The evolutionary dynamics of secondary metabolite gene clusters in plant model species

Yosapol Harnvanichvech[1,*], **Hernando Suarez-Duran**[1], **and Marnix Medema** [1]

[1]Bioinformatics Group, Wageningen University, 6708 PB, Wageningen, The Netherlands
[*]corresponding: yosapol.harnvanichvech@wur.nl

ABSTRACT

Triterpenes are one of the most structurally diverse plant natural products. They play important roles in ecological and survival strategies, contributing to protection against biotic and abiotic stresses. To date, the experimental characterization has shown that the diverse compounds of triterpene in plants were synthesized from groups of chromosomally clustered enzymatic genes, known as biosynthetic gene clusters (BGCs). However, how gene clusters are formed, and how plants make highly-diverged gene clusters is merely understood. Our analyses reveal an evolutionary insight into triterpene clusters by comparison with 14 genomes across the Brassicaceae family and 3 genomes from closely related families to Brassicaceae. We predicted a total of 158 triterpene clusters with more than 100 novel triterpene cluster candidates, which allowed for a chance to study the evolutionary dynamics. Our evolutionary analysis and phylogenetic reconstruction showed that the number of triterpene biosynthetic gene clusters have no correlation to the genome size. We further show that the triterpene clusters are more likely to be evolved from tandem array duplication rather than whole-genome duplication events. We first reveal that the putatively functional triterpene cluster is most likely to be formed by the Oxidosqualene cyclases (OSCs) genes that particularly are found in an early diverging OSCs domain. We highlight that the triterpene clusters are rapidly evolving even among closely related species, which is in contrast to the feature of prokaryotic gene clusters. We further identified that biosynthetic genes in the triterpene clusters are evolutionarily more conserved compared to non-biosynthetic genes, indicating a sign of strong purifying selection. Collectively, our work provides a new footprint of the diversification of the triterpene clusters in plants.

Keywords: Biosynthetic gene clusters (BGCs), Triterpene clusters, Brassicaceae family

INTRODUCTION

Plant secondary metabolites (also referred to as "specialized metabolites") are well-known to be synthesized for plant survival strategies[1,2]. These compounds typically play a role in the interaction of their own and environment, and may also be involved in the developmental processes[3,4]. In bacteria and fungi, secondary metabolites compounds are often produced from a group of enzymatic genes physically clustered in the chromosome, called biosynthetic gene clusters (BGCs)[5]. These BGCs contain enzyme-coding genes necessary for the synthesis of the corresponding metabolite's core structure and side-chain modifications that introduce specificity to the molecule[6,7]. In plants, it has not always been clear how secondary metabolite BGCs are related to secondary metabolic pathways. Recently, the first BGC in plants has been found in Maize, synthesizing a cyclic hydroxamic acid 2,4-dihydroxy-1,4-benzoxazin-3- one (DIBOA) compound[8]. This was considered as a successive step to opening the door to find an origin of a large community of plants secondary metabolites. Further investigation of plants' BGCs were described in *Avena* spp. for avenacin[9], *Oryza sativa* for phytocassane[10], and *Arabidopsis thaliana* for thalianol[11]; in total, up to twenty of biosynthetic gene cluster from different plant species have been reported at the time of writing[4]. It is clear that the number of identified plant BGCs is rapidly growing, suggesting that these forms of gene organization for synthesis of secondary metabolite compounds in plants may be common[12].

Among a wide variety of secondary metabolite compounds, one of the largest and most diverse groups is constituted by Triterpenes, with more than 20,000 reported to date[13]. In triterpene biosynthesis, all triterpene compounds are derived from a simple linear isoprenoid substrate 2,3-oxidosqualene. These substrate is then cyclized by oxidosqualene cyclase enzymes (OSCs) to generate a triterpene scaffold. This cyclization step from OSC enzymatic reaction is the first committed step of the triterpene biosynthetic pathway[14–16]. Currently, over a 100 different triterpene scaffolds converted from OSCs are known from diverse plant species[17]. In addition, the resulting triterpene scaffolds are often followed by a concerted series of modification steps using tailoring enzymes, such as cytochrome P450s, sugar transferases, and acyltransferases[14,18]. This results in a massive structural diversification of the triterpene compounds. Until now, many triterpene BGCs in plants have already been identified in several species, for example those for the biosynthesis of thalianol and marneral in *Arabidopsis thaliana*[11] , avenacin A-1 in *Avena* spp.[9], and a triterpene cluster in *Lotus japonicus*[19]. However, this is a small percentage from the number of triterpene structures that have been identified in plants, and insight into the formation of triterpene BGCs and their evolutionary relationships in plants lineages are still unknown[20].

Brassicaceae (Cruciferae) is one of the most important plant families for human life: it comprises over 3,700 species including model species in experimental biology (e.g., *Arabidopsis thaliana*) and economic crops (e.g., Cauliflower, Arabis, Broccoli, Cabbages)[21]. This family is referred to as an outstanding evolutionary model in plant family from two main reasons. First, the family contains extensive genome resources, with over 20 publicly available genome sequences from GenBank database (`http://www.ncbi.nlm.nih.gov/genomes/`); and second, the model plant *A. thaliana* provide an impressive information of gene annotation[22]. In addition, Brassicaceae experienced the most recent whole-genome duplications, called alpha duplication, which open up opportunities to explore its early evolutionary history[23,24].

Here, we integrate genomic and phylogenetic approaches to study the evolutionary dynamics of the triterpenes clusters from seventeen plant genomes (14 of Brassicaceae family and 3 additional genomes from others families). We show a clear evidence of how triterpene clusters have been formed. We first present the new evidence of rapid evolution of triterpene clusters, which have been found even in closely related species. We propose that both biosynthetic genes and non-biosynthetic genes are influenced to rapid evolution of the triterpene clusters. In addition, we reveal that the biosynthetic genes show more evolutionary conservation which can be interpreted as a sign of stronger selection, compared to non-biosynthetic genes

RESULTS AND DISCUSSION

Identification of triterpene gene clusters

To uncover the evolution dynamics of triterpene gene clusters, we first identified the triterpene gene clusters candidate using automated pipeline, plantiSMASH. As a result, we found a total of 158 triterpene clusters with more than 100 novel triterpene cluster candidates from fourteen genomes of Brassicaceae family and three additional plant genomes from other families, which originated in ancient duplication events. The number of detected triterpene clusters per genome ranged from 3 in *Carrica papaya* to 18 in *Brassica juncea* as seen in Supplementary Table 1. To see if the number of triterpene clusters was correlated with the genome size, the ratio of number of clusters to genome size were calculated. The result showed that there was no strong correlation between genome size and the number of triterpene clusters among taxa or lineages with the R-value of 0.065 (Figure 1A). We also compared this ratio among the taxa from the different whole genome duplication events and found that the ratios were not different among the different duplication events as shown in Figure 1B. This may suggest that the number of triterpene gene clusters in Brassicaceae family were evolved independently from whole genome duplication events. However, when considering only within the Brassicaceae clade, we found that the averaged values of plants from the most recent Brassicaceae lineage, Brassicaceae lineage I, had the higher average ratio (0.05) than the earlier diverging Brassicaceae lineage, (lineage II at 0.02) (Figure 1B). These may suggest that the speed of triterpene cluster formation may be specific to the different Brassicaceae lineage.

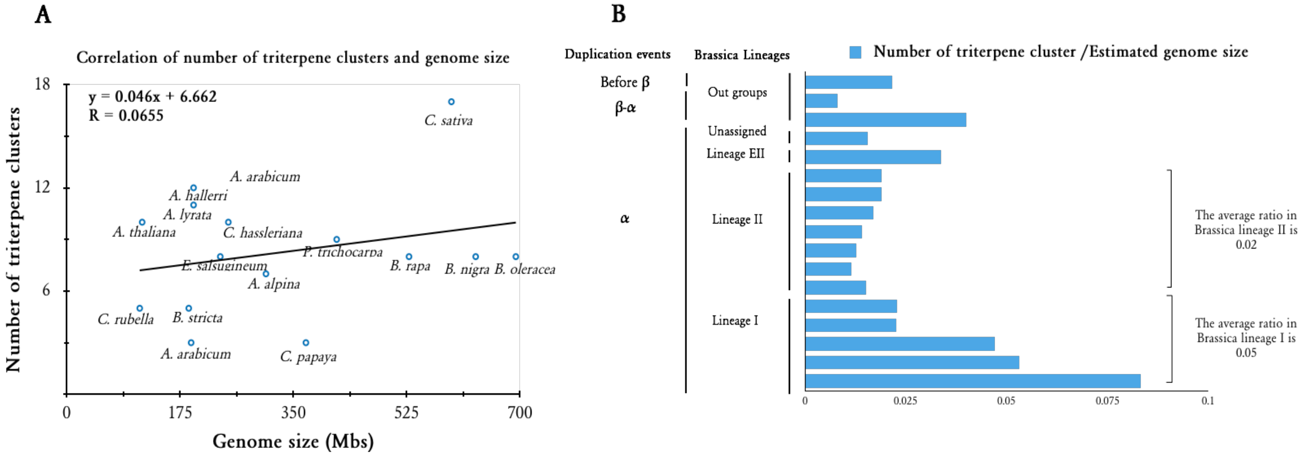

Figure 1. **(A)** Graph showing the correlation between the number of triterpene (Y axis) and genome size (X axis). **(B)** Bar chart show the average ratio of Brassicaceae lineage I had the higher average ratio (0.05) than the earlier diverging Brassicaceae lineage II (0.02).

Triterpene clusters are likely to be formed with the recently derived OSCs domain

The reaction catalyzed by OSCs defines the first committed step to determine triterpene biosynthesis[14]. To understand the evolutionary origin of triterpene biosynthetic gene clusters, the architecture of all triterpene gene clusters, including cluster-less genomic neighborhoods of OSC genes, were mapped onto the phylogeny of OSCs domains from seventeen plant genomes. From the phylogenetic reconstruction, we first spotted that the composition of triterpene clusters in closely related species are similar, but not identical, as shown in Figure 2B. Unexpectedly, we cannot observe syntenic triterpene clusters in closely related species, suggesting that the triterpene clusters were more likely formed by tandem duplication[20] rather than by whole genome duplication events. We further found that the accessions in the late diverging groups tend to have OSCs genes in a close physical distance (Figure 2C), whereas the accessions in the more recently derived group often have individual OSCs genes or multiple OSCs genes farther apart from each other (Figure 2A). We suggested that, in the early-diverging clade, a close physical distance of OSCs genes was associated to fewer genes that encode triterpene modifying enzymes in the triterpene clusters. The closed physical distance would produce excessive amount of intermediates, such as 2,3-oxidosqualene, which can be used in alternative pathways instead of triterpene biosynthesis[25,26]. This overproduction of these intermediate compounds would result to a negative catalysis, which may be the driving force of OSCs genes in order to recruit the triterpene modifying enzymes. Meanwhile, we further highlighted that OSC genes in the recently diverging groups are mostly clustered with at least three triterpene-modifying genes, which are considered to be a putatively functional triterpene cluster[5].

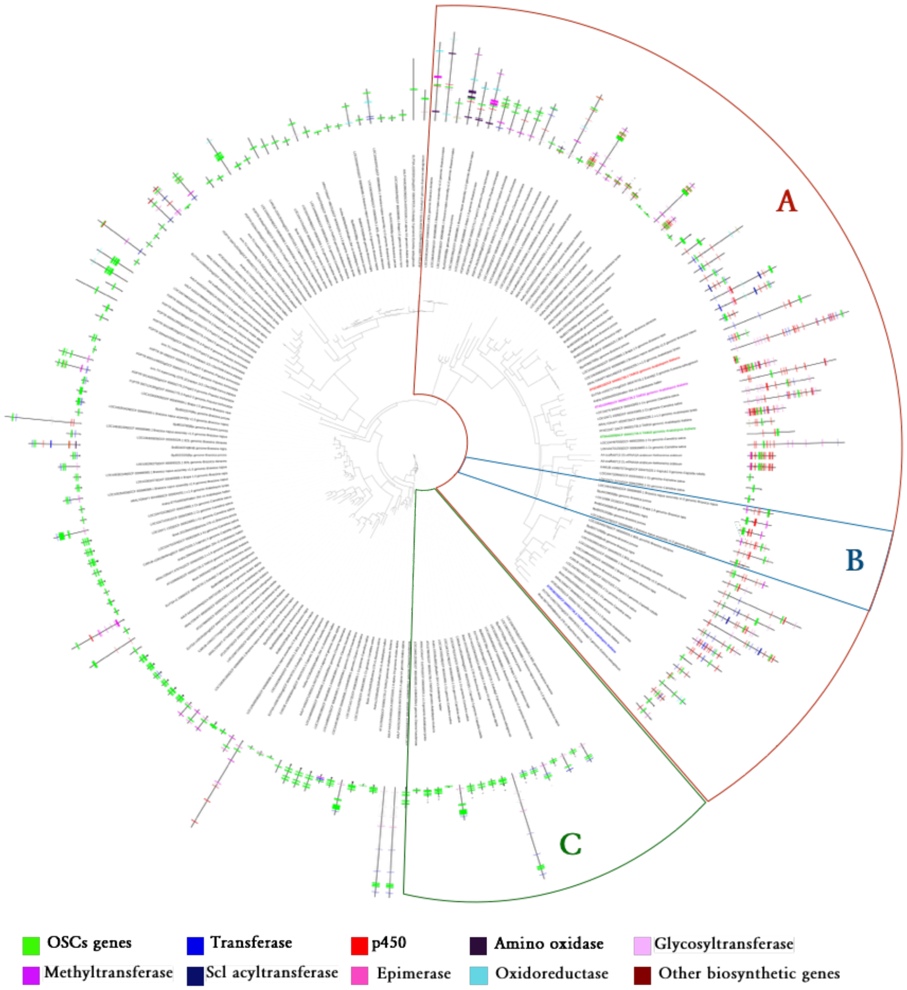

Figure 2. Phylogenetic relationships of OSCs domain mapped to biosynthetic genes from seventeen plant genomes. **(A)** The clade delimited in blue lines indicates highly similar gene clusters from closely related species, but no identical pairs were found. **(B)** The composition of the genomic neighborhoods surrounded by the red lines shows the early diversifying of OSCs domain are most likely to recruit the triterpene modifying enzymatic genes. In contrast to the clade surrounded by the green lines. **(C)** The composition of the genomic neighborhoods shows the late diversifying of OSCs domain are rarely to be formed with the triterpene modifying enzymatic genes.

Triterpene clusters in Brassicaceae are rapidly evolving

To further understand how triterpene clusters diverged in Brassicaceae, core biosynthetic genes of the earliest diverging clade were used to build a dissimilarity index heatmap of gene clusters. The dissimilarity values were ranged from 0 to 1, which indicates low-dissimilarity to high-dissimilarity respectively. Overall, the heatmap of the dissimilarity index (Figure 3) showed a high diversification of gene clusters among taxa, with a slightly lower value among the closely related species. The example can be observed in *Brassica* spp. and *Arabidopsis* spp. In the *Brassica* spp. (including *B. napus*, *B. juncea*. *B. rapa*, *B. nigra* and *B. oleracea*), seven triterpene clusters showed about 40 percent of dissimilarity among the gene clusters (Figure 3C). While the gene composition in clusters was identical, many genes were fragmented and re-organized. Another evidence of rapid evolution of the gene cluster could be highlighted in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Both species have the identified gene clusters, encoding the biosynthesis of thalianol, which contained the same biosynthetic gene composition, OSC, CYP708, CYP705 and BAHD acyltransferase. Because the thalianol clusters have been recently recruited at 17.9 ±4.8 Mya[27], and these clusters have been only found in *Arabidopsis* spp., we expected to see a very low dissimilarity of the thalianol clusters within *Arabidopsis thaliana* and *Arabidopsis lyrata*, but the cluster dissimilarity showed about 30 percent differences between the thalianol clusters (Figure 3B). To further understand the sources of dissimilarity in thalianol clusters, the amino acid sequences from each clusters' genes were used to calculate the sequence similarity index by using Fitch matrix in SeqINR package[28] and displaying it as a heatmap. The result of this can be seen in Figure 4A, where yellow indicates the low similarity and blue indicates the high similarity. The BAHD acyltransferase showed about 50 percent similarity, while the OSC, CYP708, and CYP705 showed about 90 percent similarity (Figure 3A). The rapid evolution among *Arabidopsis* spp. could be further supported by the non-synonymous (dN) and synonymous (dS) ratio, indicating that BAHD acyltransferase genes are under strong selection and therefore are more divergent (Figure 4B)[29]. However, OSC, CYP708, and CYP705 sequences are highly conserved, implying that these compositions may have an important, yet unknown, advantages to the cluster[11]. Our results showed that even within the same gene clusters, the sequence similarity of genes compositions can be different. We therefore suggested that after the triterpene clusters were formed by tandem array duplication events, each gene composition in the clusters could be conserved or independently derived depending on evolutionary driving force in each gene composition.

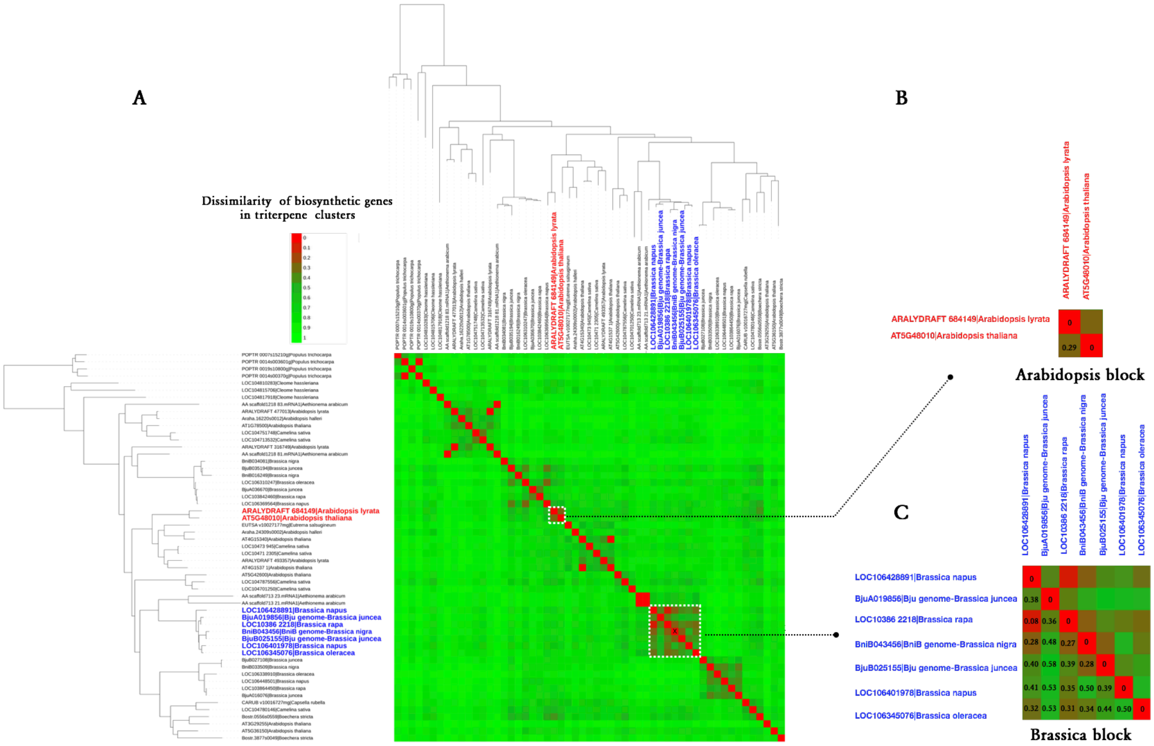

Figure 3. Heatmap of clusters dissimilarity index of biosynthetic genes in triterpene clusters. Thought, **Figure 3A** shows an overall cluster dissimilarity index of the triterpene clusters. *Arabidopsis thaliana* and *Arabidopsis lyrata* (labelled in red) and *Brassica* spp. (labelled in blue) were selected as example to determine the diversifying of biosynthetic genes in the triterpene clusters in closely related species. **Figure 3B** shows the values of cluster dissimilarity between *Arabidopsis thaliana* and *Arabidopsis lyrata*. **Figure 3C** shows the values of cluster dissimilarity between *Brassica* spp.

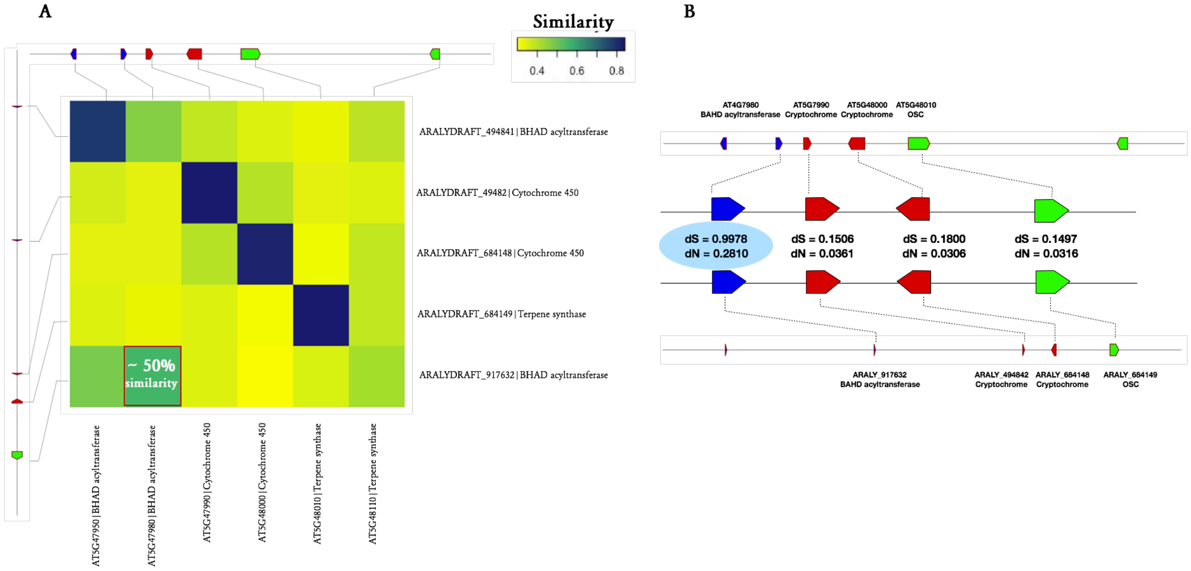

Figure 4. **(A) heatmap of similarity of gene composition in the thalianol cluste**r. Throughout, BAHD acyltransferase in the thalianol clusters shows about 50 percent similarity, while other genes show around 90 percent similarity. **(B) The comparison between dN/dS in thalianol cluster of** *Arabidopsis thaliana* and *Arabidopsis lyrata*. The BAHD acyltransferase (indicated in blue color) show the highest number of dS among the gene composition in the thalianol cluster, representing a strong selection. (The dN/dS was adapted from Reference number 28)

Both biosynthetic genes and non-biosynthetic genes show evolutionarily conservation

The previous results in Figure 3 demonstrated that the biosynthetic genes showed a high dissimilarity index among taxa, but the value was decreased when considering in the closely related species. To understand how the non-biosynthetic genes contributed the differences among the clusters, the dissimilarity index of non-biosynthetic genes were also calculated. Overall, the pattern of dissimilarity in non-biosynthetic genes were similar to that of the core-biosynthetic gene, showing a generally high dissimilarity of gene clusters among the taxa and slightly lower dissimilarity index among the closely related species (Figure 5A). However, when considering the dissimilarity of non-biosynthetic genes and core-biosynthetic genes, non-biosynthetic genes showed the higher dissimilarity value. An example can be observed in *Brassica* spp. (Figure 5C) which showed relatively high dissimilarity among the congeneric taxa, as observed in core-biosynthetic genes (Figure 3C). The average dissimilarity among seven accessions from *B. napus*, *B. juncea*, *B. rapa*, *B. nigra* and *B. oleracea* was 80 percent, which was higher than 40 percent in core-biosynthetic genes. Such a difference was more clearly observed in *Arabidopsis thaliana* and *Arabidopsis lyrata*, showing about 90 percent of dissimilarity in non-biosynthetic genes of the thalianol clusters, compared to only 30 percent in the core-biosynthetic genes (Figure 5B). These revealed that non-biosynthetic genes are more diverged than biosynthetic genes, suggesting that non-biosynthetic genes are evolutionarily less conserved. To further confirm the result of non-biosynthetic genes dissimilarity, the number of non-biosynthetic genes in *Arabidopsis thaliana* and *Arabidopsis lyrata* were calculated. In *Arabidopsis thaliana*, a total of 13 non-biosynthetic genes were detected in the thalianol clusters (Figure 6A), while the number of non-biosynthetic genes rose to 21 in *Arabidopsis lyrata* (Figure 6B). This analysis revealed that the non-biosynthetic genes in the clusters are less conserved, which can be interpreted as a sign of stronger selection compared to the biosynthetic genes, however, we did not calculate the evolving speed in biosynthetic genes and non-biosynthetic genes in the clusters. It is worth noting that neighboring genes in eukaryotes genomes including plants, fungi and animals, tend to be more frequently co-expressed than would be expected simply by chance[30]. Therefore, we propose that the non-biosynthetic genes in the clusters may be functionally related to the biosynthetic genes, corresponding to gene architecture or co-regulation.

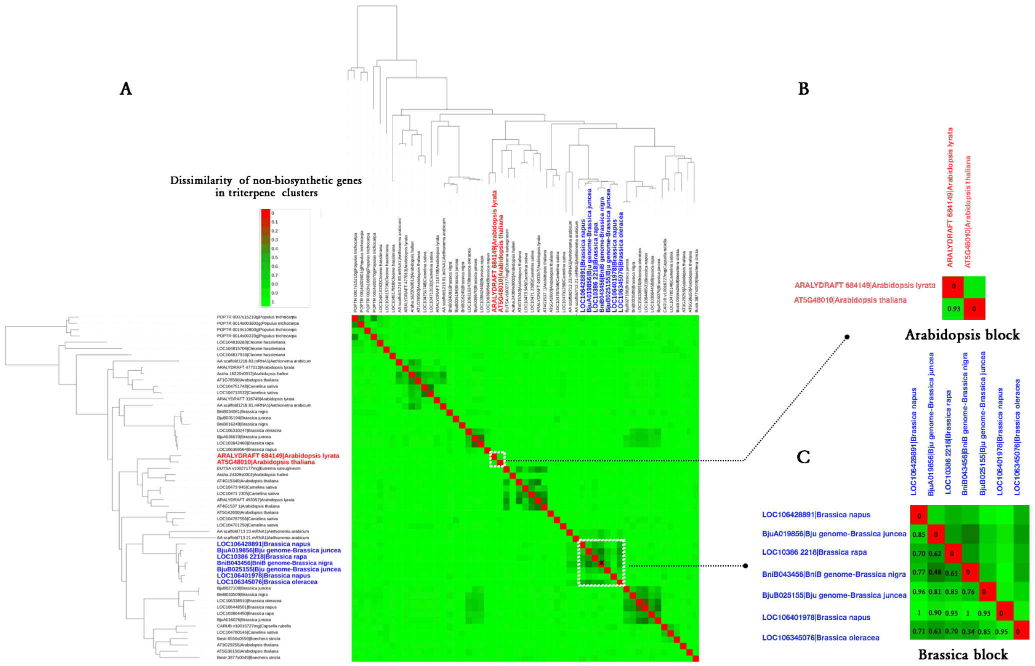

Figure 5. Heatmap of clusters dissimilarity index of non-biosynthetic genes in triterpene clusters. Thought, **Figure 5A** shows an overall cluster dissimilarity index of the triterpene clusters. *Arabidopsis thaliana* and *Arabidopsis lyrata* (labelled in red) and *Brassica* spp. (labelled in blue) were selected as example to determine the diversifying of non-biosynthetic genes in the triterpene clusters in closely related species. **Figure 5B** shows the values of cluster dissimilarity between *Arabidopsis thaliana* and *Arabidopsis lyrata*. **Figure 5C** shows the values of cluster dissimilarity between *Brassica* spp.

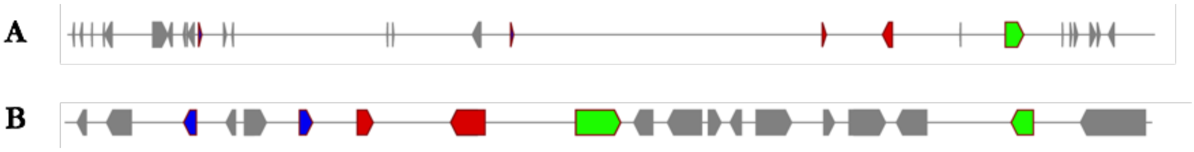

Figure 6. The thalianol clusters including biosynthetic genes and non-biosynthetic genes. **(A)** The thalianol cluster in *Arabidopsis lyrata* **(B)** The thalianol cluster in *Arabidopsis thaliana*: Gray color indicate non-biosynthetic genes, while blue, red and green represent BAHD acyltransferase, cytochrome P450 and oxidosqualene cyclase gene, respectively.

CONCLUSION

Clusters of genes that encode the biosynthetic enzyme for secondary metabolites are a conspicuous feature of bacterial genomes[31]. However, recent evidence shows that the feature of gene clustering for secondary metabolic pathways has also been found in plants[4]. One of the major unanswered questions about biosynthetic gene clusters in plants lie in understanding how gene clusters are formed, and how plants make highly diverged gene clusters[4]. Our evolution analysis and phylogeny reconstruction reveal that triterpene clusters are most likely originated from tandem array duplication rather than whole-genome duplication, supporting the previous findings in *Arabidopsis thaliana*[20], even though compositions of genes in tandem array duplication process may have an ancestral genetic composition from whole-genome duplication[32]. Putatively functional triterpene clusters are most likely to be formed by the OSCs genes that particularly contained an early diverging OSCs domain, suggesting that the OSCs domain may influenced to the recruitment of triterpene modifying enzymatic genes. Triterpene clusters are rapidly evolving, which is in contrast to the features of prokaryotes where the gene clusters are highly conserved in closely related species[33, 34]. In either case of triterpene clusters, both biosynthetic genes and non-biosynthetic genes are rapidly changing. These suggest that after the triterpene clusters have formed, each gene compositions have its own evolutionary driving force and induce an independent evolution. Notably, the biosynthetic genes are evolutionarily more conserved comparing to non-biosynthetic genes, indicating the sign of strong purifying selection in biosynthetic genes, while non-biosynthetic genes in the triterpene clusters may functionally related to gene clusters architecture, corresponding to co-regulation of gene expression. In summary, our analyses provide a new insight into a fundamental understanding of triterpene cluster formation and open up the possibility to shed light on triterpene clusters in plant genomes.

Methods

Sequence collection

Seventeen plant genomes sequences and corresponding annotations in scaffold and chromosome level were retrieved from three main available genome databases: 1) The National Center for Biotechnology Information (https://www.ncbi.nlm.gov/genome), 2) Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html), and 3) Brassica database (http://brassicadb.org/brad). Fourteen genomes from Brassicaceae were chosen to represent alpha duplication. The genome of *Cleome hassleriana* represented the beta event, and finally the genomes *Populus trichocarpa* (Sillicaceae) and *Carica papaya* (Caricaceae) represented the gamma duplication. Information for the taxa included in this study is listed in Supplementary Table 1.

Identification of triterpene gene clusters from plantiSMASH

The triterpene biosynthetic gene clusters were detected by plantiSMASH program version 1.1[35] with the Pfam database version 31.0[36], using parameters: taxon plants and clusterblast. In order to compare the detected gene clusters to the known gene clusters, knownclusterblast parameter was also included. The cluster rule of plantiSMASH program was modified to define only gene clusters that contained squalene cyclase domain in C-terminal and N-terminal which served as a key step of triterpene biosynthesis. The cut-off value was set at 10 kb to define hits space of locating clusters of signature gene profile Hidden Markov Models (pHMM). Flanking accessory genes were detected with an extension value of 2 kb on each side of the last signature gene pHMM. Minimum unique domains value was set at 1. Detected gene clusters output were written in Genbank format, json format and html format. An official release of plantiSMASH program version 1.1 is available at http://plantismash.secondarymetabolites.org.

Oxidosqualene cyclase domain (OSCs) domain phylogeny construction

Amino acid sequence of oxidosqualene cyclase domain (OSCs) in C-terminal and N-terminal were retrieved from a json-format file from the plantiSMASH output. The sequences were then written into the multi-fasta format, and aligned to the Profile Hidden Markov Models (pHMM) of squalene cyclase domain (SQHop_cyclase _C.hmm and SQHop _cyclase_N.hmm). Automated pipeline of domain sequence extraction was performed using a python script (extract_domains.py).

To reconstruct phylogenetic relationship, a protein sequence alignment in the multi-fasta format was converted to the phylip format using a perl script (Fasta2phylib.pl). The phylogeny was estimated by the Maximum-likelihood method in RAxML version 8.0.0[37]. The best model was calculated by PROTGAMMAAUTO parameters, using 1,000 pseudoreplications for bootstrapping.

Visualization of triterpene gene clusters

The phylogenetic tree of triterpene domains was visualized in Interactive Tree of Life (ITOL) platform (http://itol.embl.de/)[38], genecluster.js file from plantiSMASH output was used to extract gene locations and domain types, using a python script (protein_format.py). Data was converted to follow the protein template format, provided in http://itol.embl.de/gallery.cgi. The phylogenetic tree with bootstrap value was added to ITOL platform with protein template. The most recent phylogenetic clade with at least three biosynthetic domain were selected to analyze gene clusters similarity. Gene clusters that did not meet criteria was discarded from the phylogenetic tree using ape library in a R Script (Droptip_script.R).

Gene cluster dissimilarity from BIGSCAPE

A total of 56 locus tags from plantiSMASH's output in Genbank format from the most recent phylogenetic clade were used to calculate gene clusters dissimilarity in the BIGSCAPE program, with default settings. Biosynthetic genes and non-biosynthetic genes in the triterpene clusters were analyzed separately. Three main values of shared domain content (Jaccard index), similarity of domain sequences (DDS index), and shared domain-pair content (Adjacency index) were calculated and combined to calculate a distance matrix. The latest update of BIGSCAPE is available at https://git.wageningenur.nl/yeong001/BGC_networks.

To visualize a pattern of cluster similarity from a large matrix data, BIGSCAPE output (all_mix _c1.00.network) was written to the heatmap template format (http://itol.embl.de/gallery.cgi) using a python script (find _cluster.py). The data was used to generate a heatmap of similarities among triterpene gene clusters.

Comparing the similarity of gene compositions from the same biosynthetic gene clusters in closely related species

To examine the patterns of cluster dissimilarity in more details, a subset of specific gene clusters was selected manually, based on gene compositions and structural components. Amino acid sequence from all genes in a particular gene cluster were extracted using a python script. Then, the sequences were aligned using MAFFT program version 7[39], and were calculated the sequence similarity matrix using R program with SeqINR package version 3.3-6[40]. The automated pipeline is available in a python script (extract_clusters.py).

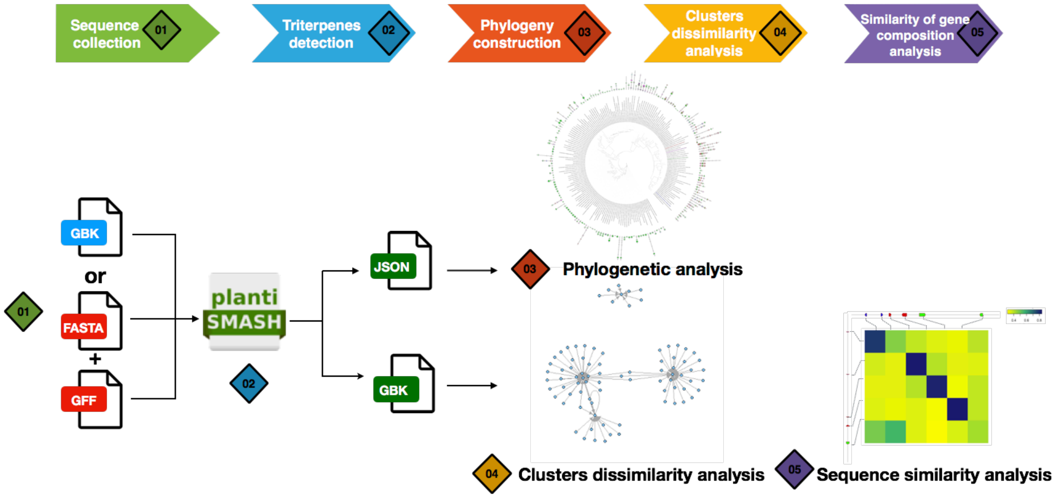

Figure 7. The pipeline approach for evolutionary dynamics study.
Input file in Genbank format or FASTA format with GFF format are used for triterpene identification using plantiSMASH. The output from plantiSMASH in JSON format are further extracted by a python script and reconstruct the phylogeny. Meanwhile, the output in Genbank format are further used for cluster dissimilarity analysis and sequence similarity analysis. Source code and data are available at https://github.com/harnv001/Master_Thesis.

ACKNOWLEDGEMENTS

My study and this work was supported by Anandamahidol foundation. To finish on this research, I am grateful to Dr. Marnix Medema and Hernando Suarez-Duran for their assistances, professional guidance and critical reading, Prof. Dick de Ridder for his valuable comments, Prof. Anne Osbourn, Dr. Zhenhua Liu, Charlie Owen and staffs at John Innes Centre for helpful discussion. I also thank Dr. Ekaphan Kraichak for providing feedback and academic support. Technical assistances were supported by Jorge Navarro-Munoz, Satria Kautsar, Joris van Steenbrugge, Mehmet Aakdel, Wasin Ponchewin, Raul Wijfjes, Miguel Correa Marrero, Janani Durairaj, Jarno Persoon and other bioinformatics staffs at Wageningen University. Most importantly, I would like to thank my parents for giving me freedom to pursue my own passions.

References

1. Wink, M. Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochem.* **64**, 3–19 (2003). DOI 10.1016/S0031-9422(03)00300-5.

2. Zhao, N., Wang, G., Norris, A., Chen, X. & Chen, F. Studying Plant Secondary Metabolism in the Age of Genomics. *CRC. Crit. Rev. Plant Sci.* **32**, 369–382 (2013). URL http://www.tandfonline.com/doi/abs/10.1080/07352689.2013.789648. DOI 10.1080/07352689.2013.789648.

3. De-La-Cruz Chacón, I., Riley-Saldaña, C. A. & González-Esquinca, A. R. Secondary metabolites during early development in plants. *Phytochem. Rev.* **12**, 47–64 (2013). DOI 10.1007/s11101-012-9250-8.

4. Nützmann, H. W., Huang, A. & Osbourn, A. Plant metabolic clusters – from genetics to genomics. *New Phytol.* **211**, 771–789 (2016). DOI 10.1111/nph.13981.

5. Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015). URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84939557642&partnerID=40&md5=fecb9988ce40a134045804ae076726c8. DOI 10.1038/nchembio.1890.

6. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014). URL http://dx.doi.org/10.1016/j.cell.2014.06.034. DOI 10.1016/j.cell.2014.06.034. NIHMS150003.

7. Osbourn, A. Gene clusters for secondary metabolic pathways: an emerging theme in plant biology. *Plant Physiol.* **154**, 531–5 (2010). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2949040&tool=pmcentrez&rendertype=abstract. DOI 10.1104/pp.110.161315.

8. Frey, M. *et al.* Analysis of a chemical plant defense mechanism in grasses. *Sci. (80-.).* **277**, 696–699 (1997).

9. Qi, X. *et al.* A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc Natl Acad Sci U S A* **101**, 8233–8238 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15148404%5Cnhttp://www.pnas.org/content/101/21/8233.full.pdf. DOI 10.1073/pnas.0401301101.

10. Shimura, K. *et al.* Identification of a biosynthetic gene cluster in rice for momilactones. *J. Biol. Chem.* **282**, 34013–34018 (2007). DOI 10.1074/jbc.M703344200.

11. Field, B. & Osbourn, A. E. Metabolic diversification—independent assembly of operon-like gene clusters in different plants. *Sci.* **320**, 543–547 (2008). URL http://science.sciencemag.org/content/320/5875/543. DOI 10.1126/science.1154990. http://science.sciencemag.org/content/320/5875/543.full.pdf.

12. Nutzmann, H. W. & Osbourn, A. Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* **26**, 91–99 (2014). DOI 10.1016/j.copbio.2013.10.009.

13. Hill, R. A. & Connolly, J. D. Triterpenoids. *Nat. Prod. Rep.* **30**, 1028–1065 (2013). URL http://xlink.rsc.org/?DOI=C3NP70032A. DOI 10.1039/C3NP70032A.

14. Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P. & Osbourn, A. Triterpene biosynthesis in plants. *Annu. Rev. Plant Biol.* **65**, 225–57 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24498976. DOI 10.1146/annurev-arplant-050312-120229.

15. Kingsolver, J. G. & Huey, R. B. Introduction: the evolution of morphology, performance, and fitness. *Integr. Comp. Biol.* **43**, 361–366 (2003). DOI 10.1093/icb/43.3.361.

16. Boutanaev, A. M. *et al.* Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci.* **112**, E81–E88 (2015). URL http://www.pnas.org/content/112/1/E81%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/25502595%5Cnhttp://www.pnas.org/content/112/1/E81.full%5Cnhttp://www.pnas.org/content/112/1/E81.full.pdf. DOI 10.1073/pnas.1419547112.

17. Xu, R., Fazio, G. C. & Matsuda, S. P. T. On the origins of triterpenoid skeletal diversity. *Phytochem.* **65**, 261–291 (2004). DOI 10.1016/j.phytochem.2003.11.014.

18. Andre, C. M. *et al.* Multifunctional oxidosqualene cyclases and cytochrome P450 involved in the biosynthesis of apple fruit triterpenic acids. *New Phytol.* **211**, 1279–1294 (2016). DOI 10.1111/nph.13996.

19. Takos, A. M. *et al.* Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in Lotus japonicus and suggests the repeated evolution of this chemical defence pathway. *Plant J.* **68**, 273–286 (2011). DOI 10.1111/j.1365-313X.2011.04685.x.

20. Field, B. *et al.* Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc Natl Acad Sci U S A* **108**, 16116–16121 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21876149%5Cnhttp://www.pnas.org/content/108/38/16116.full.pdf. DOI 10.1073/pnas.1109273108.

21. Mun, J.-H. *et al.* Sequence and structure of Brassica rapa chromosome A3. *Genome Biol.* **11**, R94 (2010). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2965386&tool=pmcentrez&rendertype=abstract. DOI 10.1186/gb-2010-11-9-r94.

22. Meinke, D. W. Arabidopsis thaliana: A Model Plant for Genome Analysis. *Sci. (80-.).* **282**, 662–682 (1998). URL http://www.sciencemag.org/cgi/doi/10.1126/science.282.5389.662%5Cnhttp://www.sciencemag.org/content/282/5389/662.abstract. DOI 10.1126/science.282.5389.662.

23. Huang, Z. *et al.* Retention, Molecular Evolution, and Expression Divergence of the Auxin/Indole Acetic Acid and Auxin Response Factor Gene Families in Brassica Rapa Shed Light on Their Evolution Patterns in Plants. *Genome Biol. Evol.* **8**, 302–16 (2016). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4779605&tool=pmcentrez&rendertype=abstract. DOI 10.1093/gbe/evv259.

24. Edger, P. P. *et al.* The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl. Acad. Sci.* **112**, 8362–8366 (2015). DOI 10.1073/pnas.1503926112.

25. Lodeiro, S. *et al.* An oxidosqualene cyclase makes numerous products by diverse mechanisms: A challenge to prevailing concepts of triterpene biosynthesis. *J. Am. Chem. Soc.* **129**, 11213–11222 (2007). DOI 10.1021/ja073133u.

26. Retey, J. Enzymic Reaction Selectivity by Negative Catalysis or How Do Enzymes Deal with Highly Reactive Intermediates? *Angew. Chemie Int. Ed. Engl.* **29**, 355–361 (1990). DOI 10.1002/anie.199003551.

27. Ossowski, S. *et al.* Linked references are available on JSTOR for this article : The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. *Sci. (80-.).* **327**, 92–94 (2016).

28. Fitch, W. M. An improved method of testing for evolutionary homology. *J. Mol. Biol.* **16**, 9–16 (1966). URL http://linkinghub.elsevier.com/retrieve/pii/S0022283666802589. DOI 10.1016/S0022-2836(66)80258-9.

29. Osbourn, A. E. & Field, B. Operons. *Cell. Mol. Life Sci.* **66**, 3755–3775 (2009). DOI 10.1007/s00018-009-0114-3.

30. Wada, M. *et al.* Prediction of operon-like gene clusters in the Arabidopsis thaliana genome based on co-expression analysis of neighboring genes. *Gene* **503**, 56–64 (2012). URL http://dx.doi.org/10.1016/j.gene.2012.04.043. DOI 10.1016/j.gene.2012.04.043. NIHMS150003.

31. Ballouz, S., Francis, A. R., Lan, R. & Tanaka, M. M. Conditions for the evolution of gene clusters in bacterial genomes. *PLoS Comput. Biol.* **6**, e1000672 (2010).

32. Fang, L., Cheng, F., Wu, J. & Wang, X. The Impact of Genome Triplication on Tandem Gene Evolution in Brassica rapa. *Front Plant Sci* **3**, 261 (2012).

33. McAdams, H. H., Srinivasan, B. & Arkin, A. P. The evolution of genetic regulatory systems in bacteria. *Nat. Rev. Genet.* **5**, 169–178 (2004).

34. Lee, J. M. & Sonnhammer, E. L. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* **13**, 875–882 (2003).

35. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **320**, 543–547 (2017). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx305. DOI 10.1093/nar/gkx305.

36. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–285 (2016).

37. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma.* **30**, 1312–1313 (2014). DOI 10.1093/bioinformatics/btu033. bioinformatics/btu033.

38. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw290. DOI 10.1093/nar/gkw290.

39. Yamada, K. D., Tomii, K. & Katoh, K. Application of the MAFFT sequence alignment program to large data - Reexamination of the usefulness of chained guide trees. *Bioinforma.* **32**, 3246–3251 (2016). DOI 10.1093/bioinformatics/btw412.

40. Fitch, W. M. An improved method of testing for evolutionary homology. *J. Mol. Biol.* **16**, 9 – 16 (1966). URL http://www.sciencedirect.com/science/article/pii/S0022283666802589. DOI http://dx.doi.org/10.1016/S0022-2836(66)80258-9.

41. Tuskan, G. A. & Torr, P. The Genome of Black Cottonwood ,. *Sci. (80-.).* **1596**, 1596–1605 (2007). DOI 10.1126/science.1128691.

42. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nat.* **452**, 991–996 (2008). URL http://www.nature.com/doifinder/10.1038/nature06856. DOI 10.1038/nature06856.

43. Cheng, S. *et al.* The Tarenaya hassleriana Genome Provides Insight into Reproductive Trait and Genome Evolution of Crucifers. *Plant Cell* **25**, 2813–2830 (2013). URL http://www.plantcell.org/cgi/doi/10.1105/tpc.113.113480. DOI 10.1105/tpc.113.113480.

44. Haudry, A. *et al.* An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* **45**, 891–898 (2013). URL http://www.nature.com/doifinder/10.1038/ng.2684. DOI 10.1038/ng.2684.

45. Yang, R. *et al.* The Reference Genome of the Halophytic Plant Eutrema salsugineum. *Front. Plant Sci.* **4**, 1–14 (2013). URL http://journal.frontiersin.org/article/10.3389/fpls.2013.00046/abstract. DOI 10.3389/fpls.2013.00046.

46. Schranz, M. E., Windsor, A. J., Song, B.-h., Lawton-Rauh, A. & Mitchell-Olds, T. Comparative Genetic Mapping in Boechera stricta, a Close Relative of Arabidopsis. *Plant Physiol.* **144**, 286–298 (2007). URL http://www.plantphysiol.org/cgi/doi/10.1104/pp.107.096685. DOI 10.1104/pp.107.096685.

47. Johnston, J. S. *et al.* Evolution of genome size in Brassicaceae. *Ann. Bot.* **95**, 229–235 (2005). DOI 10.1093/aob/mci016.

48. Slotte, T. *et al.* The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **45**, 831–835 (2013). URL http://www.nature.com/doifinder/10.1038/ng.2669. DOI 10.1038/ng.2669.

49. Kagale, S. *et al.* The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nat. Commun.* **5**, 1–11 (2014). URL http://www.nature.com/doifinder/10.1038/ncomms4706. DOI 10.1038/ncomms4706.

50. Akama, S., Shimizu-Inatsugi, R., Shimizu, K. K. & Sese, J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. *Nucleic Acids Res.* **42** (2014). DOI 10.1093/nar/gkt1376.

51. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011). URL http://www.nature.com/doifinder/10.1038/ng.807. DOI 10.1038/ng.807.

52. Swarbreck, D. *et al.* The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res.* **36**, 1009–1014 (2008). DOI 10.1093/nar/gkm965.

SUPREMENTARY

Table 1. The summary of genome samples

Taxa	Duplication events	Brassica lineages	Genome size (Mbp)	Assembly level	References
Populus trichocarpa	Before β	-	417.287	chromosome	[41]
Carica papaya	Before β	-	369.76	scaffold	[42]
Cleome hassleriana	$\beta - \alpha$	-	249.93	scaffold	[43]
Aethionema arabicum	α	Unassigned	192.48	scaffold	[44]
Eutrema salsugineum	α	EII	237.5	chromosome	[45]
Boechera stricta	α	II	264	scaffold	[46]
Arabis alpina	α	II	308.03	chromosome	[47]
Brassica juncea	α	II	1068	scaffold	[47]
Brassica napus	α	II	1132	chromosome	[47]
Brassica nigra	α	II	632	scaffold	[47]
Brassica oleracea	α	II	632	chromosome	[47]
Brassica rapa	α	II	529	chromosome	[47]
Capsella rubella	α	I	219	scaffold	[48]
Camelina sativa	α	I	750	chromosome	[49]
Arabidopsis halleri	α	I	196.24	scaffold	[50]
Arabidopsis lyrata	α	I	207	scaffold	[51]
Arabidopsis thaliana	α	I	120	chromosome	[52]