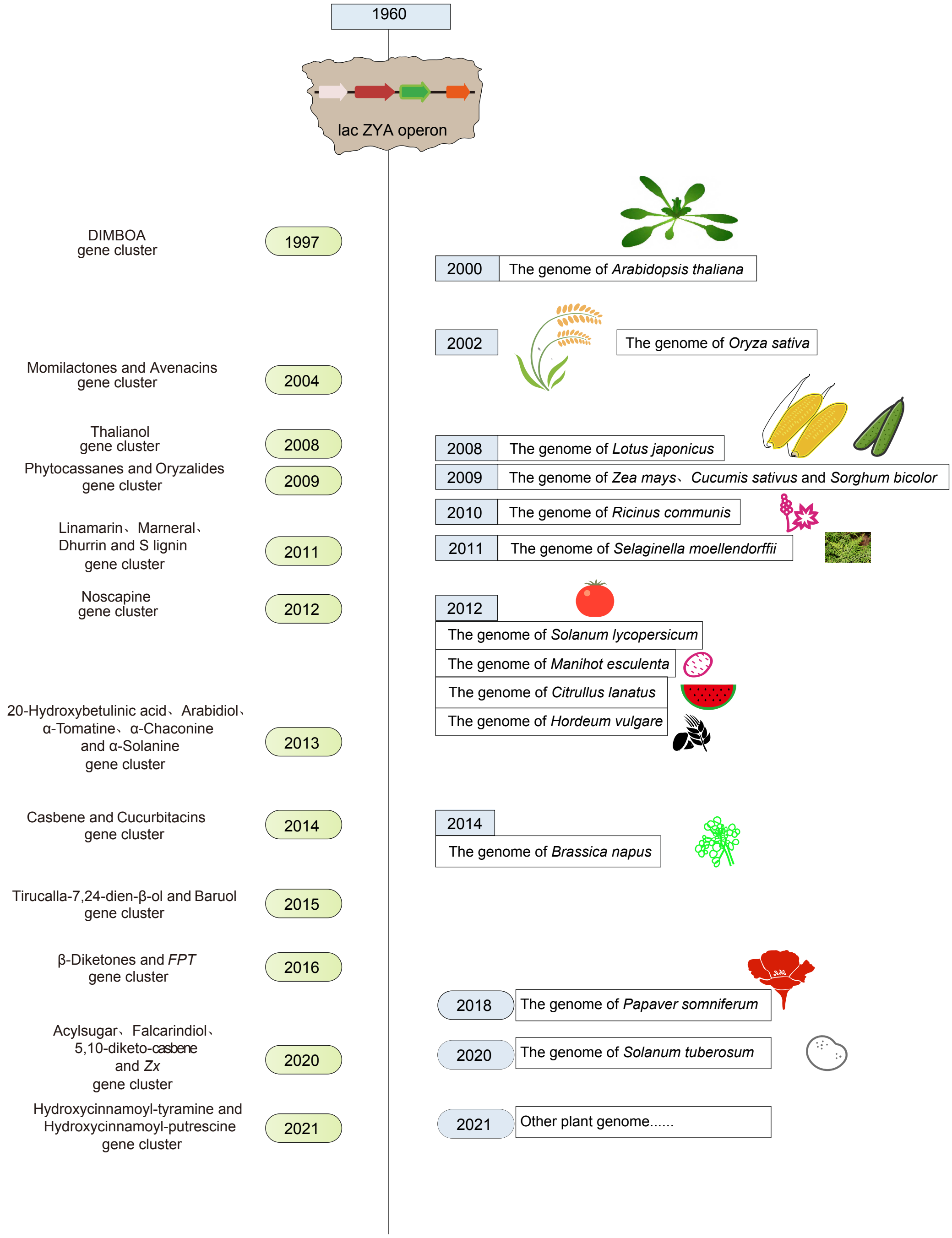

1960
lac ZYA operon
1997
DIMBOA gene cluster
2000 The genome of *Arabidopsis thaliana*
2002 The genome of *Oryza sativa*
2004
Momilactones and Avenacins gene cluster
2008
Thalianol gene cluster
2009
Phytocassanes and Oryzalides gene cluster
2008 The genome of *Lotus japonicus*
2009 The genome of *Zea mays*、*Cucumis sativus* and *Sorghum bicolor*
2010 The genome of *Ricinus communis*
2011 The genome of *Selaginella moellendorffii*
2011
Linamarin、Marneral、Dhurrin and S lignin gene cluster
2012
Noscapine gene cluster
2012
The genome of *Solanum lycopersicum*
The genome of *Manihot esculenta*
The genome of *Citrullus lanatus*
The genome of *Hordeum vulgare*
2013
20-Hydroxybetulinic acid、Arabidiol、α-Tomatine、α-Chaconine and α-Solanine gene cluster
2014
Casbene and Cucurbitacins gene cluster
2014
The genome of *Brassica napus*
2015
Tirucalla-7,24-dien-β-ol and Baruol gene cluster
2016
β-Diketones and *FPT* gene cluster
2018 The genome of *Papaver somniferum*
2020
Acylsugar、Falcarindiol、5,10-diketo-casbene and *Zx* gene cluster
2020 The genome of *Solanum tuberosum*
2021
Hydroxycinnamoyl-tyramine and Hydroxycinnamoyl-putrescine gene cluster
2021 Other plant genome......

Figure 2

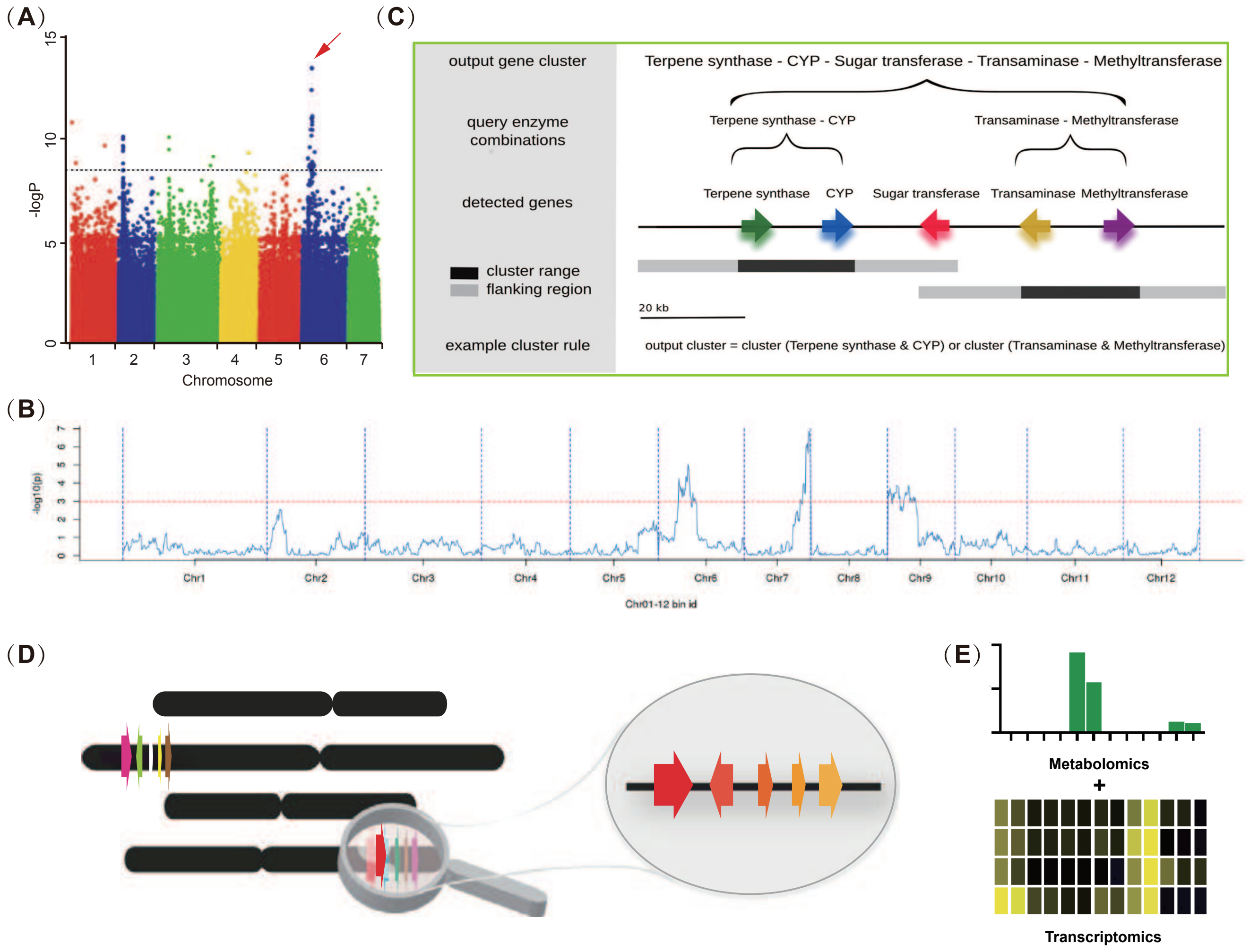
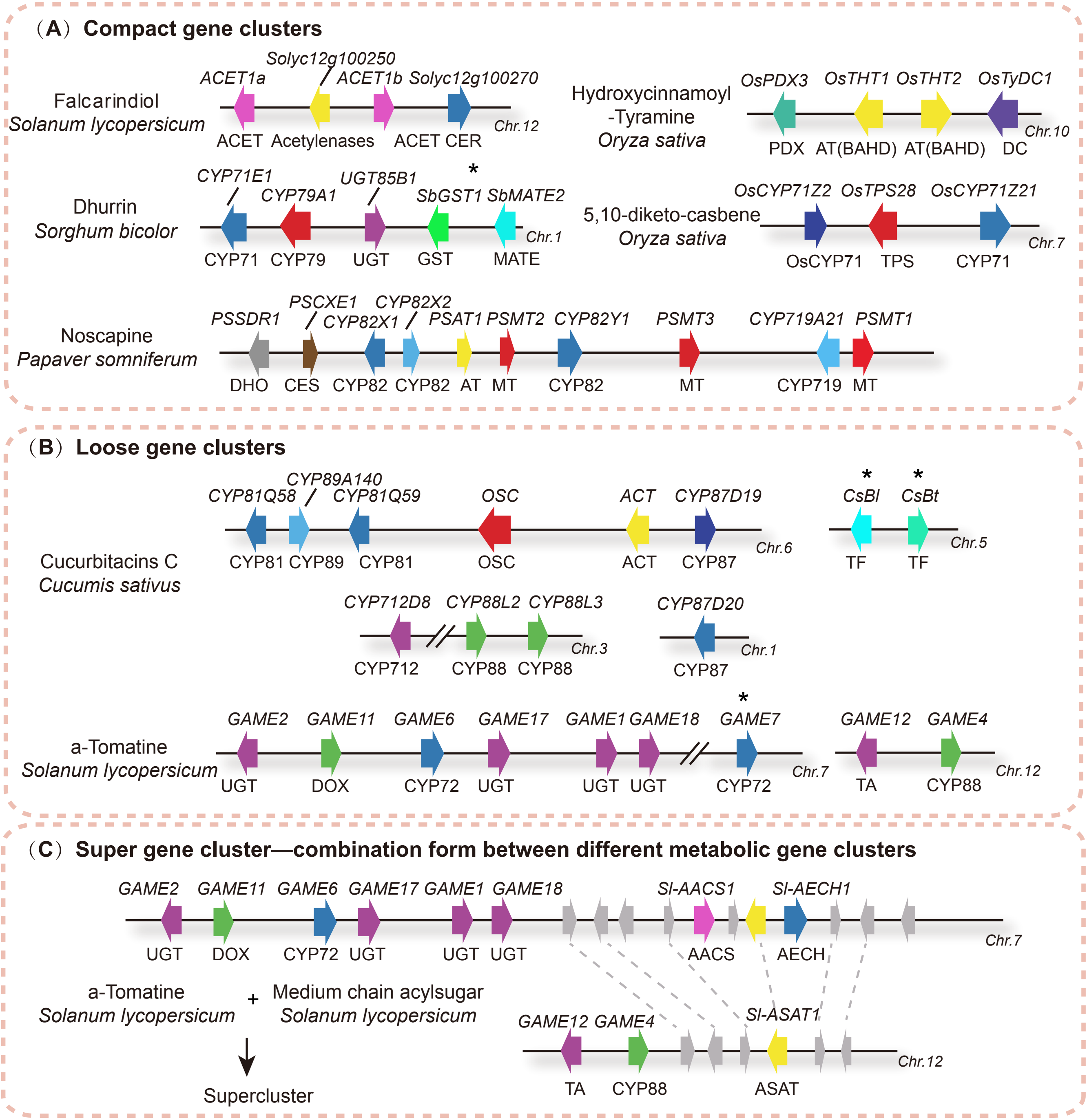
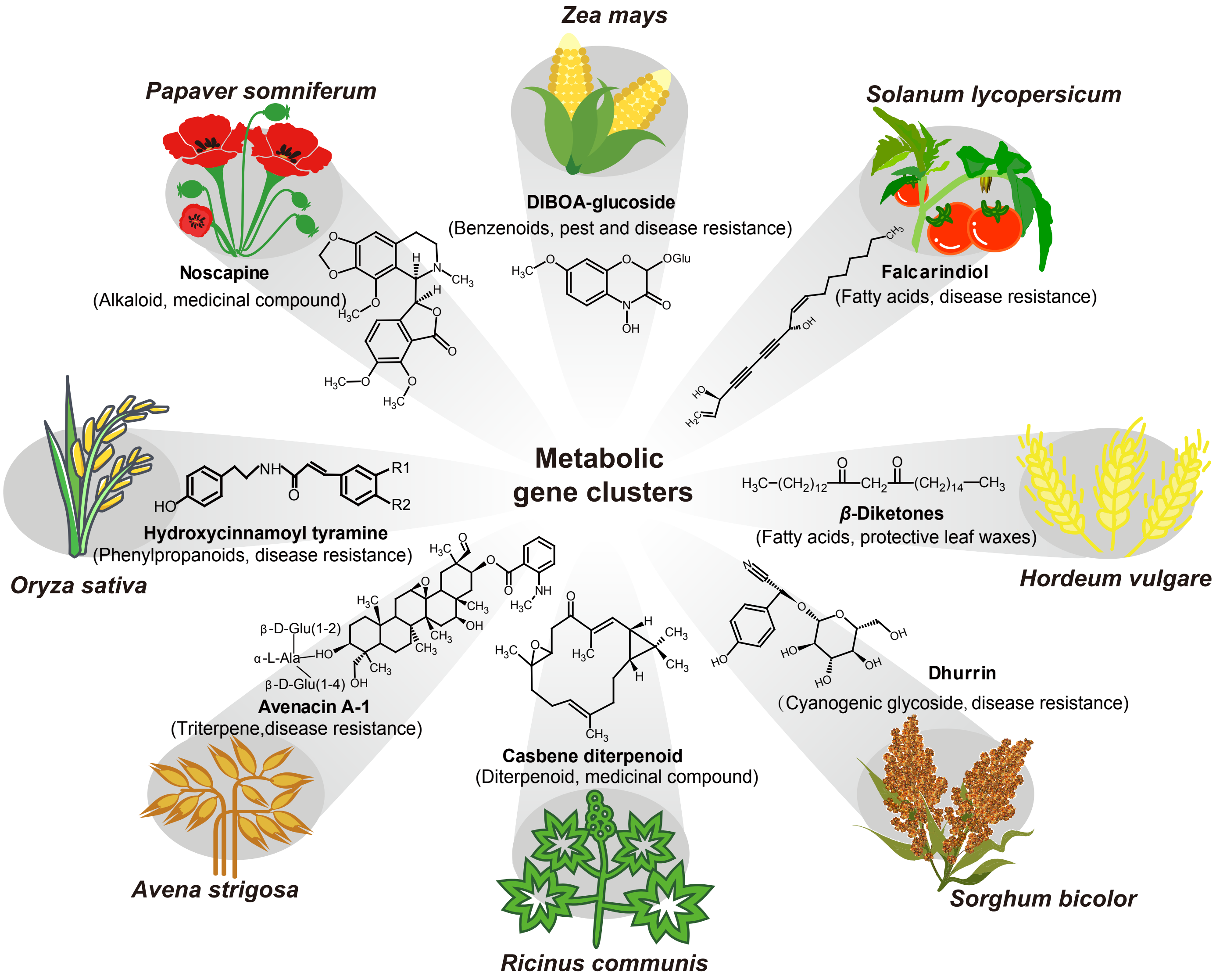

(A)
15
10
5
0
-logP
1
2
3
4
5
6
7
Chromosome
(C)
output gene cluster
Terpene synthase - CYP - Sugar transferase - Transaminase - Methyltransferase
query enzyme combinations
Terpene synthase - CYP
Transaminase - Methyltransferase
detected genes
Terpene synthase
CYP
Sugar transferase
Transaminase
Methyltransferase
cluster range
flanking region
20 kb
example cluster rule
output cluster = cluster (Terpene synthase & CYP) or cluster (Transaminase & Methyltransferase)
(B)
-log10(p)
0
1
2
3
4
5
6
7
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr01-12 bin id
(D)
(E)
Metabolomics
+
Transcriptomics

Figure 3



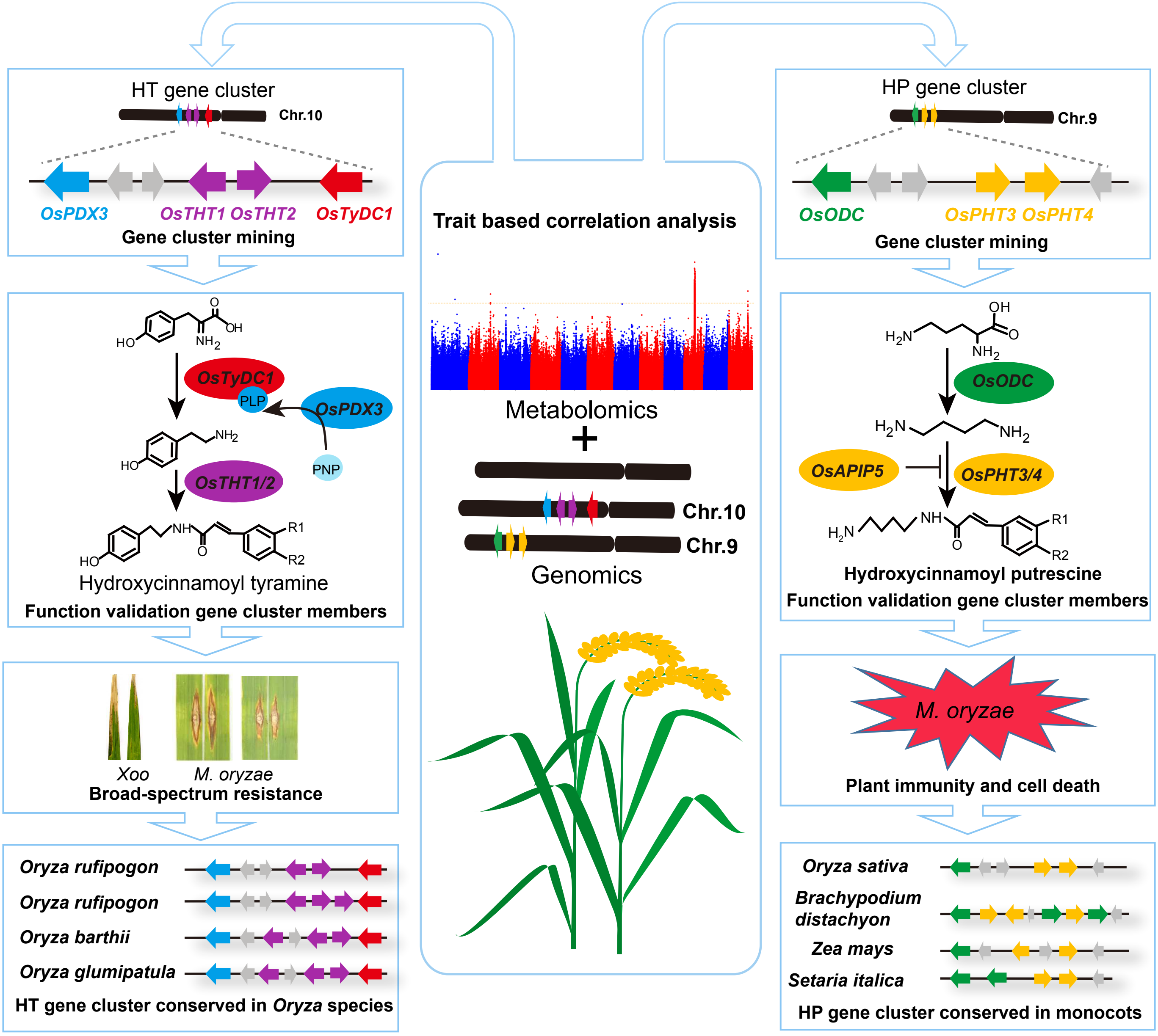
（A） Compact gene clusters
Falcarindiol
Solanum lycopersicum
ACET1a
Solyc12g100250
ACET1b
Solyc12g100270
ACET
Acetylenases
ACET
CER
Chr.12
Hydroxycinnamoyl
-Tyramine
Oryza sativa
OsPDX3
OsTHT1
OsTHT2
OsTyDC1
PDX
AT(BAHD)
AT(BAHD)
DC
Chr.10
Dhurrin
Sorghum bicolor
CYP71E1
CYP79A1
UGT85B1
*
SbGST1
SbMATE2
CYP71
CYP79
UGT
GST
MATE
Chr.1
5,10-diketo-casbene
Oryza sativa
OsCYP71Z2
OsTPS28
OsCYP71Z21
OsCYP71
TPS
CYP71
Chr.7
Noscapine
Papaver somniferum
PSSDR1
PSCXE1
CYP82X1
CYP82X2
PSAT1
PSMT2
CYP82Y1
PSMT3
CYP719A21
PSMT1
DHO
CES
CYP82
CYP82
AT
MT
CYP82
MT
CYP719
MT
（B） Loose gene clusters
Cucurbitacins C
Cucumis sativus
CYP81Q58
CYP89A140
CYP81Q59
OSC
ACT
CYP87D19
CYP81
CYP89
CYP81
OSC
ACT
CYP87
Chr.6
*
*
CsBl
CsBt
TF
TF
Chr.5
CYP712D8
CYP88L2
CYP88L3
CYP712
CYP88
CYP88
Chr.3
CYP87D20
CYP87
Chr.1
a-Tomatine
Solanum lycopersicum
GAME2
GAME11
GAME6
GAME17
GAME1
GAME18
*
GAME7
UGT
DOX
CYP72
UGT
UGT
UGT
CYP72
Chr.7
GAME12
GAME4
TA
CYP88
Chr.12
（C） Super gene cluster—combination form between different metabolic gene clusters
GAME2
GAME11
GAME6
GAME17
GAME1
GAME18
Sl-AACS1
Sl-AECH1
UGT
DOX
CYP72
UGT
UGT
UGT
AACS
AECH
Chr.7
a-Tomatine
Solanum lycopersicum
+
Medium chain acylsugar
Solanum lycopersicum
Supercluster
GAME12
GAME4
Sl-ASAT1
TA
CYP88
ASAT
Chr.12

Figure 4



Zea mays
DIBOA-glucoside
(Benzenoids, pest and disease resistance)
Papaver somniferum
Noscapine
(Alkaloid, medicinal compound)
Solanum lycopersicum
Falcarindiol
(Fatty acids, disease resistance)
Metabolic gene clusters
Hydroxycinnamoyl tyramine
(Phenylpropanoids, disease resistance)
Oryza sativa
β-Diketones
(Fatty acids, protective leaf waxes)
Hordeum vulgare
Avenacin A-1
(Triterpene,disease resistance)
Avena strigosa
Casbene diterpenoid
(Diterpenoid, medicinal compound)
Ricinus communis
Dhurrin
(Cyanogenic glycoside, disease resistance)
Sorghum bicolor

Figure 5



HT gene cluster
Chr.10
OsPDX3
OsTHT1 OsTHT2
OsTyDC1
Gene cluster mining
HO
O
OH
NH2
OsTyDC1
PLP
OsPDX3
PNP
NH2
OsTHT1/2
NH
R1
R2
Hydroxycinnamoyl tyramine
Function validation gene cluster members
Xoo
M. oryzae
Broad-spectrum resistance
Oryza rufipogon
Oryza rufipogon
Oryza barthii
Oryza glumipatula
HT gene cluster conserved in Oryza species
Trait based correlation analysis
Metabolomics
+
Chr.10
Chr.9
Genomics
HP gene cluster
Chr.9
OsODC
OsPHT3 OsPHT4
Gene cluster mining
H2N
OH
O
NH2
OsODC
OsAPIP5
OsPHT3/4
Hydroxycinnamoyl putrescine
Function validation gene cluster members
M. oryzae
Plant immunity and cell death
Oryza sativa
Brachypodium distachyon
Zea mays
Setaria italica
HP gene cluster conserved in monocots

Figure 6

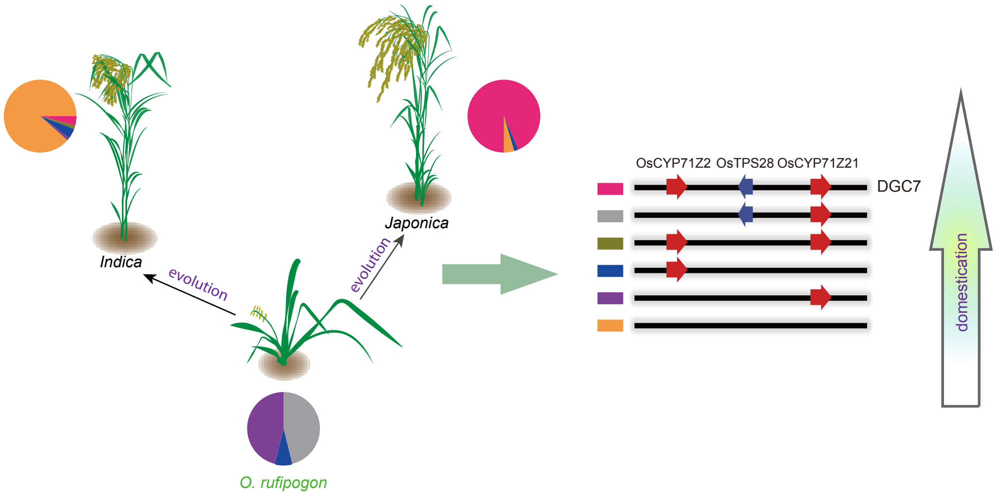
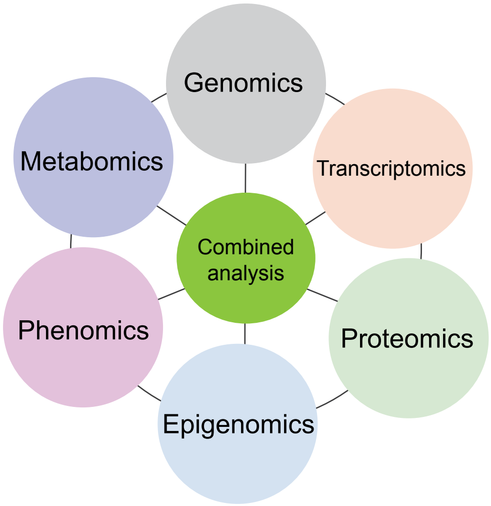

Indica
Japonica
evolution
evolution
O. rufipogon
OsCYP71Z2
OsTPS28
OsCYP71Z21
DGC7
domestication

Figure 7


Genomics
Metabomics
Transcriptomics
Combined analysis
Phenomics
Proteomics
Epigenomics