Comparative analysis of biosynthetic gene clusters (BGCs) in *Arabidopsis thaliana* and *Cleome violacea*

Yuanyuan Ma (930906536040)

Chair Group: Bioinformatics

Supervisors: Marnix Medema,

M. Eric Schranz, Hernando Suárez-Duran

March 2019

Abstract

Just like operons in bacterial genomes, some biosynthetic genes are grouped together in specific regions in eukaryotic genomes. These operon-like features are called biosynthetic gene clusters (BGCs). Since BGCs could help to discover unknown metabolic pathways in microbes and novel enzymes in plants, more and more computational approaches can predict secondary metabolite BGCs, such as antiSMASH and plantiSMASH. In this project, *Arabidopsis thaliana* was used to explore the relationship between BGCs and evolutionary events such as whole-genome duplication (WGD) and polyploidization. WGD broadly and frequently appears in the evolutionary history of Brassicaceae, it is widely accepted that the *Arabidopsis* genome includes at least three ancient WGD, called the α (the most recent), β and γ (the most ancient) WGD events. Previous studies have identified and reported the syntenic blocks in *Arabidopsis* genome, and we used the homologous genes in the syntenic blocks, called anchor genes, to define the regions syntenic to predicted BGCs. After comparing BGCs and sister regions, there is no difference of conservation in biosynthetic and non-biosynthetic genes. On the other hand, terpene clusters are more likely located in dynamic regions than other type of BGCs. Additionally, we identified only two saccharide BGCs (cluster 18 and cluster 28) located in the same syntenic block in *Arabidopsis* genome. We also explored the genome of *Cleome violacea* to *Arabidopsis thaliana*. Cleomaceae is sister-family of Brassicaceae and occurred α whole-genome triplication and shared β duplication with Brassicaceae. We identified one alkaloid BGC (cluster 30) in *Arabidopsis* that was also present in *Cleome* (cluster 1), meaning the structure has been conserved before the species' divergence 14.5 to 86 Mya.

Introduction

Genes encoding specialized metabolic pathways are regularly grouped together in gene clusters among bacterial genomes. Most of these gene clusters are organized as operons (Osbourn, 2010). Operons do not often appear in eukaryotes, but the order of some eukaryotic genes is not random and two or more genes that encode a biosynthetic pathway in a genome are physically clustered together in specialized regions. These regions are called biosynthetic gene clusters (BGCs) (Medema et al., 2015). These clusters have indeed operon-like features, such as physical clustering and coregulation, and studying them may help to discover unknown metabolic pathways and novel enzymes in plants (Rutledge & Challis, 2015; Schläpfer et al., 2017). Therefore, analysis of BGCs is relevant for answering questions about metabolism and evolution of plants. Currently, BGC prediction is becoming an emerging genome mining process in the plant kingdom. Hundreds of plant BGCs have been already predicted, however, the number of functionally characterized BGCs is about 30, and the evolutionary history of most remain unknown (Nützmann et al., 2016). Around two decades before, the first plant gene cluster was discovered in Maize (*Zea mays*) (Frey et al., 1997). Afterwards, more gene clusters have been found in plants, such as oat, rice, *Arabidopsis thaliana*, tomato and potato, and have been characterized in more details (Boycheva et al., 2014).

With the development of genome mining tools, more and more computational approaches can predict secondary metabolite BGCs, such as antiSMASH (Blin et al., 2017) and plantiSMASH (Kautsar et al., 2017), which is an extension of antiSMASH specially built for plant genome mining. As the result of this project, *A. thaliana* BGCs are available as precomputed results in the plantiSMASH web server. This tool identified 45 secondary metabolite clusters in *A. thaliana*, predicted to synthesize several secondary metabolite types, including alkaloid, lignan, polyketide, saccharide and terpene (Kautsar et al., 2017).

Recent study compared the characteristics of two gene clusters (the thalianol cluster and the marneral cluster) from *A. thaliana* and examined the evolutionary events and the cluster formation. The results indicate that the two clusters were formed after the α whole-genome duplication and are located in a dynamic region, which no longer can be identified any synteny within the genome, in *A. thaliana* chromosome 5 (Field et al., 2011). Furthermore, polyploidy or whole-genome duplication (WGD) widely appears in the evolutionary history of plants (Wendel, 2000). During ~40 million years of Brassicaceae evolution, WGD occurred frequently (Schranz et al., 2006). Studies indicated that the polyploidy event nearly occurred when Brassicaceae diverged from the sister family, Cleomaceae (Schranz et al., 2006). The event increased genome repetition and led to much more new combination between homologous

chromosomes (Lysak et al., 2016). *A. thaliana* was the first completely genomic sequenced plant (Arabidopsis Genome Initiative, 2000), and a whole-genome duplication was reported in the same publication (Figure 1)(Vision et al., 2000). It is broadly accepted that the *Arabidopsis* genome includes at least three ancient duplication events, which are α (the most recent), β and γ (the most ancient) events (Bowers et al., 2003). As a result of chromosomal rearrangements, chromosomal collinear genomic regions (syntenic blocks) appeared in *A. thaliana* genome. The syntenic blocks were formed during the ancient WGDs. Of the genes that are in the syntenic blocks, 28.6% of them have a retained pair (syntelog) and the rest of them lost one copy (Thomas et al., 2006). In this project, clusters that are located in the syntenic regions and dynamic regions will be analysed. We will also explore and analyse the regions (sister regions) syntenic to the BGCs, which are identified through anchor genes. This may examine the BGCs' formation during the evolutionary events.

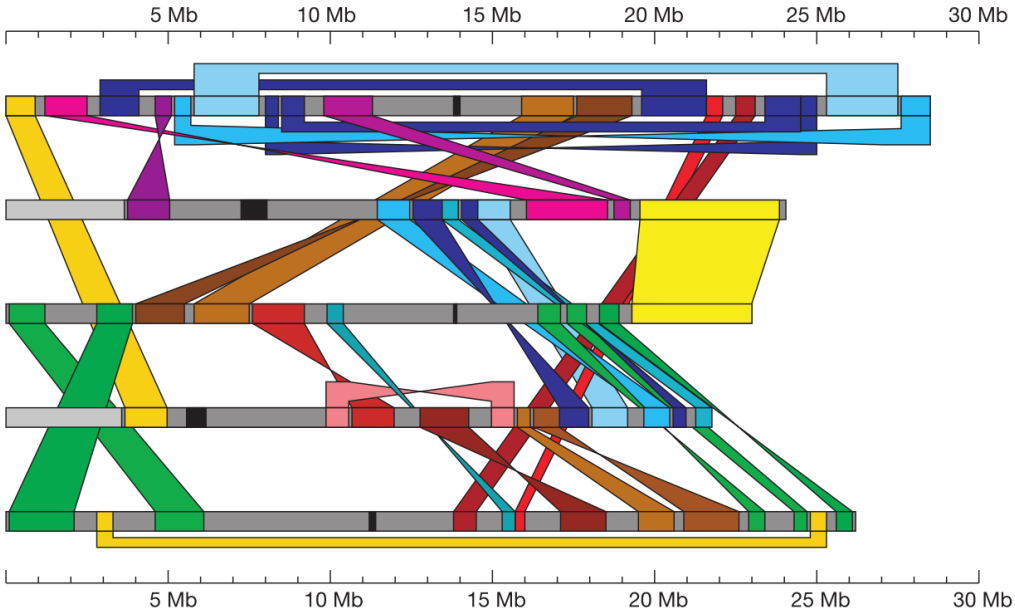

Figure 1. Segmentally duplicated chromosomes in A. thaliana genome (Arabidopsis Genome Initiative, 2000).

Cleomaceae is a sister family of Brassicaceae (Hall et al., 2002), and it has been used as important model to study ecology and evolution as well, for example: floral morphology (Patchell, Roalson, & Hall, 2014), the evolution of C4 photosynthesis (Patchell et al., 2014) and comparative genome analysis (Schranz & Mitchell-Olds, 2006). Cleomaceae is a small family that contains 18 genera and more than 200 species (Patchell et al., 2014). *Cleome* is the largest genus in the family, and contains about 200 species (Hall et al., 2002). Since WGD occurs in virous flowering plants throughout the evolutionary history, *Cleome* is not an exception (Soltis et al., 2009). Studies indicated that a genomic triplication occurred in *Cleome* and the specific polyploidization (*Cs-α*) is independent of the Brassicaceae α WGD events, and *Cleome* and *Arabidopsis* share the *At-β* duplication (Schranz & Mitchell-Olds, 2006; Barker et al., 2009). We used PlantiSMASH to identify BGCs in *Cleome violacea* genome and compared clusters predicted in *A. thaliana* and *C. violacea*. This may help to investigate the syntenic regions and identify any conserved BGCs between two species.

The objective of this project is to compare predicted *A. thaliana* BGCs to their sister regions and homologous BGCs in *C. violacea*. Due to the ancient polyploidy events, it is interesting to explore the formation of distinct BGCs in *A. thaliana* and identify metabolite pathways of conserved BGCs. Additionally, we are curious about whether biosynthetic genes are more conserved than non-biosynthetic genes in BGCs and do conserved anchor genes more often belong to certain protein families. Brassicaceae is sister family to Cleomaceae (Hall et al., 2002). Both families occurred polyploidy during the evolutionary history and they shared *At-β* duplication with each other. Therefore, we would like to compare clustered regions in *Arabidopsis* and *Cleome* genome, and try to examine the formation and

conservation of BGCs between two species. The comparative genomic approach is critical for understanding both the evolution of genes and genomes. Consequently, the following research questions will be addressed:

Research question 1: To what extent are BGCs and their constituent genes in *A. thaliana* conserved when compared to their sister regions?

Subquestions: Comparing to the biosynthetic and non-biosynthetic genes, which type of genes are more conserved? Comparing to anchor and non-anchor genes, are anchor genes enriched in genes coding for a particular protein family?

Research question 2: What are the patterns of gene gain/loss in the evolution of BGCs and their sister regions, when compared with the corresponding syntenic regions in *C. violacea*?

Subquestions: Comparing BGCs and sister regions of *A. thaliana* to *C. violacea* genome, is there any homologous gene found? What genes are they? What do the syntenic regions look like?

Results and Discussion

BGCs identification in *Arabidopsis thaliana* and *Cleome violacea*

PlantiSMASH predicted 45 BGCs in *A. thaliana* (Figure 2A), and 26 BGCs in *C. violacea* (Figure 2B) (Supplementary data 1&2). Since ancient WGDs occurred in Brassicaceae, a great number of chromosomal regions in *Arabidopsis* are homoeologous (syntenic and paralogous) with each other (Thomas et al., 2006). Pedersen et al. (2007) determined these syntenic and conserved regions by identifying anchor genes. We used these anchor genes to ensure whether BGCs are located in the syntenic blocks. For 38 BGCs of *A. thaliana* relatively syntenic regions could be identified elsewhere in *Arabidopsis* genome. These syntenic regions are called "sister regions" and show high conservation to the BGCs. Only two BGCs (cluster 18 and cluster 28) are syntenic with each other (Figure 2C). However, 7 BGCs are located in the dynamic regions that could not find any synteny in the genome (Figure 2D). Interestingly, 4 of them are terpene clusters (Figure 2E). The rest of 9 terpene clusters are located in the syntenic regions. Additionally, 26 BGCs have been identified in *C. violacea* genome, 10 of them have homologous genes with *Arabidopsis* clusters that locate in the syntenic regions (Figure 2F).

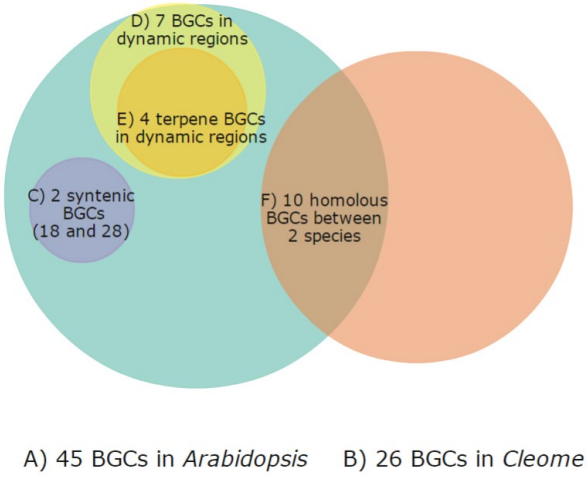

Figure 2. Venn diagram: Relationship of the BGCs in A. thaliana and C. violacea. A) 45 BGCs predicted by PlantiSMASH in A. thaliana genome. B)26 BGCs predicted by PlantiSMASH in C. violacea genome. C) Cluster 18 and cluster 28 in Arabidopsis that are syntenic of each other. D) 7 BGCs that located in the dynamic regions of A. thaliana genome. E) 4 terpene BGCs that located in the dynamic regions of A. thaliana genome. F) 10 BGCs in C. violacea genome that have homologous genes in syntenic Arabidopsis BGCs.

BGCs in dynamic regions

Terpenes are the largest class of plant secondary metabolites. They play importantly ecological function in plant defencing against biotic and abiotic stresses (Tholl & Lee, 2011). Since we found that 4 of 7 BGCs (cluster 22, 24, 39 and 40) that are located in the dynamic regions are terpene clusters, we examined whether terpene BGCs appeared significantly more in dynamic regions or not. Therefore, Fisher's exact test was used. Based on the statistically significant test, P-value equals to 0.1676. On the other hand, 2 saccharide BGCs (cluster 7 and 23) and a putative BGC (cluster 45) are also located in the dynamic regions. Based on the Fisher's exact test, the P-value of 2 saccharide BGCs that are significantly located in dynamic regions is 0.6808. Therefore, terpene clusters are more enriched in the BGCs that are located in the dynamic regions than saccharide BGCs. These may indicate the evolutionary history of terpene BGCs, which more likely formed after the ancient WGD events.

BGCs and sister regions in *Arabidopsis*

We compared BGCs that are located in the syntenic chromosomal regions with each individual sister region. In order to better study BGCs and sister regions, we extended the studied regions around each BGC until the first anchor gene flanking downstream and upstream. We identified the syntelogs to these flanking anchor genes and sued them to define the loci syntenic to each BGC. We aligned amino acid sequences of BGCs against their sister regions. To examine the similarity of syntenic BGCs and sister regions, we calculated the similar value. In total, 45 *Arabidopsis* BGCs contain 633 genes. 84 of them are anchor genes and 33 of these conserved genes are biosynthetic gene. However, we would like to explore whether these conserved genes are more likely biosynthetic genes or not. Therefore, we applied the Fisher's exact test to answer this question. The statistical p-value is 0.6273, which suggests that there is no significant difference in types of anchor genes. The result indicates that both biosynthetic and non-biosynthetic genes in BGCs show evolutionary conservation.

To further investigate what kinds of genes are enriched to be anchor genes, we identified the protein families encoded in the 633 genes in BGCs and performed enrichment analysis of these protein families. In total, 816 protein families were identified in the 633 gene. 140 protein families were identified in anchor genes. We used hypergeometric distribution to calculate the p value and analyse enrichment of each protein family that were identified in anchor genes (see methods). As shown in supplementary table 1, ferritin, methyltransferase, lyase, UDPGT and Pkinase have lower p value and appear more than one time in the cluster genes. This indicates that aforementioned protein families are significantly enriched (α = 0.1) in the anchor genes of BGCs. On contrast, p450, transferase and terpene synthase present more often in non-anchor genes.

Based on the anchor genes analysis, we have identified two saccharide BGCs (18 and 28) were syntenic with each other (Figure 3). The syntenic region of cluster 18 is flanked by two anchor genes (AT2G40110 and AT2G40380). Addtionally, the cluster region is between AT2G40180 and AT2G40300. There are three biosynthetic genes (AT2G40190, AT2G40230 and AT2G40280) and four anchor genes (AT2G40240, AT2G40270, AT2G40280 and AT2G40300) in this BGC. Locus AT2G40280, putative methyltransferase PMT23, is both biosynthetic and anchor gene. On the other hand, its syntenic region (cluster 28) is flanked by AT3G55890 and AT3G56110, and located between loci AT3G55960 and AT3G56100. Cluster 28 contains four biosynthetic genes (AT3G55970, AT3G56000, AT3G56060 and AT3G56080) and four anchor genes (AT3G56030, AT3G56050, AT3G56080 and AT3G56090). Gene AT3G56080, probable methyltransferase PMT22, is also biosynthetic and anchor gene. Figure 3 indicates that four groups of loci share higher similarity with its anchor gene. Methyltransferases in two clusters are not only biosynthetic genes but also anchor genes.

Previous study indicates that cluster 18 is a cluster of ABA-regulated genes (Wang et al., 1999). ABA is a terpenoid plant hormone and regulates numberous plant developmental processes, such as seedling, gernimation, stress tolerance, stomatal closure, flowering and pathogen responses (Finkelstein, 2013; Nambara & Marion-Poll, 2005). In cluster 28, locus AT3G55970 is jasmonate-regulated protein. Jasmonates are essential plant growth and development regulators as well. The hormones play important roles in biotic and abiotic stress responses (Delker, Zolman, Miersch, & Wasternack, 2007). Therefore,

we suggests that the two conserved saccharide BGCs were more likely formed during the ancient Brassicaceae WGD events. The clustered genes could be all related to the plant hormone synthesis.

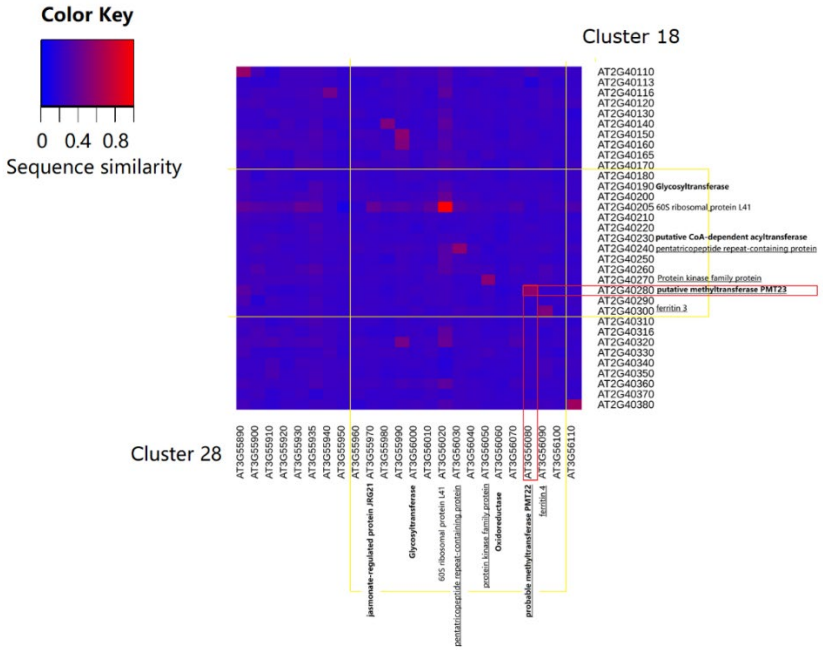

Figure 3 Comparison of cluster 18 and cluster 28. The similarity values are ranged from 0 to 1. Gene names with underline are anchor genes. Bold gene names are biosynthetic genes. Yellow squared regions are BGCs.

Cleome violacea and *Arabidopsis thaliana* shared WGD during the evolutionary history

Cleomaceae is a sister family of Brassicaceae which diverged after the *At-β* duplication of Brassicaceae (Schranz & Mitchell-Olds, 2006; Barker et al., 2009). In this project, we used the *Cleome violacea* genome to compare the synteny of clusters in *Arabidopsis* and *Cleome*. We used plantiSMASH to identify the BGCs in *Cleome* genome and 26 BGCs were reported (Supplementary data 2). We identified the syntenic regions of two species and found 10 clusters in *Cleome* that had homologous genes with *Arabidopsis* BGCs. To further analyse the relationship of these couple clusters, we performed multiple sequence alignment and built heatmap of them. The most conserved aligned results indicated that two alkaloid BGCs, cluster 1 of *Cleome* and cluster 30 of *Arabidopsis*, showed high similarity with each other (Figure 4). The two BGCs contain almost same genes and are related to the same metabolic pathway. These may indicate that the structure of clusters has been conserved before the species' divergence. As shown in figure 4, the copper amine oxidase genes have three copies in cluster 30 but only one copy in *Cleome* genome. Based on protein domains analysis, we identified that the three copies of genes indeed are two copper amine oxidase

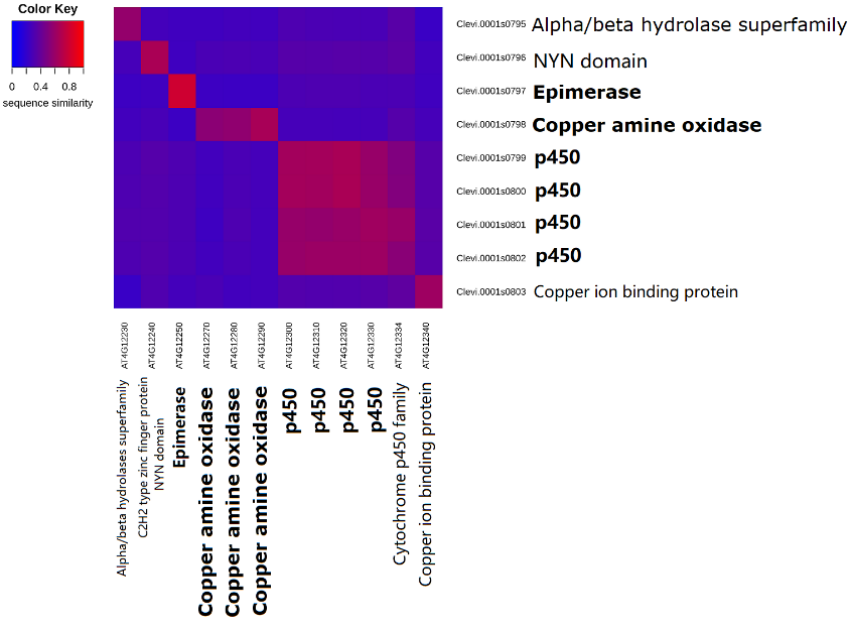

Figure 4 Heatmap of cluster 1 in Cleome and cluster 30 in Arabidopsis. The bold genes belong to biosynthetic genes in BGCs.

genes. This suggests that copper amine oxidase gene duplication occurred after the divergence of Brassicaceae from Cleomaceae

Since we identified such conserved BGCs in two plant genomes, we were curious about the comparative results with syntenic regions respectively in *Cleome*. Therefore, we searched the syntenic regions of cluster 1 in *Cleome* genome and aligned them. One of the corresponding syntenic region is flanked by two anchor genes (Clevi.002s0218 and Clevi.002s0236) (Supplementary figure 1A). Although some genes were syntenic, cluster 1 of *Cleome* was not conserved with the respective syntenic region. We also compared cluster 1 of *Cleome* with sister region of cluster 30 in *Arabidopsis* (Supplementary figure 1C). As shown in the result, the genes encoding cytochrome P450s were conserved in these two regions.

Conclusion

Biosynthetic genes that catalyse secondary metabolites are usually grouped together in specific regions in bacterial genomes (Osbourn, 2010). Recent studies show evidences that more and more clustered regions have been found in plant kingdom (Nützmann et al., 2016). These grouped regions are called biosynthetic gene clusters (Medema et al., 2015). PlantiSMASH is one of the most popular tools to identify BGCs in plants genome (Kautsar et al., 2017). Our study aims to explore the relationship between BGCs and evolutionary events such as whole-genome duplication (WGD) and polyploidization with *Arabidopsis thaliana* and *Cleome violacea*. Therefore, we predicted BGCs in *Arabidopsis* and *Cleome* with plantiSMASH. WGD widely appears in the evolutionary history of plants (Wendel, 2000). During ~40 million years of Brassicaceae evolution, WGD occurred frequently (Schranz et al., 2006). Previous studies indicated that the polyploidy event nearly occurred when Brassicaceae diverged from Cleomaceae (Schranz et al., 2006). Pedersen et al. (2007) have identified and reported the conserved blocks in *Arabidopsis* genome, and we used the homologous genes (anchor genes) to define the regions syntenic to predicted BGCs. These syntenic regions are called sister regions. After comparing BGCs and sister regions, there is no difference of conservation in biosynthetic and non-biosynthetic genes. Protein families, ferritin, methyltransferase, lyase, UDPGT and Pkinase, are more enriched than other protein families in anchor genes. Additionally, we only identified two saccharide BGCs (cluster 18 and cluster 28) located on the same syntenic block in *Arabidopsis* genome. On the other hand, importantly ecological BGCs, terpene clusters, are more likely to locate in the dynamic chromosomal regions than other type of clusters. Furthermore, we also explored the BGCs of *Cleome violacea* to *Arabidopsis thaliana*. Cleomaceae is sister-family of Brassicaceae and occurred α whole-genome triplication and shared β duplication with Brassicaceae. We identified one alkaloid BGC (cluster 30) in *Arabidopsis* that was also present in *Cleome* (cluster 1), meaning the structure has been conserved before the species' divergence 14.5 to 86 Mya.

To further examine the evolutionary history of BGCs in Brassicaceae, it is a good idea to compare the BGCs between *Arabidopsis* and other crucifer plants. Previous studies reported the 24 conserved Ancestral Crucifer Karyotype (ACK) genomic blocks of among Brassicaceae species (Schranz et al., 2006; Lysak et al., 2016). Based on ACK genomic blocks, we purpose to find more conserved BGCs among Brassicaceae.

Materials and Methodology

Sequence collection

The *A. thaliana* genome sequences and corresponding annotations were retrieved from two main databases: 1) The Arabidopsis Information Resource (TAIR) (https://www.arabidopsis.org/) (Berardini et al., 2015), 2) The National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.gov/genome/). The *C. violacea* genome was obtained from prof.dr. M. Eric Schranz. Specific information of syntenic blocks in *A. thaliana* genome was defined by supple. table 1 in Pedersen's study (2007). And the sister regions were identified by anchor genes showed in the aforementioned table.

BGCs identified by plantiSMASH

PlantiSMASH (Kautsar et al., 2017) pre-calculated biosynthetic gene clusters of *A. thaliana* with the Pfam database. 45 BGCs were identified by the automatic tool. We also used PlantiSMASH to detecte *Cleome* BGCs with *Cleome* genome sequences in FASTA format and annotation file in GFF3 format. 26 BGCs were predicted in *C. violacea* genome. The GBK files from plantiSMASH's results were used to retrieve amino acid sequences of loci.

Multiple sequence alignment and heatmap

To compare the BGCs and their sister regions, multiple sequence alignment was applied by MUSCLE (Edgar, 2004) with protein sequences FASTA format files. Multiple sequence alignments were displayed in heatmap by R script. The similar values were calculated by R packages "SeqINR" (Charif & Lobry, 2007) based on the aligned results, and displayed the value with heatmap.2 in R packages "gplots" (Warnes et al., 2009) The similarity rates of two loci were arranged from 0 to 1, which respectively indicated low similarity in blue blocks and high similarity in red blocks.

Protein families' identification and enrichment analysis

Protein families of anchor and non-anchor genes were identified by HMMer (Eddy, 1992) and web server Pfam 32.0 (Finn et al., 2016). Python script was written to analyse enrichment of each protein family under hypergeometric distribution. In this project, we used cumulative density function (hypergeom.cdf) in the python packages SciPy (Jones, 2014) to calculate the p value.

Synteny identification

Synteny between *A. thaliana* and *C. violacea* was identified by tool SynFind in CoGe platform (Lyons & Freeling, 2008). We uploaded the genome of *C. violacea* which was in scaffold level and queried syntenic regions against *A. thaliana* genome. The conserved regions were shown in linkage.

Acknowledgement

I highly appreciate my supervisors, Dr. Marnix Medema, Prof. Dr. M Eric Schranz and Hernando Suarez-Duran for their professional guidance and supervising in this project. I would also be grateful to theirs encourages and full patience to me. I also thank all colleagues studied and worked in bioinformatics group. I spent a great time with them. Most importantly, I would like to thank my parents who supported me not only in finance, but also giving me freedom to live aboard and all of their love to me.

References

Barker, M. S., Vogel, H., & Schranz, M. E. (2009). Paleopolyploidy in the Brassicales: Analyses of the Cleome Transcriptome Elucidate the History of Genome Duplications in Arabidopsis and Other Brassicales. *Genome Biology and Evolution*, *1*, 391–399. https://doi.org/10.1093/gbe/evp040

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis*, *53*(8), 474–485. https://doi.org/10.1002/dvg.22877

Blin, K., Kim, H. U., Medema, M. H., & Weber, T. (2017). Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbx146

Bowers, J. E., Chapman, B. A., Rong, J., & Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, *422*(6930), 433–438. https://doi.org/10.1038/nature01521

Boycheva, S., Daviet, L., Wolfender, J.-L., & Fitzpatrick, T. B. (2014). The rise of operon-like gene clusters in plants. *Trends in Plant Science*, *19*(7), 447–459. https://doi.org/10.1016/J.TPLANTS.2014.01.013

Charif, D., & Lobry, J. R. (2007). SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis (pp. 207–232). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-35306-5_10

Delker, C., Zolman, B. K., Miersch, O., & Wasternack, C. (2007). Jasmonate biosynthesis in Arabidopsis thaliana requires peroxisomal β-oxidation enzymes – Additional proof by properties of pex6 and aim1. *Phytochemistry*, *68*(12), 1642–1650. https://doi.org/10.1016/J.PHYTOCHEM.2007.04.024

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Field, B., Fiston-Lavier, A.-S., Kemen, A., Geisler, K., Quesneville, H., & Osbourn, A. E. (2011). Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proceedings of the National Academy of Sciences*, *108*(38), 16116–16121. https://doi.org/10.1073/pnas.1109273108

Finkelstein, R. (2013). Abscisic Acid synthesis and response. *The Arabidopsis Book*, *11*, e0166. https://doi.org/10.1199/tab.0166

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279–D285. https://doi.org/10.1093/nar/gkv1344

Hall, J. C., Sytsma, K. J., & Iltis, H. H. (2002b). Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data. *American Journal of Botany*, *89*(11), 1826–1842. https://doi.org/10.3732/ajb.89.11.1826

Hall, J. C., Sytsma, K. J., & Iltis, H. H. (2002a). Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data. *American Journal of Botany*, *89*(11), 1826–1842. https://doi.org/10.3732/ajb.89.11.1826

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). PlantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research*, *45*(W1), W55–W63. https://doi.org/10.1093/nar/gkx305

Lyons, E., & Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*, *53*(4), 661–673. https://doi.org/10.1111/j.1365-313X.2007.03326.x

Lysak, M. A., Mandáková, T., & Schranz, M. E. (2016a). Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Current Opinion in Plant Biology*, *30*, 108–115. https://doi.org/10.1016/J.PBI.2016.02.001

Lysak, M. A., Mandáková, T., & Schranz, M. E. (2016b). Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Current Opinion in Plant Biology*, *30*, 108–115. https://doi.org/10.1016/J.PBI.2016.02.001

Nambara, E., & Marion-Poll, A. (2005). ABSCISIC ACID BIOSYNTHESIS AND CATABOLISM. *Annual Review of Plant Biology*, *56*(1), 165–185. https://doi.org/10.1146/annurev.arplant.56.032604.144046

Nützmann, H.-W., Huang, A., & Osbourn, A. (2016). Plant metabolic clusters - from genetics to genomics. *New Phytologist*, *211*(3), 771–789. https://doi.org/10.1111/nph.13981

Osbourn, A. (2010). Gene clusters for secondary metabolic pathways: an emerging theme in plant biology. *Plant Physiology*, *154*(2), 531–535. https://doi.org/10.1104/pp.110.161315

Patchell, M. J., Roalson, E. H., & Hall, J. C. (2014). Resolved phylogeny of cleomaceae based on all three genomes. *Taxon*, *63*(2), 315–328. https://doi.org/10.12705/632.17

Pedersen, B., Lyons, E., Thomas, B. C., Rapaka, L., & Freeling, M. (2007). G-Boxes, Bigfoot Genes, and Environmental Response: Characterization of Intragenomic Conserved Noncoding Sequences in Arabidopsis. *The Plant Cell Online*, *19*(5), 1441–1457. https://doi.org/10.1105/tpc.107.050419

Rutledge, P. J., & Challis, G. L. (2015). Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nature Reviews Microbiology*, *13*(8), 509–523. https://doi.org/10.1038/nrmicro3496

Schranz, M. E., Lysak, M. A., & Mitchell-Olds, T. (n.d.). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. https://doi.org/10.1016/j.tplants.2006.09.002

Schranz, M. E., & Mitchell-Olds, T. (2006). Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *The Plant Cell*, *18*(5), 1152–1165. https://doi.org/10.1105/tpc.106.041111

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., … Soltis, P. S. (2009). Polyploidy and angiosperm diversification. *American Journal of Botany*, *96*(1), 336–348. https://doi.org/10.3732/ajb.0800079

Tholl, D., & Lee, S. (2011). Terpene Specialized Metabolism in Arabidopsis thaliana. *The Arabidopsis Book*, *9*, e0143. https://doi.org/10.1199/tab.0143

Thomas, B. C., Pedersen, B., & Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research*, *16*(7), 934–946. https://doi.org/10.1101/gr.4708406

Vision, T. J., Brown, D. G., & Tanksley, S. D. (2000). The origins of genomic duplications. *In Arabidopsis. Science*, *290*(December), 2114–2117.

Wang, M. L., Belmonte, S., Kim, U., Dolan, M., Morris, J. W., & Goodman, H. M. (1999). A cluster of ABA-regulated genes on Arabidopsis thaliana BAC T07M07. *Genome Research*, *9*(4), 325–333. https://doi.org/10.1101/GR.9.4.325

Wendel, J. F. (2000). Genome evolution in polyploids BT - Plant Molecular Evolution. *Plant Molecular Biology*, *42*(1), 225–249. Retrieved from http://www.springerlink.com/index/10.1007/978-94-011-4221-2_12%5Cnpapers3://publication/doi/10.1007/978-94-011-4221-2_12

Supplementary data

Supplementary data 1. BGCs of *Arabidopsis thaliana* predicted by PlantiSMASH

http://plantismash.secondarymetabolites.org/precalc/results/Arabidopsis_thaliana/

Supplementary data 2. BGCs of *Cleome violacea* predicted by PlantiSMASH

http://plantismash.secondarymetabolites.org/upload/eb36c124-2af7-4989-9f89-281979026a1c/index.html

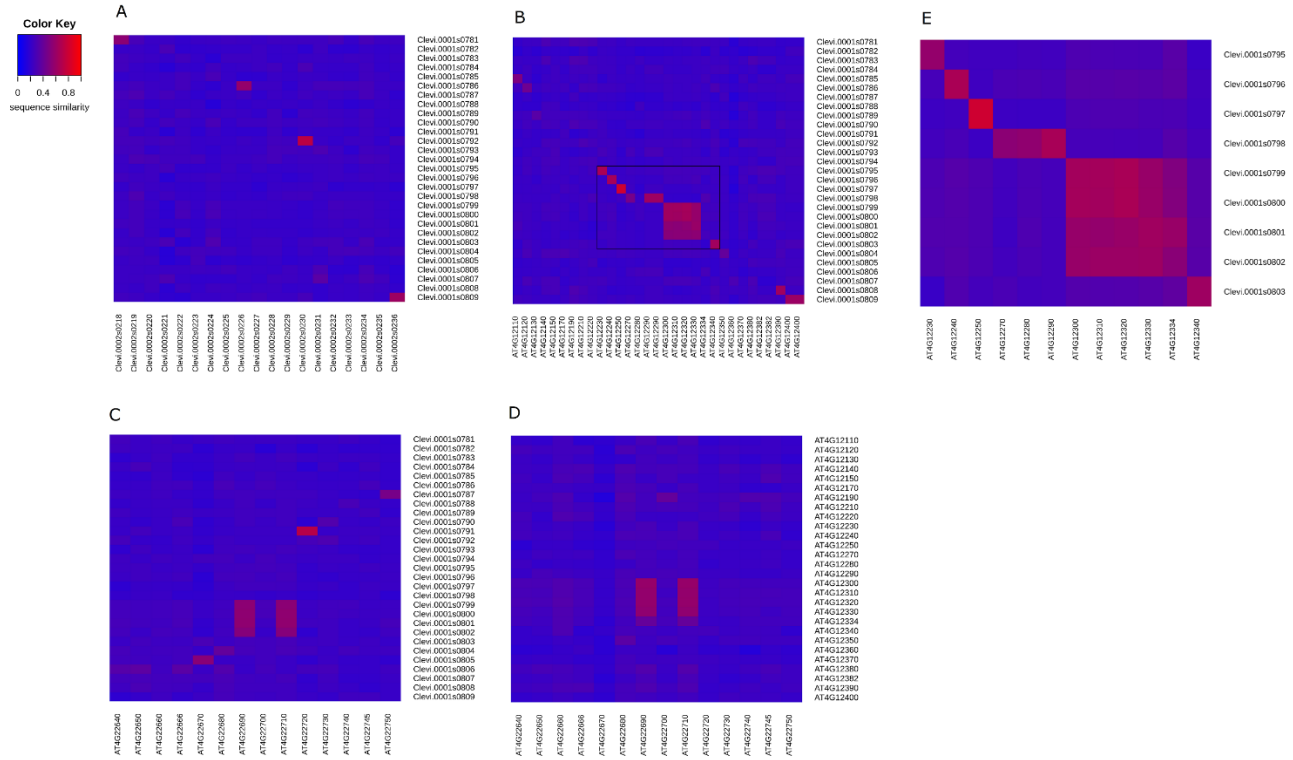

Figure 1 Heatmap: A) Syntenic region of cluster 1 in Cleome vs its sister region. B) Syntenic region of cluster 1 in Cleome vs cluster 30 in Arabidopsis. C) Syntenic region of cluster 1 in Cleome vs cluster 30 sister region in Arabidopsis D) Syntenic reg

Table 1 Enrichment analysis of anchor genes

Protein Family	Number in anchor genes	Number in all genes	p_value
Ferritin	2	2	0
Methyltransf_29	3	3	0
Lyase_aromatic	2	2	0
Acetyltransf_1	2	3	0.004961
Methyltransf_25	4	9	0.009742
Methyltransf_11	4	10	0.016823
PRA1	1	2	0.029261
LRRNT_2	1	2	0.029261
PrmA	1	2	0.029261
polyprenyl_synt	1	2	0.029261
SOUL	1	2	0.029261
FR47	1	2	0.029261
Homeobox_KN	1	2	0.029261
X8	1	2	0.029261
Glyco_tran_28_C	1	2	0.029261
LRR_8	1	2	0.029261
LRR_1	1	2	0.029261
Ank	1	2	0.029261
zf-B_box	1	2	0.029261
UDPGT	8	28	0.036594
Ank_4	1	3	0.077863
Kunitz_legume	1	3	0.077863
Homeobox	1	3	0.077863
Ank_3	1	3	0.077863
Reticulon	1	3	0.077863
EF-hand_1	1	3	0.077863
EF-hand_5	1	3	0.077863
EF-hand_7	1	3	0.077863
EF-hand_6	1	3	0.077863
EF-hand_8	1	3	0.077863
Str_synth	1	3	0.077863
SGL	1	3	0.077863
Pkinase_Tyr	4	15	0.097247

Pkinase	4	15	0.097247
Methyltransf_7	1	4	0.13839
PPR_long	2	8	0.143222
PPR_2	2	8	0.143222
PPR_3	2	8	0.143222
PPR_1	2	8	0.143222
PPR	2	8	0.143222
DIOX_N	3	13	0.169325
Dirigent	2	9	0.188953
adh_short_C2	1	5	0.205378
Methyltransf_31	1	5	0.205378
GDP_Man_Dehyd	1	5	0.205378
adh_short	1	5	0.205378
RmlD_sub_bind	1	5	0.205378
LRR_4	1	5	0.205378
KR	1	5	0.205378
2OG-FeII_Oxy	3	14	0.207131
AMP-binding	1	6	0.274864
Terpene_synth	1	8	0.410981
Epimerase	1	8	0.410981
Terpene_synth_C	1	8	0.410981
Transferase	1	18	0.843365
p450	2	45	0.990459