Comparative Genomics and Trait Evolution in Cleomaceae, a Model Family for Ancient Polyploidy

Erik van den Bergh

Comparative Genomics and Trait Evolution in Cleomaceae, a Model Family for Ancient Polyploidy

Erik van den Bergh

Propositions

1. To enjoy the evolutionary playground of polyploidy the hurdle of diploidization must first be passed.
   (this thesis)

2. Cleomaceae species provide a case study for the effects of polyploidy on numerous traits, such as C4 photosynthesis, glucosinolates and many others.
   (this thesis)

3. The accessibility of open data is as important as its availability.

4. Fabricated 'truths' in spite of scientific observations and conclusions show a lack of understanding, not a lack of goodwill.

5. What brings people together is not a good speech, but a good listen.

6. Parenthood is about guiding the next generation and forgiving the last.

Propositions belonging to the thesis entitled:
"Comparative genomics and trait evolution in Cleomaceae, a model family for ancient polyploidy"
Erik van den Bergh
Wageningen, May 17th 2017

COMPARATIVE GENOMICS AND TRAIT EVOLUTION IN CLEOMACEAE, A MODEL FAMILY FOR ANCIENT POLYPLOIDY

ERIK VAN DEN BERGH

Thesis committee

Promotors

Prof. Dr M.E. Schranz
Professor of Biosystematics
Wageningen University & Research

Prof. Dr Y. van de Peer
Professor in Bioinformatics and Genome Biology
Ghent University, Belgium

Other members

Prof. Dr D. de Ridder, Wageningen University & Research

Dr F.P. Lens, Naturalis, Leiden

Dr M.G.M. Aarts, Wageningen University & Research

Dr A.T. Groot, University of Amsterdam

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

COMPARATIVE GENOMICS AND TRAIT EVOLUTION IN CLEOMACEAE, A MODEL FAMILY FOR ANCIENT POLYPLOIDY

Erik van den Bergh

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 17 May 2017
at 1.30 p.m. in the Aula.

Erik van den Bergh
Comparative genomics and trait evolution in Cleomaceae, a model family for ancient polyploidy, 106 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2017)
With references, with summaries in English and Dutch

ISBN: 978-94-6343-170-5

DOI: http://dx.doi.org/10.18174/412208

CONTENTS

Summary **6**

Samenvatting **8**

Introduction **10**

Chapter 1 **20**

The genome of Tarenaya hassleriana provides insights into reproductive trait and genome evolution of crucifers

Chapter 2 **41**

Gene and Genome Duplications and the Origin of C_4 Photosynthesis: Birth of a Trait in the Cleomaceae

Chapter 3 **53**

Flower power and the mustard bomb: Comparative analysis of gene and genome duplications in glucosinolate biosynthetic pathway evolution in Cleomaceae and Brassicaceae

Chapter 4 **67**

Anthocyanins and flower colour in Cleome, identification of genetic variation underlying floral colouring patterns

General Conclusion **81**

References **84**

Acknowledgements **103**

SUMMARY

As more and more species have been sequenced, evidence has been piling up for a fascinating phenomenon that seems to occur in all plant lineages: paleopolyploidy. Polyploidy has historically been a much observed and studied trait, but until recently it was assumed that polyploids were evolutionary dead-ends due to their sterility. However, many studies since the 1990's have challenged this notion by finding evidence for ancient genome duplications in many genomes of current species. This lead to the observation that all seed plants share at least one ancestral polyploidy event. Another polyploidy event has been proven to lie at the base of all angiosperms, further signifying the notion that ancient polyploidy is widespread and common. These findings have led to questions regarding the apparent disadvantages that can be observed in a first generation polyploid. If these disadvantages can be overcome however, duplication of a genome also presents an enormous potential for evolutionary novelty. Duplicated copies of genes are able to acquire changes that can lead to specialization of the duplicated pair into two functions (subfunctionalization) or the development of one copy towards an entirely new function (neofunctionalization).

Currently, most research towards polyploidy has focused on the economically and scientifically important Brassicaceae family containing the model plant *Arabidopsis thaliana* and many crops such as cabbage, rapeseed, broccoli and turnip. In this thesis, I lay the foundations for the expansion of this scope to the Cleomaceae, a widespread cosmopolitan plant family and a sister family of Brassicaceae. The species within Cleomaceae are diverse and exhibit many scientifically interesting traits. They are also in a perfect position phylogenetically to draw comparisons with the much more studied Brassicaceae. I describe the Cleomaceae and their relevance to polyploid research in more detail in the Introduction. I then describe the important first step towards setting up the genetic framework of this family with the sequencing of *Tarenaya hassleriana* in Chapter 1.

In Chapter 2, I have studied the effects of polyploidy on the development of C4 photosynthesis by comparing the transcriptome of C3 photosynthesis based species *Tarenaya hassleriana* with the C4 based *Gynandropsis gynandra*. C4 photosynthesis is an elaboration of the more common C3 form of photosynthesis that concentrates CO2 in specific cells leading to decreased photorespiration by the RuBisCO and higher photosynthetic efficiency in low CO2 environments. I find that polyploidy has not led to sub- or neofunctionalization towards the development of this trait, but instead find evidence for another important phenomenon in postpolyploid evolution: the dosage balance hypothesis. This hypothesis states that genes which are dependent on specific dosage levels of their products will be maintained in duplicate; any change in their function would lead to dosage imbalance which would have deleterious effects on their pathway. We show that most genes involved in photosynthesis have returned to single copy in *G. gynandra* and that the changes leading to C4 have mostly taken place at the expression level confirming current assumptions on the development of this trait.

In Chapter 3, I have studied the effects of polyploidy on an important class of plant defence compounds: glucosinolates. These compounds, sometimes referred to as 'mustard oils', play an important role in the defence against herbivores and have radiated widely in Brassicaceae to form many different 'flavors' to deter specific herbivores. I show that in Cleomaceae many genes responsible for these compounds have benefited from the three rounds of polyploidy that *T. hassleriana* has undergone and that many duplicated genes have been retained. We also show that more than 75% is actively expressed in the plant, proving that the majority of these duplications has an active function in the plant.

Finally, in Chapter 4 I investigate a simple observation made during experiments with *T. hassleriana* in the greenhouse regarding the variation in flower colour between different individuals: some had pink flowers and some purple. Using LC-PDA mass spectrometry we find that the two colours are caused by different levels of two anthocyanin pigments, with cyanidin dominating in the purple flowers and

pelargonidin being more abundant in pink flowers. Through sequence comparison and synteny analysis between *A. thaliana* and *T. hassleriana* we find the orthologs of the genes involved in this pathway. Using a Genotyping by Sequencing method on a cross between these two flower colours, we produce a collection of SNP markers on the reference genome. With these SNPs, we find two significant binary trait loci, one of which corresponds to the location of the F3'H ortholog which performs the conversion of a pelargonidin precursor to a cyanidin precursor.

In the General Conclusion, I combine all findings of the previous chapters and explain how they establish part of a larger species framework to study ancient polyploidy in angiosperms. I then put forth what these findings can mean for possible future research and the directions that are worth to be explored further.

SAMENVATTING

Naarmate meer en meer soorten gesequenced worden, stapelt het bewijs zich op voor een fascinerend fenomeen dat lijkt voor te komen in alle plantenfamilies: paleopolyploïdie. Polyploïdie is historisch gezien een veel geobserveerde en bestudeerde eigenschap, maar tot recentelijk was de aanname dat polyploïden evolutionair gezien doodlopende lijnen waren vanwege hun steriliteit. Echter, veel studies sinds de jaren ’90 hebben deze aanname betwist met het vinden van bewijs voor zeer oude genoomduplicaties in vele genomen van huidige soorten. Dit heeft geleid tot de observatie dat alle zaadplanten op zijn minst één voorouderlijke polyploïdieronde delen. Daarbovenop is bewezen dat er een tweede polyploïdie aan de basis ligt van alle angiospermen, wat aannemelijk maakt dat voorouderlijke polyploïdie wijdverspreid en algemeen is. Deze vindingen leiden ook tot vragen over de ogenschijnlijke nadelen die geobserveerd kunnen worden in de eerste generatie van polyploïde organismen. Als deze nadelen echter kunnen worden overwonnen, leidt het dupliceren van het genoom tot een enorme hoeveelheid potentieel voor nieuwe evolutionaire mogelijkheden. Gedupliceerde kopieën van genen kunnen veranderingen verzamelen die kunnen leiden tot specialisatie van een genenpaar in twee nieuwe functies (subfunctionalisatie) of de ontwikkeling van één kopie tot een compleet nieuwe functie (neofunctionalisatie).

Op dit moment is het meeste onderzoek naar polyploïdie gericht op de economisch en wetenschappelijk belangrijke Brassicaceae familie, die de modelplant *Arabidopsis thaliana* omvat maar ook vele oogstgewassen zoals kool, raapzaad, broccoli en raap. In deze thesis leg ik de basis voor het uitbreiden van dit perspectief naar de Cleomaceae, een kosmopolitische plantenfamilie en een zusterfamilie van de Brassicaceae. De soorten in de Cleomaceae zijn divers en tonen vele wetenschappelijk interessante eigenschappen. Phylogenetisch gezien liggen ze ook in een perfecte positie om een vergelijking te maken met de veel meer bestudeerde Brassicaceae. Ik beschrijf de belangrijke eerste stap tot het opzetten van dit vergelijkende raamwerk met het sequencen van *Tarenaya hassleriana* in Hoofdstuk 1.

In Hoofdstuk 2 heb ik de effecten van polyploïdie op de ontwikkeling van C4 fotosynthese bestudeerd door het transcriptoom van de C3 soort *Tarenaya hassleriana* the vergelijken met de C4 soort *Gynandropsis gynandra*. C4 fotosynthese is een uitbreiding van de meer algemene C3 vorm van fotosynthese. C4 fotosynthese concentreert CO2 in specifieke cellen wat leidt tot minder fotorespiratie door RuBisCO en een verhoogde efficiëntie in omgevingen met een lage CO2 concentratie. Ik observeer in dit hoofdstuk dat polyploïdie niet heeft geleid tot sub- of neofunctionalisatie in de ontwikkeling van C4 fotosynthese maar in plaats daarvan vind ik bewijs voor een ander belangrijk fenomeen in post-polyploïde evolutie: de dosering-balans hypothese. Deze hypothese stelt dat genen die afhankelijk zijn van specifieke doserings niveaus van hun product behouden zullen blijven als twee duplicaten; elke verandering in hun functie zou leiden tot doserings-imbalans wat negatieve gevolgen zou hebben voor de eigenschap waarvan zij deel uitmaken. We laten ook zien dat de meeste genen die betrokken zijn bij fotosynthese teruggekeerd zijn naar één kopie in *G. gynandra* en dat de veranderingen die geleid hebben tot de ontwikkeling naar C4 het meest hebben plaatsgevonden in de expressieniveaus. Dit bevestigt de huidige opvattingen over de ontwikkeling van deze eigenschap.

In Hoofdstuk 3 heb ik de effecten van polyploïdie bestudeerd op een belangrijke klasse van plantaardige verdedigingsstoffen: glucosinolaten. Deze verbindingen, die soms ‘mosterd-olieën’ genoemd worden, spelen een belangrijke rol in de verdediging tegen herbivoren en zijn wijd uitgewaaierd in de Brassicaceae om vele verschillende ‘smaken’ te vormen voor de afweer tegen verschillende herbivoren. Ik laat zien dat in Cleomaceae vele genen die coderen voor deze verbindingen baat hebben gehad bij de drie rondes polyploïdie die *T. Hassleriana* heeft ondergaan en dat veel gedupliceerde genen zijn behouden. We laten ook zien dat meer dan 75% van deze genen actief tot expressie wordt gebracht, waarmee we bewijzen dat de meerderheid van deze duplicaten een actieve functie in de plant hebben.

Tot slot onderzoek ik in Hoofdstuk 4 een simpele observatie die werd gemaakt tijdens experimenten met *T. hassleriana* in de kas met betrekking tot de variatie in bloemkleur tussen verschillende individuen: sommige hadden roze bloemen en sommige paarse. Met het gebruik van LC-PDA massaspectrometrie vinden we dat de beide kleuren veroorzaakt worden door verschillende niveaus van twee anthocyaninepigmenten, waarbij cyanidine overheerst in de paarse bloemen en pelargonidine meer aanwezig is in de roze bloemen. Door middel van sequentievergelijking en syntenie analyse tussen *A. thaliana* en *T. hassleriana* vinden we de orthologen van de genen die betrokken zijn bij deze eigenschap. Met een Genotyping by Sequencing methode op een kruising tussen deze twee bloemkleuren produceren we een verzameling enkel nucleotide polymorfismen (SNPs) op het referentie genoom van *T. hassleriana*. Met deze SNPs vinden we twee significante binaire eigenschaps loci, waarvan één overeenkomt met de locatie van een ortholoog van F3'H, een gen dat de omzetting verzorgt van een precursor van pelargonidine naar een cyanidine precursor.

In de Algemene Conclusie combineer ik alle vindingen uit de voorafgaande hoofdstukken en leg ik uit hoe zij een groter soortenraamwerk vormen wat gebruikt kan worden om voorouderlijke polyploïdie in angiospermen te bestuderen. Ik stel daarna voor wat deze vindingen kunnen betekenen voor mogelijk toekomstig onderzoek en de richtingen die het waard zijn om verder te worden onderzocht.

INTRODUCTION

POLYPLOIDY & GENE DUPLICATION

Polyploidy is the state of an organism that has more than two copies of a genome in its cells. By contrast, most organisms are diploid, meaning there are exactly two copies of its genome in each cell corresponding to two sets of chromosomes (2n). Polyploidy can be further divided into allo- and autopolyploid: allopolyploidy is a state of polyploidy that has occurred due to hybridization between two different species, resulting in genome copies of both organisms to be present in one cell line. Autopolyploidy occurs when a genome is multiplied within the same species.

Through biological history, polyploidy has sometimes been presented as a fluke occurrence with no significant effects for the fitness and evolution of organisms. For example, in his scathing review of "alleged biosystematics studies", W.H. Wagner states that polyploids, apomicts, inbreeders and hybrids must be seen as evolutionary dead ends whereas the normal, diploid, outbreeding species continue to evolve (Wagner, 1970). The prevailing view was that evolution strictly took place in small, constant steps and that polyploids and hybrids would never lead to the development of a successful lineage. Considering the fact that hybrids and polyploids are often sterile due to issues with chromosome pairing this is not an unreasonable assumption. However, in his seminal book "Evolution by gene duplication" Susumu Ohno hypothesized that natural selection will only police and restrict small mutations and that evolutionary novelty must come from redundant genes that escape from the pressure of natural selection to form a new gene (Ohno, 1970). In his book he describes many processes that have been found in many natural systems in today's research: neo- and subfunctionalization (Monson, 2003; Flagel and Wendel, 2009; Freeling, 2009; Glasauer and Neuhauss, 2014), the process by which duplicate genes achieve new and altered functions, duplication of regulatory genes leading to novel expression patterns and subsequent altered phenotypes and the gene balance effect, where genes that are dosage sensitive can suffer from duplicated alleles (Edger and Pires, 2009; Birchler and Veitia, 2012; Conant et al., 2014; Spillane and McKeown, 2014). All of these effects will be described in more detail in the following sections.

Polyploidy is a common occurrence in crop species such as wheat (The International Wheat Genome Sequencing Consortium (iwgsc), 2014), rye (Martis et al., 2013), members of the cabbage family (Blanc and Wolfe, 2004; The Brassica rapa Genome Sequencing Project Consortium et al., 2011), cotton (Li, Fan, et al., 2015), banana (D'Hont et al., 2012), apple (Velasco et al., 2010), and many others. This reflects the human selection that has taken place on these crops, which tends to favor crops with the highest yield. Here, a phenomenon known as 'hybrid vigor' comes into play; the effect that hybrid crosses (either intraspecies or sufficiently removed interspecies) tend to show faster growth, height and yield than both of the parents. The underlying mechanisms of this effect are still partly unknown, but it is clear that the (unintentional) creation of hybrids and the associated heterosis has played an important role in the amount of polyploidy found in many crops today (Ladizinsky, 1998).

For example, in the cotton genus (*Gossypium*) four species have been domesticized throughout human history and subsequently hybridized through unintentional and intentional means, resulting in a hybrid spectrum with 8 separate genome groups (A-G, and K) (Renny-Byfield et al., 2015). In this particular case, a genome group is based on observations of pairing behavior of chromosomes and the fertility of interspecific hybrids. The crop that currently dominates 90% of the worlds cultivated cotton, *G. hirsutum* is a product of a coincidental natural allopolyploidy produced by a crossing between the A and D genomes from Africa and the Americas, respectively. However, it has been speculated that the the A-genome cotton was carried across the Pacific by man as a seed and subsequently hybridized with wild cotton in Australia (Stephens, 1947), producing the current hybrid with superior fiber and flowering time. Problematic is that *Gossypium* shows a strong tendency for intergenomic gene transfer, making

the molecular dating of polyploid events difficult. The dating of polyploidy events is of great importance and great strides have been made in this regard, as we shall see later on.

When dealing with polyploidy in relation to genes, it is important to establish definitions for genes that are related and the manner in which they are related. Classically, there are three terms to describe these relationships: homolog, ortholog and paralog. Homolog is a superterm indicating that two genes have a common ancestor. Homologous genes can be further subdivided into orthologs, where a speciation event has resulted in two separate genes and paralogs, where a gene duplication has resulted in two separate genes. Of course, these are not the only scenarios possible, as genes keep duplicating and speciation events keep happening over time. For example, when duplicated ancestor genes are separated through a speciation event and subsequently duplicated once more, the duplicated genes within one species are referred to as in-paralogs and between the two organisms are referred to as outparalogs. One special class of paralogs must also be mentioned: when a paralog is a result of ancient polyploidy it is, fittingly, referred to as “ohnolog”.

ANCIENT POLYPLOIDY IN PLANTS

Polyploidy can have two forms: recent polyploidy and ancient polyploidy which has taken place in an ancestor species. One of the first scientists to propose ancestral polyploidy was Susumu Ohno, who (correctly) deduced that the transition of the primitive Tunicate-like creatures to fish must have been a result of polyploidy (Dehal and Boore, 2005). Through measurements of chromosome sizes in various living relatives he assumed that the large variation in genome sizes in teleost fish, especially when compared to *Tunicatum* are caused by polyploidy. In recent studies, polyploidy has been shown to be present in the ancestor of teleost fish and is referred to as the teleost genome duplication (TGD) (Glasauer and Neuhauss, 2014). In later work Ohno proposed that another polyploidy event must have taken place before the amphibian transition to land (Ohno, 1973) resulting in all land animals as we

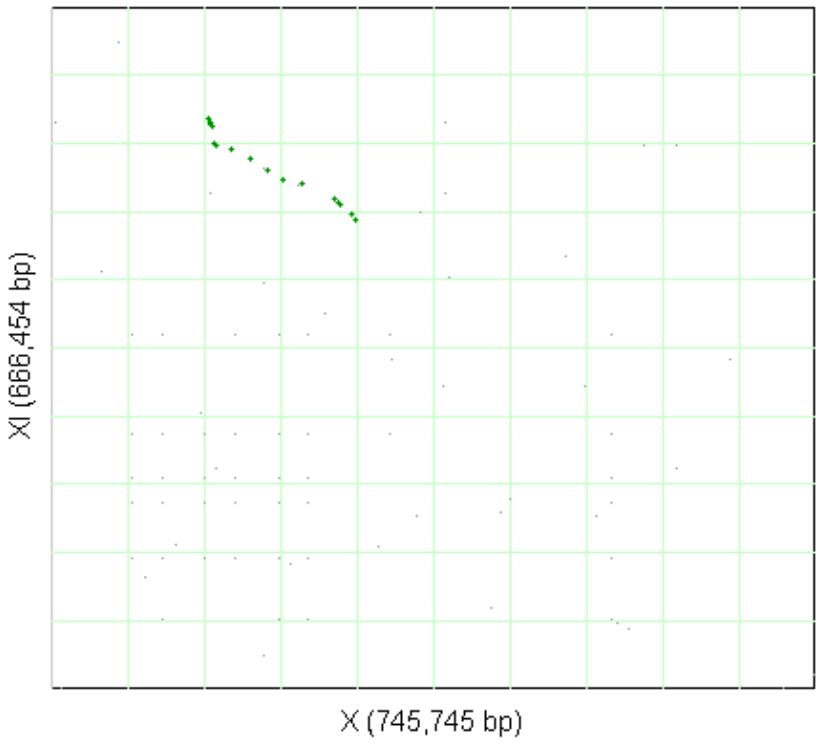


Figure 1. Dot plot of similar gene pairs between S. cerevisiae chromosome X (X-axis) and chromosome XI (Y-axis). The syntenic region at the top left is referred to as block 41 in Wolfe et al, 1997.

know them today. However, “As the well-entrenched chromosomal sex determining mechanism was acquired, it became impossible for reptiles to utilize polyploidization as a means of obtaining new gene loci.” (Ohno, 1970). The hypothesis that two rounds of polyploidy were at the base of all land animals and reptiles is now referred to as the 2r hypothesis and picked up significant interest from the 1990’s onwards as genetic tools allowed researchers more and more detail. Clear evidence to support this hypothesis accumulated, leading to a now widespread acceptance of the notion of two rounds of polyploidy at the base of the vertebrate tree.

The acceptance of the 2R hypothesis was helped by techniques developed in two non-animal model species: yeast (Saccharomyces cerevisiae) and the model plant *Arabidopsis thaliana*. In his seminal paper, Kenneth Wolfe showed convincingly that yeast had experienced a whole genome duplication (WGD) in its history (Wolfe and Shields, 1997). Key observations were made not through the comparison of gene content, but gene position. This led to the observation of chromosomal regions that have many anchor genes in a similar position, a phenomenon known as synteny.

For example, when comparing chromosome X and XI in yeast we can create a dot plot by putting one chromosome on the X axis and one on the Y axis. When an ohnolog gene pair is found between these two chromosomes we use the position on both chromosomes as an X,Y coordinate pair and draw a dot at this coordinate. When repeated for every gene, the resulting figure gives an overview of the gene positional similarity between these two chromosomes (Figure 1)

By identifying the structural similarity of these two chromosomes and finding similar syntenic regions across the yeast genome Wolfe and his colleagues convincingly showed that yeast had undergone at least one round of ancient polyploidy. A detailed map of all the syntenic regions in yeast was constructed and each continuous syntenic region was given a block number, ranging from block 1 to 55.

To strengthen the theory that these blocks were not the product of widespread segmental or any other type of regional chromosomal duplication, another factor proved important. Several key genes in yeast of which the orthologs were known in other species, were compared in terms of sequence identity (Wolfe and Shields, 1997). This allowed a simple type of molecular dating through the ratio of synonymous mutations between these genes (Ks). It was shown that the molecular age of most gene pairs was very similar (specifically between a Ks of 0.5 - 0.9) meaning that all blocks were approximately the same age and must thus be a product of the same event. More details on the molecular dating of polyploidy will be discussed further on.

After having shown the ancestral polyploidy of *Saccharomyces cerevisiae*, a collaboration between Kenneth Wolfe and Guillaume Blanc lead to another influential paper on polyploidy in *Arabidopsis* (Blanc et al., 2003). In it, they define 108 syntenic conserved blocks that are a product of ancient polyploidy. Through more detailed analysis of the molecular age of the blocks based on gene pair Ks, the blocks were classified in two age groups: a ‘recent’ group which covers the entire genome and a much more sparsely distributed ‘old’ group, belonging to an older polyploidy event. Less than 6 months later, John Bowers together with his colleagues released a study showing widespread synteny in *Arabidopsis* (Bowers et al., 2003). They too find an older set of blocks nested within the more recent blocks and couple all found blocks to the three major polyploidy events in the *Arabidopsis* lineage: the α, β and γ events.

The α, β and γ events range from Brassicaceae specific, to Brassicales specific, to Rosid specific respectively. The α event is the most recent, and is estimated to have occurred 56 Ma (Kagale et al., 2014). It is one of the most well studied polyploidy events in plants, especially in *Arabidopsis*. The syntenic blocks originally identified by Wolfe have been restudied and improved (Bowers et al., 2003; Schranz et al., 2006; Lysak et al., 2016) and have been reconciled into a potential ancestral crucifer

karyotype. This was done based on not only blocks found in *Arabidopsis* but on multiple versions of the syntenic regions found in *A. lyrata*, *Capsella rubella*, *Brassica rapa* and later refined using more and more sequenced Brassicaceae.

The α event has also served as a model for trait and gene family evolution after polyploidy. In *Aethionema arabicum*, a Brassicaceae that shares At-α it has been shown that ohnolog glucosinolate genes have been preferentially retained conferring novel types of glucosinolates (Hofberger et al., 2013); a direct result of polyploidy which had been predicted earlier (Schranz et al., 2011). Another class of genes, NUCLETOTIDE BINDING/LEUCINE-RICH REPEATS (NBS-LRR) that provide rapidly evolving variable resistances have been strongly retained following At-α, providing a large evolutionary playground that can rapidly adapt existing genes as new resistances are necessary (Hofberger et al., 2014).

The α event provides a model for the study of species radiation after polyploidy. Polyploids have been shown to have lower diversification rates and higher extinction rates than diploids (Mayrose et al., 2011, 2015; Soltis et al., 2014) seemingly contradicting the observation that polyploidy is common in the ancestors of all plant lineages. Polyploids have been shown to have lower reproductive fitness, deleterious long-term effects due to reproductive biology shifts, and higher rates of (deleterious) mutations (Arrigo and Barker, 2012). How then, are we to believe that these 'hopeful monsters' stood at the beginning of all higher order life on Earth?

The answer is that the benefits of polyploidy most likely outweigh the drawbacks, especially in times of abiotic and biotic bottlenecks. The proof of this lies in the fact that all modern plants are ancient polyploids (Jiao et al., 2011) as summarized in Figure 2. Below we describe the hypothesis that the polyploidy event at the base of the rosids provided benefits causing a large radiation during the K-Pg mass extinction.

MOLECULAR DATING OF GENE AND GENOME DUPLICATIONS

When studying polyploid history and ancestry estimation of the age of a polyploidy event is vital information. Synonymous substitution rate (Ks) has been mentioned before and it will be explained in more detail here. When a gene is translated from RNA to protein, it is well known that the third base in the coding triplet is less specific than the first two (leading to it being referred to as the 'wobble base'). Evolutionary speaking, this has the effect that mutations in the third base of a triplet are less likely to have a phenotypic effect and when it does not it is referred to as a synonymous substitution. Synonymous substitution rates on their own or as a ratio of nonsynonymous : synonymous mutations

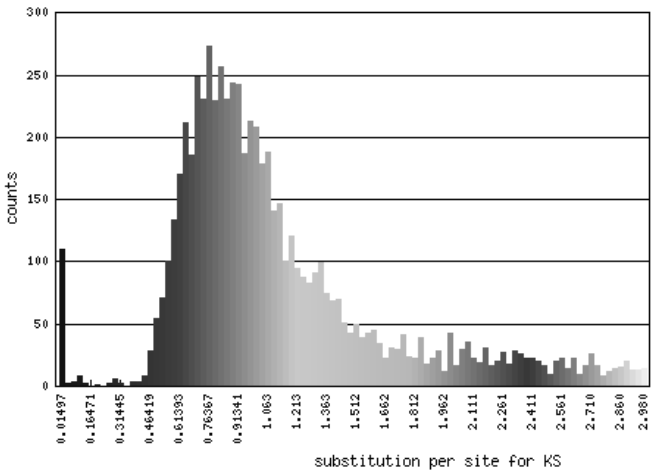


Figure 3. Histogram of binned Ks counts for *Arabidopsis* interspecies paralogs.

(Ka/Ks) have been used for a long time as a measure of selective pressure on protein coding genes (Yang and Bielawski, 2000). This is based on the assumption that synonymous mutations will accumulate in a

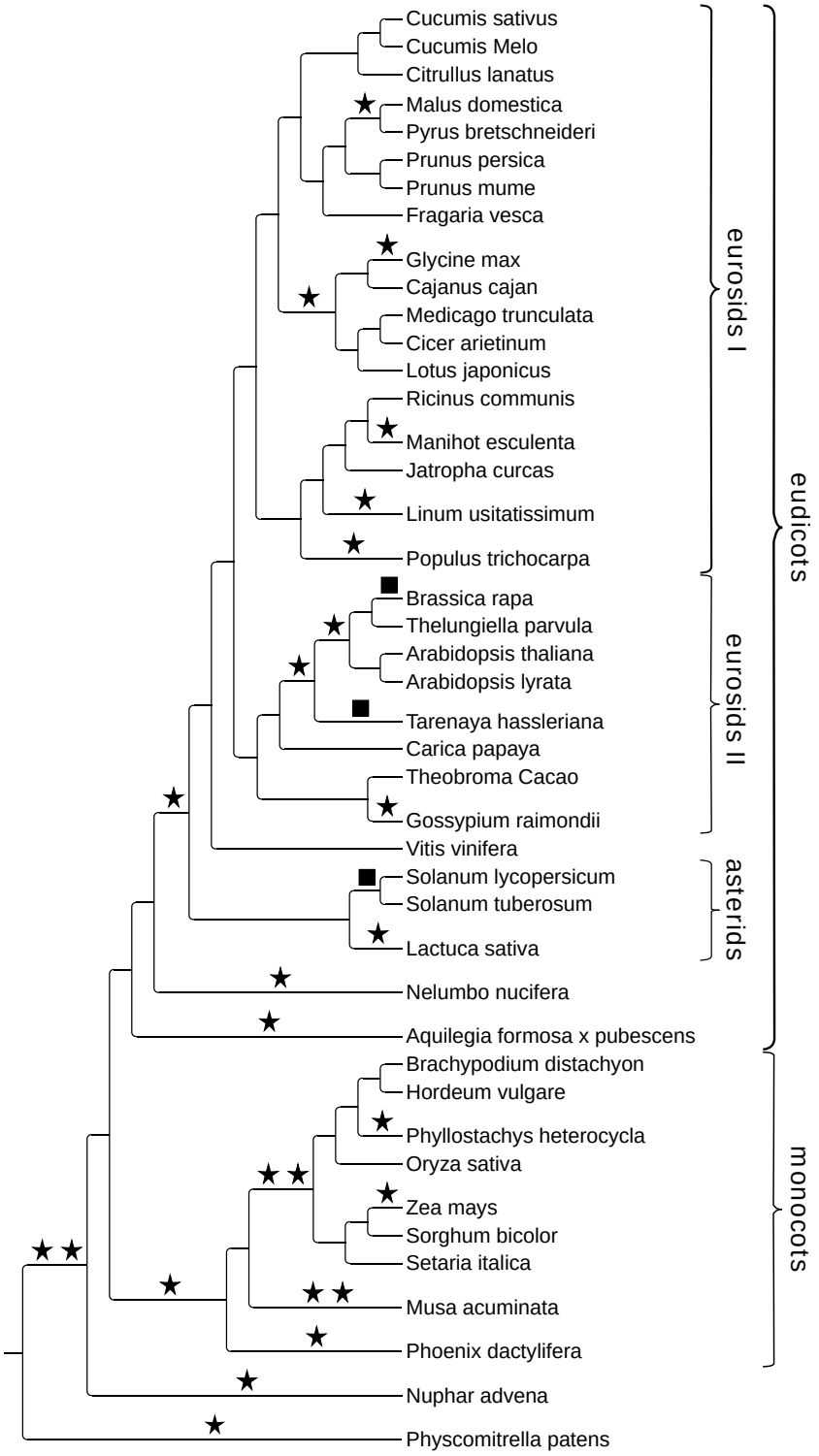


Figure 2. Phylogenetic tree based on the most current APG phylogeny of most sequenced green plants to date. Polyploidy events are indicated by stars (tetraploidy) or squares (hexaploidy). Adapted from Vanneste et al., 2014.

more or less constant fashion over time, whereas nonsynonymous mutations have been maintained due to natural selection.

When estimating Ks a number of models are available, some dating back to the 1980's. These more naïve approaches count the amount of synonymous and nonsynonymous sites and subsequently do a simple estimation of the likelihood of multiple substitutions per site. This does not take into account the unequal chances of transitions (A <-> G or C <-> T) versus transversions (all other mutations), leading to a structural overestimation of Ka and underestimation of Ks (Li et al., 1985). More sophisticated maximum likelihood (ML) methods that utilize models of codon substitution and usage bias have been developed (Goldman and Yang, 1994; Yang and Nielsen, 2000), but these methods are of course only as well performing as their underlying statistical model meaning that they do not always outperform the naïve method (Yang and Bielawski, 2000).

Based on the assumption that synonymous substitutions can act as a 'molecular clock' Ks has been used as an age estimation of syntenic regions resulting from ancient polyploidy in e.g. yeast and *Arabidopsis* as described above. Ks can also be used to estimate polyploid history without structural information about the genome, which is the case for many 'gene space' sequenced organisms today. The process is described in detail by Blanc and Wolfe (Blanc and Wolfe, 2004) and shall be summarized here. Paralogs in a genome are identified through BLASTN/P or other pairwise sequence comparison methods. The paralogous gene pairs are subsequently aligned and their Ks is calculated using one of the methods described above, usually dependent on time and the available computational power. Gene pairs are then binned, counted and plotted in a histogram, resulting in a figure like Figure 3. If no polyploid history is present, a small peak should be present at a low Ks (< 0.05) representing recent tandem and other single gene duplicates. This peak should gradually lower as genes diverge and asymptotically start approaching a count of 0 as the Ks increases. However, as we can see in Figure 3 peaks are present at certain Ks values. These peaks represent a large number of gene pairs that have a similar molecular age, which can be assumed to be the result a large gene birth event, i.e. polyploidy. These peaks will become less pronounced over time as genes diverge and are no longer recognizable as paralogs. In the case of *Arabidopsis* in Figure 3 we can clearly see the At-α event around a Ks of 0.8 and the β event around Ks 1.3.

Translating Ks into an actual age estimation in terms of years is a very debatable exercise. The assumption of Ks being relatively constant over time usually only hold for interspecies comparisons or within close relatives. To reconcile the intra-lineage substitution rates with each other when making wide species phylogenies, statistical models are used for these rates. Many advanced models have been proposed over the years and most recently the 'relaxed clock' model has gained in popularity, forming a middle ground between assuming a completely independent substitution rate per lineage and assuming a strict clockwise substitution rate across lineages (Drummond et al., 2006). In this model the rates in each branch are estimated by drawing from a parametric distribution such as a lognormal or exponential distribution, based on the mean substitution rate found in the parent branch.

Using a relaxed clock model, the polyploidy events that are known to have occurred across many plant lineages have been dated, as can be seen in work by Vanneste and Maere (Vanneste et al., 2014; Lohaus and Van de Peer, 2016). Interestingly, these datings show that many polyploidy events seem to have taken place at the Cretaceous-Paleogene (K-Pg) boundary; a mass extinction event estimated to have happened 66 Ma. The cause of this extinction is still uncertain, and two hypotheses include a massive impact from a meteorite that caused many years of debris blocking out the sun (Alvarez et al., 1979) or a group of volcanoes called the Deccan Traps experiencing massive eruptions spouting debris and sulphuric aerosols in the air blocking out the sun (Keller, 2014). The dating of these polyploid events

around this event have led the hypothesis that polyploids may have had an advantage due to hybrid vigor and greater availability of genes for natural selection (Fawcett et al., 2009).

DIPLOIDIZATION, NEO- & SUBFUNCTIONALIZATION

After the establishment of polyploidy in a population a process called diploidization takes place in which a tetraploid 'decays' to a diploid lineage. The mechanisms involved in this process of duplicate gene loss and retention are still unclear and are referred to as fractionation. One factor that complicates the study of this process is the fact that genetic drift may cause all copies of certain loci to become fixed into one of the parental alleles, whereas others might be maintained as separate alleles. Attempts at molecular dating of these two hypothetical loci will show different divergence times due to the asymmetric nature of this process (Wolfe, 2001). A second process after polyploidy is the loss and retention of duplicate genes. Four general outcomes have been hypothesized based on observations made in nature. The first is subfunctionalization, in which each duplicated copy takes on part of the function of the parental gene (Force et al., 1999). Another possibility is the partitioning of functions, such as enzymes with multiple substrates where each daughter gene specializes in one substrate (Wagner, 2000). A more population oriented effect is the retention in duplicate of genes in which a mutation would give a dominant negative phenotype, thus retaining both copies with very little change (Gibson and Spring, 1998). Lastly, dosage effects would make genes that form enzyme complexes be retained in similar copy number as mutations would lead to dosage interruptions which will have deleterious effects (Birchler and Veitia, 2007, 2012).

Other factors that may influence the loss and retention of duplicates are chromosomal properties. One hypothesis holds that if the two duplicated genomes are referred two as subgenomes, one subgenome may have 'dominant' properties leading to fractionation that is biased towards retention on the 'dominant' subgenome. Genome dominance has been hypothesized to have been a factor in the polyploid aftermath of Brassica rapa (Chinese cabbage) (The Brassica rapa Genome Sequencing Project Consortium et al., 2011) and is proposed to have influenced many paleopolyploids (Garsmeur et al., 2014). Garsmeur et al. also suggest that allotetraploids are more likely to show genome dominance effects and biased fractionation than autotetraploids. Algorithms that facilitate the study of these effects in silico have been developed e.g. Quota-align, which scans pairwise syntenic blocks and filters them based on an a priori expectation of genome quota (1:2 for diploid:tetraploid, 2:2 for 2:3 for tetraploid:hexaploid, etc.) (Tang et al., 2011).

NON-POLYPLOID GENE DUPLICATION

When studying the effects of genome duplication it must be considered that paralogous gene pairs can have their origin in other methods of duplication than polyploidy. Single gene duplication events occur constantly in all genomes and can occur due to two processes. First, tandem duplication is caused by errors in DNA replication and results in local clusters or arrays of neighboring homologs. In *Arabidopsis*, 15% of all protein coding genes is organized in one of these tandem arrays (Rizzon et al., 2006). The other single gene duplication mechanisms is gene transposition duplication and occurs when transposon activity duplicates a gene in another part of the genome. Transposition through recombination has been shown to increase after a polyploidization in *Arabidopsis* (Pecinka et al., 2011) and in the *Arabidopsis* genome 14% of protein coding genes has transposed at least once in its lineage (Woodhouse et al., 2011).

CLEOMACEAE

Because most study towards polyploidy in plants has focused on *Arabidopsis* and the Brassicaceae family there is a potential for bias. *Carica papaya* forms a good outgroup to mitigate this as it has undergone no paleopolyploidy except for the event shared by all angiosperms. However, to unravel common mechanics after polyploidy and distinguish them from Brassicaceae specific evolutionary phenomena

an independent outgroup with a similar polyploid history is needed. In this thesis, we present the first steps towards using *Cleomaceae* to fulfil this role.

Cleomaceae are an herbaceous and shrubby plant family with palmately compound leaves. Like Brassicaceae, the flowers have four petals and six stamens. Unlike Brassicaceae the flowers are zygomorphic, a trait that has a single origin in Cleomaceae (Patchell et al., 2011a) and may have been influenced by pollinator interactions. Flower morphology is grand and diverse, leading to a wide variety of pollinators, including bats (Machado et al., 2006). Fruits are dehiscent siliques of which a loop-like woody placenta remains on the plant after the silique valves fall off. Their habitat is cosmopolitan, and consists of the global temperate region where they can often be found in tropical habitats (Figure 4) (Stevens, 2001b). In contrast to Brassicaceae, no self-incompatibility systems exist in Cleomaceae (Cane, 2008). The cleome crown group species radiated 43-38 Ma (Cardinal-McTeague et al., 2016). Historically, it was placed under the Capparaceae but has later been reclassified to be an independent family (Sánchez-Acebo, 2005). Phylogentically it remains a complex clade, with many revisions and replacements; the namesake genus Cleome is widely scattered across the tree (Iltis and Cochrane, 2007; Hall, 2008; Feodorova et al., 2010; Cardinal-McTeague et al., 2016). Two species are being developed to serve as models for the family: *Tarenaya hassleriana* and *Gynandropsis gynandra*.

Tarenaya hassleriana is a species which is widely grown for its ornamental qualities and it has many commercial cultivars such as ‘Violet Queen’, ‘Rose queen’ and ‘Helen Campbell’. It has been rediscovered in the 2000’s by researchers as an outgroup for Brassicaceae in the study towards polyploidy and its effects on the genome and trait evolution (Schranz and Mitchell-Olds, 2006; Barker et al., 2009).

Gynandropsis gynandra (also known as *Cleome gynandra*) is a crop plant in southern Africa, where it is used as a pot herb or side dish and is grown mostly during the dry season (http://africanorphancrops.org/cleome-gynandra/, accessed 28/8/2016). In southeast Asia, where it is known as *phak sian* it is most commonly used to make a pickle of the leaves using rice water known as *phak sian dong*, which is eaten over pork soup or grilled fish (http://frynn.com/ผักเสี้ยน/, accessed 28/8/2016). As a dry season crop is part of the African orphan crops project and research has been done on its transcriptome (Külahoglu et al., 2014) and sequencing efforts are under way (Schranz et al., in preparation).

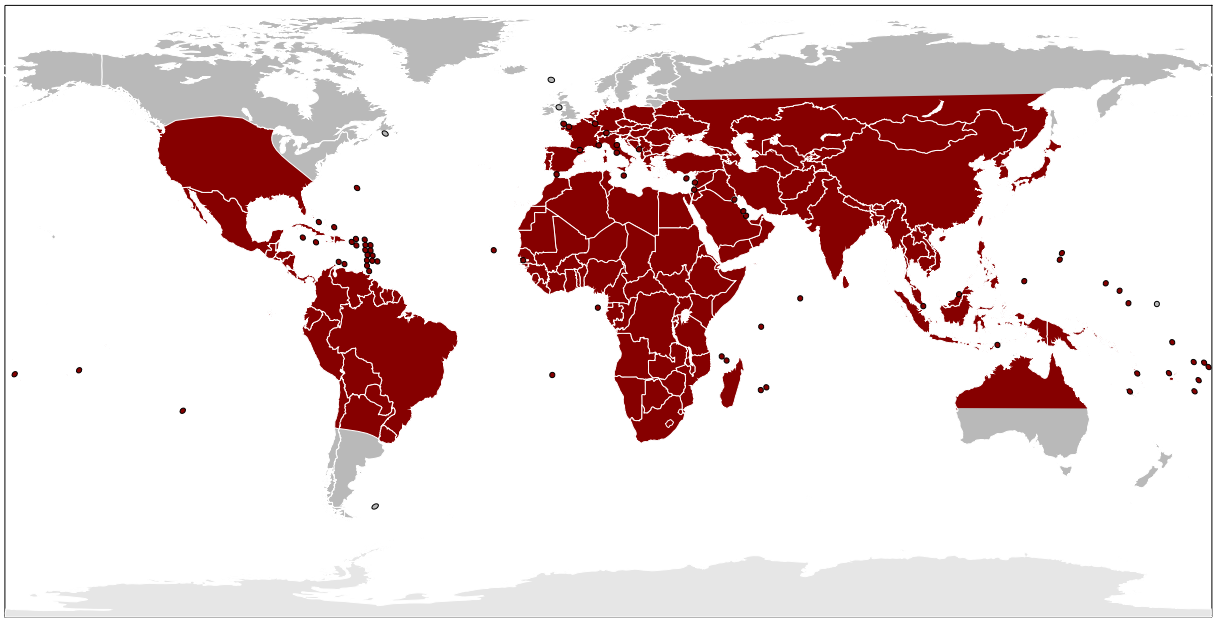


Figure 4. Global distribution of Cleomaceae

Besides being a model for the study of polyploidy, Cleomaceae can provide insight in the development of C4 photosynthesis. This form of photosynthesis is a more efficient form of photosynthesis in conditions of high temperature or low CO2. Cleomaceae contain a number of C4 species and an intermediate C3/C4 species, *C. paradoxa* (Koteyeva et al., 2011). Classically *Flaveria* species are the species of choice for the study of this trait, but the Cleome family provides a framework that benefits from the short phylogenetic distance to the genetically well characterized Brassicaceae enabling rapid orthologous identification and annotation prediction of genes (Brown et al., 2005).

CLEOMACEAE AND THE EFFECTS OF ANCIENT POLYPLOIDY IN THE BRASSICALES (THIS THESIS)

In this thesis, a number of studies are described relating to polyploidy and Cleomaceae in general and the overlap of these two subjects. In Chapter 1, the genome sequence of *Tarenaya hassleriana* is described and analyzed setting the stage for an in-depth genetic analysis of this organism. In it, the genome size and chromosomal make-up is described, together with *de novo* and orthology based annotations of more than 24000 genes. The polyploid history of the Cleomaceae is reaffirmed, showing the Cleomaceae specific Th-α event and the β and γ events. Syntenic similarity with *Arabidopsis* and other Brassicaceae are shown and the implications for ohnologs in these blocks are studied, with the Brassicaceae specific S-locus as an example.

In Chapter 2 the effects of the paleopolyploid history of Cleomaceae on their photosynthetic evolution is is studied. Ohnologs and other duplicated genes are identified in *Tarenaya hassleriana* and *Gynandropsis gynandra* using expressed sequences. These are classified into age groups using syntenic and Ks methods and the influence of the various polyploidy and recent gene duplication events is determined; this is further confirmed by expressed gene analysis. In Chapter 3 a similar approach is used to determine the postpaleopolyploidic effects on glucosinolate evolution. Glucosinolates are an important herbivore defense mechanisms that is also referred to as the 'mustard bomb'. It is a trait that is present in most higher plants but which has radiated significantly in Brassicaceae evolving into many different 'flavors'. It is shown that although glucosinolates are not as varied in Cleomaceae, polyploidy and gene duplication has had a significant effect on the evolution of this trait in this plant family.

Finally, in Chapter 4 we discuss the details of flower colour in *Tarenaya hassleriana* and perform genetic mapping of this trait based on mass spectrometry analysis of anthocyanins found in two colour variants.

We conclude that flower colour variation is based on a few genes that have been similarly identified in other species.

CHAPTER 1

THE GENOME OF TARENAYA HASSLERIANA PROVIDES INSIGHTS INTO REPRODUCTIVE TRAIT AND GENOME EVOLUTION OF CRUCIFERS

Shifeng Cheng[1*], Erik van den Bergh[2*], Peng Zeng[1], Xiao Zhong[1], Jiajia Xu[3], Xin Liu[1], Johannes Hofberger[2], Suzanne de Bruijn[4, 10], Amey S. Bhide[5], Canan Kuelahoglu[6], Chao Bian[1], Jing Chen[1], Guangyi Fan[1], Kerstin Kaufmann[4, 10], Jocelyn C. Hall[7], Annette Becker[5], Andrea Bräutigam[6], Andreas P.M. Weber[6], Chengcheng Shi[1], Zhijun Zheng[1], Wujiao Li[1], Mingju Lv[3], Yimin Tao[3], Junyi Wang[1], Hongfeng Zou[1, 8], Zhiwu Quan[1,8] Julian M. Hibberd[9], Gengyun Zhang[1,8], Xin-Guang Zhu[3], Xun Xu[1], & M. Eric Schranz [2]

1 BGI-Shenzhen, Shenzhen, China.

2 Biosystematics Group, Wageningen University, the Netherlands.

3 Plant Systems Biology Group, Partner Institute of Computational Biology, Chinese Academy of Sciences/Max Planck Society, YueYang Road, Shanghai 200031, China

4 Molecular Biology Group, Wageningen University, the Netherlands

5 Plant Developmental Biology Group, Institute of Botany, Justus-Liebig-University Giessen Heinrich-Buff-Ring 38, 35392 Giessen, Germany

6 Institute of Plant Biochemistry, Center of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-University, D-40225 Düsseldorf, Germany

7 Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2E9

8 BGI-Shenzhen, Chinese Ministry of Agriculture, Key Lab of Genomics, Shenzhen, China.

9 Department of Plant Sciences, Downing Street, University of Cambridge, Cambridge, UK CB2 3EA, UK

10 Institute for Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Strasse 24-25, Haus 20

[*]These authors contributed equally to the manuscript

ABSTRACT

The Brassicaceae, including Arabidopsis and Brassica crops, are unmatched among plants in their wealth of genomic and functional molecular data and has long served as a model for understanding genome, gene, and trait evolution. However, the genome of a phylogenetic outgroup that is essential to infer directionality of evolutionary change has been lacking. We therefore sequenced the genome of the spider flower (Tarenaya hassleriana) from the Brassicaceae sister-family, the Cleomaceae. By comparative analysis of the two lineages we show that genome evolution following independent ancient polyploidy and gene duplication events affect reproductively important traits, including floral development and self-incompatibility systems. We found an ancient genome triplication in Tarenaya (Th-α) that is independent of the Brassicaceae-specific duplication (At-α) and nested Brassica (Br-α) triplication and we reconstruct the gene content of the last common ancestor. To showcase the potential of sister lineage genome analysis, we investigated the retention and synteny of floral developmental genes and show Brassica retains twice as many floral MADS genes as Tarenaya that likely contribute to Brassica morphological diversity. Especially the class B floral homeotic genes show differences in patterns of gene duplication, retention and differential expression in Tarenaya compared to Brassicaceae that may partially explain differences in floral morphology. To unravel the evolutionary origin of Brassicaceae-specific Self-Incompatibility (SI) system, we performed synteny analysis between families and find the critical SRK receptor gene is derived from a lineage-specific tandem duplication. The genome of Tarenaya hassleriana will facilitate future research aimed at elucidating the evolutionary and functional history of Brassicaceae genes and pathways.

INTRODUCTION

Studies of the model-plant *Arabidopsis* and its close-relatives in the family Brassicaceae have provided fundamental insights into the processes and patterns of plant evolution and function (Koornneef and Meinke, 2010; Hu et al., 2011; The Brassica rapa Genome Sequencing Project Consortium et al., 2011). By comparative analyses to crop species these results have had profound influences on plant improvement and production. For example, knowledge about the control and evolution of plant reproductive traits, such as floral and fruit development and self-incompatibility systems, can be directly related to plant fitness and yield(Diepenbrock, 2000; Li et al., 2013; Rahman and McClean, 2013). The Brassicaceae have also been a model for the understanding the dynamics and impacts of ancient polyploidy (genome doubling), with the entire family having undergone a whole genome duplication (named At-α) and the crop Brassicas having had an additional genome triplication (Br-α) (Blanc et al., 2003; Schranz and Mitchell-Olds, 2006; Thomas et al., 2006; The Brassica rapa Genome Sequencing Project Consortium et al., 2011). Genes retained in multiple copies due to these ancient polyploidy events, in addition to more recent tandem-duplications, have played important roles in the evolution and regulation of key-traits (Edger and Pires, 2009; Flagel and Wendel, 2009). However, the polyploid history of the Brassicaceae also complicates synteny and evolutionary inferences to distantly related crop species.

To more fully exploit the fundamental trait and genome insights garnered from Brassicaceae systems and improve synteny analyses to more distant crops, we report the genome sequencing and analysis of *Tarenaya hassleriana* from the sister-family Cleomaceae. Currently papaya (*Carica papaya*), a member of the order Brassicales, is the closest relative with a complete genome sequence; however, these two lineages diverged more than 70Mya (Ming et al., 2008). The Cleomaceae is the phylogenetic sister-family to the Brassicaceae with the two lineages having diverged only ~38Mya (Schranz and Mitchell-Olds, 2006). Brassicaceae and Cleomaceae share many traits in common (Hall et al., 2002; Iltis et al., 2011), such as a preponderance of herbaceous species, the same general floral ground plan (four sepals, four petals, six stamens and two fused carpels) and a replum in the mostly dehiscent fruits, referred to as capsules. There are also a number of key differences. Most of the 300 Cleomaceae species are

restricted to the semi-tropics and arid desert regions and lack a genetic pollen-pistil self-incompatibility system, whereas most of the 3700 Brassicaceae species largely radiated into cold temperate regions and possess a genetically regulated self-incompatibility (SI) system (Guo et al., 2011). Another striking distinction is in the floral symmetry: Cleomaceae have mostly monosymmetric flowers and Brassicaceae have mostly disymmetric flowers (Endress, 1999; Patchell et al., 2011b). Cleomaceae also exhibit greater variation in the basic floral plan with increases in stamen number, petal dimorphisms, and stalks to the ovary, whereas Brassicaceae exhibit greater diversity in fruit morphology and dehiscence capabilities (Franzke et al., 2011). Comparative analyses can be used to elucidate the genomic basis of these differences. The Cleomaceae species we have sequenced is *Tarenaya hassleriana*, often referred to as the Spider Flower, which is widely grown as an ornamental species and used as an educational model (Marquard and Steinback, 2009). This species was formerly named *Cleome hassleriana* (often erroneously labeled as *C. spinosa*), but the genus *Cleome* has undergone recent taxonomic revisions (Iltis and Cochrane, 2007).

Brassicaceae and Cleomaceae have undergone independent ancient polyploidy events. At least five ancient polyploidy events have occurred in the evolutionary history of *Arabidopsis* (Bowers et al., 2003; van de Peer et al., 2009) four of which are shared with Cleomaceae: ζ near the origin of seed plants (Jiao et al., 2011); ε near the origin of angiosperms (Jiao et al., 2011); the ancient hexaploidy At-γ shared by nearly all eudicots (Jaillon et al., 2007; Vekemans et al., 2012); and At-β restricted to part of the order Brassicales as it is lacking from the papaya genome (Ming et al., 2008). The most extensively studied ancient polyploidy event is the more recent At-α genome duplication (Bowers et al., 2003; Schranz and Mitchell-Olds, 2006) and is shared by all Brassicaceae species (Eric Schranz et al., 2012). The crop genus *Brassica* has all the ancient polyploidy events in common with *Arabidopsis*, but also has undergone an additional and more recent whole genome triplication (hexaploidy) event (Br-α) after its split with *Arabidopsis* around ~17Myr ago (The Brassica rapa Genome Sequencing Project Consortium et al., 2011). Limited Bacterial Artificial Chromosome and transcriptome sequencing revealed that *Tarenaya* lacks the At-α event and that it underwent an independent ancient genome triplication (Th-α) (Schranz and Mitchell-Olds, 2006). Thus, *Tareneya* provides a unique opportunity to contrast genome evolution from a common ancestor comparing three genomic equivalents in *Tarenaya*, two in *Arabidopsis* and six in *Brassica,* and furthermore to contrast two independent ancient genome triplications (Th-α vs. Br-α). We not only compare these polyploidy events and more recent tandem-duplication events, but also show how they contributed to the genes regulating key reproductive traits (van de Peer et al., 2009; Cardenas et al., 2012).

RESULTS

GENOME SEQUENCING AND INTEGRATION WITH PHYSICAL MAP

The *Tarenaya hassleriana* genome is relatively small (~290 Mb: 2n=20) and within the range of sequenced Brassicaceae species (*A. thaliana* 157Mb (Bennett et al., 2003); *A. lyrata* 207Mb (Hu et al., 2011); *Schrenkiella parvula* (formerly *Thellungiella parvula*) 140Mb (Dassanayake et al., 2011)*; Eutrema halophilum* (formerly *Thellungiella halophila*) 239Mb (Wang et al., 2010); *B. rapa* 485Mb (The Brassica rapa Genome Sequencing Project Consortium et al., 2011)). To generate a high-quality draft genome assembly, we utilized both sequenced paired-end libraries and constructed a BAC-based WGP physical map. We used the Illumina next-generation sequencing platform to generate ~70.2 Gb (245X genome-depth) raw data of paired-end reads ranging from 90 - 100bp (**Supplementary Table 1)** from seven libraries with varying insert sizes (350bp - 20Kb). Sequence data was filtered, yielding ~40 Gb of high-

Table 1. Summary of the genome sequencing, assembly, and annotation.

Assembly			
	N50 (size/number)	N90 (size/number)	Total sizes
Contigs	21.58 kb/2761	2.7 kb/13591	222 Mb
Scaffolds	551.9 kb/98	64.8 kb/622	256.5 Mb
Superscaffolds	1.26Mb/40	7.4kb/1014	273Mb
Annotation			
	Glean	RNA-Seq supported	Homologous with *Arabidopsis*
Gene	28917	20337	24245
	LTR	DNA transposons	Total size
TE sizes, Mb (%)	97.28 (38.19)	11.8 (4.62)	110 (43.3)

quality sequence (~139X coverage) (**Supplementary Table 2**). Assembly was done using SOAPdenovo (Li et al., 2010) (version 2.21) (**Supplementary Table 3**). Remapping ESTs to the assembly showed that >96% of the genic regions were covered (**Supplementary Table 4**). The physical map was made using the Keygene Whole Genome Profiling (WGP™) fingerprinting technique (van Oeveren et al., 2011). In total 192,000 BAC-based Illumina sequence tags were generated from 19,200 BACs from two libraries (*EcoRI* and *MseI*) with an average insert size of approximately 125kb (giving a total of 16X genome equivalents each) (**Supplementary Table 5**). High-quality WGP tag sequences were used to build a high-stringency map assembly using modified FPC software (Engler et al., 2003) to generate 786 contigs (**Supplementary Table 6**). We integrated the WGP physical map scaffolds with the SOAPdenovo sequence scaffolds to produce 71 super-scaffolds through BLAST mapping of the WGP anchors (tags). We evaluated the quality of the integration between physical map and superscaffolds by manually checking the ordering and orientation of connected scaffolds by analyzing collinearity relative to *A. lyrata* (Example shown in **Supplementary Figure 1**), confirming that all *Tarenaya* superscaffolds having extensive and extended synteny. The final assembly statistics of the integrated dataset are summarized (**Table 1, Supplementary Table 7).** With this integration, the N50 was increased by more than 2.6-fold (N50=1.26 Mb) due to the merger of most of the *de novo* assembled scaffolds into superscaffolds.

GENE ANNOTATION

Gene annotation was conducted using a pipeline that integrates *de novo* gene prediction, homology-based alignment, as well as utilization of the RNA-seq data. In total, we conducted >4Gb of RNA-seq, of which 77.4% of reads could be confidently mapped onto the genome (**Supplementary Table 8**). To analyze the overall gene expression patterns and to provide basic gene expression information, transcriptomes were generated for several *T. hassleriana* tissues (**Supplementary Table 9**). A principal component analysis showed that stamen, root and seed profiles separate most, a result similar to the

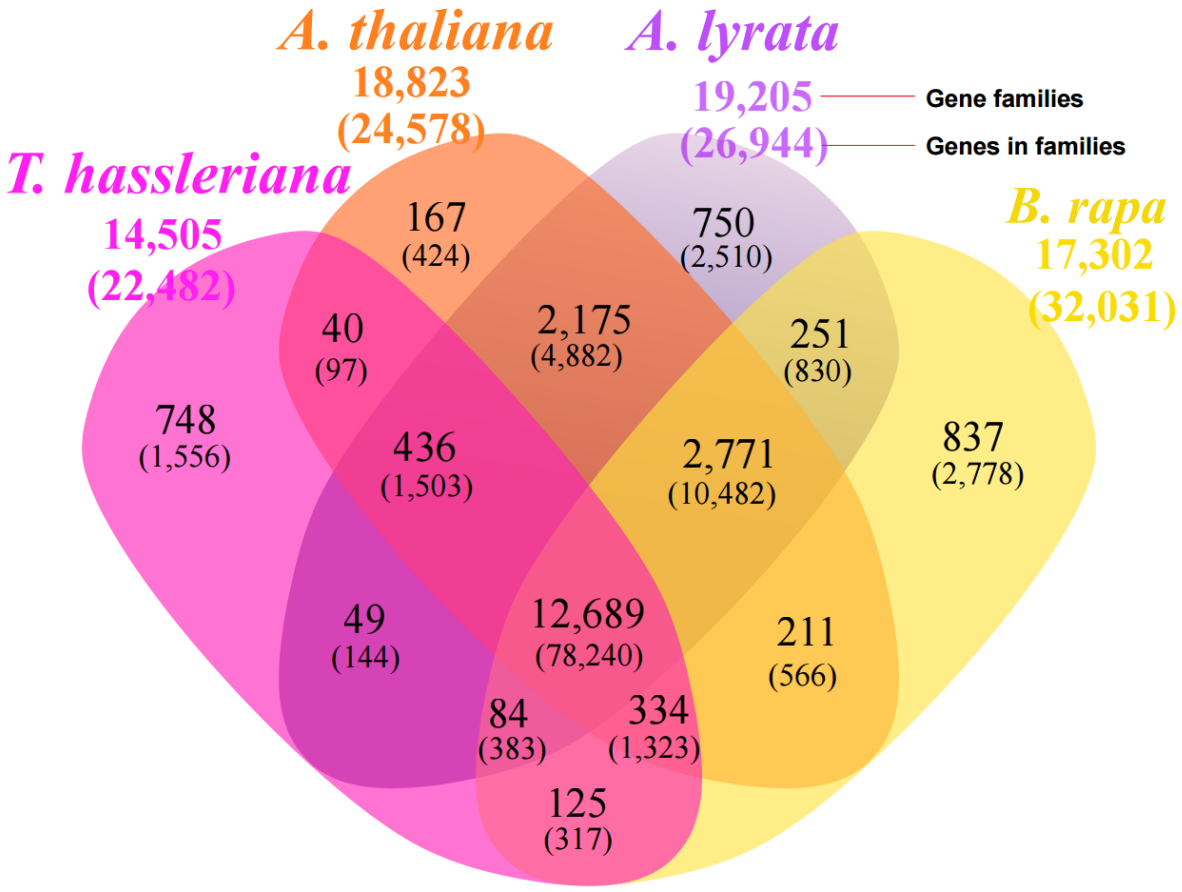


Figure 1**.** Venn diagram illustrating the shared and unique gene families from *Tarenaya hassleriana* (Cleomaceae), *Arabidopsis thaliana, Arabidopsis lyrata* and *Brassica rapa* (Brassicaceae). In total we predicted 28,917 well-supported gene models for *T. hassleriana* of which 22,482 could be placed into one of the 14,505 gene families. 87.5% of gene families found in *Tarenaya* were present in all three Brassicaceae species, 7.4% in one or two Brassicaceae species and only 5% of gene families were unique to *Tarenaya* with many of these associated genome-specific retrotransposons. Thus, comparative functional and evolutionary analysis of well-characterized *Arabidopsis* and Brassicaceae genes is feasible using *Tarenaya* as an out-group.

pattern detected in *A. thaliana* where these three tissue types separate most over the first two components (**Supplementary Figure 2**) (Schmid et al., 2005).

The prediction of gene models by various techniques was summarized (**Supplementary Table 10)**, with a large range in the number of predicted genes. To be conservative, we used the intersection of the various gene prediction techniques which resulted in the identification of 28,917 highly supported gene models with an average transcript length of 2,216bp, coding sequence size of 1,169bp and 5.27 exons per gene, both similar to that observed in *A. thaliana* and *B. rapa* (**Supplementary Figure 3**). A total of 92.9% of gene models have a homolog match or conserved motif in at least one of the public protein databases (including Swissprot (McMillan and Martin, 2008) 71.1%, TrEMBL (Boeckmann et al., 2003) 92.5%, InterPro (Zdobnov and Apweiler, 2001) 74.9%, KEGG (Kanehisa and Goto, 2000) 55.2%, and GO (Ashburner et al., 2000) 55.7%) (**Supplementary Table 11**) and 97.1% are represented among the public EST collections or *de novo* Illumina mRNA-Seq data. In addition to protein-coding genes, we also identified 220 microRNA (miRNA), 862 tRNA and 685 small nuclear (snRNA) genes in the *T. hassleriana* genome (**Supplementary Table 12**). Orthologous clustering of proteomes predicted for *T. hassleriana* and three Brassicaceae species (*A. thaliana, A. lyrata* and *B. rapa*) revealed 15,112 genes in 12,689 families in common (**Figure 1**). 20,926 *T. hassleriana* genes clustered with at least one of the 3 genomes.

1,556 *Tarenaya*-specific genes in 748 families were identified, most of which were enriched for genes of unknown function and for which 34% have EST supported annotation.

More than 44% of the *T. hassleriana* genome was composed of repetitive elements. Both *de novo* repeat identification and homology-based methods were applied to predict transposable elements (**Supplementary Table 13**). The majority of repetitive sequences were Class I long-terminal repeat (LTR) retrotransposons composing 36.6% of the genome compared to 27.1% in *B. rapa*. The overall lower percentages of annotated transposons in *Brassica* are likely due to its lower genome sequence coverage, because the *B. rapa* genome is nearly 200Mb larger than *Tarenaya*. Most of the repeats were located in the intergenic regions.

COMPARATIVE ANALYSIS OF ANCIENT POLYPLOIDY EVENTS

The Brassicaceae-Cleomaceae system allowed us to compare genome evolution after several rounds of independent ancient polyploidy **(Figure 2**). We confirmed that the Cleomaceae polyploidy event (Th-α) occurred independently of, and more recently than, the Brassicaceae-specific duplication event detected in *Arabidopsis* and Brassica (At-α). We also detected the nested *B. rapa* ancient hexaploidy (triplication of the genome) event (Br-α), and show that it is of approximately the same age as Th-α. For all three taxa we detected the diffuse signal of the older and shared events (At-β, At-γ, ε and ζ) (**Figure 2**). The Th-α and Br-α events are of approximately the same age and they represent independent ancient hexaploidy events. Using whole genome intra-genomic dot-plots of *Tarenaya* we showed many triplicated blocks (**Supplementary Figure 4)** and analysis of syntenic depth with QuotaAlign (Tang et al., 2011) shows that 49.4% of genes are found at 3x coverage (**Supplementary Table 14**). To illustrate the

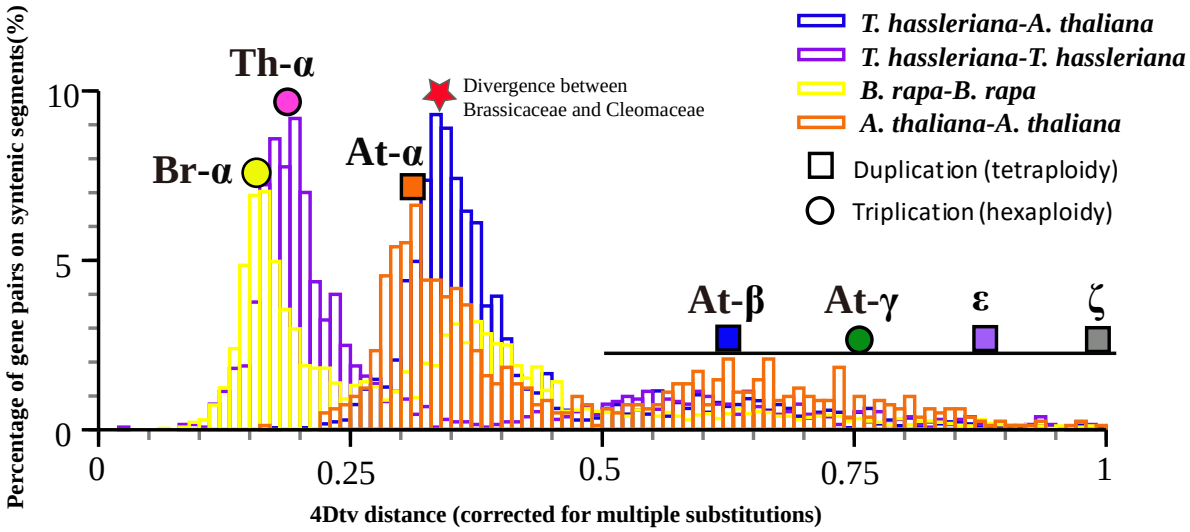


Figure 2**.** Relative timing of the polyploidy events and lineage splitting based on divergence of 4-fold degenerate sites (4DTv) for duplicated genes within *A. thaliana*, *B. rapa* and *T. hassleriana*, and orthologous genes between *A. thaliana* and *T. hassleriana*. All plots detect broad overlapping peaks between 0.5 and 1.0, representing shared older polyploidy events (At-β, At-γ, ε and ζ). The divergence of the Brassicaceae-Cleomaceae lineages is seen by the differentiation of *Arabidopsis* and *Tarenaya* homologs at the peak centered at ~0.35 (highlighted by red star). The divergence between paralogs from the At-α duplication event occurred slightly after the lineage splitting and is detected by the peaks centered at ~0.3 for both *Arabidopsis* and *Brassica*. The At-α peak is lacking from *Tarenaya* proving At-α is Brassicaceae-specific. Nearly overlapping distributions between 0.15 and 0.25 were detected for *Brassica* and *Tarenaya* representing the independent Br-α and Th-α ancient hexaploidy events, respectively.

homologous relationships and the evolutionary history of triplicated/duplicated segments in Cleomaceae and Brassicaceae, we integrated intra- and inter-genomic analyses (**Figure 3**). We analyzed synteny relationships both within and between genomes (**Supplementary Figures 4-7**). For *T. hassleriana*, 86%, 83%, and 85% of the protein-coding genes were homologous to the genes in *A. thaliana, A. lyrata* and *B. rapa* genomes, respectively (**Supplementary Table 15**). By making these comparisons of *Tarenaya* vs. *A. thaliana, A. lyrata*, and *B. rapa* genomes (**Figure 3**), we found significant, 3:2, 3:2, 3:4, 3:5 and 3:6 homologous patterns, respectively, which is consistent with the polyploid history of the species. To illustrate this pattern, we have highlighted two ancestral blocks (A1 and A2) (**Figure 3,** two small insets). Note that the three *Tarenaya* blocks show almost perfect collinearity whereas for example one of two *Arabidopsis* regions is broken across two chromosomes, suggesting a Brassicaceae-specific rearrangement(s) after At-α. Since synteny analysis has been extensively carried out within Brassicaceae, we also show our results with the collinear blocks color-coded according to the current Brassicaceae conventions (Schranz and Mitchell-Olds, 2006) (**Supplementary Figure 8**).

We inferred the putative 'A ancestor' (pre-At-α) shared by *A. thaliana*, and *A. lyrata*, the 'B ancestor' of *B. rapa* (pre-Br-α but post-At-α ancestral genome state), and 'T ancestor' of *T. hassleriana* (pre-Th-α ancestral genome state) (see **Methods**). We compared our identified homologous replicated blocks within and between genomes in Brassicaceae species and *T. hassleriana* by comparison to earlier analysis of conserved At-α blocks (Blanc et al., 2003; Thomas et al., 2006). First, we reconstructed our version of the pre-At-α ancestor ('A' ancestor) of *A. thaliana* (version TAIR9), which resulted in 64 ancestral regions involving 19,976 protein-coding genes (**Supplementary Table 14**). The corresponding relationships to the blocks identified by Wolfe and colleagues (Blanc et al., 2003) (that we refer to as "Wolfe blocks") are illustrated in **Supplementary Figure 9**. We utilized the same method on the minimized genomes of *A. lyrata, B. rapa*, and *T. hassleriana*, and generated 61, 71, and 87 conserved ancestral blocks covering 24,373; 25,646; and 20,680 protein-coding genes, respectively, to represent the postulated 'A', 'B', and 'T' ancestors (**Supplementary Figures 10-12**). When we compare the reconstructed 'T' and 'A' ancestor genomes we find a clear 1:1 relationship, supporting our conclusion that the "ancestral genome" of the Brassicaceae and Cleomaceae was conserved before the independent duplication (At-α) and triplication (Th- α) events, respectively. Since *B. rapa* underwent a nested and specific triplication following the At-α, we see a 1:2 pattern when we compare the inferred 'T" to 'B' ancestors. A comparison of conserved ancestral genomic blocks across species is shown in **Supplementary Figure 13**. We then traced the extent of gene retention and fractionation in homologous blocks after polyploidy events. We partitioned the two sub-genomes of *Arabidopsis*, three sub-genomes

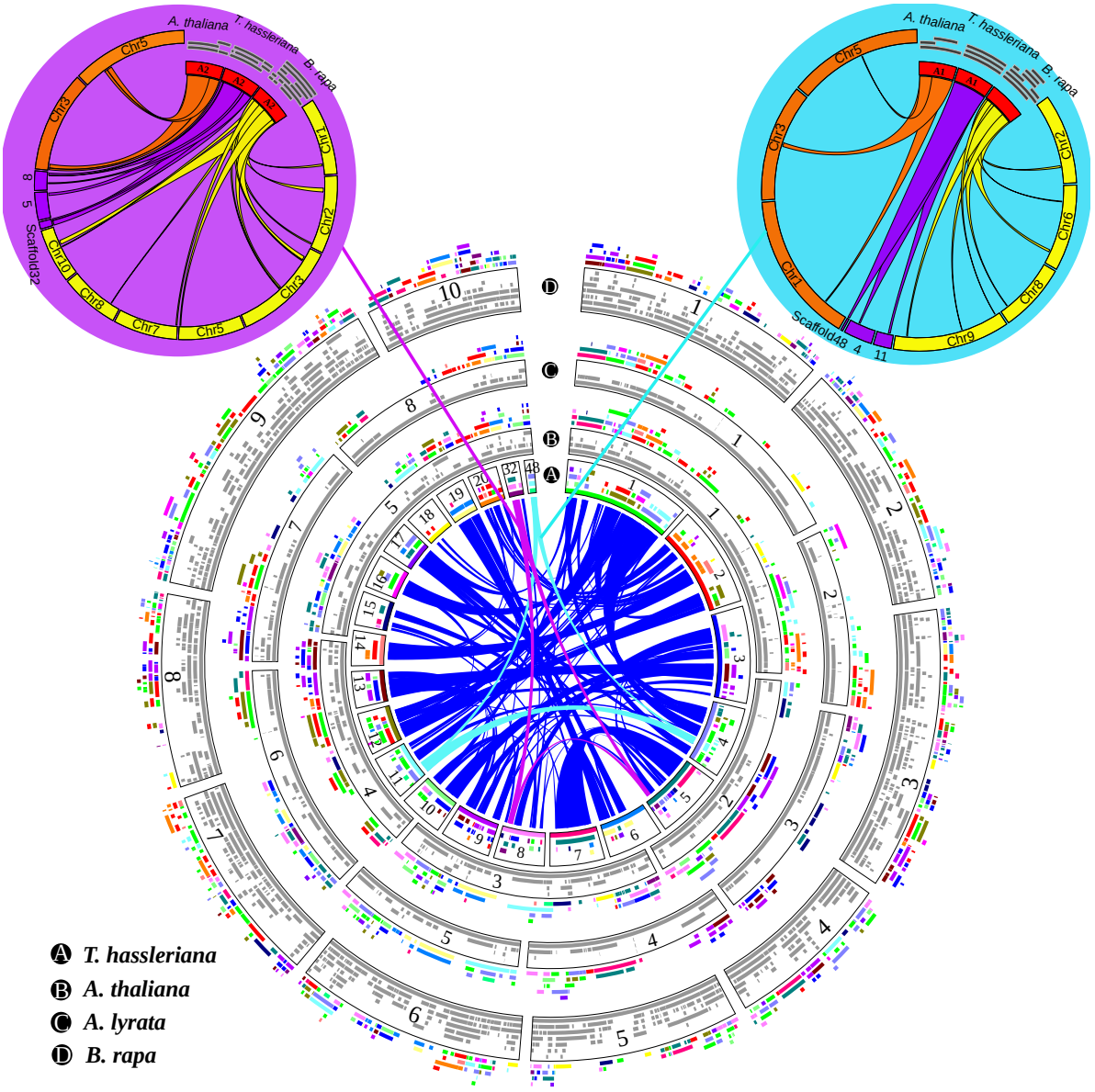


Figure 3. Homologous genome blocks within and between genomes for Cleomaceae and Brassicaceae. The largest 20 (plus additional 2 smaller scaffolds) color-coded super-scaffolds of *T. hassleriana* are taken as the reference, such that any region homologous to the 22 scaffolds is colored accordingly. **A.** Self-alignment of *Tarenaya* super-scaffolds, with the inner circle showing links of syntenic blocks. Over 47% of the genome is found in three copies, supporting the conclusion that it experienced an ancient hexaploidy event (Th-α=triplication). Rings within a genome (inner grey bars) and blocks homologous to *Tarenaya* (outer color-coded bars) for completed Brassicaceae genomes: **B.** *A. thaliana,* **C.** *A. lyrata,* and **D**. *B. rapa*. The inner grey bars show a clear pattern related to the ancient polyploidy events of the Brassicaceae ((At-α= duplication) and nested Brassica-specific lineage (Br-α=triplication)). The color-coded outer rings of homology relative to *Tarenaya* show a complex pattern due to the independent polyploidy events between families. The two small insets illustrate clear examples of the 3 *Tarenaya* to 2 *Arabidopsis* to 6 Brassica genome equivalents due to ancient polyploidy events.

of Brassica, and three sub-genomes of *T. hassleriana*, respectively, by comparing with the reference 'A' ancestor of *A. lyrata* (**Supplementary Figure 14-17**). These improvements to understanding genome evolution after independent ancient polyploidy events of Brassicaceae species will facilitate more synteny analyses to more distant crop species.

COMPARATIVE ANALYSIS OF TYPE II MADS BOX GENES

The development of the four floral organ types and later the fruits are regulated by Type II MADS-domain proteins (Smaczniak et al., 2012) as described by the ABCDE model (Theißen, 2001). The types of MADS-box genes controlling development are remarkably well conserved across eudicots, with the molecular mechanisms of their action extensively studied in *Arabidopsis*. We found that the *Tarenaya* genome contains representatives of all the major Type II MADS box genes described in *Arabidopsis* and Brassica (**Supplementary Figure 18**). We concentrated on the retention of the MADS-box genes derived from At-α, Br-α and Th-α and compared this with the polyploid origins of additional duplicates (At-β, At-γ, ε, T (tomato (The Tomato Genome Consortium, 2012) and Pt-α (poplar) (Tuskan et al., 2006)) (**Figure 4**). Theoretically, the At-α, Br-α and Th-α events should have given rise to 2 *Arabidopsis*, 6 Brassica and 3 *Tarenaya* gene copies (syntelogs) from a single ancestral gene. Of the eleven MADS-box gene clades involved in floral, fruit and inflorescence development that were likely present in the most recent common ancestor of Brassicaceae and Cleomaceae shown in Figure 4, we found that only three duplicate pairs are in fact maintained in *Arabidopsis* due to At-α: *AP1/CAL* (A-function), *SHP1/SHP2* (D-function) and *SEP1/SEP2* (E-function). Thus, *Arabidopsis* has only 3 of 11 possible replicates (27.3% syntelog retention). This implies that during early Brassicaceae evolution (before the split of *Arabidopsis*-Brassica) there were 14 gene lineages. From these 14 lineages then there would be 28 possible additional syntelogs in Brassica due to Br-α. We find a remarkable 19 of these additional gene copies (67.8% syntelog retention). This includes all three possible copies maintained for the following 7 Brassica gene families: *SEP4*, *SEP3*, *AGL79*, *FUL, AP1, PI,* and *SHP1.* The only two genes to return to single copy in *Brassica* after Br-α are *CAL* and *SHP2*. When we also consider the At-α gene retention, then a maximum of 4 of 6 gene copies are found in the *SEP1/2* clade, *AP1/CAL* clade, and the *SHP1/2* clade. *Tarenaya* would be expected to have a maximum of 22 additional retained syntelogs due to Th-α, however, we only recover 6 (27.3% syntelog retention) with no cases where all three possible copies are maintained (we do not count the additional tandem duplicate of *ThAP3* here which is discussed below). Thus, we find more than double the syntelog retention in Brassica than in *Tarenaya*, despite the fact that both are ancient hexaploids of approximately the same age. The greatest differential in gene copy retention between Brassica and *Tarenaya* is for the *AGL79* clade (3 vs. 1) and the *SHP1/2* clade (4 vs. 1). The *SHATTERPROOF* genes in Brassicaceae regulate various traits during carpel and fruit formation (Colombo et al., 2010). The single-copy nature of *SHP* homolog in *Tarenaya* is thus notable, since this is the only gene that is duplicated in *Arabidopsis* due to At-α but has returned to single-copy in *Tarenaya* (**Supplementary Figure 19**). In Cleomaceae fruit morphology is less diverse than in Brassicaceae, and we hypothesize that the retention of *SHP* genes plays an important role in the morphological variability of Brassicaceae.

COMPARATIVE COLLINEARITY ANALYSIS OF FLORAL DEVELOPMENTAL REGULATORS

To assess the contributions of ancient polyploidy and tandem gene duplications to floral regulatory gene diversification, we conducted a more detailed analysis of gene synteny and expression patterns. Almost all floral MADS-box genes show conserved synteny between Brassicaceae and Cleomaceae. For example, the A-class genes show very stable duplicate retention; the loci containing the *AP1* and *CAL* homologs in *Tarenaya*, *Arabidopsis* and *Brassica* are syntenic to one another with little evidence of local rearrangements (Supplementary Figure 20).

In contrast, the B-class (*PI* and *AP3*) genomic regions show a more dynamic pattern (Figure 5). The split between *PI* and *AP3* is a very old duplication due to the angiosperm ε polyploidy event (Figure 4), with almost no detectable collinearity between these regions (Figure 5). Comparison of B-class *Tarenaya* genomic regions with Brassicaceae allowed us to detect two B-class duplication events: A recent *Tarenaya* tandem duplication of *AP3* (Th02920 and Th02921) and an older *PI* duplication, likely due to

At-β, which has been lost from Brassicaceae but still retained in *Tarenaya* (Th17298) (Figure 5 and Supplementary Figures 21-22).

Strikingly we also detected two gene transposition events: a Brassicaceae-specific *AP3* transposition and a shared transposition event of one *PI* gene before the split of Brassicaceae and Cleomaceae (Figure 5). The Brassicaceae-specific *AP3* transposition event also involved the flanking *EMBRYO DEFECTIVE 1967 (EMB1967)* gene (At3g54350) containing two conserved domains: N-terminal region of micro-spherule protein (MCRS_N) and Forkhead-associated (FHA). The *Tarenaya AP3* gene is similarly flanked by a Forkhead-associated protein (Th02919) that has its highest match to *EMB1967,* however, the orientation of *AP3* and the Forkhead-genes is inverted between species (Supplementary Figure 23). In general, B-class genes are functionally highly conserved across angiosperms, whereas A-class gene function appears to be less conserved (Litt and Kramer, 2010). However we have shown that it is in fact the B-class genes that have undergone transposition events.

Members of the *TCP* gene family play an important role in the transition from polysymmetric to monosymmetric flowers whenever examined (reviewed in (Busch and Zachgo, 2009; Jabbour et al., 2009; Rosin and Kramer, 2009)), including monosymmetric Brassicaceae (Busch and Zachgo, 2007; Busch et al., 2012). Cleomaceae floral morphology, especially in petal and stamen position, numbers and asymmetry, is quite variable. However, the role of *Tarenaya TCP* homologs in monosymmetry has not yet been fully characterized. We find a pattern of conservation of genomic collinearity around the *TCP1* locus between species (Supplementary Figure 24). *Arabidopsis* contains only a single *TCP1* locus

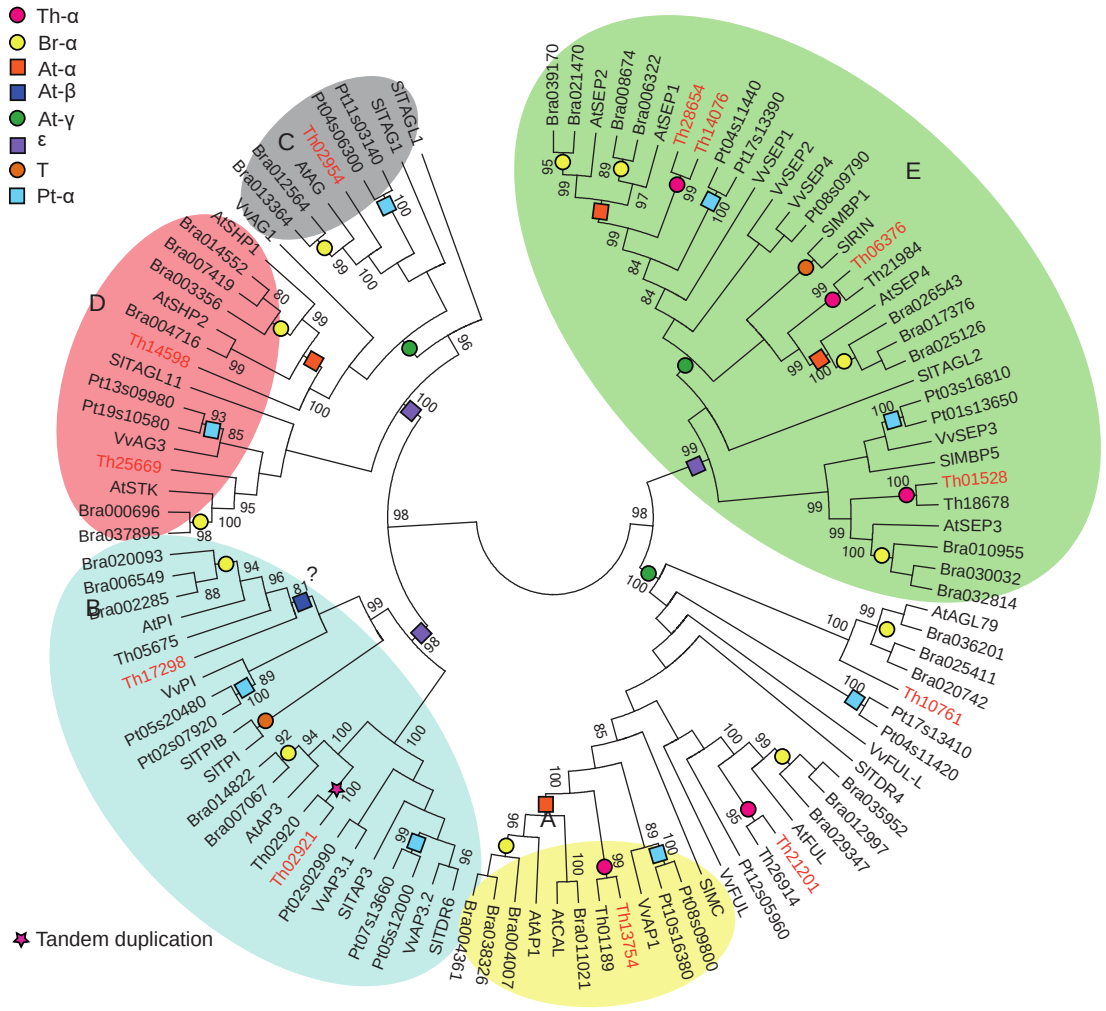


Figure 4. Phylogenetic tree of Type-II MADS-box transcription factor genes involved in floral organ specification. The major floral MADS-box genes are clustering into five groups corresponding to the five main functional types (AP1-like genes: shown in yellow, AP3/PI-like genes: B-type shown in blue, AG-like genes shown in gray, STK-like genes shown in red and SEP-like genes in green) according to the ABC(DE) model of floral development. Species: *Arabidopsis thaliana* (At), grape (*Vitis vinifera*, Vv), tomato (*Solanum lycopersicum*, Sl), Poplar (*Populus trichocarpa* (Pt), *Brassica rapa* (Br) and *Tarenaya hassleriana* (Th). *Tarenaya* genes are indicated in bold red. The colored squares (duplication events) and circles (triplication events) placed on nodes represent gene-lineage expansion(s) that can be associated with particular ancient polyploid events: Th-α, At-α, Br-α, At-β, At-γ, ε, T (identified by tomato genome sequencing) and Pt-α (identified by Poplar genome sequencing). Type-II MADS-box genes are often retained after ancient polyploidy events. We determined 27.3% syntelog (homolog generated by a polyploidy event) retention after At-α, 27.3% syntelog retention after Th-α, and a much higher 67.8% syntelog retention after Br-α, despite the fact that Th-α and Br-α triplications are of approximately the same age. The *Tarenaya* B-class genes show unusual patterns in that the *AP3* homologs (Ch02920 and Th02921) represent a recent tandem-duplication which is rare for floral MADS-box genes and there are two copies of *PI* homologs that are likely due to At-β with one lineage being lost in Brassicaceae. Tree constructed using maximum-likelihood with 1000-replicate bootstrap values >80 presented, visualized topology-only.

(At1g67260), as does *A. lyrata*. Due to Br-α, *Brassica* has three syntenic copies of *TCP1*. We also can detect the syntenic regions in Brassicaceae species due to At-α (Supplementary Figure 24) but find no At-α derived homologs of *TCP* genes, suggesting the loss of a *TCP1* syntelog occurred early in Brassicaceae evolution. In *Tarenaya* we find three genomic regions derived from Th-α, with two copies

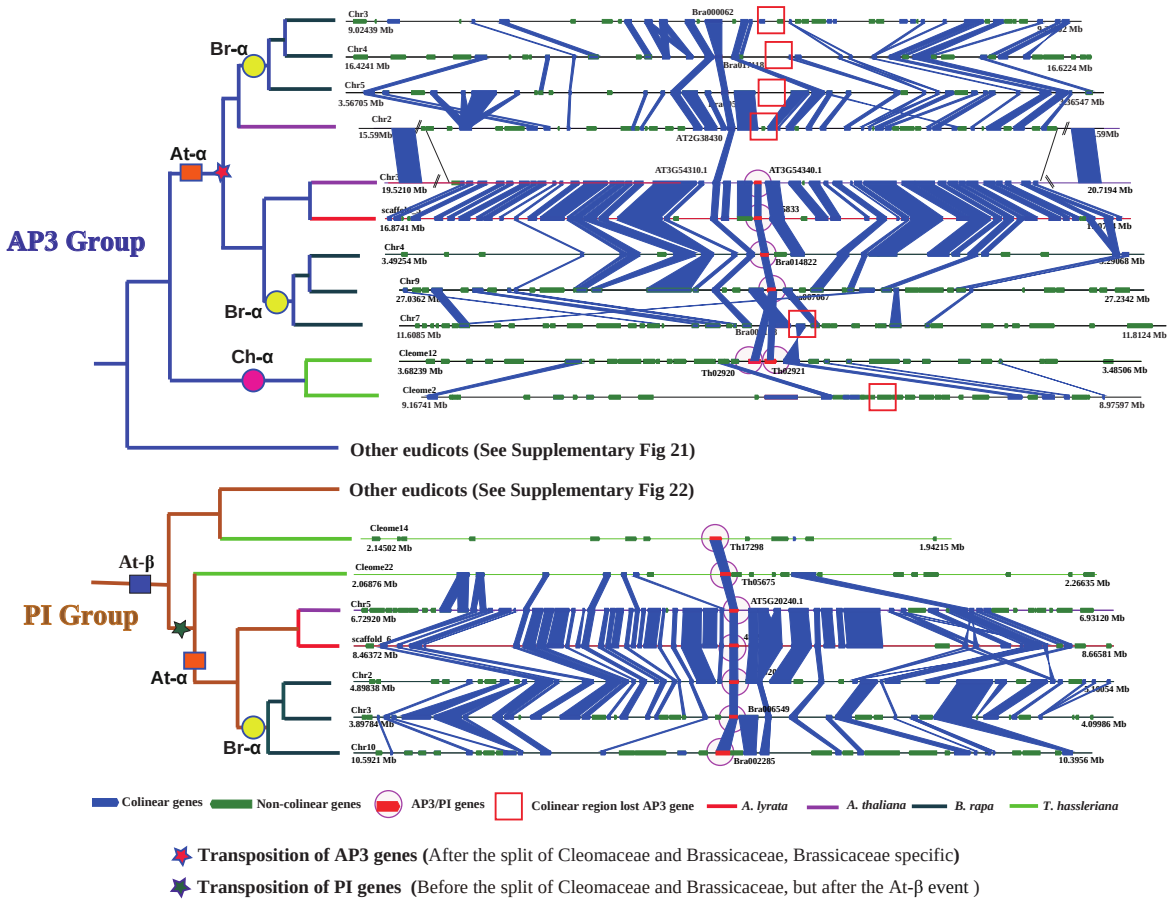


Figure 5. Collinearity analysis of B-class Type II MADS box gene (*AP3* and *PI*) homologs reveals unusual patterns of gene loss, lineage-specific transpositions and local tandem duplications**.** The placement of ancient polyploid events giving rise to gene duplicates are shown on appropriate nodes (At-β, At-α, Br-α and Th-α). For the *AP3* group genes in Brassicaceae (shown by red bars), there is only a single locus retained in *A. thaliana* and *A. lyrata* and two retained Brassica syntelogs derived from Br-α. Collinear homeologous regions derived from At-α are detectable in Brassicaceae genomes, however the *AP3* synteologs were lost (regions highlighted in red boxes). *Tarenaya hassleriana* has an unusual tandem duplication of *AP3* genes in one of two homeologous regions derived from Th-α. The *AP3* genes and the neighboring fork-head gene (*EMBRYO DEFECTIVE 1967)* are the only genes syntenic to the AP3 Brassicaceae region (Supplemental Figure 21). The Cleomaceae AP3 region is syntenic with AP3 regions of all other eudicot genomes analyzed (Supplemental Figures 21-22). Thus, we conclude that there was a lineage-specific transposition of AP3 and the neighboring fork-head locus in the Brassicaceae. There is only a single copy of PI genes (red bars) in *A. thaliana* and *A. lyrata* and all three Br-α derived synteologs in Brassicaceae. There is no detectable homeologous region in Brassicaceae derived from At-α. In *Tarenaya* we detect one syntenic PI gene and region to the Brassicaceae, but also a second region that is syntenic to other eudicots (Supplemental Figure 23). We conclude that these two *Tarenaya* PI genes were generated due to the At-β ancient duplication event with the subsequent transposition of one locus into the region collinear between Brassicaceae and Cleomaceae, and loss of the non-transposed locus from only the Brassicaceae lineage. The differences in genomic context and gene expression (Supplemental Figure 24) may contribute to shifts in floral morphology and symmetry between families.

of *TCP1* intact (Th21666 and Th24587) (Supplementary Figure 24). The correlation between multiple copies of *TCP* members and monosymmetry has been noted across angiosperms (Rosin and Kramer, 2009).

EXPRESSION OF A- AND B-CLASS HOMOLOG GENES

The expression of *T. hassleriana* homologs of *A. thaliana* major floral regulators was also analyzed with qRT-PCR. The two putative *T. hassleriana* homologs of *CAL*/*AP1 (ThCAL/AP1-1*, Th01189 and *ThCAL/AP1-2*, Th13754) show similar expression in all bud stages, but *ThCAL/AP1-1* with only half of the transcript abundance of *ThCAL/AP1-2. ThCAL/AP1-1* shows highest expression in sepals and approximately 10 times lower expression in petals. Expression of neither homolog was detectable in stamens, petals, gynoecium, capsules, roots or leaves. *ThPI-1* (Th05675), the homolog of *PI* in *A. thaliana* is collinear with the Brassicaceae gene order and is expressed mainly in petals and stamens, with less expression in younger and higher expression in older stages (Supplementary Figure 25). The second *PI* homolog of *T. hassleriana*, *ThPI-2* (Th17298) is expressed at a much lower rate than *ThPI-1*, ranging from around 50 % of the *ThPI-1* expression in stamens to only 10% of the *ThPI-1* expression in late bud states.

The second two putative floral homeotic B function genes, *ThAP3-1* (Th02920) and *ThAP3-2* (Th02921) are highly similar in coding, 3', and 5' UTR sequence and both are homologous to the *AP3* gene in *A. thaliana*. *ThAP3-1* is expressed throughout the observed stages of bud development. In petals and stamens at anthesis, it is the most highly expressed gene of the MADS-box genes analyzed (Supplementary Figure 25). In petals, it is expressed at an approximately 200% and in stamens around 400% higher level than *ThCAL*, *ThPI-1*, and *ThAP3-2*. Expression of *ThAP3-2* in buds is around one third lower than that of *ThAP3-1*, and differential expression between both genes is detected in petals and stamens. While *ThAP3-1* shows higher expression in stamens than in petals, *ThAP3-2* has a stronger expression in petals than in stamens (Supplementary Figure 25). The divergence in expression of the B-class genes, along with the aforementioned gene transpositions, is indicative of the likely role in B-class gene functional differentiation and the regulation of the different floral morphologies between families.

EVOLUTION OF BRASSICACEAE SELF-INCOMPATIBILITY LOCUS

Many Brassicaceae species possess a pollen-pistil recognition system that confers self-incompatibility (SI) through the rejection of self-pollen (Boyes et al., 1997). This system is based on the interaction of the stigmatically expressed S-receptor kinase (coded by the *SRK* gene) with a small polymorphic peptide that is coded by the S-locus cysteine-rich protein (*SCR*) gene, both located on the so-called S-locus of the genome. Many *SCR* alleles are likely derived from (partial)-duplication and/or gene conversion from *SRK* alleles (Koornneef and Meinke, 2010; Guo et al., 2011). The cysteine rich protein kinase (*CRK*) genes and *ARK* genes that belong to the S-locus are part of a larger family of receptor-like protein-kinases (*RLK*) genes, which have been shown to be mostly involved in oxidative stress and pathogen response (Chen et al., 2004; Wrzaczek et al., 2010). It should be noted that most ecotypes of *A. thaliana* are self-compatible due to pseudogenization of the *SRK* and/or *SCR* genes, a pattern seen in other self-compatible crucifers, but still an exception amongst Brassicaceae (Nasrallah et al., 2004). *T. hassleriana* does not possess a self-incompatibility system, but it still contains an S-like locus that contains functional genes and it is likely that this part of the genome is close to the ancestral state of this locus for Brassicaceae.

The *SRK* and *ARK* genes on the S-locus are characterized by specific variations on the following protein domain compositions: B_lectin (B), S_locus_glyco (S), PAN-2 (Pa), Pkinase_Tyr (Pk) and Duf3403 (D1)

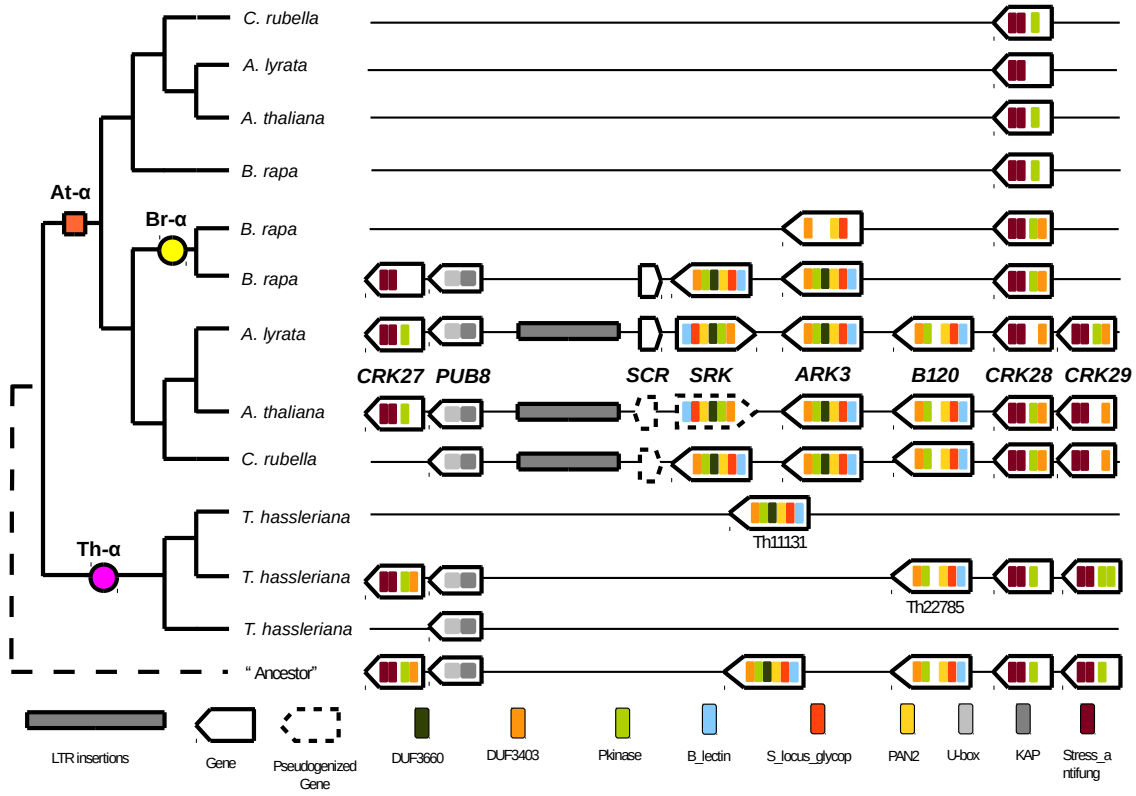


Figure 6. Synteny and protein domain analysis of the Brassicaceae Self-Incompatibility (SI) like regions with Cleomaceae and inference of the ancestral genomic region. All regions presented here are drawn in more detail (including genetic coordinates) in Supplemental Figure 27. Genes are marked by block arrows and color coded according to their protein domain composition as listed in the legend. In the case of pseudogenes, the block arrows have a dashed border. To find protein domain composition in pseudogenes, the longest ORF was translated *in silico* (see Methods). Gene orientation is shown by block arrows pointing left for gene orientation towards the 5' end and pointing right for gene orientation towards the 3'end. The At-α duplication and the Br-α and Th-α triplications have been marked on the tree with an orange box (At-α) and yellow and purple circles (Br-α and Th-α, respectively). Each branch corresponds to a subgenome resulting from such a polyploid event. Theoretically, *B. rapa* should have six subgenomes but only the regions showing synteny are listed here for clarity. The bottom branch represents a hypothetical layout of this genome region in the common ancestor of these species before the At-α, Br-α and Th-α polyploid events. From our results, we conclude that an *ARK3*-like gene underwent a Brassicaceae-specific tandem gene duplication generating the key SI receptor *SRK*.

and/or DUF3660 (D2 (Zhang et al., 2011). Using the Pfam database (Punta et al., 2012), we found that the S locus region in *C. rubella, B. rapa*, *A. lyrata* and *A. thaliana* as well as the homologous region in *T. hassleriana* mostly contains *SRK* genes with a B-S-Pa-D1-Pk-D2 protein domain structure, followed by the B-S-Pa-Pk-D1/2 protein domain structure which is more common across all SRK families (Figure 6 and Supplementary Figure 26). Three syntenic regions containing most of the genes of the S-locus were found in *Tarenaya*. One of these contained a gene with the exact B-S-Pa-Pk-D1/2 protein domain structure: Th11131. Our analysis of domain structure further found that another gene, Th22785, had a B-S-Pa-Pk protein domain structure, an architecture shared with many angiosperm genes but not with SI specific SRK alleles. We have also found two homologs of the SI-modifier gene, *Pub8*, in two of the three *Tarenaya* syntenic regions. One of the homologs, Th22784, is adjacent to the Th22785 locus. The other homolog, Th25331 is contained in the syntenic region that completely lacks any S_locus_glyco Pfam containing proteins. Interestingly, we do not find a *Pub8* homolog in close-proximity to Th11131.

However, based on the alignment of all three regions we can assume that the single copy ancestral region contained homologs to *Pub8*, *ARK3* and *B120* (Figure 6).

We conclude that the syntenic *Tarenaya* gene, Th11131, which is most closely related to *ARK3* with which it shares protein architecture, is similar to the ancestral version of SI-locus genes in Brassicaceae. We hypothesize that the origin of the S-locus is due to a local rearrangement/duplication of an *ARK3*-like gene and subsequent expansion and diversification of the S-locus. *ARK3* in *Arabidopsis* is not expressed in the stigma, so potentially the regulatory domains needed for tissue-specific expression may be derived from the co-current duplication of elements from neighboring genes.

DISCUSSION

The completed genome of *Arabidopsis thaliana* was a major milestone in plant biology and has provided a key-tool for elucidating plant gene function, genome structure and evolutionary patterns (Meinke et al., 1998). Subsequently, there has been a great effort and interest to sequence other crucifer species to leverage the knowledge gained from *Arabidopsis* to other species in a comparative context. To date there are more completed crucifer genomes than any other plant family, including the crop *Brassica rapa* (The Brassica rapa Genome Sequencing Project Consortium et al., 2011)*,* and ambitious plans to sequence many more (such as the Brassica Map Alignment Project (Pires et al., 2013)). We have sequenced *Tarenaya hassleriana,* the first genome from a phylogenetic out-group to the Brassicaceae: its sister-family the Cleomaceae. We have shown that the vast majority of genes in *Tarenaya* have clear homologs within Brassicaceae. We provide several examples of how this sister-group genome can be used to elucidate patterns of gene, genome, and trait evolution within the Brassicaceae. Specifically we focused on independent ancient polyploidy events, floral MADS-box, and Self-Incompatibility gene evolution. The genome of *Tarenaya hassleriana* will facilitate future research into the evolutionary and functional history of *Arabidopsis* genes and pathways.

While it has long been known that there are numerous recent polyploid plants (Jiao et al., 2011), with the arrival of the genomics era it has become clear that there also is extensive evidence for ancient polyploidy across the tree of life (Soltis and Soltis, 1999). Most ancient plant polyploid events that have been identified are ancient tetraploidy events (duplications such as At-α), but there are at least four published genome analyses of ancient plant hexaploidy (triplication events): at the base of the eudicots (At-γ) (Vision et al., 2000), in tomato (T)(2012), in Brassica (Br-α) (The Brassica rapa Genome Sequencing Project Consortium et al., 2011) and this report of the ancient genome triplication in the *Tarenaya* (Tr-α). The Br-α and Tr-α events are of approximately the same age, allowing us to contrast independent ancient hexaploidy events from closely related lineages. The analysis of the retention of replicated genes (syntelogs) of the Type II MADS-box genes provides a compelling example of what can be deduced by the comparison of these independent ancient triplications. We found that Brassica retains nearly twice as many Type-II MADS box genes as does *Tarenaya.* Genes retained after polyploidy often are dosage-sensitive gene complexes whereby interacting partners must be maintained in the proper ratios (Edger and Pires, 2009). Considering the wealth of phenotypic diversity seen in the crop Brassicas, we hypothesize that this great enrichment of morphological regulators in Brassica derived from Br-α may play a significant role.

Comparative analyses can also be used to identify important gene transposition and deletion events. Type-II MADS box genes are remarkably resistant to gene transpositions and thus their collinearity is highly conserved across all angiosperms (Type I MADS-box genes are highly prone to transposition, but their functions are less known). When comparing *Tarenaya* to Brassicaceae we have found almost all Type-II MADS-box genes are collinear, except for the B-class homologs of *AP3* and *PI*. Specifically, we establish that there has been a Brassicaceae-specific transposition *AP3*, and a rare tandem duplication of a floral MADS-box gene of *AP3* homologs in *Tarenaya*. The transposition of *AP3* also involved a

neighboring Forkhead gene, which maybe co-regulated and important for *AP3* function. We further demonstrate that *Tarenaya* has two homologs of *PI,* one that is syntenic with other eudicots and for which the locus is lost from Brassicaceae, and one that is syntenic with Brassicaceae *PI* homologs. Both *PI* and *AP3* in *Tarenaya* have diverged in expression levels. MADS-box B-class gene homologs in the *APETALA3* (*AP3*) lineage as well as *TCP* members have been implicated in the establishment of monosymmetric flowers in monocots, including orchids (Tsai et al., 2004, 2008; Mondragon-Palomino and Theissen, 2009; Bartlett and Specht, 2010; Preston and Hileman, 2012), *PI* genes have contributed to floral diversification in Asterids (Viaene et al., 2009) and B-class genes have contributed to *Aquilegia* floral diversification (Kramer et al., 2007). Thus, it is possible that both B-class and TCP genes may impact floral monosymmetry in Cleomaceae. Furthermore, B-class gene diversification has also been implicated in controlling floral gender shifts (Ackerman et al., 2008) and could similarly have diversified in Cleomaceae.

By comparison of Brassicaceae genomes to *Tarenaya* we establish that the *SRK* gene in the S-locus occurred via a local rearrangement/duplication of an *ARK3*-like gene and subsequent expansion and diversification in Brassicaceae. In *Tarenaya* we can clearly identify syntenic regions that contain homologous functional genes, including an *ARK3* and *CRK* homologs. The exact function of *ARK3* is not known, but based on gene expression analysis is thought to function during development of the sporophyte, perhaps in processes related to organ maturation ,the establishment of growth pattern transitions (Dwyer et al., 1994) and/or involvement in pathogen responses (Pastuglia et al., 2002). However, it is not expressed in the stigma. Further research on Brassicaceae and Cleomaceae *ARK3* homologs is needed to establish the function of these genes. Interestingly, while Cleomaceae species do not have a SI-system, many species including *T. hassleriana* are polygamous (trimonoecious) and can have flowers on the same inflorescence with different genders: male sterile, female sterile, or complete (Stout, 1923; Cruden and Lloyd, 1995; Machado et al., 2006), providing an alternative mechanism to reduce inbreeding.

Our results demonstrate the utility of the *Tarenaya* genome to complement Brassicaceae genetic research to understand the function and evolution of genes and traits. The *Tarenaya* genome will also pave the way for further studies of Cleomaceae traits not found in Brassicaceae, such as the evolution of C_4 photosynthesis (Brown et al., 2005; Marshall et al., 2007).

METHODS

Data access: The genomic reads of *T. hassleriana*, as well as RNA sequencing data, have been deposited into NCBI Short Read Archive (SRA) under accession number SRA058749 and GSM1008474. The information for the raw reads data can be found in Supplementary Table 1. The genome sequence and annotation data set have been deposited into NCBI (Project ID: PRJNA175230 (superscaffolds), the accession number is AOUI00000000).

Sample preparation, Library construction, Genome sequencing and assembly: The purple-flowered *Tarenaya hassleriana* (Purple Queen) line selected for sequencing (ES1100) was first inbred by hand-pollination and floral bagging for four generations. Earlier generations of this line have previously been used for both BAC-library construction and limited BAC-sequencing (Schranz and Mitchell-Olds, 2006) and transcriptome sequencing and analysis (Barker et al., 2009), however the material was referred to as being from *C. spinosa*. The two species are morphologically very similar, with only slight differences in stem-spine morphology pubescence of sepals, ovary and capsules and flower color; *C. spinosa* has only white flowers and the sepal and ovary are glandular-pubescent, whereas *T. hassleriana* can be white, pink or purple and the sepals and ovary are glabrous (have no pubescence). Thus, many commercial seed providers erroneously label their *T. hassleriana* material as *C. spinosa*.

We extracted DNA from leaves *of T. hassleriana* and constructed seven pair-end libraries with insert sizes of 350bp, 500bp, 800bp, 2K, 5K, 10K and 20K. Illumina Hiseq 2000 was then applied to sequence those DNA libraries and in total 70.22 Gb raw data was generated. Low quality reads, reads with adaptor sequences and duplicated reads were filtered and the remained high quality data was used in the assembly. SOAPdenovo2.21 was applied to assemble the genome in the procedure of contig construction, scaffold construction and gap closure. After gap closure, the assembly was broken down into contigs again according to the position of Ns in the assembly. Then those contig sequences were subjected to evaluation of contig assembly.

Whole genome profiling and physical map construction: We prepared the BAC library construction using leaf material of *Tarenaya hassleriana*. Two BAC libraries were subsequently generated, the first using HindIII (CLEH library) and the second using EcoRI (CLEE library). Average insert sizes were 145 kb for the CLEH library and 130 kb for the CLEE library. The vector used for library construction was pCC1BAC (Epicentre). For each library, 9,600 clones were picked and arrayed into 384 well plates. Together the two libraries equal approximately 8.8 genome equivalents (4.6 GE CLEH library & 4.2 GE CLEE library) at an estimated haploid genome size of 300 Mbp. Whole Genome Profiling was performed according to the methods detailed in (van Oeveren et al., 2011).The resulting FPC map was used in further analysis (Supplementary Tables 5 and 6).

FPC map assembly and integration with *de novo* assembled scaffolds: Sequence-based physical BAC maps were assembled using an improved version of FPC software (Keygene N.V.), capable of processing sequence-based BAC fingerprint (WGP) data instead of fragment mobility information as used in the original FPC software (Soderlund et al., 1997). The scaffolds from the SOAPdenovo assembly were then mapped to the physical contigs using nucleotide blast (Altschul et al., 1997). Hits were used only when they had a 100% identity match to the anchors. Subsequent filtering was performed to eliminate anchors with multiple hits and to establish superscaffold strand direction. The scaffolds were then ordered according to the mapped anchors and reassembled into superscaffolds.

RNA-seq: A mixed sample from five tissues (buds, leaves, petioles, stems and flowers) from *Tarenaya* flowering plant was used to isolate RNA. Total RNA was extracted using Trizol (Invitrogen). The isolated RNA was then treated by RNAse-Free DNase, and then subsequently treated using Illumina mRNA-Seq Prep Kit following the manufacturer's instruction. The insert size of the RNA libraries was about 200 bp, and the sequencing was done using Illumina GA II. Raw reads were filtered if there were adaptor contaminations and low quality (>10% bases with unknown quality). After filtering, all RNA reads were mapped back to the reference genome using Tophat Version 1.3.3 (Trapnell et al., 2009), implemented with bowtie66 Version 0.12.7 (Langmead, 2010) and assembled the transcripts according to the genome using Cufflinks (Version 1.1.0) (Trapnell et al., 2012). Single libraries from eight different tissues were isolated with the Qiagen RNeasy plant minikit and treated with RNase free DNase. Illumina TruSeq Libraries were constructed according to the manufacturer's suggestions and sequenced. Raw reads were filtered to remove adaptors and low quality bases and mapped to the predicted coding sequences using Cufflinks.

Plant genome sources: In all analyses where plant genomes are used, source database and version information can be found in Supplementary Table 16.

Gene annotation: We predicted gene models following several steps: A) *De novo* gene prediction. We performed *de novo* predictions on repeat masked genome assembly. We used AUGUSTUS (Version 2.03) (Stanke and Morgenstern, 2005), GlimmerHMM (Version 3.02) (Majoros et al., 2004) and SNAP (Version 2.0) to do the *de novo* annotation. B) Homology gene prediction. We mapped the protein sequences from *A. thaliana, B. rapa, C. papaya, G. max, T. cacao* and *V. vinifera* to the *Tarenaya* genome using tblastn, by an E-value cutoff of 10^{-5}, and then Genewise (Version 2.2.0) (Birney et al., 2004) was used for

gene annotation. C) RNA aided annotation. We mapped all the RNA reads back to the reference genome by Tophat (Version 1.0.14 (Trapnell et al., 2009), implemented with bowtie Version 0.12.5) and assembled the transcripts according to the genome using Cufflinks (Version 0.8.2). All the predictions were combined using GLEAN to produce the consensus gene sets.

The tRNA genes were identified by tRNAscan-SE (Lowe and Eddy, 1997). For rRNA identification, we first downloaded the *Arabidopsis* rRNA sequences from NCBI (http://www.ncbi.hlm.nih.gov/guide/dna-rna). Then rRNAs in the database were aligned against the *Tarenaya* genome using blastn to identify possible rRNAs. Other ncRNAs, including miRNA, snRNA, were identified using INFERNAL (Nawrocki et al., 2009) by searching against the Rfam database.

Gene family clustering: We used OrthoMCL (version 1.4) (Li et al., 2003) with default parameters followed by an all-vs-all BLASTP (E-value<=1e-5) process, to apply to the protein sequence datasets from six plant species. We removed Splice variants from the data set (usually the longest protein sequence prediction is kept) and filtered the internal stop codons and incompatible reading frames. After getting all gene families, we classified the families according to the presence or absence of genes for specific species and determined which gene families were species-specific or genus-specific.

A total of 184204 sequences from *Arabidopsis thaliana, Arabidopsis lyrata, Brassica rapa, Carica papaya, Vitis vinifera* and *Tarenaya hassleriana* were clustered into 24591 gene families. 9395 contained sequences from all six genomes, 2492 from Brassicaceae (*Arabidopsis thaliana, Arabidopsis lyrata, Brassica rapa*), 1176 from plants as out-groups only bearing the At-γ event (*Carica papaya* and *Vitis vinifera*) and 748 clusters were specific to *Tarenaya*. Of the 28917 protein-coding genes predicted for *Tarenaya*, 22,482 were clustered in a total of 14,505 groups. The 748 *Tarenaya*-specific clusters contained 1556 genes of which 529 have at least one INTERPRO domain. Singletons make up a total of 6435 genes of which 2926 have at least one INTERPRO domain. Interestingly, many gene families contracted observed in *Tarenaya* show bigger genes than others, more exons, and there is more TEs insertion in these contraction gene families. However, more GO annotation is enriched for these contraction gene families. These results indicated that these contraction gene families may be functional constraint.

Repeat annotation: Repeats of *Tarenaya* genome were identified by a combination of homology-based and *de novo* approaches. In the homology-based method, we used databases of known repetitive sequences to search against the genome assembly, in this way RepeatMasker-3.2.9 and RepeatProteinMask software (Chen, 2004) were used to build the homology database and search the repeat sequence.

Furthermore, in the denovo approach, three denovo software packages (Piler-DF-1.0) (Edgar and Myers, 2005),RepeatScout-1.0.5 and LTR-FINDER-1.0.5 (Xu and Wang, 2007)) were utilized to build *de novo* repeat database of the *Tarenaya* genome. We then used RepeatMasker to identify repeats using both the repeat database we've built and Repbase. At last we combined the *de novo* prediction, the homolog prediction of TEs according to the position in the genome.

Phylogenetic analysis and species divergence time estimation: We constructed the ML phylogenetic tree of *T. hassleriana* and other plant genomes using whole genome 4-fold degenerate sites among species. *Oryza sativa* and *Sorghum bicolor* were taken as the monocot out-groups. The following steps are taken: Firstly, we extracted all the single copy gene families from the OrthoMCL clustering results. Secondly, we run multiple sequence alignment for each single copy gene family using the protein-coding sequences. Thirdly, for each aligned gene family, we did the CDS back-translation of the protein multiple alignments from the original DNA sequences using in-house Perl scripts, and then we extracted the 4-fold degenerate sites of orthologous genes in all single-copy gene families (concatenated into one

supergene for each species). The branch length represents the neutral divergence rate. The substitution model (GTR+gamma+I) and Mrbayes (Ronquist et al., 2012) was used to reconstruct the phylogenetic tree.

Synteny and collinearity analysis: First, we did homology search within and between species by BLASTP (E-value threshold 1e-7, top 20 hits). We removed tandem gene families and weak matches using in-house Perl scripts for further analysis. Tandem gene families were defined as clusters of genes within 10 intervening genes from one another, and we kept the longest model to represent each family. For the weak matches, we retained only top BLAST hits by applying a C-score threshold of 0.8 (C-score(A, B) = score(A, B)/max(best score of A, best score of B)) (Putnam et al., 2007).

Then, based on these filtered BLAST results, the whole genome-wide sequence alignments within and between genomes using genes as anchors, which was to search syntenic blocks, were conducted by an in-house pipeline implementing Dynamic Programming (Parameters: score_of_match: 50, penalty of mismatch: -5, penalty of indel:-5, penalty_of_extension_indel: -2, block_size: >=5 gene pairs, gap_between_neighbor_blocks: 30 genes). The time to running each whole-genome sequence alignment using genes as anchors by Dynamic programming is about 5~6 hours. At the same time, we also used i-Adhore 3.0 (Proost et al., 2012) (http://bioinformatics.psb.ugent.be/software) to identify syntenic and collinearity blocks (gap_size= 30, cluster_gap= 35, q_value=0.75, prob_cutoff=0.01, anchor_points=5, alignment_method=gg4, level_2_only=false, table_type=family) within and between genomes, and we found all the syntenic blocks identified by i-adhore were contained in our results identified by Dynamic programming method, however for the later, is more sensitive and accurate. All the dot plot figures were plotted using SVG package implemented perl scripts (Supplementary Figures 1, 4-7 and 9-12).

Ancestral genomes reconstruction: Based on the paralogous duplicates within each genome, we created four minimized genomes independently for *A. thaliana, A. lyrata, B. rapa* and *T. hassleriana*, by condensing local duplications to one gene, removing transposons, and including only genes within blocks defined by retained pairs. Each of the minimized genome represents the ancestral state predate the recent polyploidy event. At the same time, we compared these four ancestral state genomes with the Ken Wolfe's 45 ancestral blocks *of A. thaliana* (Supplementary Figure 9-12).

Partitioning of the *T. hassleriana* genome into three subgenomes following the recent polyploidy event: To avoid the potentially confounding results with the independent ancient polyploidy events, we take the ancestor genome of A. lyrata ('A' ancestor) as the reference, and identify the collinear blocks using i-adhore 3.0 by aligning the four proteomes (*A. thaliana, A. lyrata, B. rapa, T. hassleriana*) against the 'A' ancestor genome. From the last common ancestor (The 'A' ancestor represents this genome state) of these four species both of *A. thaliana and A. lyrata* experienced a WGD event (At-α), *B. rapa* experienced one WGD event (At-α) and an additional whole genome triplication event (Br-α), and *T. hassleriana* experienced an independently whole genome triplication event (Th-α), respectively. So, we can observe obviously 2:1, 2:1, 3/4/5/6:1, and 3:1 quota ratios for *A. thaliana, A. lyrata, B. rapa, T. hassleriana* genomes. For *T. hassleriana*, the triplicated blocks identified were chained into three subgenomes compared with the 'A' ancestor using dynamic programming. The main criteria are that the chained triplicated blocks are: 1) non-overlapping in the *T. hassleriana* genome; 2) have no more than 10% overlap between their orthologous 'A' ancestor regions (results were similar using 0% overlap in 'A' ancestor); 3) maximize coverage of the *T. hassleriana* genome (annotated gene space). A similar strategy was taken for other genomes based on their polyploidy level from the last recent common ancestor, as shown in Supplementary Figure 14-17. For *T. hassleriana*, a total of 16770 (63.2%) *Tarenaya* genes are in the triplicated blocks compared with the 'A' ancestor, 688 (2.6%) genes are in the duplicated blocks that indicates maybe another copy is lost after the triplication event, and only 135 (0.5%) have one

syntenic orthologs which means a few part of the trios lost two copies simultaneously. However, 8913 (33.6%) *Tarenaya* genes are not in any replicated blocks, which indicates that specie-specific deletion in *A. lyrata* or specie-specific gains in *Tarenaya* after their divergence along evolutionary time.

Interproscan and gene functional annotation: We used INTERPROSCAN version 4.5 to scan protein sequences against the protein signatures from InterPro (Hunter et al., 2012) (version 22.0) to infer functions for the protein-coding genes. We did so for the entire target proteomes involved in our Main text analysis, including *Arabidopsis thaliana, Arabidopsis lyrata, Tarenaya hassleriana, Brassica rapa, Solanum lycopersicum, Vitis vinifera, Prunus persica, Populus trichocarpa,* and *Carica papaya*. InterPro integrates protein families, domains and functional sites from different databases: Pfam, PROSITE, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, and PANTHER. INTERPROSCAN integrates the searching algorithms of all these databases. In total, INTERPROSCAN identified 93038 protein domains of 4733 distinct domain types. 75% of the genes (21829 out of 28917 genes in total) have been assigned with at least one domain.

Genomic analysis for Reproductive traits: For the syntenic and protein domain analysis of SI-genes we used the genes annotated in *A. thaliana, C. rubella* and *A. lyrata* as published in an extensive study into S-locus variation in *Arabidopsis* species (Koornneef and Meinke, 2010). We then sought homologs using top BLAST hits of these genes against *T. hassleriana* and *B. rapa*. We confirmed all homology candidates by manual inspection of the alignment in dot-plots generated in MAFFT (Katoh and Frith, 2012) to confirm synteny of candidate regions. After compiling a definitive list of syntenic regions we ran the genes from these regions through the PFAM online protein domain analysis program (Punta et al., 2012). Figure 6 was manually compiled from the results of this program.

Quantitative Reverse Transcription PCR (qRT-PCR): For the q RT-PCRs total RNA was isolated from roots, leaves, three bud stages (1-5 mm, 5-10 mm, 10-25 mm length), sepals, petals, stamens, carpels at anthesis and 3 stages of siliques (10 mm, 10-30 mm, 30-50 mm length) with the GeneJET™ Plant RNA purification mini kit (Fermentas GmbH, St.Leon-Rot Germany). First strand cDNA was synthesized using 500 ng total RNA with the RevertAid™ H Minus First Strand cDNA Synthesis Kit (Fermentas, St.Leon-Rot, Germany) using random hexamer primers.

The qRT-PCR experiments were performed according to the MIQE guidelines (Bustin et al., 2009). Exon spanning primers were generated using PerlPrimer 1.1.21 (Marshall, 2004). A primer efficiency test was carried out and the primers were tested with genomic DNA to ensure cDNA specificity. Standard dose response (SDR) curves were constructed for all genes by using serial dilutions (1:50 to 1:50,000) of 10-25 mm long bud cDNA template to calculate amplification efficiency. The qRT-PCR assay was performed with the LightCyler®480 II (Roche, Mannheim Germany) and the data analyzed with the LCS480 1.5.0.39 software. Each reaction was composed of 10µl of 2x DyNAmo™ Flash SYBR® Green Mastermix (Biozym Scientific GmbH, Oldendorf Germany), 2 µl each of 10 µM forward and reverse primers, 1 µl H_2O and 5 µl of 1:100 diluted template cDNA. Each reaction was performed in biological duplicates and technical triplicates along with water and RNA controls for each primer pair. The *GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C SUBUNIT* (*GAPC1*) and *ELONGATION FACTOR 1-ALPHA* (*ELFA*) genes served as internal controls. The following PCR program was used: 7 min at 95^0C; 45 cycles of 10 s at 95^0C, 15 s at 60^0C, 15 s at 72^0C, followed by a melting curve of 5s at 95^0C, 1 min at 65^0C and 30 s at 97^0C. The quantification cycles (C_q) were calculated according to the second derivative maximum algorithm. The raw C_q data was analyzed according to the Comparitive C_q method ($\Delta\Delta C_q$) (Schmittgen and Livak, 2008). Gene expression was first normalized relative to the expression of the two reference genes in the respective tissues. The expression was further normalized with the expression of the reference genes in 10-25 mm long buds which acted as an inter-assay calibrator. The relative expression was then calculated with reference to the expression of *T. hassleriana CAL* in stage 3 buds.

ACKNOWLEDGMENTS

This work was supported by following funding sources to BGI-Shenzhen: State Key Laboratory of Agricultural Genomics, Guangdong Provincial Key Laboratory of core collection of crop genetic resources research and application (2011A091000047), Shenzhen Engineering laboratory of Crop Molecular design breeding, National Natural Science Funds for Distinguished Young Scholar (30725008). E.v.B., J. H. and M.E.S. were supported by the Netherlands Organization for Scientific Research (NWO VIDI Grant 864.10.001 and NWO Ecogenomics Grant 844.10.006). K.K. was supported by an SK grant from the Alexander-von-Humboldt foundation. S.d.B was supported by an NWO Experimental Plant Sciences graduate school 'master talent' fellowship. A.P.M.W. appreciates support from the Deutsche Forschungsgemeinschaft (grants WE 2231/9-1; EXC 1028).

AUTHOR CONTRIBUTIONS

M. E. S., G. Z., J.M.H. and X. Z. designed the project. G. Z., M. E. S and S. C. led the sequencing and analysis. J. C. and G. F. did the SOAPdenovo genome assembly. J. C. and X. Z. did the annotation. P. Z., S. C., E. v.d. B., M. E. S. and C. B. did the evolutionary analysis. S. C., J. X., E. v.d. B., and J. H. constructed the physical map. M. E. S., S. C, and E. v.d. B. did the reproductive trait evolution analysis. K.K and S.d.B did the phylogenetic analysis of MADS-box genes. C.K., A.B., and A.P.M.W. conducted tissue-specific transcriptomic analyses of *T. hassleriana*. . A. B. and A. B. did the qRT-PCR analysis of MADS-box genes. J.C.H. did analysis of the *TCP* gene family. M. E. S., E. v.d. B and S. C. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

SUPPLEMENTAL DATA

All supplemental figures, tables and data can be found in the published version of this article online: http://dx.doi.org/10.1105/tpc.113.113480 (Last accessed 30th January 2017).

CHAPTER 2

GENE AND GENOME DUPLICATIONS AND THE ORIGIN OF C_4 PHOTOSYNTHESIS: BIRTH OF A TRAIT IN THE CLEOMACEAE

Erik van den Bergh[1], Canan Külahoglu[2], Andrea Bräutigam[2], Julian M. Hibberd[3], Andreas P.M. Weber[2], Xin-Guang Zhu[4], M. Eric Schranz[1]

1 Biosystematics, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

2 Institute of Plant Biochemistry, Center of Excellence on Plant Sciences(CEPLAS), Heinrich-Heine-University, D-40225 Düsseldorf, Germany

3 Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom

4 Plant Systems Biology Group, Partner Institute of Computational Biology, Chinese Academy of Sciences/Max Planck Society, Shanghai 200031, China

ABSTRACT

C4 photosynthesis is a trait that has evolved in sixty-six independent plant lineages and increases the efficiency of carbon fixation. The shift from C3 to C4 photosynthesis requires substantial changes to genes and gene functions effecting phenotypic, physiological and enzymatic changes. We investigate the role of ancient Whole Genome Duplications (WGD) as a source of new genes in the development of this trait and compare expression between paralog copies. We compare Gynandropsis gynandra, the closest relative of Arabidopsis that uses C4 photosynthesis, with its C3 relative Tarenaya hassleriana that underwent a WGD named Th-α. We establish through comparison of paralog synonymous substitution rate that both species share this paleohexaploidy. Homologous clusters of photosynthetic gene families show that gene copy numbers are similar to what would be expected given their duplication history and that no significant difference between the C3 and C4 species exists in terms of gene copy number. This is further confirmed by syntenic analysis of Tarenaya hassleriana, Arabidopsis thaliana and Aethionema arabicum, where syntenic region copy number ratios lie close to what could be theoretically expected. Expression levels of C4 photosynthesis orthologs show that regulation of transcript abundance in T. hassleriana is much less strictly controlled than in G. gynandra, where orthologs have extremely similar expression patterns in different organs, seedlings and seeds. We conclude that the Th-α and older paleopolyploidy events have had a significant influence on the specific genetic makeup of Cleomaceae versus Brassicaceae. Because the copy number of various essential genes involved in C4 photosynthesis is not significantly influenced by polyploidy combined with the fact that transcript abundance in G. gynandra is more strictly controlled, we also conclude that recruitment of existing genes through regulatory changes is more likely to have played a role in the shift to C4 than the neofunctionalization of duplicated genes.

KEYWORDS

Plant Genome Evolution; Synteny; Cleomaceae; Brassicaceae; Bioinformatics; Whole Genome Duplication; Paleopolyploidy; C_4 Photosynthesis

INTRODUCTION

Over sixty lineages of both monocot and eudicot angiosperms have evolved a remarkable solution to maximize photosynthesis efficiency under low CO_2 levels, high temperatures and/or drought: C_4 photosynthesis (Sage et al., 2011). The evolution of this modified photosynthetic pathway represents a wonderful example of convergent evolution. While the changes necessary for the transition from C_3 to C_4 photosynthesis are numerous, the trait has a wide phylogenetic distribution across angiosperms, with 19 different plant families across the globe known to contain one or multiple members capable of C_4 photosynthesis (Sage, 2004). Much research on eudicot C_4 has focused on *Flaveria* species (Asteraceae), which contains not only C_4 species but also a number of C_3/ C_4 intermediates (Ku et al., 1991). With the emergence of genomics and the choice of *Arabidopsis thaliana* as the genomics standard model organism, species in the Cleomaceae, a sister-family to the Brassicaceae (containing *Arabidopsis* and Brassica crops) have been proposed for genetic studies of C_4 (Brown et al., 2005; Marshall et al., 2007)).

C_4 plants spatially separate the fixation of carbon away from the RuBisCO active site by using phospho*enol*pyruvate carboxylase, an alternate carboxylase that does not react with oxygen. As a consequence they are more efficient under permissive conditions (Zhu et al., 2010). The typical C_4 system is characterized by a morphological change: so-called Kranz anatomy (Edwards et al., 2004). In this anatomy, specialized mesophyll (M) cells surround enlarged bundle sheath (BS) cells, with the leaf veins internal to the BS. Generally, the veination in C_4 leaves is increased (McKown and Dengler, 2014). This internal leaf architecture physically partitions the biochemical events of the C_4 pathway into two main phases. In the first phase, dissolved HCO_3^- is assimilated into C_4 acids by phospho*enol*pyruvate carboxylase (PEPC) in the mesophyll cells. In the second phase, these acids diffuse into the chloroplast loaded bundle sheath (BS) cells, where they are decarboxylated and the released CO_2 is fixed by

RuBisCO. The increased CO_2 concentration in the BS cells allows carbon fixation by RuBisCO to be much more efficient by reducing photorespiration. Two subtypes of the C_4 biochemical pathway are defined, based on the most active C_4 acid decarboxylase that liberates CO_2 from C_4 acids in the bundle sheath: NADP-malic enzyme (NADP-ME), NAD-malic enzyme (NAD-ME); a facultative addition of phosphoenolpyruvate carboxykinase (PEPCK) activity can be present in either subtype (Wang et al., 2014). The subtypes are used as a classification scheme for C_4.

The process of carboxylation and decarboxylation costs more energy than the simpler C_3 form of photosynthesis, but it diminishes photorespiration. In conditions of low atmospheric CO_2 pressure, photorespiration causes a major loss in photosynthetic output and the elaborate concentrating mechanisms of C_4 photosynthesis circumvent this (Ehleringer et al., 1997).

All genes important for the C_4 pathway are expressed at relatively low levels in C_3 leaves (Bräutigam et al., 2011). The mechanism for recruitment of these genes into the C_4 pathway remains to be elucidated. For some ancestral C_3 genes changes in *cis*-regulatory elements, while in others changes in *trans* generate M and BS cell specificity (Hibberd and Covshoff, 2010; Brown et al., 2011; Kajala et al., 2012), indicating variation in the mechanisms underlying gene recruitment into the C_4 pathway. It has been proposed that gene duplication and subsequent neofunctionalization of one gene copy has facilitated the alterations in gene expression that underlie the evolution of C_4 photosynthesis (Monson, 1999, 2003). Gene duplication is proposed to be a (pre)condition for the evolution of C_4 because it allows the organism to maintain the original gene while a duplicate version can acquire beneficial changes. This can lead to significant changes in metabolism without the deleterious effect of modifications to essential genes. A recent study that compared convergent evolution of photosynthetic pathways with parallel evolution concluded that duplications are not essential for the development of C_4 biochemistry, but rather changes in expression and localization of specific genes (Bräutigam et al., 2011; Külahoglu et al., 2014). However, this study highlighted just the number of C_4 genes and did not take into account the age and mechanism of gene duplications.

The modifications necessary for the anatomical changes from C_3 to C_4 photosynthesis are not well established. Recent work has shown that the SCARECROW (SCR) gene that is responsible for vein formation in roots, can produce proliferated bundle sheath cells as well as other changes that can be coupled to the shift to the Kranz anatomy (Slewinski et al., 2012). Further work supports this relation by describing the role that the upstream interacting partner of SCR, SHORT-ROOT (SHR) plays in the variations in anatomy seen in various C_4 species (Wang et al., 2013; Slewinski et al., 2014).

Gene duplicates must be further refined by the mechanism by which they arise; either as single gene tandem duplication or Whole Genome Duplication (WGD). Tandem duplications occur frequently, but the duplicates are often lost again resulting in a constant birth-death cycle of duplicate genes (Cannon et al., 2004). Second, there is Whole Genome Duplication (WGD) or polyploidy, where all genes are simultaneously duplicated. After duplication there are often dramatic changes in the plant genomic structure, a process referred to as diploidization in which most genes return to single copy. However, the genes that are maintained in duplicate after WGD often have important functions in enzyme

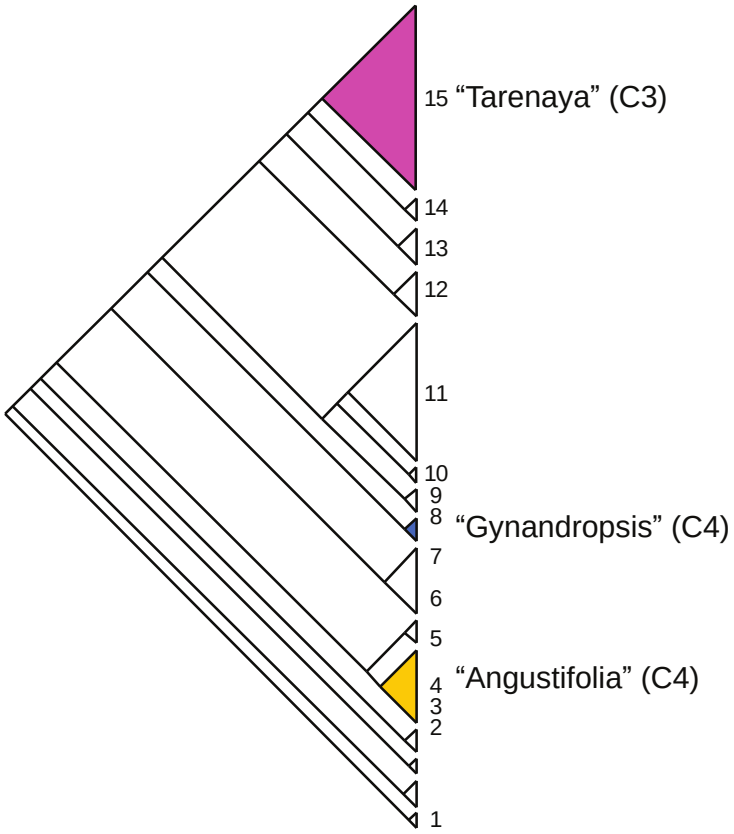


Figure 1. Simplified phylogeny of Cleomaceae. Clades are numbered following the most recently published Maximum Likelihood phylogeny of Cleomaceae (Feodorova et al., 2010). Clade 15 containing *T. hassleriana* is marked in pink. Clade 8 containing *G. gynandra* is marked in blue. Clade 5 (Yellow) contains the other origin of C_4 in Cleomaceae, with *C. angustifolia* and C_4/C_3 intermediate *C. paradoxa*.

complexes (e.g. to maintain proper gene balance (Edger and Pires, 2009)) or can diversify and evolve new gene functions (e.g. neo-functionalization).

The contribution of WGD to photosynthesis-related genes has been studied in soybean, barrel-medic, *Arabidopsis*, and sorghum (Wang et al., 2009; Coate et al., 2011). The polyploid and non-polyploid duplicated gene retention in *Glycine max, Medicago truncatula* and *Arabidopsis* for four classes of photosynthesis-related genes was compared: the Calvin-Benson-Bassham-cycle (CBBC), the light-harvesting complex (LHC), photosystem I (PSI) and photosystem II (PSII). It was found that photosystem genes were more dosage sensitive, with more duplicates derived only from WGD whereas CC gene families were often larger with more non-polyploid duplicates retained. In *Sorghum bicolor*, a recent WGD was reported to be an important origin of C_4 specific genes. Several key C_4 genes of this crop were found to be collinear with genes that function in C_3 photosynthesis when compared to maize and rice. Here, we combine the approaches of these two studies to examine the evolution of photosynthesis and C_4-related genes in C_3 and C_4 Cleomaceae species.

Gynandropsis gynandra (Fig. 1, blue clade) belongs to the NAD-ME C_4 photosynthesis sub-type (Voznesenskaya et al., 2007; Feodorova et al., 2010) and is an important South-East Asian and African dry-season leafy vegetable (sometimes referred to as Phak-sian or African cabbage), and is closely related to horticultural C_3 species *Tarenaya hassleriana* (Fig. 1, pink clade). Both species are easily cultivated in the greenhouse, and a robust phylogenetic framework for Cleomaceae species is emerging (Brown et al., 2005; Marshall et al., 2007; Marquard and Steinback, 2009). There are two other independent origins of the C_4 within the Cleomaceae, *Cleome angustifolia* and *Cleome oxalidea* (Fig. 1,

yellow clade), identified by carbon isotope discrimination (Marshall et al., 2007; Feodorova et al., 2010). Because of the economic importance and ease of growth, the C_4- C_3 contrast between *G. gynandra* and *T. hassleriana* makes this system most attractive and tractable. Both species also have relatively small genome sizes (*T. hassleriana* = 292 Mb and *G. gynandra* ≈ 1 Gb). *Tarenaya hassleriana* underwent a WGD named Th-α (Barker et al., 2009) but it is not yet known whether this event is shared with all or a subset of other Cleomaceae.

In this study we compare C_3 *T. hassleriana* of the Cleomaceae with C_4 *G. gynandra* of the same family. We use the knowledge of Brassicaceae gene functions to identify the important photosynthetic genes in both species and address the following questions: Does *G. gynandra* share the Th-α event? What is contribution of duplicate genes to photosynthesis and C_4-related gene families? And finally, what is the role of gene duplicates from WGD compared to continuous small-scale duplications?

METHODS

TRANSCRIPTOME SEQUENCING AND ASSEMBLY

All transcriptome data was used directly from the Cleomaceae transcript atlas (Külahoglu et al., 2014). In the atlas, *T. hassleriana* genes were used as a reference to map transcripts from both species to Cleomaceae "unigenes" indicated by the gene name coined in the published *T. hassleriana* genome (Cheng et al., 2013). For gene quantification we used default BlatV35 parameters (Kent, 2002) in protein space for mapping, counting the best matched hit based on e-value for each read uniquely.

HOMOLOG SELECTION

A TBlastX (Altschul et al., 1997; Camacho et al., 2009) search of transcriptomes of *T. hassleriana* and *G. gynandra* was performed with default parameters (no evalue cutoff) to have a maximum number of hits for subsequent filtering. To filter paralogs and orthologs from these results, CIP/CALP filtering was used (Murat et al., 2012). Cumulative Identity Percentage (CIP) is defined as the sum of the number of matching nucleotides for each high-scoring segment pair (HSP) of a pair of genes divided by the total lengths of those HSPs. Cumulative Alignment Length Percentage (CALP) is defined as the sum of the alignment lengths of all HSPs of a matching gene pair divided by the total length of the query sequence. Both of these values give a reliable estimation of the similarity of two genes and is a more accurate method than evalue or bit score threshold filtering. A CIP/CALP threshold of 50/50 was chosen as a suitable cutoff point for orthology and/or paralogy.

KS/4DTV CALCULATION OF PARALOG PAIRS

Paralogs identified with CIP/CALP filtering were aligned using Exonerate (Slater and Birney, 2005) with the coding2coding model parameter, using a custom output format through the "roll your own" parameter. The exact command line used was: "`exonerate -m c2c seq1.fasta seq2.fasta --ryo \"%Pqs %Pts\\n" --showalignment false --verbose 0`". The output from this command was fed into CodeML from the PAML package using standard parameters (Codonfreq = 2, kappa = 2, omega = .4). Output from PAML (Yang, 1997) was parsed using custom Perl scripts to read the synonymous substitution rate (Ks) and the fourfold transversion rate (4dtv). This workflow is identical to the established paralog identification pipeline Duppipe (Barker et al., 2010) using updated tools and more stringent selection using CIP/CALP.

HOMOLOG CLUSTERING

Photosynthesis genes were selected from known functionally annotated *Arabidopsis* genes. Gene identifiers used for each family are listed hereafter and in Table 2. βCA: AT1G23730, AT1G58180, AT1G70410, AT3G01500, AT4G33580, AT5G14740. MDH (cytosolic): AT1G04410, AT5G43330, AT5G56720. MDH (mitochondrial): AT1G53240, AT2G22780, AT3G15020, AT3G47520, AT5G09660. MDH (peroxisomal): AT1G53240, AT2G22780, AT3G15020, AT3G47520, AT5G09660. MDH (plastidic):

AT1G53240, AT2G22780, AT3G15020, AT3G47520, AT5G09660. NAD-ME: AT2G13560, AT4G00570. NADP-ME: AT1G79750, AT2G19900, AT5G11670, AT5G25880. PEPC: AT1G21440, AT1G53310, AT2G42600, AT3G14940. PPCK: AT1G08650, AT3G04530, AT3G04550, AT4G37870, AT5G28500, AT5G65690. These genes were then used as a BLAST database and queried with *T. hassleriana* and *G. gynandra* atlas unigenes. Hits were then filtered using a 50/50 CIP/CALP cutoff. Using custom Perl scripts, the hits of these hits were picked up, iterating recursively until convergence (no new hits found). All unique genes resulting from this process form a family cluster.

SYNTENY ANALYSES

Tha genes were used as a query in the CoGe Synfind (Lyons and Freeling, 2008) program using the following parameters: Comparison algorithm: Last, Gene window size: 40, Minimum number of genes: 4, Scoring Function: Collinear, Syntenic depth: unlimited. As query genomes, the following were used: *Aethionema arabicum* VEGI unmasked v2.5, *A. thaliana* Col-0 TAIR unmasked v10.02 and *Tarenaya hassleriana* BGI; Eric Scranz Lab; Weber lab unmasked v5.

RESULTS

EVIDENCE OF WGD IN BOTH SPECIES CONFIRMING A SHARED EVENT

Using the transcript sets of *Gynandropsis gynandra* and *Tarenaya hassleriana*, paralogs were matched to each other by BLAST search and CIP/CALP filtering. In total, 55014 paralogs were found: 26883 in *T. hassleriana* covering 49% of transcript space and 28131 in *G. gynandra* covering 48% of transcript space. Of all paralog pairs, Ks and fourfold transversion substitutions (4dtv) were determined and binned to establish an evolutionary time distribution (Figure 2). In both species a large gene birth event has taken place around Ks = 0.4 (Figure 2 between Ks = 0.25 and Ks = 0.5), which corresponds to the Ks window established earlier for the Th-α hexaploidy event (Barker et al., 2009). The same analysis was performed using 4dtv values and results were extremely similar. Enumerating the paralogs that fall within the Th-α peak, we see that 15785 gene pairs in *T. hassleriana* are retained from the Th-α paleohexaploidy, or

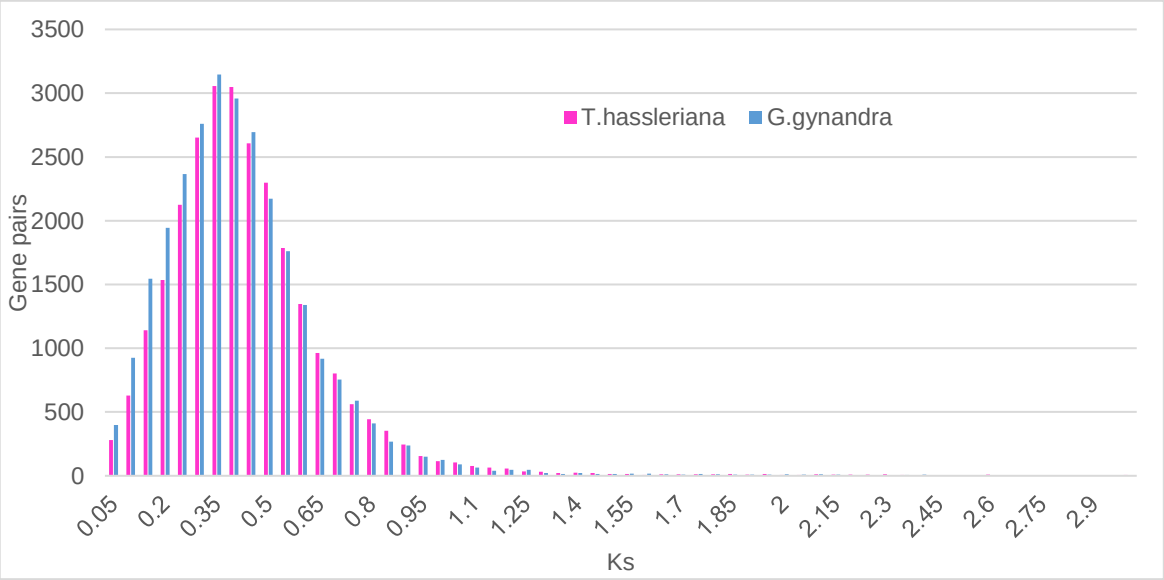


Figure 2. Histogram showing the amount of gene pairs per Ks bin for *T. hassleriana* (pink) and *G. gynandra* (blue). The peak at around Ks = 0.45 is an indication of a massive gene birth event and is considered evidence of paleopolyploidy. Both species have an extremely similar peak, indicating that this is a shared polyploidy event. The Ks values of these peaks corresponds with Ks values found earlier for the Th-α hexaploidy event, indicating that this event has occurred before divergence of *T. hassleriana* and *G. gynandra*.

~29% of the total transcriptome. For *G. gynandra,* 16096 gene pairs fall within the Th-α window, or around 27% of all transcripts.

DUPLICATE LOSS AND RETENTION IN ESSENTIAL C_4 FAMILIES

We examined six gene families that are essential in C_4 photosynthesis in detail: NAD malic enzyme (NAD-ME), NADP malic enzyme (NADP-ME), β carbonic anhydrase (βCA), malate dehydrogenase (MDH), phospho*enol*pyruvate carboxylase (PEPC) and phospho*enol*pyruvate carboxykinase (PPCK). Using *Arabidopsis* genes as a reference, homologous clusters were created using a CIP/CALP cutoff of 50/50. 146 homologous pairs could be placed in a cluster across the three species comprising 105 unique genes (Table 1); 40 in *A. thaliana,* 57 in *T. hassleriana* and 49 in *G. gynandra*. In most cases both Cleomaceae species have around 1.5 times the number of genes of *A. thaliana* except, interestingly, the NADP-ME family where numbers are almost the same in all species. Also of note is that *T. hassleriana* has 16% more C4 related genes in total than *G. gynandra* (57 over 49).

All genes of one species in a cluster were then aligned to each other and the Ks value of each pairing was established and subsequently binned with a stepsize of Ks = 0.15 (Figure 3). At the Ks corresponding to the Th-α hexaploidy, both *T. hassleriana* and *G. gynandra* show a relative increase of gene pairs with this amount of synonymous substitutions. *A. thaliana* at the Ks of its older At-α event shows a similar, if slightly lower increase. Even longer ago in evolutionary time at the Ks corresponding to the β event *T. hassleriana* has retained ~20% of C_4 related genes, where the other species show 2% and 0% retention for *G. gynandra* and *Arabidopsis thaliana*, respectively. The final confirmed paleohexaploidy that all three species share, the ancient γ event at Ks = 2.4, has contributed substantially to the genetic makeup of all three species. In *A. thaliana* the number of relations that stem from the γ paleohexaploidy is 23%, with both Cleomaceae at 15% and 21% for *T. hassleriana and G. gynandra*, respectively.

Table 1. C4 photosynthesis homolog cluster sizes in *A. thaliana*, *T. hassleriana* and *G. gynandra*. Both Cleomaceae species have around 1.5 times the number of genes of *A. thaliana* except the NADP-ME and NAD-ME families where numbers are lower than average in the Cleomaceae species resulting in a similar amount of homologs in each species for these two gene groups.

	A. thaliana	*T. Hassleriana*	*G. gynandra*
βCA	6	10	7
MDH (cyt.)	3	6	6
MDH (mit.)	5	6	6
MDH (per.)	5	8	6
MDH (plast.)	5	6	6
NAD-ME	2	3	3
NADP-ME	4	4	3
PEPC	4	8	6
PPCK	6	6	6
Total	40	57	49

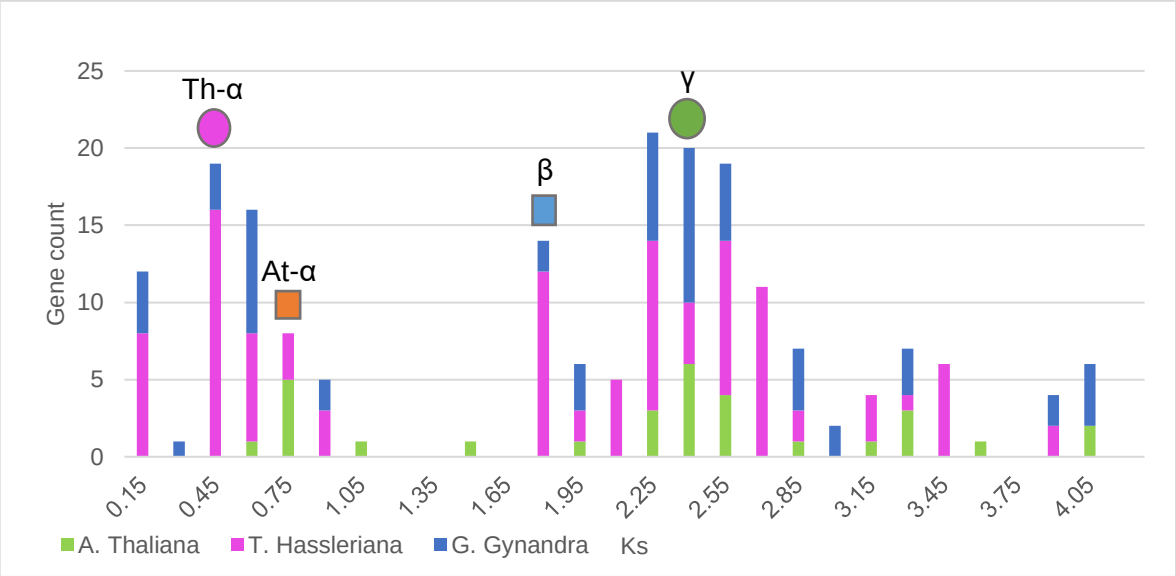


Figure 3. Histogram showing Ks values of homolog gene clusters associated with C4 photosynthesis: MDH, NAD-ME, NADP-ME, PEPC and βCA. Gene duplication events are marked at their associated Ks value and colored according to earlier publication (Barker, Vogel et al. 2009); a square indicates a duplication (tetraploidy), a circle indicates a triplication (hexaploidy). The contribution of the Th-α (pink circle) and the At-α (orange square) on photosynthesis related gene copy number can be seen at Ks = 0.45 and Ks = 0.6 respectively. The β event at Ks = 1.8 (blue square) has contributed substantially to the expansion of gene copy number in *T. hassleriana*. Further in evolutionary time, around Ks = 2.4, the γ event (green circle) that is also shared by all three species has contributed equally to the polyploid presence in photosyntenic orthologs.

SYNTENIC COPY NUMBER VARIATION

Syntenic analyses of the previously mentioned gene families was performed using CoGe Synfind.(Lyons and Freeling 2008). Each *T. hassleriana* c4 related ortholog was used as a query with *T. hassleriana, Arabidopsis thaliana*, *Aethionema arabicum* (Haudry et al., 2013) as a basal representative of Brassicaceae. Thus for the *T. hassleriana* : *A. thaliana* : *Aethionema arabicum* ortholog ratio we would theoretically expect 3 (Th-α) : 2 (At-α) : 2. Query results were enumerated and the average number of regions per family was determined (Figure 4). For many families, the average is comparable to the 3:2:2 ratio, which is also represented by the average ratio (Figure 4, rightmost set of bars) being 3.6 : 2.1 : 2.5. The exception is the NAD-ME family, which has seen more than expected retention with an orthologs ratio 4.3 : 3.3 : 4.3. The PEPC family also seems slightly under-retained in Brassicaceae, with a ratio of 3.3 : 1.3 : 1.6. Unfortunately, syntenic data is impossible to obtain without a sequenced genome so data syntenic regions of *G. gynandra* will have to be obtained in future work.

REGULATION OF PHOTOSYNTHETIC HOMOLOG EXPRESSION

Both Cleomaceae have substantially more copies of photosynthetic genes (Figure 4). Using the Cleomaceae expression atlases (Külahoglu et al., 2014), the expression of separate copies was compared in the C_3 and the C_4 species. In the expression atlas, the *T. hassleriana* coding sequence was used as a reference to map expression in both *T. hassleriana* and *G. gynandra* to a single Cleomaceae 'unigene'. Expression was quantified in nine different tissues including three developmental series: development from young to mature leaf (six stages), root, stem, stamen, petal, carpel, sepal, a seedling developmental series (three stages) and a seed time series (three stages).

For the photosynthetic gene families(NAD-ME, NADP-ME, PEPCK, PEPC, MDH, CA), homolog selection resulted in a data set of 43 unigenes with expression data for both Cleomaceae species. Expression levels were normalized and compared amongst photosynthetic gene families, examples of which are plotted for NAD-ME and βCA (Figure 5). Immediately noticeable is the highly similar expression profiles

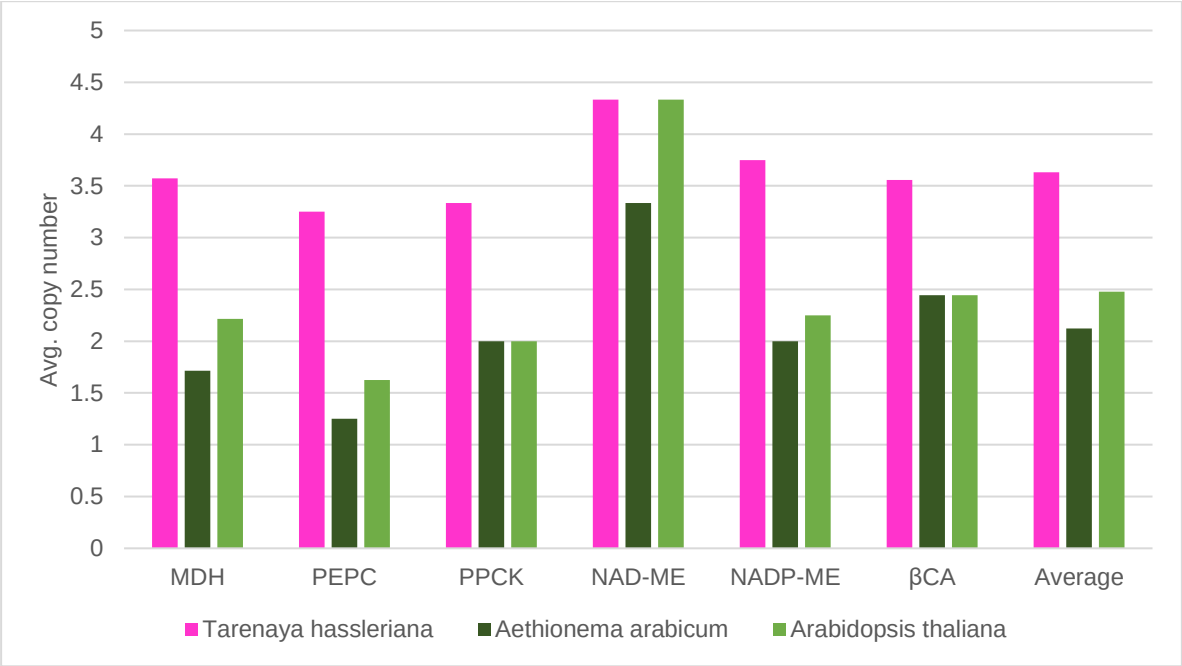


Figure 4. Histogram showing average syntenic region copy number for *T. hassleriana, A. thaliana* and *Aethionema arabicum*. Because *A. arabicum* and *A. thaliana* both share a paleotetraploidy, the expected ratio of syntenic regions for *T. hassleriana* : *A. thaliana* : *Aethionema arabicum* is 3 : 2 : 2. In most cases, syntenic regions follow this distribution which is also reflected in the average ratio of all families being 3.6 : 2.1 : 2.5 (rightmost bars). The exception is NAD-ME, where the average region number in both *A. arabicum* and *A. thaliana* is as high as *T. hassleriana*.

of *G. gynandra* when compared to the more chaotic profiles of *T. hassleriana*. This is observed in all except one gene family. *G. gynandra* has 176 expressed unigenes with a highly correlated expression pattern (Pearson correlation > .95) whereas in *T. hassleriana* 87 unigenes share a highly correlated expression pattern (Pearson correlation > .95).

The expression pattern that is observed in *G. gynandra* in the β-CA family also correspond to their *A. thaliana* highest ranking match (Table 2). The cluster consisting of C.spinosa_00253, C.spinosa_13896, C.spinosa_18526 and C.spinosa_10164 for example all match highest to *A. thaliana* gene β carbonic anhydrase 4 (AT1G70410). The cluster consisting of C.spinosa_07642 and C.spinosa_13410 both map to carbonic anhydrase 1 (AT3G01500). A similar pattern is present in NAD-ME where the cluster of C.spinosa_03046 and C.spinosa_09126 both map to NAD-ME1 (AT2G13560) and the C.spinosa_12536 singleton maps to NAD-ME2 (AT4G00570).

DISCUSSION AND CONCLUSIONS

In this study, we have analyzed the transcriptomes of the C_3 *T. hassleriana* and C_4 *G. gynandra* to address the potential contribution of WGD and recent gene duplicates to the evolution of photosynthesis and C_4-pathway related genes. The initial comparison of *T. hassleriana* and *G. gynandra* was performed to identify the differential expression of key-genes involved in the NAD-ME C_4 biochemical pathway. However, it did not consider the role of gene duplicates. We show that very distinct patterns will occur when the duplication history is taken into account.

We could confirm the Th-α hexaploidy that has been found in *T. hassleriana* using an independent transcriptome dataset. We also find that *G. gynandra* shares this WGD with *T. hassleriana*, further establishing the occurrence of WGD in this lineage. Based on the phylogenetic position of both species in Cleomaceae, the Th-α duplication took place at least before the divergence of the two species which means that it is shared across Cleomaceae lineages 8-15 according to the latest phylogeny of the family

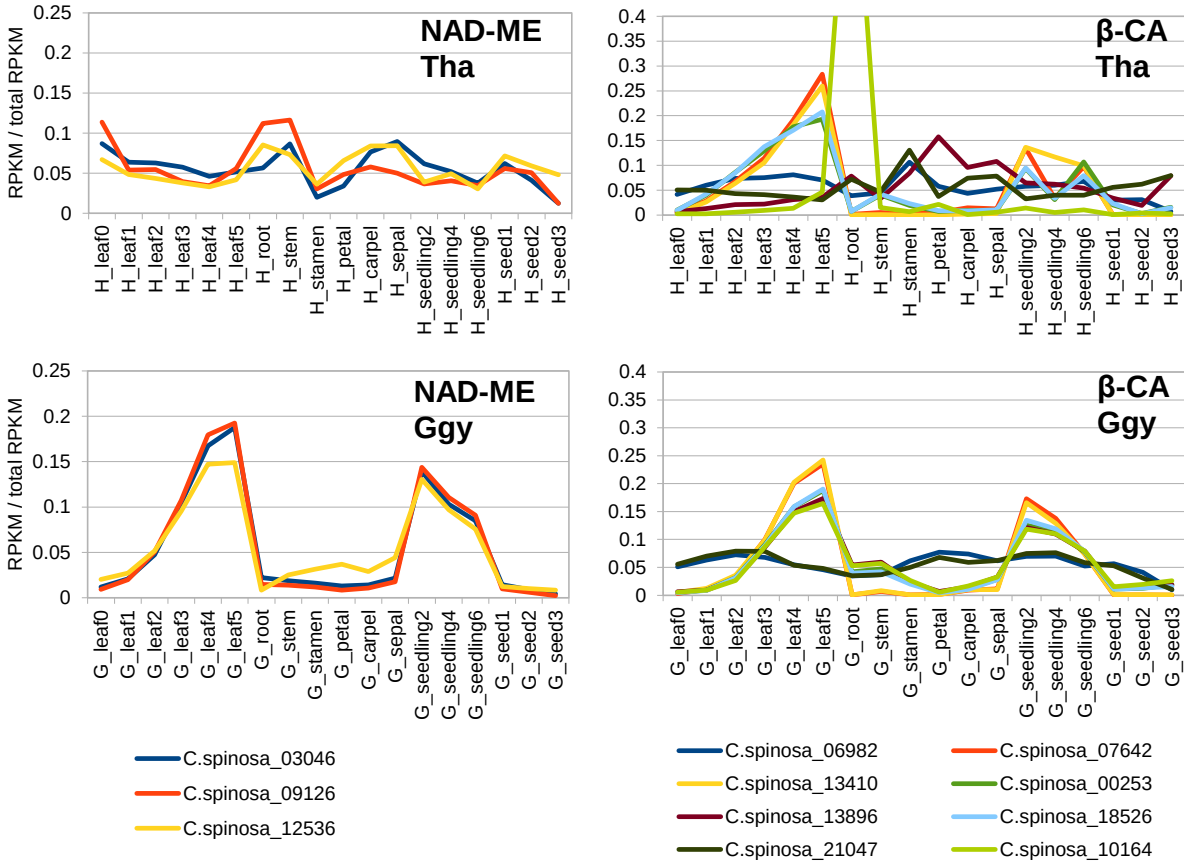


Figure 5. Canalization in expression of NAD malic enzyme (top and bottom left) and β carbonic anhydrase (top and bottom right) homologs in *T. hassleriana* and *G. gynandra*. Top left: NAD-ME expression in *T. hassleriana*. Top right: βCA expression in *T. hassleriana*. Bottom left: NAD-ME expression in *G. gynandra*. Bottom right: βCA expression in *G. gynandra*. (Mapped) gene names and associated colors are displayed, see Materials and Methods for more details on the mapping of *G. gynandra* transcripts to *T. hassleriana* genes. Note that leaf0–leaf5 as well as seedling2–seedling6 and seed1–seed3 are time series of the same organ, with the leaf and seedling gradient being two days separated by stage. Transcription levels in *G. gynandra* (lower graphs) are more strictly regulated across organs, seeds and seedlings. The chaotic patterns in *T. hassleriana* (upper graphs) results in half the genes having a Pearson correlation > 0.95 compared to *G. gynandra*.

(Feodorova et al., 2010). Dating this polyploidy event in terms of absolute age is always a difficult task, however, here we find that the Ks rate of *G. gynandra* is extremely similar if not identical to *T. hassleriana.* Assuming then that mutation rates between these two species are the same, we can reaffirm the previous date estimation of Th-α at 13.7 mya (Barker et al., 2009).

The influence of the Th-α WGD event on photosynthetic gene composition is apparent, both in ortholog number as well as in syntenic region copy number for both species. From absolute orthologs numbers we can see that there is no increased retention between Cleomaceae species and even a slightly lower rate of retention in *G. gynandra.* This indicates that both species have experienced similar evolutionary constraints for a significant amount of time. Also we need to consider that genes sharing a similar sequence, do not necessarily have to share the same function. Even using strict CIP/CALP filtering which has been proved to be an accurate measure for the prediction of true orthologs (Murat et al., 2012), differential expression either in time, localization or regulation can substantially change the function of a gene. This is especially the case for genes in the core C_4 photosynthesis pathway, where many C_3 genes have been recruited into new functions (Gowik et al., 2004; Hibberd and Covshoff, 2010).

When establishing Ks values of deeper ortholog nodes of photosynthesis genes, a large proportion of genes seems to have been retained from the γ duplication. For a trait that is likely to be highly dosage sensitive (Coate et al., 2011), we expect that gene loss will be rare and that remnants from this old paleohexaploidy are still present. However, considering the time that has passed since the γ paleohexaploidy event and on the basis of absolute gene copy numbers some gene loss has taken place predating the transition from C_3 to C_4.

The evolutionary importance of WGD events is made clear from the dominant presence of retained Th-α genes in both Cleomaceae species. However, certain questions remain: Can we couple this importance to the evolution of specific traits or in this case, C_4 photosynthesis? This is an old discussion, dating back to the works of Ohno who was the first to suggest that the massive radiation of vertebrates was caused by a whole genome duplication in the ancestor (Ohno et al., 1968). An earlier study on the evolution of photosynthesis in soybean, showed that the Calvin-Benson-Bassham cycle (CBBC) and the light harvesting complex (LHC) gene families show a greater expansion from single gene duplications than both photosystem groups. This is explained by the increased dosage sensitivity of photosystem genes: if some subunits are expressed differently due to duplications while others are not, this is deleterious for the system as a whole (Coate et al., 2011). This acts as a conservation mechanism for gene copy number that does not affect the more loosely connected enzyme collection of the CBBC and LHC genes.

In *G. gynandra,* where the expression of C_4 genes is tightly linked in clusters we would expect a high retention of orthologs. However, this dependency on transcriptional regulation has not lead to an increased retention of photosynthetic genes, as evidenced by lower copy numbers for all C_4 gene families when compared to *T. hassleriana.* It is not likely that neofunctionalization of genes after polyploidy has played a major role in the shift to C_4 photosynthesis. The much more stringent transcriptional regulation of C4 cycle genes in *G. gynandra* when compared to *T. hassleriana* as evidenced in this study is in accordance with the alternative hypothesis, which states that this process was mainly due to recruitment of existing genes in transcriptional space as suggested by several authors (Brown et al., 2011; Gowik and Westhoff, 2011; Kajala et al., 2012; Williams et al., 2014).

We still have much to learn regarding the development of C_4 photosynthesis. When studying this exceptional trait, we must always consider the genetic history of the species in question. Here, we give evidence that duplications, on a large scale and small, contribute to trait evolution. The exact mechanisms behind the recruitment of these genes into new biochemical pathways however are still largely unknown. Current sequencing efforts for *G. gynandra* will significantly aid in finding the detailed mechanisms of gene and C_4 photosynthesis evolution. The *Cleome* genus provides an excellent model system for unravelling the evolutionary origin and workings of C_4 photosynthesis and hopefully will enable us to harvest the fruits of our knowledge on this remarkable form of plant energy conversion.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Cleome transcriptome sequencing, processing, assembly and quantification was done by CK. AB and APMW provided comments on handling highly expressed duplicates as well as proofreading the manuscript. EvdB performed the bioinformatic analyses. EvdB and MES prepared the manuscript. JMH and XZ proofread and edited the manuscript.

ACKNOWLEDGEMENTS

The work of EvB and MES was funded by NWO Vernieuwingsimpuls Vidi Grant number 864.10.001. APMW acknowledges support by the Deutsche Forschungsgemeinschaft (SPP 1529; EXC 1028).

Table 2. List of *Arabidopsis* genes used as representatives of C4 photosynthesis families. ATG identifiers correspond to identifier following the ATG system from the *Arabidopsis* Information Resource

Gene family	ATG Identifiers
βCA	AT1G23730
	AT1G58180
	AT1G70410
	AT3G01500
	AT4G33580
	AT5G14740
MDH (cytosolic)	AT1G04410
	AT5G43330
	AT5G56720
MDH (mitochondrial)	AT1G53240
	AT2G22780
	AT3G15020
	AT3G47520
	AT5G09660
MDH (peroxisomal)	AT1G53240
	AT2G22780
	AT3G15020
	AT3G47520
	AT5G09660
MDH (plastidic)	AT1G53240
	AT2G22780
	AT3G15020
	AT3G47520
	AT5G09660
NAD-ME	AT2G13560
	AT4G00570
NADP-ME	AT1G79750
	AT2G19900
	AT5G11670
	AT5G25880
PEPC	AT1G21440
	AT1G53310
	AT2G42600
	AT3G14940
PPCK	AT1G08650
	AT3G04530
	AT3G04550
	AT4G37870
	AT5G28500
	AT5G65690

SUPPLEMENTAL DATA

All supplemental figures, tables and data can be accessed through the published version of this article online: http://dx.doi.org/10.1016/j.cpb.2014.08.001 (Last accessed 30th January, 2017).

CHAPTER 3

FLOWER POWER AND THE MUSTARD BOMB: COMPARATIVE ANALYSIS OF GENE AND GENOME DUPLICATIONS IN GLUCOSINOLATE BIOSYNTHETIC PATHWAY EVOLUTION IN CLEOMACEAE AND BRASSICACEAE

Erik van den Bergh[1,2], Johannes A. Hofberger[1,3], and M. Eric Schranz[1]

1 Biosystematics Group, Wageningen University & Research Center, Droevendaalsesteeg 1, 6708PB, Wageningen, the Netherlands;

2 Computational Biology, Earlham Institute, Norwich Research Park, Norwich, NR4 7UH, United Kingdom;

3 Big Data Analytics, Detecon Co., Ltd., 100600 Beijing, PR China

ABSTRACT

Glucosinolates (GS) are a class of plant secondary metabolites that provide defense against herbivores and may play an important role in pollinator attraction. Through coevolution with plant-interacting organisms, glucosinolates have diversified into a variety of chemotypes through gene sub- and neofunctionalization. Polyploidy has been of major importance in the evolutionary history of these gene families and the development of chemically separate GS types. Here we study the effects of polyploidy in Tarenaya hassleriana (Cleomaceae) on the genes underlying GS biosynthesis. We established putative orthologs of all gene families involved in GS biosynthesis through sequence comparison and their duplication method through calculation of synonymous substitution ratios, phylogenetic gene trees, and synteny comparison. We drew expression data from previously published work of the identified genes and compared expression in several tissues. We show that the majority of gene family expansion in T. hassleriana has taken place through the retention of polyploid duplicates, together with tandem and transpositional duplicates. We also show that the large majority (>75%) is actively expressed either globally or in specific tissues. We show that MAM and CYP83 gene families, which are crucial to GS diversification in Brassicaceae, are also recruited into specific tissue expression pathways in Cleomaceae. We conclude that many GS genes have expanded through polyploidy, gene transposition duplication, and tandem duplication in Cleomaceae. Duplicate retention through these mechanisms is similar to A. thaliana, but based on the expression of GS genes, Cleomaceae-specific diversification of GS genes has taken place.

Key words: *Arabidopsis* thaliana; Brassicaceae; Cleomaceae; gene duplication; gene family evolution; glucosinolates; polyploidy; Tarenaya hassleriana

INTRODUCTION

In the arms race between plants and herbivores, plants deploy a diverse arsenal of secondary metabolites. For example, plants in the order Brassicales have specialized glucosinolate (GS) compounds, often referred to as mustard oil “bombs”. Some GSs act as preformed front-line defenses against herbivores, whereas others can be induced and deployed upon herbivory (Halkier and Gershenzon, 2006). Alternatively, volatile breakdown products of GSs can also play a role in attracting pollinators, such as evening visitor bats (Bestmann et al., 1997). Chemically, GSs are sulfur-rich, amino acid-derived secondary metabolites that split into four “flavors”, each depending on the chemical class of amino acid precursors. The first group is derived from aliphatic amino acids such as Ala, Leu, Ile and Val. The second group is referred to as benzenic because its members derive from Phe and Tyr. The third group comprises GSs that are based on Trp. The last group derives from methionine exclusively. All four of these flavors are generated through the same basic biosynthesis pathway, which has three stages: chain elongation (for Phe-derived glucosinolates), core structure formation, and side chain substitution (Sønderby et al., 2010). Chain elongation starts with deamination by a branched-chain amino acid transaminase (BCAT), producing a 2-oxo acid. This molecule then cycles through a three-step process: condensation with acetyl-CoA with methylthioalkylmalate synthases (MAMs), isomerization by isopropylmalate isomerases (IPMIs) and oxidative decarboxylation by isopropylmalate dehydrogenases (IPM-DHs). Once the chain has reached the correct length, it is once again transaminated by BCAT and enters core structure formation. This phase starts off with conversion to aldoximes by the cytochrome P450 family 79 (CYP79) gene family (CYP79B2/3 for Trp, CYP79A2 for Phe and CYP79F1/2 for Met). The aldoximes are then oxidized by cytochrome P450 family 83 (CYP83) (CYP83B1 for Trp, CYP83A1 for all aliphatic derived amino acids) and subsequently converted to thiohydroximates by an S-alkylthiohydroximate lyase encoded by the SUR1 gene. These are S-glycosylated by the uridine diphosphate glycosyltransferase 74 (UGT74) gene family (UGT74B1 for Phe, UGT74C1 for Met). Finally, the intermediates are sulfated by sulfotransferases (SOT16, 17 and 18) to become glucosinolates. In the third and last phase, secondary modifications of the core structure take place. In aliphatic

glucosinolates, S-oxygenation is performed by flavin-monooxygenases FMO-GS-OX1–5. The side-chain can then be converted from 3-butenyl to 2-hydroxy-but-3-enyl glucosinolate by GS-OH. In indolic glucosinolates, CYP81F1 has been identified as the gene that encodes the oxidizing enzyme that converts indolyl-3-methyl glucosinolate (I3M) to 4OH-I3M, 4M-I3M and 1M-I3M. This is not an exhaustive list of secondary modifications, as these "decorations" can vary immensely across nature and they are the main cause of diversity observed across the more than 120 types of GSs that have been described to date (Kliebenstein et al., 2001).

Benzenic and aliphatic GS flavors are found in few non-Brassicales eudicot groups including the Phytolaccaceae, Pittosporaceae, and Euphorbiaceae (Fahey et al., 2001). The Brassicales order (containing, for example, Carica papaya, *Gynandropsis gynandra*, and the model plant *Arabidopsis* thaliana) also possesses both types. During Brassicales evolution, two new flavors evolved. Firstly, the GS of the indolic type (that are, for example, not present in papaya but in most core-Brassicales) and second, Met-derived and chain-elongated types. The latter are particularly abundant and diverse in the Brassicaceae family, but may also occur in less-elongated forms in the Capparaceae (Rodman et al., 1996). Together with Phe-derived GS, they undergo elongation of the amino acid side chain before core structure formation and have been proposed to be derived from the core amino acid chain elongation pathway (Sawada et al., 2009).

From a genetic perspective, it has been hypothesized that one possible key mechanism behind the rapid GS diversification is due to whole-genome and gene duplications and subsequent neofunctionalization (i.e., the development of a new genetic function after gene duplication) (Ohno, 1970; Mitchell-Olds and Schmitt, 2006). The rapid expansion of GS flavors is mostly dependent on the initial side chain elongation and final "decoration", which requires different enzymes for each amino acid derivative. These enzymes are all encoded by genes that are near-identical in sequence and differ only in a small set of mutations. Hence, this system seems to be the result of classic neofunctionalization (Sharma et al., 2014).

There are three main causes for variation in gene copy number in GS gene families: first, through local tandem duplications, second, through gene transpositional duplication (GTD hereafter), and third, through whole genome duplication or ancient polyploidy (Freeling, 2009). The first mechanism is a continuous process whereby gene copies are formed through errors in DNA replication, resulting in a tandem array of duplicate gene copies. An example of a tandem array that underwent neofunctionalization is the S-locus (Cheng et al., 2013), which prevents self-fertilization in Brassicaceae. Furthermore, it has been shown that for several classes of defense-related genes such as L-type lectin receptor kinases (LecRKs) and L-type lectin domain proteins (LLPs) (Hofberger, Nsibo, et al., 2015), terpenoid synthase (TPS) (Hofberger, Ramirez, et al., 2015) and nucleotide-binding–leucine-rich repeat (NB-LRR) genes (Hofberger et al., 2014), tandem duplications have caused a large radiation and "boosted" R-gene diversities at different times during flowering plant radiation.

The second mechanism, GTD, occurs when a nontransposon gene duplicates and changes the new copy is inserted into a new genomic position potentially due to linkage with a transposon-like element (TE) and segregants contain CDS duplicates. In *A. thaliana*, between one- and three-quarters of all protein-coding genes have are been estimated to have been transposed at least once during in the Brassicales evolution (Freeling et al., 2008). Classes that were shown to exhibit an increased rate include type I MADS-box genes, F-box genes and NB-LRR genes across angiosperms (Freeling, 2009; Malacarne et al., 2012; Hofberger et al., 2014).

The third mechanism, ancient polyploidy, creates genes through the duplication of the whole genome, after which all genes are maintained in duplicate in initial generations. Polyploidy is a common occurrence in plants (Jiao et al., 2011), with many successful, well-known and studied neopolyploids well known and studied (Comai, 2005; van de Peer et al., 2009; Jiao et al., 2011). Not surprisingly, there

is widespread and growing evidence for ancient polyploidy events, including one at the origin of seed plants and at the origin of all angiosperms (Jiao et al., 2011). After polyploidy, a process called fractionation takes place in which genes return to single copy. The selection criteria behind this mechanism are not completely understood, but it has clearly been shown that it follows a nonrandom pattern (Baucom et al., 2009). An important factor is summarized as the gene balance hypothesis: Genes that are strongly dosage sensitive will experience less gene loss, because loss of a single gene will cause an imbalance in the pathway, leading to deleterious events (Edger and Pires, 2009; Birchler and Veitia, 2014). Conversely, genes that are dosage independent will experience a higher chance of fractionation throughout their pathway.

The genetic diversification of GS in Brassicaceae has been suggested by several authors to be correlated with the polyploid history of this family; all species that share the ancient paleopolyploidy event named At-γ (122 million years ago [Ma];(Kagale et al., 2014)) have indolic glucosinolates, contrary to the species that do not (Schranz et al., 2011). The At-β event, which occurred approximately 56 Ma (Kagale et al., 2014) and is shared by Brassicaceae, including its earliest split sister group Aethionema arabicum (Haudry et al., 2013) seems to correlate with chain elongation of Met-derived GS (Schranz et al., 2011). Additionally, speciation rate shifts have been observed that are consistent with the estimated age of that polyploid event (Franzke et al., 2011; Hohmann et al., 2015). Detection and rough dating of these polyploid events are mostly dependent on calculation of synonymous substitution rates (Ks). Following the assumption that genes that were duplicated at a similar time share a similar Ks, a large gene birth event such as polyploidy can be detected by comparing a large number of genes belonging to a distinct gene family sharing a similar Ks value. On the basis of this assumption, gene duplicates (ohnologs) can be "dated" to a specific polyploidy event (Blanc et al., 2003).

GS gene comparison within Brassicaceae is well established because of the level and depth of gene annotation in *A. thaliana* and other sequenced Brassicaceae species (Koenig and Weigel, 2015). For example, Hofberger et al. (2013) analyzed gene content with the *Ae. arabicum*, the first branching species in the Brassicaceae. They found that gene families had expanded through polyploidy, tandem duplications, and transpositions to 67 glucosinolate loci of which the large majority exhibited (micro) synteny. It was also shown that more than 95% of genes in *A. thaliana* and *Ae. arabicum* are remnants of these duplications.

However, from within Brassicaceae, it was impossible to study the effects of the At-α duplication completely because of the lack of an appropriate phylogenetic outgroup. This problem has been recognized, and the solution has been suggested to include analysis of species from the closest related outgroup and sister family, the Cleomaceae (Cheng et al., 2013). Cleomaceae species, herbaceous plants with mostly palmately compound leaves, are found primarily in warm temperate, desert, and tropical zones on all continents. In contrast, crown Brassicaceae are especially abundant in colder temperate climates (Stevens, 2001a). Some Cleomaceae are capable of C4 photosynthesis and, because of their phylogenetic relationship to Brassicaceae, they have been suggested as a new and emerging framework for analysis of gene and genomic structural and functional evolution (Marshall et al., 2007; Marquard and Steinback, 2009; van den Bergh et al., 2014). Moreover, Cleomaceae do not share the At-α duplication event of Brassicaceae, but have undergone a different, lineage-specific polyploidy event. This genome triplication event is termed Th-α and occurred approximately 24–13 Ma (Barker et al., 2009; van den Bergh et al., 2014). Because of the recent Th-α and At-α events in both families, these polyploidy events form an ideal comparison to study the effects of ancient genome multiplication. To date, research has been focused on Tarenaya hassleriana (formerly known as Cleome spinosa), which has recently been sequenced, together with the C4 plant *Gynandropsis gynandra*, for which sequencing efforts are currently in progress (M. E. Schranz et al., in preparation).

Compared with Brassicaceae, Cleomaceae have chemically lower degrees of GS profile variation, but do show a different response to herbivory in terms of released GS chemotypes (Riach et al., 2015). In Cleomaceae, several GS types are predominantly found: Capparaceae-specific capparin and [gluco]sinalbin which are alcoholic and aromatic derived, respectively; in the indolic-derived class, glucobrassicin and neoglucobrassicin are mostly present (Fahey et al., 2001). Lastly, cleomin is an aliphatic Cleomaceae-specific glucosinolate that is alcoholic derived and not present in Brassicaceae (Ahmed et al., 1972; Ajaiyeoba, 2000).

The release of separate chemotypes of GS in response to herbivory is consistent with the concept of plant–herbivore or –pollinator coevolution (Edger et al., 2015). According to this hypothesis, gain-of-function mutations in genes encoding for herbivore enzymes lead to the ability to neutralize novel flavors of GS, to the point where Pieridae herbivores developed the ability to lay eggs exclusively on Brassicaceae species with a distinct GS chemotype. While there has been a great focus on the specialization of Pieridae on Brassicaceae species, in fact, there are even more clades of Pieridae detected on Cleomaceae and Capparaceae (Edger et al., 2015). Furthermore, GS compounds may convey a key function in bat-pollination, which could have driven the diversification of GS chemotypes in Cleomaceae species in a way that is completely different from Brassicaceae (Bestmann et al., 1997; Johnson et al., 2009). Bats are known to be attracted to sulfur-containing compounds (von Helversen et al., 2000) and glucosinolates specifically (Bestmann et al., 1997).

In this study, we have compared expansion and contraction of key GS gene families in *T. hassleriana* (Cleomaceae) relative to *A. thaliana* (Brassicaceae). It is not unreasonable to assume a separate evolutionary history due to the specific ecological circumstances that played a role in the GS biosynthetic pathway development within the Brassicaceae. We have also examined all the aforementioned forms of gene family expansion and diversity causes: polyploidy, tandem duplication, and GTD.

Establishing orthologous relationships is only the first step toward unraveling a complete picture of GS evolution and its causes and consequences. Therefore, we used the Cleomaceae gene expression atlas (Külahoglu et al., 2014) to test the biological activity of putative ortholog, and we examined tissue specificity of ortholog expression to establish gene localization. In summary, we aim to create a detailed catalog of GS orthologs in *T. hassleriana* to create a robust framework for the study of these fascinating compounds within Cleomaceae, in general and the *T. hassleriana* species in particular, with a focus on modes of gene duplication between Brassicaceae and Cleomaceae.

MATERIALS AND METHODS

ORTHOLOG IDENTIFICATION

We extracted glucosinolate genes and associated AT gene identifiers from (Sønderby et al., 2010). Coding sequences from the TAIR 10 genome release (Lamesch et al., 2012) were used as a query in the CoGe Synfind tool (Lyons and Freeling, 2008) with settings as follows: comparison algorithm LAST (Kiełbasa et al., 2011), gene window size of 40, minimum number of genes of 4; collinear scoring function, syntenic depth unlimited. The genes were queried against *A. thaliana* itself and Tarenaya hassleriana with the Weber annotation (v5) (Cheng et al., 2013) as published in CoGe at the time of writing.

TANDEM GENE ANALYSIS

We performed a nucleotide BLAST search (Altschul et al., 1990; Camacho et al., 2009) with an e-value cutoff 1e-50 of coding sequences of *A. thaliana* TAIR10 and *T. hassleriana* v5. The associated GFF files from CoGe were converted to BED files using GNU sed and awk. These files were used with the Quota

align software package (Tang et al., 2011) to create tab delimited files containing tandem duplicates for *A. thaliana* and *T. hassleriana*.

GTD IDENTIFICATION

A reciprocal best blast hit (RBBH) method with tandem filtering was used to identify nonsyntenic orthologs. RBBH often outperforms more complex algorithms for ortholog identification (Altenhoff and Dessimoz, 2009). We performed a protein BLAST with evalue cutoff of 1e-50 and a minimum identity of 50% between the protein sets of *A. thaliana* (TAIR10) and *T. hassleriana* (v5). If the top hit was part of a gene TAR (as identified above), it was ignored. On some occasions, a cross gene RBH was observed (in a gene family of 3 or more, all combinations are the top hits of each other but not direct RBBH). In this case, a gene was also identified as part of the ortholog family.

SYNONYMOUS SUBSTITUTION WITHIN GENES

Using an in-house developed Perl script, all pairwise nonrepetitive permutations of gene families were created. The coding sequences of these pairs were used as queries in the Ks app of the JCVI bioinformatics library (Haibao Tang et al., 2015). This library uses the programs Pal2nal (Suyama et al., 2006), MUSCLE (Edgar, 2004), and PAML4 (Yang, 2007 p. 4) to align sequences, convert the alignment to protein, and then calculate the synonymous substitution (Ks) using the Yang and Nielsen (Yang and Nielsen, 1998) and Nei and Gojobori (1986) (Nei and Gojobori, 1986) models.

Additionally, to normalize altered mutation rate in specific coding genes, Ks values for syntenic genes were calculated as the average of surrounding genes in its harboring syntenic block. We extracted syntenic blocks in *T. hassleriana* and *A. thaliana* using CoGe SynMap (Lyons and Freeling, 2008), with settings as follows: alignment algorithm LAST, relative gene order, maximum distance 30, minimum number of aligned pairs 5, merge syntenic blocks with Quota align, Quota align dm 15, synonymous substitution rate calculation without log transform, tandem duplication distance 15, C-score 0.1. Genes within blocks were then taken to have the average Ks of all genes within that block.

PHYLOGENETIC GENE ANALYSIS

Genes were grouped into their gene families: CYP79, CYP83, GSTU, GSTF, GGP, SUR, UGT, SOT, CYP81, and MYB similar to (Edger et al., 2015). Carica papaya glucosinolate orthologs were identified using SynFind (as described under “Ortholog identification”). Papaya gene identifiers (Ming et al., 2008) were lifted over from the 0.3 to the 0.5 edition on CoGe so that the glucosinolate ancestor genes could be used. Genes were aligned with MAFFT (Katoh et al., 2002) using standard settings, and alignments were used to create a neighbor- joining tree out of conserved sites.

EXPRESSION ANALYSIS

Cleomaceae expression data were procured from Külahoglu et al. (2014). Found orthologs were mapped to the unigenes in the atlas through BLASTN (Altschul et al., 1990; Camacho et al., 2009) with default settings, using the top hit. Average reads per mappable million (RPKM) were calculated for mapped *T. hassleriana* glucosinolate orthologs. Gene expression profile clustering was performed using MultiExperiment Viewer (Saeed et al., 2003) through the hierarchical clustering algorithm (Eisen et al., 1998). Clustering was performed using Pearson correlation, which is standard for hierarchical clustering; a complete linkage clustering algorithm was used, which performs best with equally sized clusters as is expected with this data.

RESULTS

IDENTIFIED ORTHOLOGS BETWEEN A. THALIANA AND T. HASSLERIANA

The first step in identifying glucosinolate genes in *T. hassleriana* is the establishment of syntenic genes when compared with *A. thaliana*. Glucosinolate genes were separated into four classes: aliphatic, indolic/benzenic, cosubstrate, and transcription factor regulation. These groups encompass 65 genes in

total, representing the core glucosinolate pathway as it is known today in *A. thaliana*. Using these genes as queries, we identified 66 genes in *T. hassleriana* as syntenic orthologs (Appendix S1, see Supplemental Data with the online version of this article). In both of these groups, tandem duplications were not counted and will be discussed in detail below. Portions of genome that were syntenic to *A. thaliana* but lacked the query gene (listed in CoGe as "proxy for region"), which suggest transpositions or gene loss events were not initially considered (transposed genes that could be located in other genomic contexts are discussed below).

Classification of genes through orthology can be confounded by tandem duplicates that have separate functions, a situation that is common in the glucosinolate pathway. It must therefore be assumed that when a gene to gene orthology relationship is established any of the genes in the tandem array on both sides can be the "functional" ortholog. Confounded duplicates are the cause of GSTF being classified as having separate genes in the aliphatic and indolic pathways in *A. thaliana*, but the same genes in these pathways in *T. hassleriana*.

DATING OF T. HASSLERIANA GSL GENE BIRTHS

To establish the full duplication history of the identified GS genes, two methods were used: Synonymous substitution (Ks) analysis and gene tree comparisons. As explained in the introduction, genes can be assigned to a specific polyploidy event by matching their Ks with known Ks estimates for these events. For *A. thaliana*, the At-α window is estimated to be between 0.55 and 1. In *T. hassleriana* Th-α is slightly earlier with a Ks window between 0.25 and 0.6. The β and γ duplications are shared between these species and thus fall in the same Ks windows of 1.5–2.5 and 2.5–8, respectively.

Ks analysis places seven duplications in the γ Ks window, nine duplications in the β Ks window, and 13 triplications in the Th-α Ks window. Three triplicated genes from Th-α have retained all three of the expected paralogs (GGP, UGT74, and CYP79); the other 10 genes have retained two copies (and thus have lost one paralog). In total, 31 genes are a result of polyploid gene retention, which is 54.4% of all genes involved in the indolic and aliphatic pathways, which is a similar duplicate retention rate when compared with *A. thaliana*: 28 of 51 (49.1%) genes in the indolic and aliphatic pathways are retained in Brassicaceae from polyploidy (Hofberger et al., 2014). A different picture forms when put into the global duplication retention rate: in *A. thaliana* the overall retention is only 13.6%, whereas in *T. hassleriana* it is 50.0%. Thus, the retention rate in GS is only slightly higher than the overall retention rate; however, it must be considered that Th-α was a hexaploidy leading to higher chances of duplicates being retained. Furthermore, the genome triplication is much younger, leaving less time for fractionation to occur (Cheng et al., 2013).

To get a more detailed picture of gene duplication histories, genes were aligned and compared with those of Carica papaya, an early-branching member of the Brassicales, as the outgroup. This species does not share the β- and α-event with Cleomaceae and Brassicaceae and should thus have a single gene per gene family (Edger et al., 2015). Using this data, we established a gene duplication history of all major families in the glucosinolate pathway (Appendix S2, see online Supplemental Data). In CYP79, GGP and UGT full Th-α triplets have been conserved. Additional groups that have retained genes from the Th-α triplication are IPMI, IPMDH, BCAT, CYP83, GSTF, and FMO.

GLUCOSINOLATE ORTHOLOG FAMILY EXPANSION AND CONTRACTION

The identification of orthologs together with their duplication age allows us to give a complete overview of the evolution of these gene families, which is described next.

COMMON PATHWAY

Both the indolic/benzenic and aliphatic pathways share genes up to CYP79, and this common section will be discussed first. Tarenaya hassleriana has three methylthioalkylmalate synthases (MAM)

orthologs (Fig. 1B) compared with the single TAR of *A. thaliana* (Fig. 1A, lower half). The Th2v24105 TAR corresponds to that TAR in *A. thaliana* and is discussed in further detail in the tandem duplicate section. The isopropylmalate isomerase (IPMI) enzyme family, an additional β duplication retained gene can has caused this family to consist of four genes in *T. hassleriana* (Fig. 1B, lower half). Next in the pathway is the isopropylmalate dehydrogenase (IPMDH), which has experienced many transpositions in *A. thaliana* (Fig. 1A, lower half), but is more "well behaved" in *T. hassleriana*, where its two copies are syntenic to one another (Fig. 1B, lower half). In both species, IPMDH are remnants of their lineage-specific α-events, but it must be considered that the Th-α event was a paleohexaploidy, whereas At-α was a paleotetraploidy. Thus, the third Th-α copy of this gene was lost in *T. hassleriana*. The last common gene is BCAT3, which has double the copy number in *T. hassleriana*. Remarkably, polyploid expansion has been the same in both species; the copy number increase in *T. hassleriana* is due to two transpositions of the Th2v26622 gene.

ALIPHATIC PATHWAY

In the aliphatic specific part of the pathway, CYPY79 is a key variable gene in Brassicaceae that has been implied as a driving force in glucosinolate diversification Indeed, in *A. thaliana*, seven genes code for this enzyme, of which three are in the aliphatic pathway (Fig. 1A, third quarter). In *T. hassleriana*, a similar picture emerges when viewed in terms of copy number. However, this is an excellent example of why gene duplication history can provide a much more rich history and detail. Indeed, in copy number, both gene families are the same between species, but their duplication history is different. In *T. hassleriana*, all three genes are retained from the most recent Th-α duplication (Fig. 1B, upper half), in contrast to *A. thaliana* in which the triplet is a remnant from the two rounds of polyploidy (At-α and β) (Fig. 1A, third quarter). CYP83 by contrast shows much less variation, with only an extra GTD in *T. hassleriana* expanding the copy number to three (Fig. 1B, third quarter). Next in the pathway, GSTF has benefited from the At-α duplication in *A. thaliana* with its four gene copies both being α pairs (Fig. 1A, third quarter). In *T. hassleriana*, one α pair is present together with a GTD dated before the β polyploidy

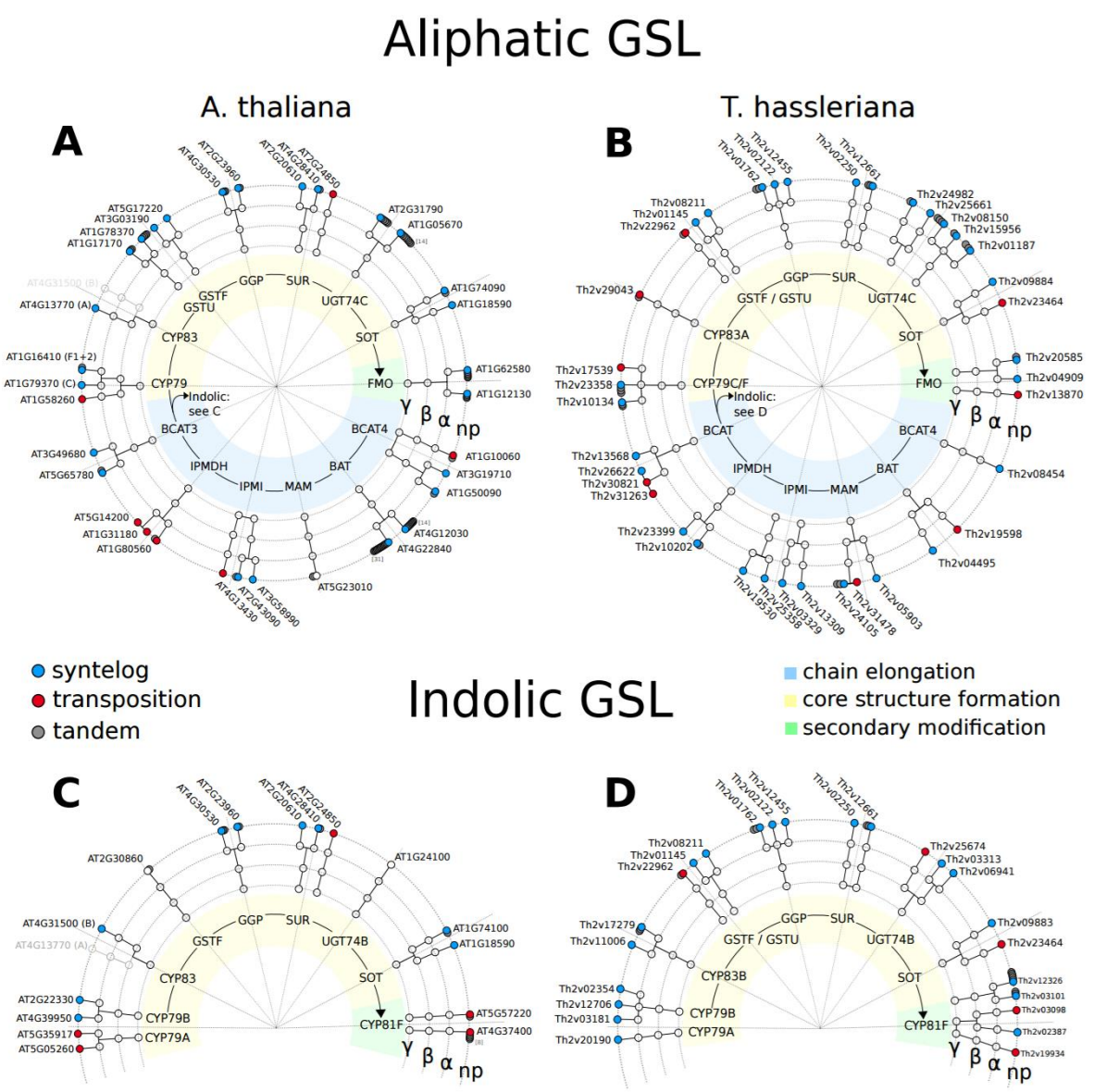


Figure 1. Radiation of glucosinolate (GS) gene families in (A, C) *Arabidopsis thaliana* and (B, D) *Tarenaya hassleriana*. Syntelogs are represented as a blue circle, transposition duplications as a red circle; tandem arrays are represented as a stack of gray circles with the amount of circles representing the size of the array. Retained duplicate genes have been marked on circles corresponding to their age: The innermost circle represents the γ polyploidy event, the inner middle circle the β event, the outer middle circle the (A, C) At-α event or the (B, D) Th-α event and the outermost circle represents more recent and tandem duplicates. (A) Aliphatic glucosinolate (GS) pathway in *A. thaliana*. Size of large tandem arrays is represented in grey square brackets. In CYP79, CYP79C and CYP79F are specified after the gene names in brackets; CYP83A and CYP83B are marked similarly. (B) Aliphatic GS pathway in *T. hassleriana*. (C) Indolic GS pathway in *A. thaliana*. Because this pathway shares the first six genes with the indolic pathway, these genes are not shown here. GSTF and UGT74B, both single copy genes, are represented with a white circle. (D) Indolic GS pathway in *T. hassleriana*. Because this pathway shares the first six genes with the indolic pathway these genes are not shown here.

(Fig. 1C, third quarter). In GGP, both organisms have retained all α copies, with the hexaploid retained triplet in *T. hassleriana* and the tetraploid pair in *A. thaliana* (Fig. 1A and B, third quarters).

Continuing down the pathway, SUR has a different duplication history in both organisms. Both SUR copies are derived from the oldest γ polyploidy event in *T. hassleriana* (Fig. 1B, last quarter), whereas in *A. thaliana* only one of these γ genes was retained and then additionally duplicated by At-α (Fig. 1A, last

quarter). The largest family in the aliphatic pathway for *T. hassleriana* is UGT74 with five genes (Fig. 1B, last quarter). While the family in *A. thaliana* consists of a single α pair, both the β and the Th-α event have caused a large radiation of genes in *T. hassleriana*. This variety in gene history stands in contrast to the next gene family, SOT, where in both species the whole family consists of one α pair (Fig. 1A and B, last quarters). Finally, the FMO family has a similar history to SOT with a single α pair in both species, with the addition of a single GTD in *T. hassleriana*.

INDOLIC/BENZENIC PATHWAY

Some genes in the indolic/benzenic pathway are shared with the aliphatic pathway, so we will only discuss the genes that are not: CYP79, GSTF, UGT74, and CYP81.

Starting off with CYP79, we immediately notice a corresponding copy number in both species, but with a different history. A Th-α triplet coupled with a single, pre-γ singleton makes up the four-gene family in *T. hassleriana* (Fig. 1D, third quarter); in *A. thaliana* this family is formed by two At-α pairs (Fig. 1C, third quarter). The next indolic/benzenic specific gene family, GSTF shows a remarkable difference in copy number between the two species: only a single gene is part of this pathway in *A. thaliana* (Fig. 1C, third quarter) and a Th-α pair, together with a single GTD form a three-gene family in *T. hassleriana* (Fig. 1D, third quarter). The *T. hassleriana*-specific expansion of UGT74C in the aliphatic pathway can also be seen in UGT74B in the indolic/benzenic pathway with three copies standing in contrast to a single gene in *A. thaliana* (Fig. 1C and D, last quarters). Again, a Th-α pair with a GTD makes up the duplication history of this family. Finally, the gene responsible for "decoration", CYP81 shows extensive radiation in *T. hassleriana* resulting in four total gene copies, which stands in contrast to the two gene copies in *A. thaliana* (Fig. 1C and D, last quarters).

TRANSPOSITIONS IN T. HASSLERIANA

Using the core 65-gene set of *A. thaliana* as queries, we identified 17 additional orthologs in *T. hassleriana* using protein similarity. This results in a total glucosinolate ortholog set of 84 genes in *T. hassleriana*, which are divided into the four groups previously mentioned as follows: 17.8% in TF regulation, 16.7% in cosubstrate, and 65.5% in aliphatic + indolic (combined). Transpositional duplicates are marked as red circles in Fig. 1A–D.

TANDEM GENE DUPLICATIONS

Tandem duplications have played an important role in glucosinolate diversification in Brassicaceae, a clear example of which is the FMO-GSOX1-4 cluster of genes, which are all located in a tandem array (TAR) on *A. thaliana* chromosome 1. We therefore identified tandem arrays around the orthologs found in *T. hassleriana*. In the whole genome, 6344 genes were organized in a TAR (20.1% of CDS), and 34 putative glucosinolate orthologs were part of a tandem array (52% of all orthologs), with all glucosinolate TARs having an average size of 2.86 genes. In contrast, in *A. thaliana*, the average TAR size for glucosinolate genes is 5.5.

Interestingly, the MAM TAR, which is functional in most Brassicaceae with MAM2 being pseudogenized in the Col-0 ecotype, seems to be present in *T. hassleriana* as well. This TAR, centered on gene Th2v24105, corresponds to the one in *A. thaliana* (online Appendix S3, panel A). One of the *T. hassleriana* genes has undergone an inversion when compared with *A. thaliana* (Appendix S3, panel A). Interestingly, this TAR has undergone additional expansion in *Ae. arabicum* (Hofberger et al., 2013), which corresponds to the same region in *T. hassleriana* (Appendix S3, panel B). This specific MAM cluster has been linked to an insect-resistance QTL rooted in glucosinolate composition, and this locus has been shown to contribute to the ecotypic variation of GS profiles in *A. thaliana*. However, no Met-derived glucosinolates have been described in Cleomaceae, so the exact function of this specific gene cluster remains unknown. Expression-wise, one MAM gene in this TAR seems to have subfunctionalized to be

expressed in the young leaf stages of days 0–4 (Fig. 2, Th2v24111), whereas the other two (Th2v24110 and Th2v24105) are expressed in the sepal.

The largest TAR is formed by the five-gene cluster around the CYP81 ortholog, Th2v12326. CYP81F2 has been suggested to have neofunctionalized toward plant innate immunity and subsequently been retained in *A. thaliana*, but lost however in the ancestral Brassicaceae species *Ae. arabicum*. The large expansion in *T. hassleriana* through tandem duplication is in line with neofunctionalization in *A. thaliana* as previously mentioned and expansion in Eutrema salsugineum where it has been suggested to have been recruited in the biosynthesis of Eutrema phytoalexins. Both Th2v12326 and Th2v12322 are expressed in the developing seed, but the others in this TAR are expressed in the root (Th2v12324, Th2v12327) or very little at all (Th2v12323).

ORTHOLOG GENE EXPRESSION

Ortholog identification provides the history and expansion/contraction of a gene family, but is not necessarily representative of current gene function. We therefore used the expression atlas available for *T. hassleriana* to analyze which orthologs are actively expressed.

Including tandems, 84 orthologs were present in the expression atlas; three were not found due to annotation differences between version 5 of the *T. hassleriana* genome and the version used during construction of the expression atlas. Expression data are divided into nine sets: leaf senescence time series (5 time points), root, stem, stamen, petal, carpel, sepal, seedling development (3 time points) and a seed development (3 time points) shown as H_leaf0-5, H_root, H_stem, H_stamen_H_petal, H_carpel, H_sepal, H_seedling2-6, and H_seed1-3, respectively, in Fig. 2. Continuous expression in all tissues except in the root and stem can be seen for a number of genes in the BCAT group (Fig. 2, 5th cluster). Interestingly, all BCAT orthologs specifically seem to be expressed throughout all tissues and time series. This class of enzymes is responsible for initiation of the chain elongation pathway but is also responsible for the synthesis of Val, Leu, and Ile as essential plant amino acids. In *A. thaliana*, 2 BCATs are part of the glucosinolate pathway (BCAT3 and BCAT4) in which BCAT4 is the initial step of the entire glucosinolate pathway (Schuster et al., 2006; Knill et al., 2008). A similar situation can be seen for IPMI, four members of which are continuously expressed in the three time series, petals, carpels, and sepals. Two IPMI members (Th2v25358 and Th2v10203), however, seem to be expressed strongly and almost exclusively in the stamen. Th2v10203 is a tandem duplicate of the IPMDH1 ortholog, whereas Th2v25358 is a syntelog of the IPMI LSU1 ortholog, both of which are part of the Met chain elongation cycle.

Remarkably, there is a very low expression for many CYP79F and CYP79C orthologs (Fig. 2; online Appendix S4). These enzymes are specific for the aliphatic pathway, and despite their abundance in copy number, an RPKM of >20 could be found in only 2.5% of tissue measurements compared with an average of 36% in all measurements. In *A. thaliana*, CYP79F has been confirmed to be active as the catalyst for short- and long-branch aliphatic amino acids, but the function of the CYP79C class of genes is currently unknown. Low expression levels in *A. thaliana* suggest that it could be responsible for low-abundance

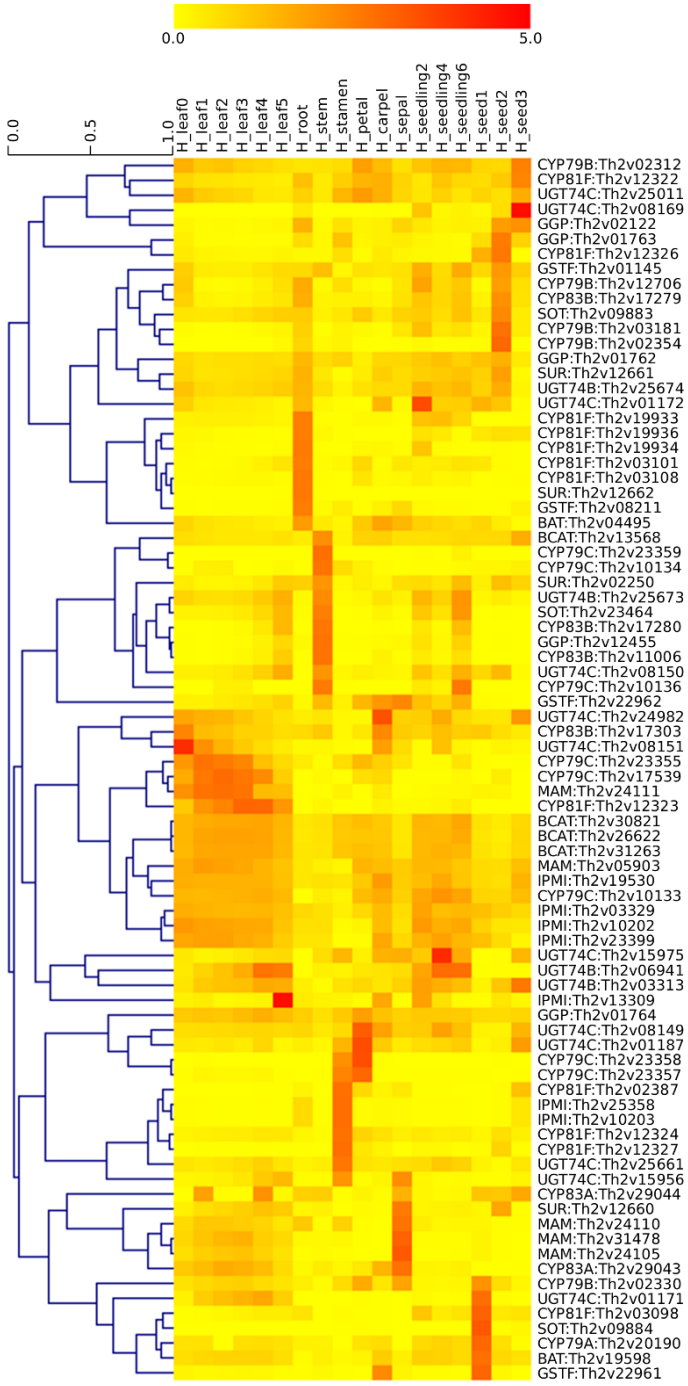


Fig. 2. Expression profiles of glucosinolate (GS) genes in T hassleriana. Tissues samples were a leaf senescence time series (five time points), root, stem, stamen, petal, carpel, sepal, seedling development (three time points), and a seed development (three time points) shown as H_leaf0-5, H_root, H_stem, H_stamen_H_petal, H_carpel, H_sepal, H_seedling2-6, and H_seed1-3 respectively. Expression values are given as RPKM values divided by the root mean square of an entire row from 0 (yellow) to 5 (red). Gene names are marked on the right as the gene family they belong to and the gene identifier separated by a colon. On the left, the hierarchical clustering tree is shown.

GS (Chen et al., 2003), which in light of the data presented here seems to be the case in *T. hassleriana* as well. Genes in the benzenic- and indolic-specific CYP79A and CYP79B group also show low expression levels, with only 11% having an absolute expression of 20 RPKM or higher. In *A. thaliana*, CYP79A is part of the Phe-derived pathway and CYP79B is part of the Trp-derived pathway.

The majority (9/13) of CYP81F orthologs are expressed strongly in the roots (Fig. 2, third cluster). CYP81F acts in the final step of the indolic glucosinolate pathways with a wide array of natural variation among *A. thaliana* ecotypes (Pfalz et al., 2009), catalyzing the step from I3M to 4MO-I3M. Interestingly, root tissues of glucosinolate-producing plants are dominated by 1 and 4MO-I3M and show only 23% I3M content in contrast to aboveground structures, where I3M is 60% of detected GS (van Dam et al., 2008).

DISCUSSION

Here, we clearly showed that the Th-α event has had significant impact on glucosinolate gene families in *T. hassleriana*. A large majority has expanded through the retention of two or all three (in three cases) paralogs that could theoretically be retained after a hexaploidy event. Interestingly though, the Cleomaceae are thought to have a fairly limited diversity in GS chemotypes with cleomin and capparin being the dominant types (Fahey et al., 2001). Through dating methods using synonymous substitution values, we have established that expanded gene pairs and triplets are retained from this recent Th-α polyploidy event. We further confirmed this by phylogenetic comparison with *A. thaliana* and C. papaya homologs. These findings are in line with expectations based on other species. In *Ae. arabicum*, an early diverging Brassicaceae species, similar retentions occurred after the (Brassicaceae-wide) At-α event. In that species, 97% of glucosinolate genes were retained duplicates, which is a figure similar to the case in *A. thaliana* (95% retained duplicates) (Hofberger et al., 2013). Here we find that 89% of *T. hassleriana* putative glucosinolate orthologs are remnants of polyploid duplication either from the most recent Th-α event or the older β- and γ-events. In addition, we show that GTD and tandem duplication have further shaped the variation of enzymes in the glucosinolate biosynthetic pathway, though the average size of TARs is not as great as in *A. thaliana*. Relatively, 52% of putative glucosinolate orthologs in *T. hassleriana* is part of a TAR, a percentage similar to the proportion in *A. thaliana* (45%) and *Ae. arabicum* (46%) (Hofberger et al., 2014).

Through expression analysis, we show that CYP81F is strongly expressed in roots, as has been observed in other species (van Dam et al., 2008) and that several putative orthologs of CYP79C genes have extremely low expression rates in all species, something that is also observed for this class of genes in *A. thaliana* (Chen et al., 2003). However, from localized expression alone, we cannot determine the exact role, if any, of a putative ortholog in the glucosinolate pathway in *T. hassleriana*. Our established set of putative orthologs is a major first step to pave the way to future experimental verification of function.

In Brassicaceae, the genes that have undergone large diversification and that are responsible for a large part of the GS chemotypes observed encode the MAM transcription factor and the CYP81 side-chain modification enzymes. In *T. hassleriana*, we see diversification of expression of the latter genes within tandem duplicates. Their expression seems specific to roots and stamens. The role of glucosinolates in below-ground defenses is discussed in more detail below. In contrast, MAM expression seems specific to the sepals. Enhanced GS content in sepals could be linked to pollinator attraction and/or bud protection especially considering the role of MAM in side-chain variation (Heidel et al., 2006). Profiling of Cleomaceae GS chemotypes would help further classify the role of MAM in this process.

Glucosinolate evolution is intricately linked to the evolution of herbivores and plant pathogens that they protect against. Not only do the “flavors” of GS vary from species to species and even from ecotype to ecotype, but genes from this pathway can even be converted into different plant defensive pathways entirely. In this study, we show that there is much variation in putative glucosinolate orthologs showing that diversifying selection must have taken place in the evolutionary history of *T. hassleriana*. Which ecological factors specifically have played a role in this, however, remains to be unraveled. Pollination by bats and moths has been shown to take place in Cleomaceae (Holloway, 1989; Fleming et al., 2009; Landry and Hebert, 2013), and both of these animals are known to use sulfur-containing volatile classes,

including specific GSs for flower identification (Bestmann et al., 1997; Johnson et al., 2009; Sun et al., 2010). Defense against herbivores has definitely played a role in the evolutionary history of *T. hassleriana* due to the herbivory on this species and the dependence of these herbivores on GS (Catling and Brownell, 1997; Renwick and Lopez, 1999).

As is clear from this study, polyploidy in addition to other forms of gene birth has contributed significantly to the genetic makeup of glucosinolate homologs. On a genetic level, the retention of duplicates can be an effect of gene dosage sensitivity, e.g., the "gene balance" hypothesis in action. Looking at several other species in which this is a confirmed and important phenomenon in gene family diversification (Freeling and Thomas, 2006; Coate et al., 2011; Birchler and Veitia, 2014), we can assume that this effect has also played a role in the development of glucosinolate gene families in *T. hassleriana*.

Furthermore, looking from an ecological perspective many Pieris species as well as other moths take to Cleomaceae and *T. hassleriana* specifically as a host plant. The identifying role that GS play in this families' oviposition strategy must mean that the underlying genetic changes and subsequent variation in glucosinolate expression profiles have played a role in the genetic development of *T. hassleriana*. However, as *T. hassleriana* is a quite recently introduced botanical cultivar in many habitats, it is unclear whether this relationship with herbivores has been influential enough in its original geographical origin to have played a major role similar to the coevolution observed in Brassicaceae.

The role of GS in root protection against microorganisms and nematodes must not be underestimated. Due to the differences in distribution of these pathogens in aboveground vs. belowground situations, then glucosinolate composition and profiles must differ in the roots compared with the leaves and stem (van Dam et al., 2008). Differential induction of GS in roots vs. shoots support this supposition (Matsuura et al., 2012). The observation in this study that CYP81F is predominantly expressed in roots suggests that the development of this gene family has been affected by different selection mechanisms (e.g., pressure from root pathogens and herbivores (Lazzeri et al., 1998) than gene families whose resulting compounds are expressed in leaves and shoots.

The complexity of glucosinolate compounds and the evolutionary history underlying it continues to fascinate and mystify. With this present study, the first steps have been taken toward the functional identification of genes that are part of the glucosinolate biosynthesis pathway for the Cleomaceae. We have also laid a foundation for the evolutionary study of this plant defense mechanism in *T. hassleriana* and Cleomaceae in general and contributed toward the study of the evolutionary forces behind the diversification of this plant chemical weaponry in all glucosinolate-producing species. We are confident that the putative orthologs defined here form a significant functional part of the glucosinolate biosynthesis pathway in *T. hassleriana* that will aid future research on GS and their evolutionary origins. Further unraveling the history behind the proliferation of the mustard oils will provide many answers that will allow us to dive into the trenches of the plant–herbivore arms race and learn to "love the bomb".

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their valuable comments. E.v.d.B. and M.E.S. are funded by NWO Vernieuwingsimpuls Vidi Grant number 864.10.001.

SUPPLEMENTAL DATA

All supplemental figures, tables and data can be accessed through the published version of this article online: http://dx.doi.org/10.3732/ajb.1500445 (Last accessed 30th January, 2017).

CHAPTER 4

ANTHOCYANINS AND FLOWER COLOUR IN CLEOME, IDENTIFICATION OF GENETIC VARIATION UNDERLYING FLORAL COLOURING PATTERNS

E. van den Bergh[1], P. de Oliveira Monteiro[2], S.J. van de Kerke[3], F. Becker[3], R.C.H. de Vos[2] and M. E. Schranz[3]

1 Computational Biology, Earlham Institute, Norwich Research Park, Norwich, NR4 7UH, United Kingdom

2 Bioscience, Plant Research International, Wageningen University & Research Center, Droevendaalsesteeg 1, 6708PB, Wageningen, the Netherlands

3 Biosystematics Group, Wageningen University & Research Center, Droevendaalsesteeg 1, 6708PB, Wageningen, the Netherlands

ABSTRACT

Flower colour is a trait that has an important biological function by guiding pollinator behaviour. Anthocyanins form the majority of the floral pigments responsible for this trait. Variation in the composition of various anthocyanins in the flower and downstream modification of these pigments result in the final colour phenotype. Flower colour variation is common even between individuals of the same species. Here, we examine *Tarenaya hassleriana*, a member of the Cleomaceae, which exhibits two distinct flower colours: purple and pink. We analyse metabolites in the flower by liquid chromatography–photodiode array detector–mass spectrometry (LC-PDA-MS) and find that pink flowers have a distinct anthocyanin profile containing predominantly pelargonidin complexes, whereas purple flowers mostly contain cyanidin complexes. We review the pathway known to lead to these compounds and map the genes in this pathway to their orthologs in *T. hassleriana*. We then create a SNP map on the *T. hassleriana* reference genome using an F2 cross between a pink and purple parent. We find 2 significant binary trait loci using these SNP markers, one of which lies extremely close to Th2v13860: a candidate gene involved in the conversion of a pelargonidin precursor to a cyanidin precursor. We conclude that changes in either sequence structure of this gene or changes in transcription factors leading to a changed expression of this locus must be responsible for the determination of flower colour in *T. hassleriana*.

INTRODUCTION

Flower colour is a distinctive plant trait responsible for the aesthetically pleasing nature of a blooming garden, field of poppies or romantic bouquet. Biologically, it is a defining feature for pollinator attraction (Hannan, 1981; Gigord et al., 2001) and can guide pollinator behaviour leading to improved reproductive fitness (Weiss, 1991; Schiestl and Johnson, 2013).

Flower colour variation is widespread amongst angiosperm species; more than 38% of orders exhibit floral colour variation (Weiss and Lamont, 1997) and even more variation has been suggested to be present when the UV spectrum is also considered (Ohashi et al., 2015). Having species unique floral signalling patterns in a mixed population will be the basis of genetic variation and selection as it prevents incompatible pollen deposition and encourages stable visitation by pollinators, leading to fitness advantages (Schiestl and Johnson, 2013). The key to the success of a colour variant is actual or perceived reward (mimicry). Convergence of specific colours and shapes is assumed to be a consequence of innate sensory preferences of pollinators. However, learned behaviour of pollinators can play an important role in the establishment of flower colour rewarding uniqueness as opposed to the conserved sensory preferences of pollinators. In a study of monkeyflower (*Mimulus*), a widely used model for flower colour variation, swapping the flower colour of two *Mimulus* species through introgressive breeding resulted in inverted pollinator visitation, highlighting the crucial nature of flower colour for pollination in specific species (Bradshaw and Schemske, 2003). By contrast, flower colour can also deter nectar and pollen robbery by incompatible visitors. For example, a South-African plant guild relying on wasp pollination has flower colours that resemble background vegetation preventing unwanted visitation from bees and birds; the wasp pollination is completely guided by olfactory signals (Shuttleworth and Johnson, 2012). Interestingly, there are examples of floral colour change in a blooming flower independent of senescence. This is usually accomplished by the accumulation or disappearance of anthocyanins pigments after pollination, further establishing the link between flower colour and pollinator attraction (Weiss, 1995; Farzad et al., 2002).

Genetic variation underlying flower colour visible to the human eye has been identified for many species and tends to occur in the anthocyanin, carotenoid and betalain groups of metabolites. Typically, differential expression of these biosynthetic pigment genes is responsible for flower colour variation as opposed to structural gene evolution (Schiestl and Johnson, 2013); for example, changes in anthocyanin

gene expression have been shown to be the driving force behind the additive evolution of flower colour in *Mimulus* (Streisfeld and Rausher, 2009).

Anthocyanins form the majority of visible floral pigments (Koes et al., 2005) and are responsible for the majority of the orange, red, purple and blue flower colours (Grotewold, 2006). Many anthocyanin biosynthesis pathway enzymes perform essential functions aside from pigment formation and are present as a single copy, which results in the pathway to be highly conserved (Holton and Cornish, 1995). Transcription factors that control anthocyanin structural genes are highly tissue specific and changes in expression would not necessarily be deleterious in other plant structures. (Quattrocchio et al., 2006). More than 70% of anthocyanin variations are due to changes in the transcription factor R2R3-MYB either through coding mutations or *cis*-regulatory mutations (Sobel and Streisfeld, 2013). Thus, evolutionary changes in flower colour will most likely be a consequence of mutations leading to changes in expression of anthocyanin structural genes (Whittall et al., 2006; Streisfeld and Rausher, 2009). Nevertheless, sequence mutations in anthocyanin genes themselves have been involved in flower colour evolution (Hoballah et al., 2007). Most likely mutations on both the structural and expression level have worked in concert to provide the situation present today in all plant species.

Considering all possible evolutionary scenarios, gaining a pigment would also be possible by the duplication of a pigment producing enzyme (and subsequent subfunctionalization) or recruitment of duplicated genes with an anthocyanin-like domain makeup into the pigment pathway through *cis*-regulatory mutations. These gain mutations have been shown to have taken place in nature by changes in the R2R3-MYB regulatory class of genes (Sobel and Streisfeld, 2013).

Anthocyanins are synthesized through the very extensive flavonoid pathway (Figure 1). The core flavonoid pathway consists of 4 genes: chalchone synthase (CHS), chalcone isomerase (CHI), flavonone 3β-hydroxylase (FHT, synonym F3H) and flavonone 3'-hydroxylase (F3'H) (Martens et al., 2010). This first group of genes is also referred to as early biosynthetic genes (EBG) (Xu et al., 2015). After conversion to dihydroflavanols by F3'H, further conversion to leucoanthocyanins by dihydroflavonol 4-reductase (DFR) takes place. The anthocyanin precursors pelargonidin, cyanidin and delphinidin are then formed by leucoanthocyanidin deoxygenase (LDOX, synonym ANS), ending with the conversion of these anthocyanidins to anthocyanins by glutathione-S-transferase (GST) (Sobel and Streisfeld, 2013). The genes in these latter steps of the anthocyanin pathway are also referred to as late biosynthetic genes (LBS). The final glucosylation modifications in the pathway by flavonoid glycosyltransferases like GST have been shown to be unique and taxa specific (Grotewold, 2006; Martens et al., 2010).

Anthocyanins expression is regulated by a regulatory complex consisting of proteins from three main families: R2R3-MYB, basic helix-loop-helix (bHLH) and WD40-repeat (WDR), which can be referred to as the MYB-bHLH-WDR (MBW) complex (Ramsay and Glover, 2005; Quattrocchio et al., 2006; Sobel and Streisfeld, 2013). Three genes have been identified as coding for the core proteins of the MBW complex in *Arabidopsis*: transparent testa 2 (TT2) or MYB123, transparent testa 8 (TT8) or bHLH042, and transparent testa glabra 1 (TTG1) of the WDR family. The TT2–TT8–TTG1 complex controls the

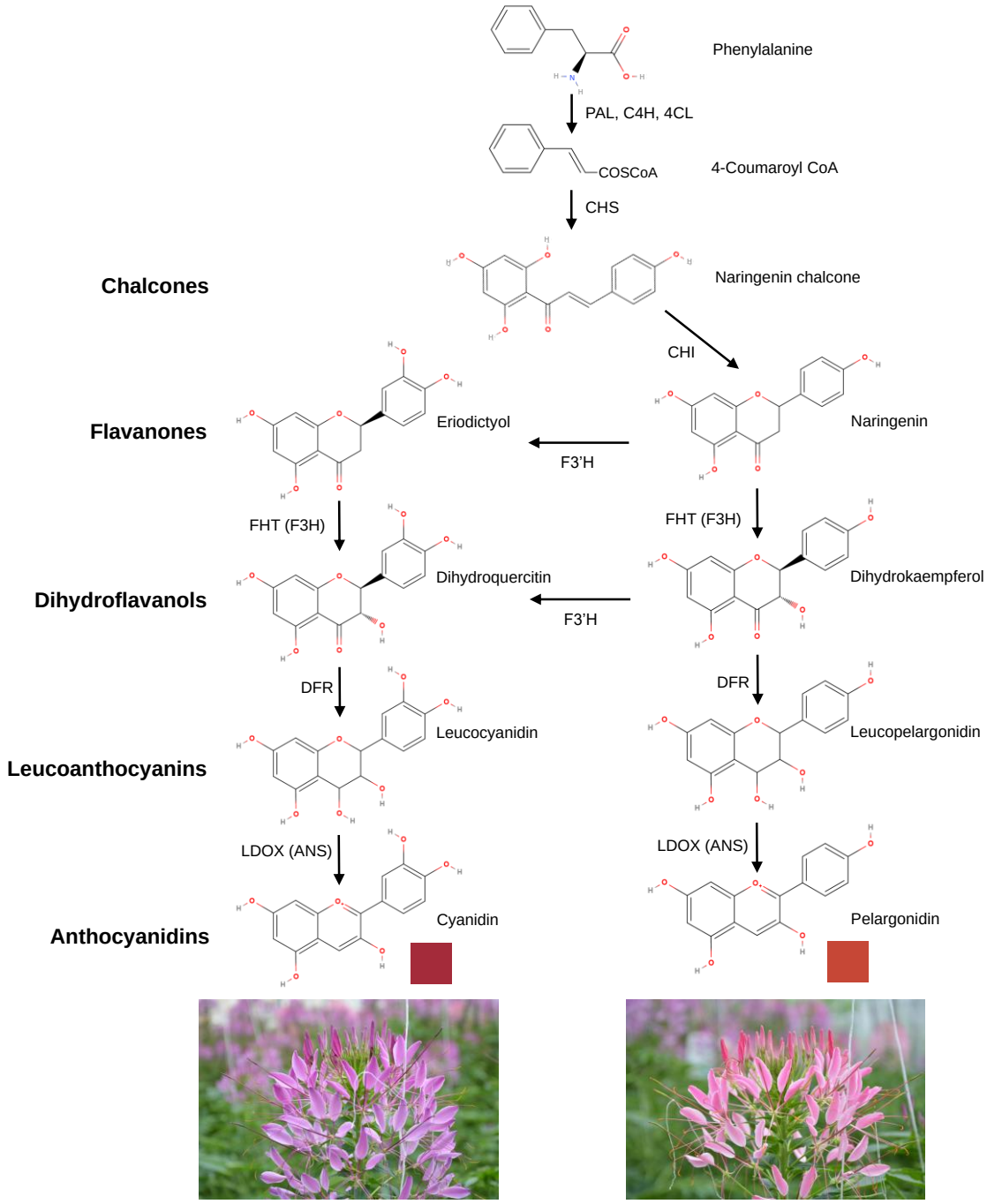


Figure 1. Flavonoid / anthocyanin pathway and resulting flower colour superimposed. The left pathway shows cyanidin synthesis and the 1100 or 'purple' flower line. The right pathway shows pelargonidin synthesis and the 1102 or 'pink' flower line. PAL: phenylalanine ammonia-lyase; C4H: cinnamate-4-hydroxylase; 4CL: 4-coumarate:CoA ligase; CHS: chalcone synthase; CHI: chalcone flavanone isomerase; F3'H: flavanone 3-hydroxylase; FHT, synonym F3H: flavonone 3β-hydroxylase; DFR: dihydroflavonol 4-reductase; LDOX, synonym ANS: leucoanthocyanidin deoxygenase.

expression of the LBG group of enzymes. Other MYBs and bHLH domain proteins can form complexes with TTG1 as well and various combinations of these proteins form complexes that are unique to different tissues. The specificity is mostly determined by small amino acid variations in the MYB protein that confer altered target gene specificity. (Xu et al., 2015).

The EBG group in *Arabidopsis* is controlled by an MBW complex consisting of four genes: purple acid phosphatase 1 & 2 (PAP1/PAP2), GLABRA 3 (GL3) and transparent testa glabra 1 (TTG1) (Zhang et al., 2003). Similar MBW complexes that control the EBG group have been identified in maize (Carey et al., 2004), *Petunia* (Spelt et al., 2000) and *Antirrhinum majus* (Schwinn et al., 2006).

Tarenaya hassleriana (formerly known as *Cleome spinosa* (Iltis and Cochrane, 2007)) is a species of the *Cleome* genus which tend to inhabit tropical habitats and are shrubby plants with palmately compound leaves (Stevens, 2001b). Members of the *Cleome* genus are important crop vegetables in tropical and subtropical regions of Africa, a notable example being *Gynandroposis gynandra* which is part of the African Orphan Crops project (African Orphan Crops Consortium website, accessed Aug 2016(Anon, n.d.) and for which sequencing efforts are underway (Schranz et al., in preparation). Cleomaceae have been identified as a model genus for studying the transition from C3 to C4 photosynthesis (Brown et al., 2005) as it contains a number of C4 species, as well as a C3 / C4 intermediate (Feodorova et al., 2010). Cleomaceae are also studied as a model outgroup for Brassicaceae regarding polyploidy, and after sequencing of the *T. hassleriana* genome (Cheng et al., 2013), a model framework has rapidly emerged (van den Bergh et al., 2014, 2016).

Flower colour in Cleomaceae is a relatively understudied trait, with a basic establishment of colour dominance (Ladd, 1983), and the discovery of a number of novel pigments (Jordheim et al., 2009), and new aspects with regard to changes in colour over the lifetime of the flower (Nozzolillo et al., 2010). However, limited genetic basis of this trait is available, mostly based on orthologous information from *Arabidopsis* and the Brassicaceae as a whole. In this study, we aim to elucidate the genetic basis of flower colour variation in *Tarenaya hassleriana*. Flavonoid pigments, in addition to all other compounds detectable, were screened through liquid chromatography (LC) coupled to both photodiode array (PDA) detection and mass spectrometry (MS) and the associated accurate signals are used in finding specific trait loci using genetic map information established through genotyping by sequencing (GBS).

MATERIALS AND METHODS

PLANT MATERIAL

A cross between *Tarenaya* line 1100 (‘purple’) and 1102 (‘pink’) (for photographs of colours see Figure 1) was made and 200 F2 plants were grown in Wageningen University and Research greenhouse (Wageningen, The Netherlands) under defined and controlled conditions of light period (16h day/8h night) and temperature (20°C day/16°C night). Two petals from 5 different flowers from each plant were collect into 15 mL plastic tubes and immediately frozen in liquid nitrogen and kept at -80°C for further analysis.

FLAVONOL EXTRACTION

Metabolite extraction was made according to the protocol established by De Vos et al. (2007) with slight modifications. Fresh flower material was ground in liquid nitrogen, 100 mg frozen powder was weighed and extracted with 1.2 mL of 99.75% methanol (MeOH) containing 0.25% of formic acid (FA). The solution was transferred into 1.5 mL eppendorf tubes, submitted to a 30-minute ultrasonic bath for metabolite extraction. After 10-minute centrifugation at maximum speed to precipitate proteins and cell debris, samples were filtered in a 96-well plate (Captiva 0.45 µm, Ansys Technologies) through vacuum pressure into a glass vial for LC-PDA-MS analysis.

LC-PDA-MS PROFILING AND FLAVONOID IDENTIFICATION

The LC-PDA-MS profiling method was performed using an Acquity HPLC-PDA system (Waters Chromatography) coupled to an LTQ-Orbitrap FTMS hybrid mass spectrometry system (THermo Scientific). Chromatographic separation and compound detection was performed according to van der Hooft et al. (2012).The PDA was set to scan absorbance spectra from 240-600 nm; the MS was set to scan masses from m/z 90-1300, in negative ionization mode. LCMS data were processed in an essentially

untargeted manner (De Vos et al., 2007) resulting in a data matrix with the relative intensities of 1150 compounds including flavonoids in all samples. For compound identification, both LC retention times, absorbance spectra from the PDA detector and accurate masses of parent ions were assessed. After this process, the compounds found were confirmed against public and in-house metabolite libraries containing accurate masses, MSMS fragments, retention times and/or absorbance spectra.

SNP IDENTIFICATION BY GENOTYPING BY SEQUENCING

DNA was isolated from snap frozen leaf material of the F2 cross described earlier using an adapted CTAB protocol from Maloof (http://malooflab.openwetware.org/96well_CTAB.html, accessed 25 January 2017). DNA was treated with RNAse overnight at 37°C with RNAse one by Promega. Quality was checked on a 1 % agarose gel and DNA quantity was checked with Pico Green. Based on this, DNA was diluted down to 20 ng/uL with MQ water.

Genotyping by sequencing (GBS) was performed in general by following the procedure described by Elshire et al. (2011). Oligonucleotides for creation of common as well as 96 barcoded ApeKI adapters were ordered at Integrated DNA Technologies and diluted to 200 µM. For each barcoded and common adapter, top and bottom strand oligonucleotides were combined to a 50uM annealing molarity in TE to 100 µL total volume. Adapter annealing was carried out in a thermocycler (Applied Biosystems) at 95°C for 2 minutes, ramped to 25°C by 0.1 degree per second, held at 25°C for 30 minutes and kept at 4°C until further analysis. Annealed adapters were further diluted to a 0.6 ng/µL concentrated working stock of combined barcoded and common adapter in 96 well microtiter plate and dried using a vacuum oven.

Of each genomic DNA sample 100 ng (10 ng/µL) was used and added to lyophilized adapter mix and dried down again using a vacuum oven.

Adapter DNA mixtures were digested using 2.5 Units ApeKI (New England Biolabs) for 2 hours at 75°C in a 20 µL volume. Digested DNA and Adapters were used in subsequent ligation by 1.6 µL (400 Units/µL) T4DNA Ligase in a 50 µL reaction volume at 22°C for one hour followed by heat inactivation at 65°C for 30 minutes. Sets of 96 digested DNA samples, each with a different barcode adapter, were combined (10µL each) and purified using a Qiaquick PCR Purification columns (Qiagen). Purified pooled DNA samples were eluted in a final volume of 10µL. DNA fragments were amplified in 50 µL volume reactions containing 2 µL pooled DNA, 25 µL KAPA HiFi HotStart Master Mix (Kapa Biosystems), and 2 µL of both PCR primers (12.5 µM). PCR cycling consisted of 98°C for 30 seconds, followed by 18 cycles of 98°C for 30 seconds, 65°C for 30 seconds, 72°C for 30 seconds with a final extension of 5 minutes and kept at 4°C. Amplified libraries were purified as above but eluted in a 30 µL volume. Of the amplified libraries, 1 µL was loaded onto a Bioanalyzer High Sensitivity DNA Chip (Agilent technologies) for evaluation of fragment sizes and 1 µL was used for quantification using Qubit (Life Technologies). Amplified library products were used for extra size selection using 2% agarose gel cassette on a blue pippin system (Sage Science) to remove fragments smaller than 300bp. Eluted size selected libraries were purified by AmpureXP beads (Agencourt). Final libraries were used for clustering on five lanes of an Illumina Paired End flowcell using a cBot. Sequencing was done using an Illumina HiSeq2000 instrument using 2*100 nt Paired End reads.

Raw sequencing data was processed using the TASSEL software package using the GBSv2 pipeline (Bradbury et al., 2007). The GBSSeqToTagDBPlugin from this package was run with the following parameters: kmerLength: 64, minKmerL: 20, mnQs: 20, mxKmerNum 100000000. Tags were dumped from the database using TagExportToFastqPlugin and mapped to the publicly available Tarenaya hassleriana reference genome (latest version at time of writing, v5) using the bwa software package (Li and Durbin, 2009) in single-ended mode (samse parameter). Positional information from aligned SAM files was stored in the TASSEL database using the SAMToGBSdbPlugin. The DiscoverySNPCallerPlugin was run using the following parameters: mnLCov: 0.1, mnMAF: 0.01. Found SNPs were scored for quality

using SNPQualityProfilerPlugin and the Average taxon read depth at SNP was used as a quality score for filtering in the next step (minPosQS parameter), these scores were written to the TASSEL database using UpdateSNPPositionQualityPlugin. Finally, the ProductionSNPCallerPluginV2 was run with the following parameters: Ave Seq Error Rate: 0.002, minPosQS: 10, mnQS: 20.

HOMOLOG IDENTIFICATION

To find homologous genes in *T. hassleriana* a nucleotide BLAST was performed using the AT5G42800, AT4G22880 and the AT5G07990 genes as queries (see Results section). The ncbi-blast+ package was used (version 2.5.0) (Altschul et al., 1990; Camacho et al., 2009) with an e-value cutoff of 1e-10. Subsequently, the top results were queried in CoGe synfind (Lyons and Freeling, 2008) using the LAST algorithm (Kiełbasa et al., 2011), with a gene window size of 30 and a minimum number of hits of 4, using the collinear scoring function.

QTL MAPPING

The r/qtl package was used for QTL mapping (Broman et al., 2003). SNP data was manually adjusted to fit the input style required by r/qtl read.cross(). The genome was scanned for QTLs using the extended Hayley-Knott regression (Feenstra et al., 2006). To establish the genome wide significance threshold a permutation test was run with n = 1000, after which the significance threshold and p-values were calculated using the summary function at a $p < 0.05$ significance threshold with an alpha value of 0.2.

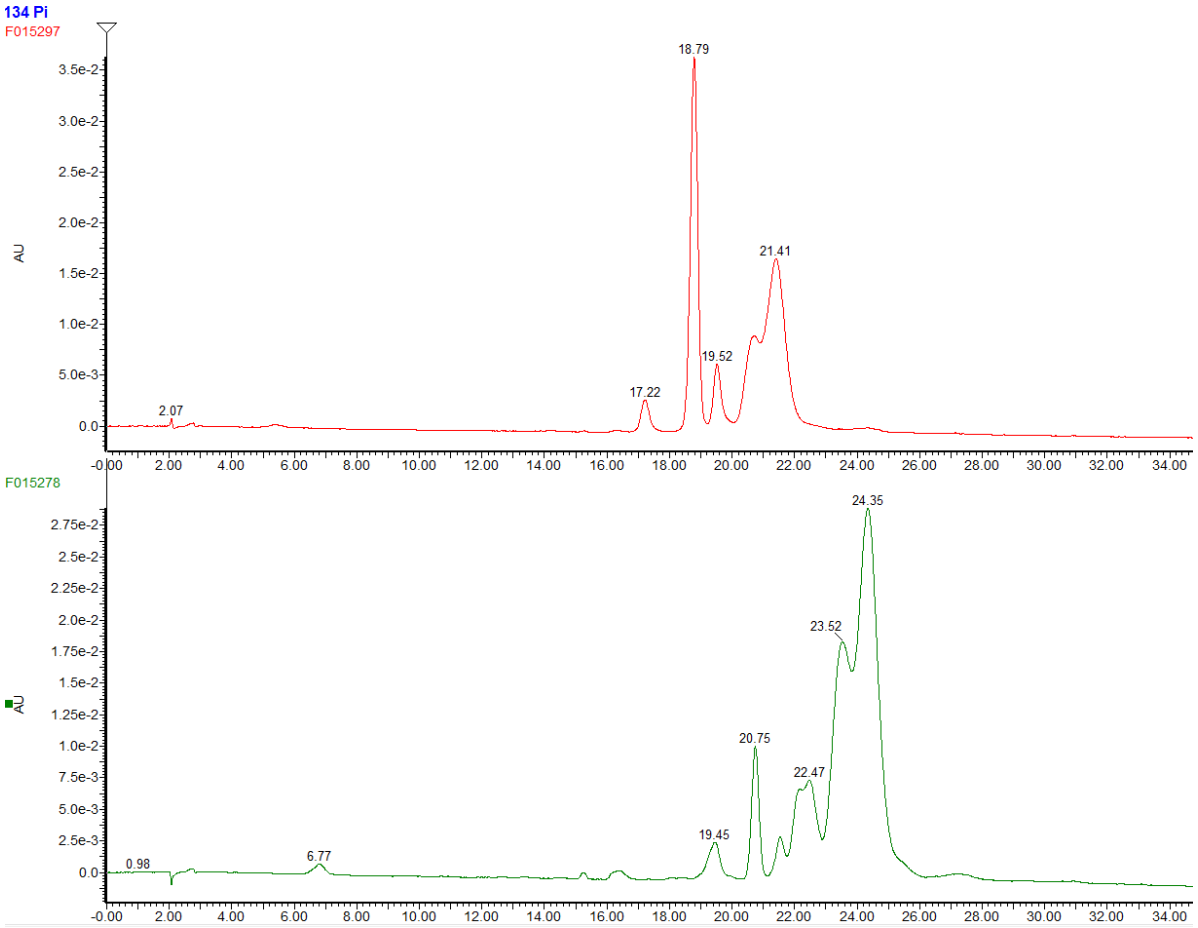


Figure 2. Anthocyanin profiles, visualized at 520 nm, of representative purple (upper trace) and pink (lower trace) flowering plant.

RESULTS

LC-PDA-MS PROFILING AND FLAVONOID IDENTIFICATION

LC-PDA-MS of purple flower extracts showed several abundance peaks absorbing at 520 nm with retention times (RT) between 16 and 22 minutes (Figure 2). The most abundant 510 nm-absorbing peak at 18.79 minutes had a mass of 919.248 and MSMS fragments corresponding to a complex cyanidin glycoside with acylation of phenylpropanoids. The second and third most abundant peaks at 20.79 and 21.41 minutes had an accurate mass of 1125.305 and 1095.295 respectively, and also corresponded to a cyanidin backbone decorated with several glycosidic and acyl moieties. Contrastingly, pink flower extracts showed abundancy peaks at slightly higher retention times, so a more apolar character, with the most abundant peaks at 24.35 and 23.52 minutes having a mass of 1079.300 and 1109.311, respectively. The MSMS fragments of these peaks correspond to a pelargonidin and a di-hexose molecule acylated with the phenylpropanoids ferulic acid and sinapic acid, respectively. The other peaks in purple flowers also corresponded to pelargonidin conjugated to glycosidic moieties. In summary, purple flowers contained pelargonidin-based anthocyanins conjugated to different sugars (hexoses and deoxyhexoses) with or without additional decoration with phenylpropanoids (coumaric acid, ferulic acid and synapic acids), while purple flowers contained cyanidin-based anthocyanins, similarly decorated with various sugar and phenylpropanoids.

Principal Component Analysis (PCA) based on 1150 compounds detected in the LC-MS profile analysis was performed, based on a mix of petals randomly taken from 10 pink and 10 purple flowering plants. The first three principal components represented 32.9%, 15.6% and 12.4% of the total metabolite variation (Figure 3) and showed clear separation between purple and pink flowers on the 1st principal component.

In the flavonoid synthetic pathway, pelargonidin is synthesized from dihydrokaempferol via the enzymes dihydroflavonol 4-reductase (DFR) and anthocyanidin synthase (ANS). These two enzymes are also used in the synthesis of cyanidin but from the precursor dihydroquercetin. The conversion from dihydrokaempferol to dihydroquercetin is performed by flavonoid-3’-hydroxylase (Figure 1). We therefore needed to find the most likely functional orthologs of these genes in *T. hassleriana*. DFR and ANS are encoded by AT5G42800 and AT4G22880, respectively. These two genes both show one nucleotide BLAST hits above the homolog standard cutoff of 1e-10. DFR matches to Th2v29825 (E-value 0.0, bit-score 977 in a 38,471,823nt database) and ANS matched to Th2v10835 (E-value 0.0, bit-score 1046 in a 38,471,823nt database). Syntenic analysis of these genes shows that neither the AT5G42800 nor the AT4G22880 gene has syntelogs in *T. hassleriana* which can indicate transpositions in either organism

In *Arabidopsis thaliana*, flavonoid-3'-hydroxylase (F3’H) is encoded by the CYP75B1 (AT5G07990) gene (Schoenbohm et al, 2000) and the abundance relative to CHS, the first step in the flavonoid pathway,

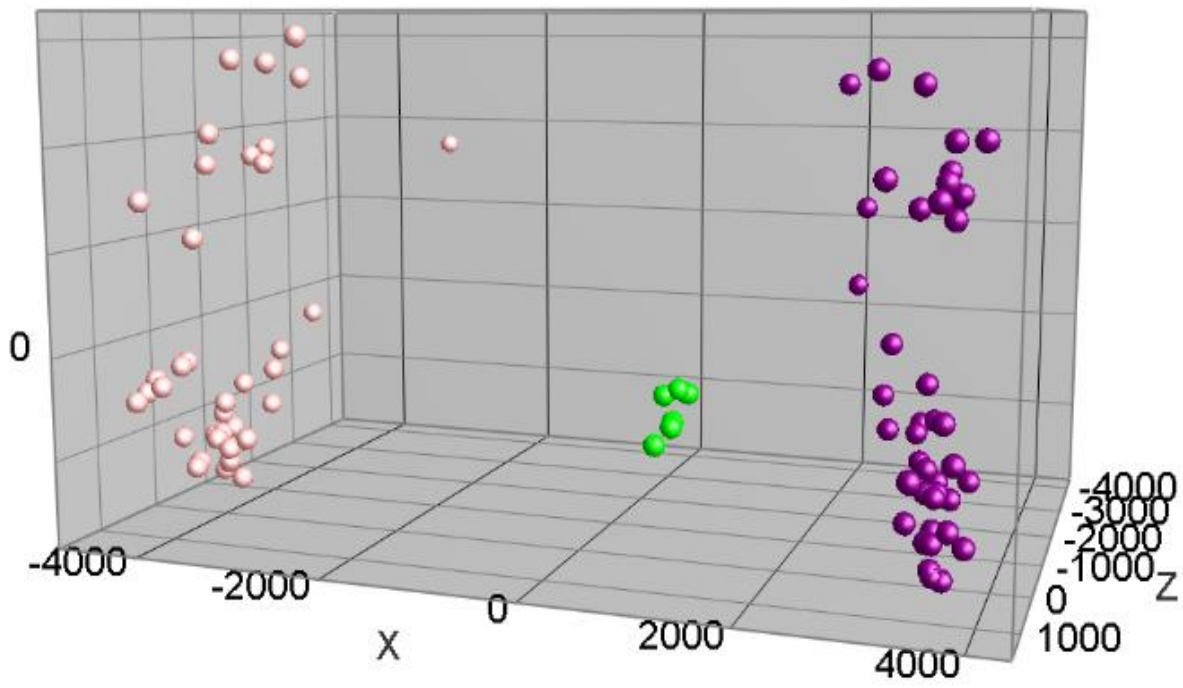


Figure 3. PCA plot based on 1150 compounds detected in the LC-MS profile analysis. Pink globes indicate pink flowers, purple globes indicate purple flowers, green globes are quality control extracts based on a mix of petals randomly taken from 10 pink and 10 purple flowering plants. X, Y and Z axes are the 1st, 2nd and 3rd principal component, respectively, representing 32.9%, 15.6% and 12.4% of the total metabolite variation, respectively.

determines the Quercetin / Kaempferol-based metabolites ratio. Sequence alignment with *T. hassleriana* genes through nucleotide BLAST shows Th2v13860 as a potential ortholog (E-value 0.0, bit-score 1045 in a 38,471,823nt database) with no other viable candidates. Syntenic analysis confirms the orthology assessment and identifies Th2v13860 as a syntelog with a synteny score of 9. Two other genomic regions in *T. hassleriana* show synteny around this gene, but not with the actual gene itself which might indicate gene loss through fractionation (https://genomevolution.org/r/j62a).

SNP CALLING ON T. HASSLERIANA REFERENCE GENOME.

Genotyping by sequencing was used to generate genotype data of an F2 population of *T. hassleriana* with mixed flower colours. Sample collection and subsequent Illumina sequencing resulted in 671,766,309 paired end reads. Read quality processing and tag matching resulted in 474697 tags which could be matched to the reference genome. SNP calling resulted in 10809 SNPs over the 786 longest scaffolds (N90 cutoff), which were reduced through quality filtering to a set of 2845 core SNPs.

SNP distribution was varied across scaffolds (Supp. Fig. 1). In these core SNPs, many SNP "islands" were present, where a group of SNPs with the same state across individuals are close (<10kb) together. These were collapsed down to a single SNP using a sliding window method, resulting in 1291 SNPs which were used in further analyses.

The state of individuals versus the read quality of SNPs was compared (Supp. Fig. 2). A large portion of SNPs are AA or BB (compared to the reference genome) homozygous. Since a segregation of 1:3 would be expected from an F2 cross, we selected only SNPs that were between 15% and 85% BB homozygous with an average read depth >5 for genetic map construction, resulting in 264 SNPs which were used for further analysis.

BTL ANALYSIS OF FLOWER COLOUR

Even though the total genetic make-up of this F2 population shows relatively low variance, the parents showed at least one specific varied trait: flower colour (purple x pink). This was also apparent in the F2 which consisted of 162 purple flowers and 52 pink flowers. This binary trait was mapped using a QTL approach, resulting in 2 significant QTLs on scaffolds 3 and 7, with 7 having the highest LOD score (Figure 4). The LOD peak on scaffold 7, corresponds to 77.0cM, which is roughly equivalent to the region surrounding the 6,094,438bp region. This is extremely close to the F3'H candidate gene found in the MCMC analysis, the Th2v13860 gene which lies at 6,121,252 – 6,124,103 on scaffold 7. The other LOD peak on scaffold 3 lies at the very end at 95.60cM which is roughly equivalent to the region around 965,703,400. Two candidate genes in this locus are Th2v12569 (an ortholog of AT1G78600 which is mapped to GO:0009718 anthocyanin-containing compound biosynthetic process) and Th2v12595 (an ortholog of AT2G16720 which is mapped to GO:1900384: regulation of flavonol biosynthetic process).

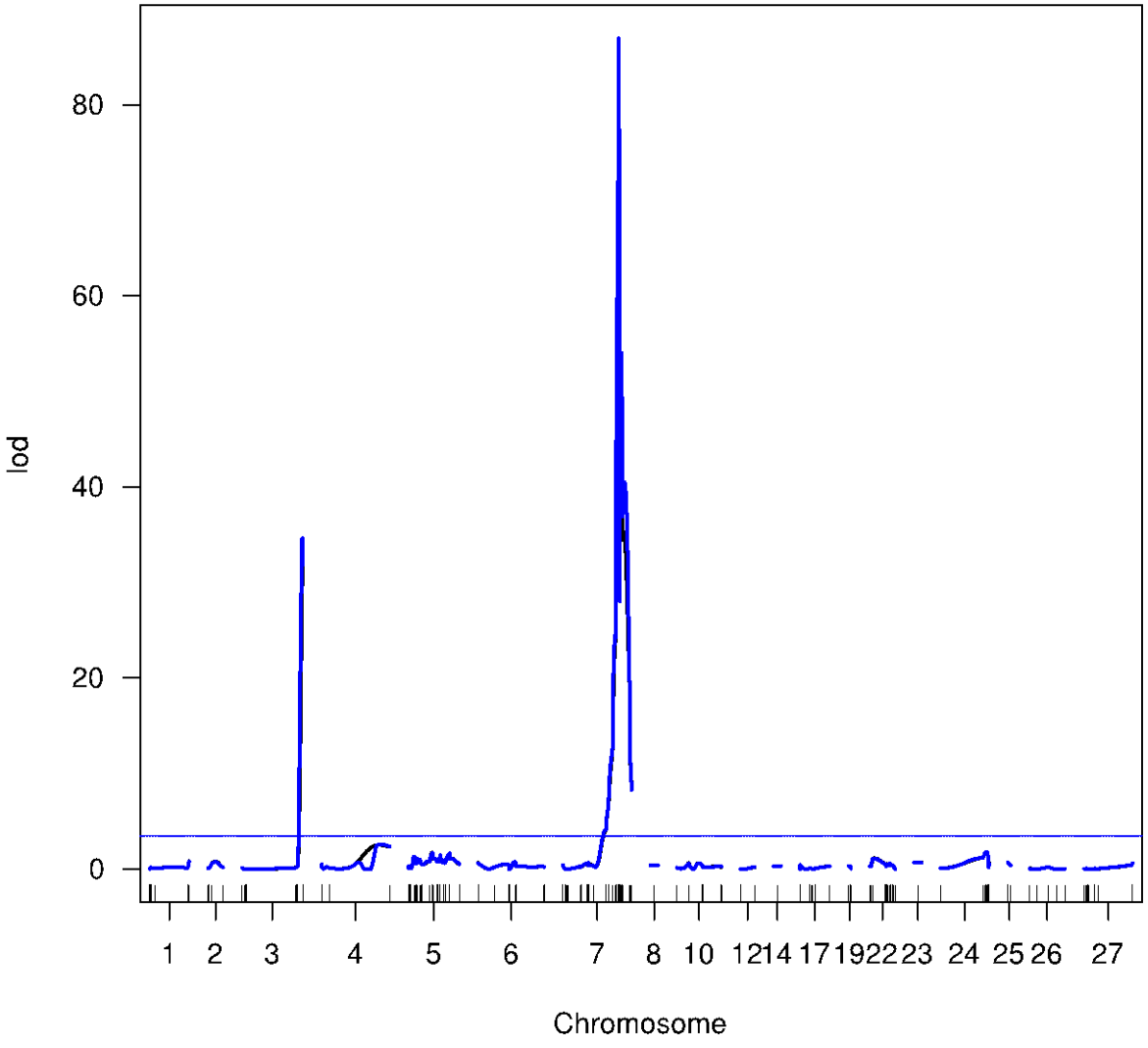


Figure 4. Binary Trait Locus lod peaks based on the extended Hayley-Knott regression. The 27 largest scaffolds of *T.* hassleriana are shown, no significant peaks were detected in smaller scaffolds. The dashed line indicates lod = 3.5 at which $p < 0.05$ as determined by permutation analysis with n=1000.

Neither of the two candidate genes found in the MCMC analysis (Th2v29825 and Th2v10835) are located on scaffold 3.

CONCLUSION AND DISCUSSION

In this study we examine the metabolite differences in an F2 population in *Tarenaya hassleriana* with limited genetic variation focusing on anthocyanin differences and flower colour by extension. We find that flavonoid profiles are markedly different between pink and purple offspring. We performed SNP analysis and confirm that genetic variation is low but with 'islands' of strong genetic variation. We identify a BTL locus for the pink/purple flower colour trait on scaffold 7 and scaffold 3, one of which confirms our earlier candidate gene on scaffold 7.

Results obtained from LCMS showed a varied mixture of components, but a clear separation of flavonoid types between the two types of flower colour. Cyanidin-based anthocyanins are strongly present in purple flowers, whereas pink flowers contain mostly pelargonidin-based ones. From a chemical perspective this is in line with the fact that cyanidin has a darker hue while pelargonidin has a lighter hue (Grotewold, 2006).

We identify three candidate genes based on the enzyme pathway that lead to the different pigment compounds and find clear homologs in *T. hassleriana*. As anthocyanins are strongly conserved throughout the plant kingdom this is expected (Campanella et al., 2014). We find only one homolog per enzyme which suggests that ohnologous copies of these genes have been lost after the most recent Cleomaceae polyploidy event (Barker et al., 2009; Cheng et al., 2013). Syntenic analysis confirms these homologs are accurate based on homologous genes surrounding the genomic location of these genes.

Genotyping by sequencing is a relatively novel method that allows for cheaper genetic variation studies than classical methods. In this case we have found that the data was not sufficient for the construction of a *de novo* genetic map; variation across contigs was too low to reliably create a genetic map, resulting in most chromosomes being >100cM in length and scaffolds being arranged in an overlapping fashion on those scaffolds. This could be due to the usage of F2 progeny instead of a RIL population, but the limits of the GBS technique are still to be determined. It has so far only been used in maize, barley and wheat (Elshire et al., 2011; Li, Vikram, et al., 2015); three species for which genetics maps were already well established, meaning that the predictive power of this technique with regards to genetic map construction is still uncertain. However, the data obtained here was still accurate and varied enough to map an abundance of individual SNPs to the current reference sequence.

Even though no genetic map could be constructed we found enough variation in SNPs after rigorous selection to determine a BTL using established QTL methods (rqtl). We find two clear binary trait loci that are far above the significance threshold of LOD = 3.5. One of the loci is very close to the candidate gene which determines the differential hydroxylation pattern of the anthocyanidin backbone, Th2v13860. Looking into the SNP markers around this area in greater detail, none lie exactly within the gene. Therefore, it is not possible to tell what exactly confers the flower colour variation from this data alone; changes can lie in transcription factors resulting in changed expression or variations in amino acid sequence resulting in an altered function of the protein itself.

Our initial measurements indicate that in pink flowers, besides the pelargonidin type of anthocyanidins also various kaempferol conjugates are present in higher abundance than in purple flowers. However, in the cyanidin-containing purple flowers quercetin conjugates are more abundant. Combining this with the fact that genetic variation points to the F3'H gene, which is responsible for the conversion of dihydrokaempferol into dihydroquercetin (Figure 1), we conclude that the effect seen in flavonols reflects the difference in anthocyanins; flavonoids and anthocyanins are both flavonoid classes branching at dihydroflavonols. This difference could be the result of lower expression of the F3'H gene

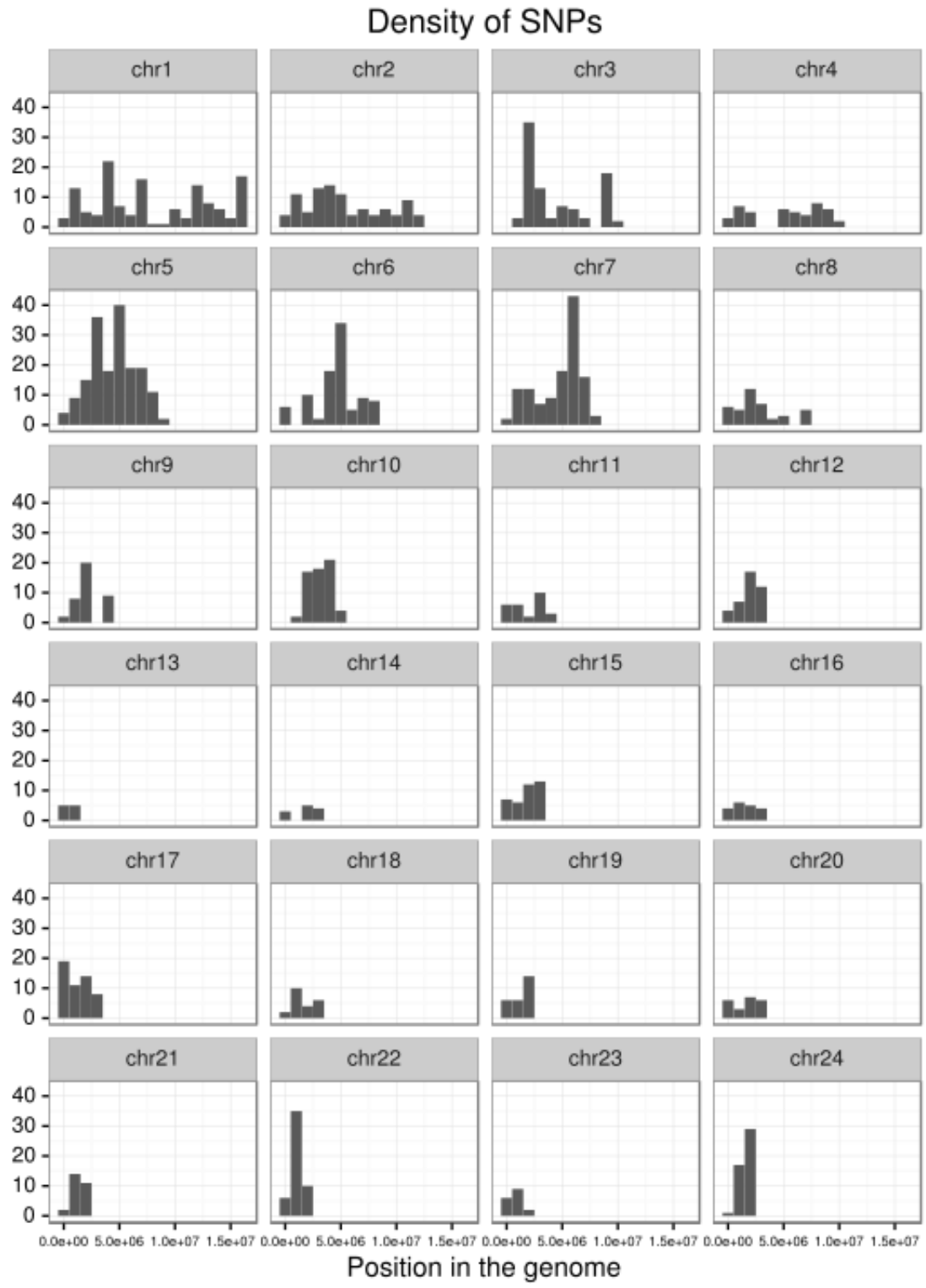
in pink flowers as compared to purple flowers (leading to a changed rate of conversion of kaempferol into quercetin) or changes in the F3'H enzyme itself leading to altered conversion rates. Individual cloning of the gene from both floral types shows SNP variation present (data not shown), but the effect of these changes is near impossible to predict *in silico*. The exact mechanism through which the F3'H enzyme affects flower colour remains to be uncovered.

Flower colour is a relatively simple trait which is largely governed by a small amount of pigments. In this study we show that with single SNP changes in genotype, the phenotype of a plant can undergo a significant change; in this case a single nucleotide change in the F3'H gene resulting in flower colour variation. The ecological effects on for example pollinator visitation patterns and resulting reproductive fitness of these changes remain to be researched further.

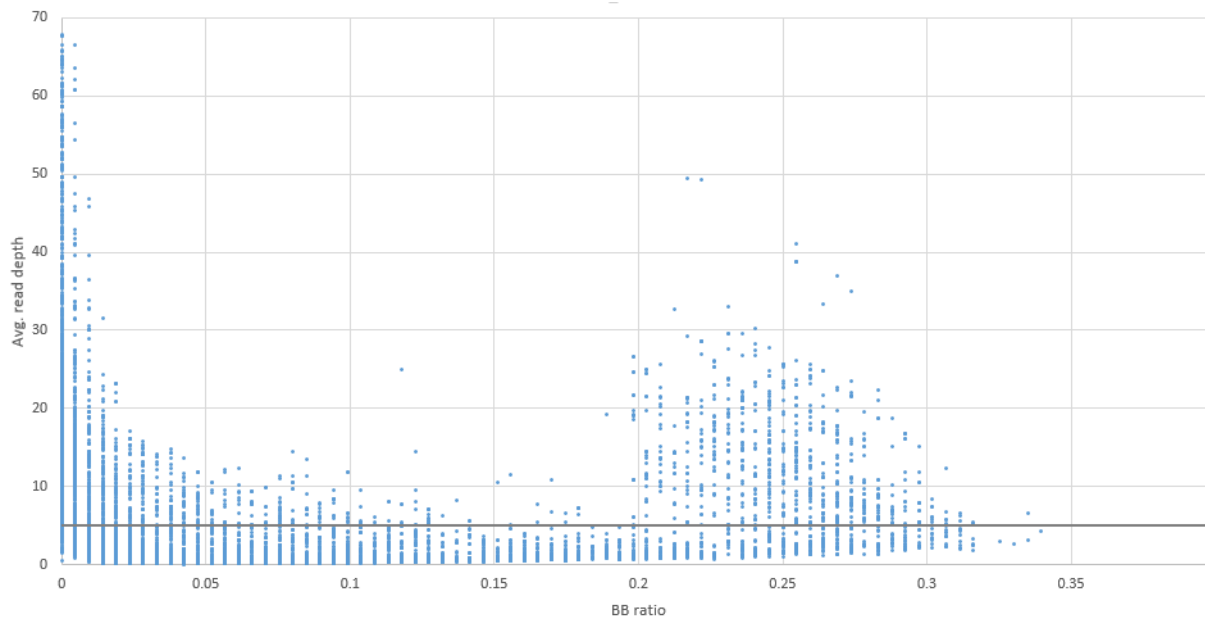
We have convincingly shown here that flower colour in *T. hassleriana* is determined by the ratio of two visible pigments, which in turn are controlled by a single enzyme. Indeed, in this case it is clear that as Isaac Newton stated in his *Mathematical Principles of Natural Philosophy*: "Nature is pleased with simplicity, and affects not the pomp of superfluous causes."

ACKNOWLEDGEMENTS
EvdB prepared GBS data and performed orthology and bioinformatic analysis and prepared the manuscript. PdOM performed flower harvesting, mass spectrometry measurements, data processing and statistical analyses. SJvdK cloned and sequenced the F3'H gene and assisted in preparation of the manuscript. FB performed DNA extraction and processing from the F2 population and cloning and sequencing of the F3'H gene. RCHdV supervised the MS measurements, data processing and statistical analyses, as well as preparation of the manuscript. MES assisted in bioinformatic and orthology analyses and preparation of the manuscript.

SUPPLEMENTAL FIGURES



Supplementary Figure 1. SNP distribution across the 24 longest scaffolds. Lengths of scaffolds in base pairs.



Supplementary Figure 2. SNP plotted according to average read depth (y-axis) and ratio of BB (non-reference) homozygosity. The large number of SNPs with a 0 BB ratio represents the low variation found in this F2 population. The grey line at read depth = 5 indicates the quality cutoff that was used to include SNP's for further downstream analysis.

GENERAL CONCLUSION

In this thesis, several important conclusions are made, some explicit and some implied. I will first give a brief overview of how the explicit conclusions relate to the current state-of-the-art and possible future research. Following that, I will attempt to clarify the more implied conclusions and how they relate to the future of data-driven science.

THE IMPORTANCE OF SPECIES FRAMEWORKS

No biological research would be structured as it is now without the use of model species. Just as the BALB/c *Mus musculus* strain has proven to be invaluable for medical research, *C. elegans* has been essential for the research on multicellular eukaryote development and *S. cerevisiae* is key for the most fundamental insights in DNA structure and genetics, so plant science has its own model species that have been of key importance in the recent explosion of genetic insights. The most important of course, is *Arabidopsis thaliana*. Most gene functional predictions are based on the knowledge derived from the study of this single plant and yet as of January 2017 30% of GO biological process annotated genes have GO:0008150 or 'biological process unknown', meaning so many genes are yet to be discovered. This lack of knowledge also echoes on in all annotation predictions based on this data, meaning that these genes will be ignored in genome wide analyses. It is important that these types of omissions are mitigated by additional research in the model species but also by considering research from additional species in which genotype-to-phenotype correlation may be more apparent for certain genes or traits.

Thus expansion of the scope beyond model species will benefit the state and quality of plant genetics research as a whole. Crop species can fulfil this role partially, such as the study towards the various cultivated Brassicaceae have benefited *Arabidopsis* research. However, crop species have often not been selected for their scientific but for their functional value. Another complicating factor is the thousands of years of human driven selection and hybridization that confound the genetic evolution history of these species. A more appropriate solution is the selection of several key species in a family or group based on phylogenetic history thus forming a larger model species framework. The sequencing and study of the 'ancestral' species of Brassicaceae, *Aethionema arabicum* is one of the first keystones in establishing such a model species framework for Brassicaceae.

To expand this framework to cover the larger Brassicales order, we have in this work sequenced a representative of the Cleomaceae, the sister family to Brassicaceae, namely *Tarenaya hassleriana* (**Chapter 1**). We expect that the species framework triangle of *Arabidopsis*, Aethionema and Tarenaya will provide a wealth of information for the study of genome structure, gene evolution and the effects of polyploidy on these processes.

POLYPLOIDY AS A DRIVING FACTOR OF TRAIT GENESIS

As described in the introduction, polyploidy has been of major importance in the evolutionary history of many widespread eukaryotic lineages on Earth. However, modelling of speciation and extinction rates in polyploids versus diploids shows that polyploids have a higher extinction rate (Arrigo and Barker, 2012), seemingly pointing to the conclusion that polyploids tend to be evolutionary dead-ends and that the historic polyploidy events that we see evidence of today are part of the lucky few that made it. This is also in line with the competitive disadvantages a polyploidy organism has when it arises in a population with its diploid ancestors. Competitive exclusion in the form of minority cytotype disadvantage as well as genetic factors such as allele masking by multiple copies, loss of self-incompatibility and the obvious meiotic difficulties arising from chromosome duplications seem to put newly born polyploids at a severe disadvantage. How then can we explain the ubiquity of polyploidy in so many lineages, especially plants?

One answer is that there are also several ecological advantages to being polyploid. Although poorly understood, many physiological traits have been shown to differ between polyploids and diploids: water relations, stress tolerance and cold tolerance amongst others (Weiss-Schneeweiss et al., 2013). These differing traits may confer advantages depending on the environment, such as in the case of *Senecio carniolicus* which grow at higher altitudes in the alps and in which individuals with different ploidy levels separate into different microhabitats (Weiss-Schneeweiss et al., 2013). These types of ecological niche-forming are an important factor in speciation and are also likely to play a role in the establishment of polyploid lineages. Separation and speciation through these and similar mechanisms can eventually lead to new traits being developed in the polyploid lineage.

In the genetics behind the process of trait evolution in polyploids, there are two main forces at play: One is neo- and subfunctionalization, where a copy of a duplicated gene acquires a new or more specialized function from the original gene. The other is dosage balance, where alterations of gene function in a duplicated copy lead to deleterious effects due to a dosage imbalance of the gene's products (Conant et al., 2014). We investigate both of these forces in **Chapter 2** and **Chapter 3** of this thesis.

In **Chapter 2** we investigate the effects of the polyploid history of *Gynandropsis gynandra* on the development of C4 photosynthesis. We compare it to *Tarenaya hassleriana*, which has the same polyploid history but does not have C4 photosynthesis. We find that both plants have a distinct history of gene loss and retention and find that the number of C4 related orthologs is conserved across the two species. Combined with differences in expression data we conclude that the recruitment of C3 genes into C4 photosynthesis must have taken place in expression space and was most likely not significantly affected by the polyploidy event shared by Cleomaceae.

In **Chapter 3** we study compounds that are vital for plant herbivore defence and have radiated extensively in Brassicaceae: glucosinolates. The evolution of the different 'flavours' of these compounds has been varied and extensive. We utilize our established species framework and compare the gene families underlying the glucosinolate biosynthetic pathway between *A. thaliana* and *T. hassleriana*. We find that the radiation and expansion of these gene families has been significantly affected by their respective polyploid history. We find that the Th-α event has had a significant impact on the expansion of these gene families, with many families having retained two or even all three ohnolog copies from the hexaploidy. Considering all three polyploidy events, 89% of glucosinolate related genes are retained ohnologs. Thus, the impact of polyploidy on the development of these gene families has been enormous.

TRAIT STUDIES IN CLEOMACEAE

Much ground has to be covered when establishing a new model system as we have in this case done with *T. hassleriana*. One part is the mapping of known genes onto common traits such as abiotic and biotic stress response, circadian rhythm and plant morphology. One of the most eye catching traits of *T. hassleriana* are the abundant and colourful flower clusters when the plant is blooming. The varieties collected in our greenhouses showed a clear trait difference: part of the flowers had a bright violet colour whereas the others were salmon pink. Flower colour is not just a coincidental visual trait but it is important for the attraction of various pollinators and variation (whether visible by the human eye or in the UV spectrum) can be advantageous to reproductive fitness (Schiestl and Johnson, 2013). Through a cross with a pink and a purple parent and subsequent creation of an F2 line, we tried to identify the genetic basis of floral colour variation in *Tarenaya hassleriana* (**Chapter 4**). We measure the levels of different pigments in the flowers and find that purple flowers contain mostly pelargonidin based anthocyanins whereas pink flowers contain mainly cyanidin based anthocyanins. Through genotyping by sequencing we then identify a QTL that locates very close to a candidate gene which synthesizes a

cyanidin precursor (F3'H). We conclude that SNP variation in or near this gene must play a role in the determination of this important floral trait.

CONCLUDING REMARKS

The significant increase in the knowledge and study of genes and genomes over the past two decades have contributed immensely to our understanding of every aspect of plant life on Earth. Our knowledge and understanding of plants not only provides us with vital insights into a prominent and influential supergroup of eukaryotes but also provides us the means to influence and shape our everyday lives. The struggle for food that is shared by all non-photosynthetic organisms has been eased for humans by humans, starting with the agricultural revolution in the Neolithic, subsequent advances in farming techniques throughout the 1st millennium CE, the increase and spread of breeding knowledge from the 16th to the 19th century, the rise of mechanisation during and after the industrial revolution starting at the end of the 19th century, the massive increase in crop yield efficiency during the 'green revolution' in the 1970's, and concluding with the hybrid vigour enabled agricultural practice of today. All these advances have led to human civilization being able to support a population of 7.4 billion at the start of 2017, but not without its fair share of issues. Current threats to food security come in the form of growing competition for land use, decreases in arable land due to unsustainable farming practices, overexploitation of fisheries (Godfray et al., 2010) against a backdrop of climate change leading to more unpredictable and extreme weather patterns (Intergovernmental Panel on Climate Change, 2014). Thus the challenge to feed the expected population of 9 billion in 2050 will require adaptive and new methods of farming the worlds existing and upcoming crops. Genetic insights into the inner workings of many of these crops will be invaluable for new breeding strategies such as marker assisted breeding and genetic modification. The study of polyploidy and the aftereffects can confer more and advanced knowledge into the processes that underlie many current breeding techniques such as hybrid vigour, a phenomenon that is currently still largely unexplained (Schnable and Springer, 2013). The underpinnings of C4, studied in this thesis, may provide a largely unexplored path towards C4 rice which could hypothetically increase rice yield efficiency by as much as 50% (Hibberd et al., 2008). Thus the framework and trait studies presented here fit well within a knowledge based improvement of global food security for all people on Earth.

REFERENCES

ACKERMAN, C.M., Q. YU, S. KIM, R.E. PAULL, P.H. MOORE, and R. MING. 2008. B-class MADS-box genes in trioecious papaya: two paleoAP3 paralogs, CpTM6-1 and CpTM6-2, and a PI ortholog CpPI. *Planta* 227: 741–53.

AHMED, Z.F., F.M. HAMMOUNDA, and M.M. SEIT ET NASR. 1972. Naturally occurring glucosinolates with special reference to those of the family Capparidaceae. *Planta Medica* 21: 35 – 60.

AJAIYEOBA, E.O. 2000. Phytochemical and Antimicrobial Studies of *Gynandropsis Gynandra* and *Buchholzia Coriaceae* Extracts. *African Journal of Biomedical Research* 3: 161 – 165.

ALTENHOFF, A.M., and C. DESSIMOZ. 2009. Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. *PLoS Computational Biology* 5: e1000262.

ALTSCHUL, S.F., W. GISH, W. MILLER, E.W. MYERS, and D.J. LIPMAN. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

ALTSCHUL, S.F., T.L. MADDEN, A.A. SCHÄFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D.J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25: 3389–3402.

ALVAREZ, L., W. ALVAREZ, F. ASARO, and H. MICHEL. 1979. Extraterrestrial cause for the Cretaceous-Tertiary extinction: Experiment and theory. *In* Applications of Space Developments: Selected Papers from the XXXI International Astronautical Congress, Tokyo, 21 — 28 September 1980, 241–271. Elsevier.

Anon. Cleome gynandra | African Orphan Crops Consortium. Available at: http://africanorphancrops.org/cleome-gynandra/ [Accessed August 25, 2016].

ARRIGO, N., and M.S. BARKER. 2012. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology* 15: 140–146.

ASHBURNER, M., C.A. BALL, J.A. BLAKE, D. BOTSTEIN, H. BUTLER, J.M. CHERRY, A.P. DAVIS, ET AL. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25: 25–29.

BARKER, M.S., K.M. DLUGOSCH, L. DINH, R.S. CHALLA, N.C. KANE, M.G. KING, and L.H. RIESEBERG. 2010. EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics. *Evolutionary Bioinformatics* 6: 143–149.

BARKER, M.S., H. VOGEL, and M.E. SCHRANZ. 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.

BARTLETT, M.E., and C.D. SPECHT. 2010. Evidence for the involvement of GLOBOSA-like gene duplications and expression divergence in the evolution of floral morphology in the Zingiberales. *New Phytologist* 187: 521–541.

BAUCOM, R.S., J.C. ESTILL, C. CHAPARRO, N. UPSHAW, A. JOGI, J.-M. DERAGON, R.P. WESTERMAN, ET AL. 2009. Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. *PLoS Genetics* 5: e1000732.

BENNETT, M.D., I.J. LEITCH, H.J. PRICE, and J.S. JOHNSTON. 2003. Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) Using Flow Cytometry Show Genome Size in *Arabidopsis* to be ~157 Mb and thus ~25 % Larger than the Arabidopsis Genome Initiative Estimate of ~125 Mb. *Annals of Botany* 91: 547–557.

VAN DEN BERGH, E., J.A. HOFBERGER, and M.E. SCHRANZ. 2016. Flower power and the mustard bomb: Comparative analysis of gene and genome duplications in glucosinolate biosynthetic pathway evolution in Cleomaceae and Brassicaceae. *American Journal of Botany* 103: 1212–1222.

VAN DEN BERGH, E., C. KÜLAHOGLU, A. BRÄUTIGAM, J.M. HIBBERD, A.P.M. WEBER, X.-G. ZHU, and M. ERIC SCHRANZ. 2014. Gene and genome duplications and the origin of C4 photosynthesis: Birth of a trait in the Cleomaceae. *Current Plant Biology* 1: 2–9.

BESTMANN, H.J., L. WINKLER, and O. VON HELVERSEN. 1997. Headspace analysis of volatile flower scent constituents of bat-pollinated plants. *Phytochemistry* 46: 1169–1172.

BIRCHLER, J.A., and R.A. VEITIA. 2012. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences* 109: 14746–14753.

BIRCHLER, J.A., and R.A. VEITIA. 2014. The Gene Balance Hypothesis: Dosage Effects in Plants. *In* C. Spillane, and C. P. McKeown [eds.], Plant Epigenetics and Epigenomics: Methods and Protocols, 25–32. Humana Press, Totowa, New Jersey, USA.

BIRCHLER, J.A., and R.A. VEITIA. 2007. The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics. *The Plant Cell* 19: 395–402.

BIRNEY, E., M. CLAMP, and R. DURBIN. 2004. GeneWise and genomewise. *Genome research* 14: 988–995.

BLANC, G., K. HOKAMP, and K.H. WOLFE. 2003. A Recent Polyploidy Superimposed on Older Large-Scale Duplications in the Arabidopsis Genome. *Genome Research* 13: 137–144.

BLANC, G., and K.H. WOLFE. 2004. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *The Plant Cell* 16: 1667–1678.

BOECKMANN, B., A. BAIROCH, R. APWEILER, M.C. BLATTER, A. ESTREICHER, E. GASTEIGER, M.J. MARTIN, ET AL. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* 31: 365–370.

BOWERS, J.E., B.A. CHAPMAN, J. RONG, and A.H. PATERSON. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438.

BOYES, D.C., M.E. NASRALLAH, J. VREBALOV, and J.B. NASRALLAH. 1997. The self-incompatibility (S) haplotypes of *Brassica* contain highly divergent and rearranged sequences of ancient origin. *Plant Cell* 9: 237–247.

BRADBURY, P.J., Z. ZHANG, D.E. KROON, T.M. CASSTEVENS, Y. RAMDOSS, and E.S. BUCKLER. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23: 2633–2635.

BRADSHAW, H.D., and D.W. SCHEMSKE. 2003. Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. *Nature* 426: 176–178.

BRÄUTIGAM, A., K. KAJALA, J. WULLENWEBER, M. SOMMER, D. GAGNEUL, K.L. WEBER, K.M. CARR, ET AL. 2011. An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *Plant Physiology* 155: 142–156.

BROMAN, K.W., H. WU, S. SEN, and G.A. CHURCHILL. 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics (Oxford, England)* 19: 889–890.

BROWN, N.J., C.A. NEWELL, S. STANLEY, J.E. CHEN, A.J. PERRIN, K. KAJALA, and J.M. HIBBERD. 2011. Independent and parallel recruitment of preexisting mechanisms underlying C4 photosynthesis. *Science* 331: 1436–1439.

BROWN, N.J., K. PARSLEY, and J.M. HIBBERD. 2005. The future of C4 research – maize, *Flaveria* or *Cleome*? *Trends in Plant Science* 10: 215–221.

BUSCH, A., S. HORN, A. MUHLHAUSEN, K. MUMMENHOFF, and S. ZACHGO. 2012. Corolla monosymmetry: evolution of a morphological novelty in the Brassicaceae family. *Molecular Biology and Evolution* 29: 1241–54.

BUSCH, A., and S. ZACHGO. 2007. Control of corolla monosymmetry in the Brassicaceae *Iberis amara*. *Proceedings of the National Academy of Sciences of the United States of America* 104: 16714–9.

BUSCH, A., and S. ZACHGO. 2009. Flower symmetry evolution: towards understanding the abominable mystery of angiosperm radiation. *Bioessays* 31: 1181–90.

BUSTIN, S.A., V. BENES, J.A. GARSON, J. HELLEMANS, J. HUGGETT, M. KUBISTA, R. MUELLER, ET AL. 2009. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry* 55: 611–622.

CAMACHO, C., G. COULOURIS, V. AVAGYAN, N. MA, J. PAPADOPOULOS, K. BEALER, and T.L. MADDEN. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.

CAMPANELLA, J.J., J.V. SMALLEY, and M.E. DEMPSEY. 2014. A phylogenetic examination of the primary anthocyanin production pathway of the Plantae. *Botanical Studies* 55: 10.

CANE, J.H.; 2008. Breeding biologies, seed production and species-rich bee guilds of *Cleome lutea* and *Cleome serrulata* (Cleomaceae). *Plant Species Biology* 23: 152–158.

CANNON, S.B., A. MITRA, A. BAUMGARTEN, N.D. YOUNG, and G. MAY. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC plant biology* 4: 10.

CARDENAS, P.D., H.A. GAJARDO, T. HUEBERT, I.A. PARKIN, F.L. INIGUEZ-LUY, and M.L. FEDERICO. 2012. Retention of triplicated phytoene synthase (PSY) genes in *Brassica napus L.* and its diploid progenitors during the evolution of the Brassiceae. *Theoretical and Applied Genetics* 124: 1215–1228.

CARDINAL-MCTEAGUE, W.M., K.J. SYTSMA, and J.C. HALL. 2016. Biogeography and diversification of Brassicales: A 103 million year tale. *Molecular Phylogenetics and Evolution* 99: 204–224.

CAREY, C.C., J.T. STRAHLE, D.A. SELINGER, and V.L. CHANDLER. 2004. Mutations in the pale aleurone color1 Regulatory Gene of the *Zea mays* Anthocyanin Pathway Have Distinct Phenotypes Relative to the Functionally Similar TRANSPARENT TESTA GLABRA1 Gene in Arabidopsis thaliana. *The Plant Cell* 16: 450–464.

CATLING, P.M., and V.R. BROWNELL. 1997. Use of *Cleome spinosa* as a larval foodplant by *Pieris rapae*. *Holarctic Lepidoptera* 4: 37.

CHENG, S., E. VAN DEN BERGH, P. ZENG, X. ZHONG, J. XU, X. LIU, J. HOFBERGER, ET AL. 2013. The *Tarenaya hassleriana* Genome Provides Insight into Reproductive Trait and Genome Evolution of Crucifers. *The Plant Cell* 25: 2813–2830.

CHEN, K., B. FAN, L. DU, and Z. CHEN. 2004. Activation of hypersensitive cell death by pathogen-induced receptor-like protein kinases from Arabidopsis. *Plant Mol Biol* 56: 271–83.

CHEN, N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4: Unit 4 10.

CHEN, S., E. GLAWISCHNIG, K. JØRGENSEN, P. NAUR, B. JØRGENSEN, C.-E. OLSEN, C.H. HANSEN, ET AL. 2003. CYP79F1 and CYP79F2 have distinct functions in the biosynthesis of aliphatic glucosinolates in Arabidopsis. *The Plant Journal* 33: 923–937.

COATE, J.E., J.A. SCHLUETER, A.M. WHALEY, and J.J. DOYLE. 2011. Comparative Evolution of Photosynthetic Genes in Response to Polyploid and Nonpolyploid Duplication. *Plant Physiology* 155: 2081–2095.

COLOMBO, M., V. BRAMBILLA, R. MARCHESELLI, E. CAPORALI, M.M. KATER, and L. COLOMBO. 2010. A new role for the SHATTERPROOF genes during Arabidopsis gynoecium development. *Developmental Biology* 337: 294–302.

COMAI, L. 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* 6: 836–846.

CONANT, G.C., J.A. BIRCHLER, and J.C. PIRES. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology* 19: 91–98.

CRUDEN, R.W., and R.M. LLOYD. 1995. Embryophytes have Equivalent Sexual Phenotypes and Breeding Systems: Why Not a Common Terminology to Describe Them? *American Journal of Botany* 82: 816–825.

VAN DAM, N.M., T.O.G. TYTGAT, and J.A. KIRKEGAARD. 2008. Root and shoot glucosinolates: a comparison of their diversity, function and interactions in natural and managed ecosystems. *Phytochemistry Reviews* 8: 171–186.

DASSANAYAKE, M., D.-H. OH, J.S. HAAS, A. HERNANDEZ, H. HONG, S. ALI, D.-J. YUN, ET AL. 2011. The genome of the extremophile crucifer *Thellungiella parvula*. *Nature Genetics* 43: 913–918.

DEHAL, P., and J.L. BOORE. 2005. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLOS Biology* 3: e314.

DIEPENBROCK, W. 2000. Yield analysis of winter oilseed rape (*Brassica napus L.*): a review. *Field Crops Research* 67: 35–49.

DRUMMOND, A.J., S.Y.W. HO, M.J. PHILLIPS, and A. RAMBAUT. 2006. Relaxed Phylogenetics and Dating with Confidence. *PLOS Biol* 4: e88.

DWYER, K.G., M.K. KANDASAMY, D.I. MAHOSKY, J. ACCIAI, B.I. KUDISH, J.E. MILLER, M.E. NASRALLAH, and J.B. NASRALLAH. 1994. A superfamily of S locus-related sequences in Arabidopsis: diverse structures and expression patterns. *Plant Cell* 6: 1829–43.

EDGAR, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

EDGAR, R.C., and E.W. MYERS. 2005. PILER: identification and classification of genomic repeats. *Bioinformatics* 21: I152–I158.

EDGER, P.P., H.M. HEIDEL-FISCHER, M. BEKAERT, J. ROTA, G. GLÖCKNER, A.E. PLATTS, D.G. HECKEL, ET AL. 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences* 112: 8362–8366.

EDGER, P.P., and J.C. PIRES. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research* 17: 699–717.

EDWARDS, G.E., V.R. FRANCESCHI, and E.V. VOZNESENSKAYA. 2004. Single-cell C4 photosynthesis versus the dual-cell (Kranz) paradigm. *Annual Review of Plant Biology* 55: 173–196.

EHLERINGER, J.R., T.E. CERLING, and B.R. HELLIKER. 1997. C4 photosynthesis, atmospheric CO2, and climate. *Oecologia* 112: 285–299.

EISEN, M.B., P.T. SPELLMAN, P.O. BROWN, and D. BOTSTEIN. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95: 14863–14868.

ELSHIRE, R.J., J.C. GLAUBITZ, Q. SUN, J.A. POLAND, K. KAWAMOTO, E.S. BUCKLER, and S.E. MITCHELL. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE* 6: e19379.

ENDRESS, P.K. 1999. Symmetry in Flowers: Diversity and Evolution. *International Journal of Plant Sciences* 160: S3–S23.

ENGLER, F.W., J. HATFIELD, W. NELSON, and C.A. SODERLUND. 2003. Locating sequence on FPC maps and selecting a minimal tiling path. *Genome research* 13: 2152–2163.

FAHEY, J.W., A.T. ZALCMANN, and P. TALALAY. 2001. The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry* 56: 5–51.

FARZAD, M., R. GRIESBACH, and M.R. WEISS. 2002. Floral color change in *Viola cornuta* L. (Violaceae): a model system to study regulation of anthocyanin production. *Plant Science* 162: 225–231.

FAWCETT, J.A., S. MAERE, and Y.V. DE PEER. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proceedings of the National Academy of Sciences* 106: 5737–5742.

FEENSTRA, B., I.M. SKOVGAARD, and K.W. BROMAN. 2006. Mapping Quantitative Trait Loci by an Extension of the Haley–Knott Regression Method Using Estimating Equations. *Genetics* 173: 2269–2282.

FEODOROVA, T.A., E.V. VOZNESENSKAYA, G.E. EDWARDS, and E.H. ROALSON. 2010. Biogeographic Patterns of Diversification and the Origins of C4 in *Cleome* (Cleomaceae). *Systematic Botany* 35: 811–826.

FLAGEL, L.E., and J.F. WENDEL. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist* 183: 557–564.

FLEMING, T.H., C. GEISELMAN, and W.J. KRESS. 2009. The evolution of bat pollination: a phylogenetic perspective. *Annals of Botany* 104: 1017–1043.

FORCE, A., M. LYNCH, F.B. PICKETT, A. AMORES, Y. YAN, and J. POSTLETHWAIT. 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics* 151: 1531–1545.

FRANZKE, A., M.A. LYSAK, I.A. AL-SHEHBAZ, M.A. KOCH, and K. MUMMENHOFF. 2011. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends in Plant Science* 16: 108–116.

FREELING, M. 2009. Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annual Review of Plant Biology* 60: 433–453.

FREELING, M., E. LYONS, B. PEDERSEN, M. ALAM, R. MING, and D. LISCH. 2008. Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Research* 18: 1924–1937.

FREELING, M., and B.C. THOMAS. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research* 16: 805–814.

GARSMEUR, O., J.C. SCHNABLE, A. ALMEIDA, C. JOURDA, A. D'HONT, and M. FREELING. 2014. Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology and Evolution* 31: 448–454.

GIBSON, T.J., and J. SPRING. 1998. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends in genetics: TIG* 14: 46–49; discussion 49–50.

GIGORD, L.D.B., M.R. MACNAIR, and A. SMITHSON. 2001. Negative frequency-dependent selection maintains a dramatic flower color polymorphism in the rewardless orchid *Dactylorhiza sambucina* (L.) Soò. *Proceedings of the National Academy of Sciences of the United States of America* 98: 6253–6255.

GLASAUER, S.M.K., and S.C.F. NEUHAUSS. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular genetics and genomics: MGG* 289: 1045–1060.

GODFRAY, H.C.J., J.R. BEDDINGTON, I.R. CRUTE, L. HADDAD, D. LAWRENCE, J.F. MUIR, J. PRETTY, ET AL. 2010. Food Security: The Challenge of Feeding 9 Billion People. *Science* 327: 812–818.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11: 725–736.

GOWIK, U., J. BURSCHEIDT, M. AKYILDIZ, U. SCHLUE, M. KOCZOR, M. STREUBEL, and P. WESTHOFF. 2004. cis-Regulatory elements for mesophyll-specific gene expression in the C4 plant *Flaveria trinervia*, the promoter of the C4 phosphoenolpyruvate carboxylase gene. *The Plant Cell Online* 16: 1077–1090.

GOWIK, U., and P. WESTHOFF. 2011. The path from C3 to C4 photosynthesis. *Plant Physiology* 155: 56–63.

GROTEWOLD, E. 2006. The Genetics and Biochemistry of Floral Pigments. *Annual Review of Plant Biology* 57: 761–780.

GUO, Y.L., X. ZHAO, C. LANZ, and D. WEIGEL. 2011. Evolution of the S-Locus Region in Arabidopsis Relatives. *Plant Physiology* 157: 937–946.

HAIBAO TANG, VIVEK KRISHNAKUMAR, and JINGPING LI. 2015. jcvi: JCVI utility libraries. Available at: http://dx.doi.org/10.5281/zenodo.31631 [Accessed October 5, 2015].

HALKIER, B.A., and J. GERSHENZON. 2006. Biology and Biochemistry of Glucosinolates. *Annual Review of Plant Biology* 57: 303–333.

HALL, J.C. 2008. Systematics of Capparaceae and Cleomaceae: an evaluation of the generic delimitations of *Capparis* and *Cleome* using plastid DNA sequence data. *Botany* 86: 682–696.

HALL, J.C., K.J. SYTSMA, and H.H. ILTIS. 2002. Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data. *American Journal of Botany* 89: 1826–42.

HANNAN, G.L. 1981. Flower Color Polymorphism and Pollination Biology of *Platystemon californicus* Benth. (Papaveraceae). *American Journal of Botany* 68: 233–243.

HAUDRY, A., A.E. PLATTS, E. VELLO, D.R. HOEN, M. LECLERCQ, R.J. WILLIAMSON, E. FORCZEK, ET AL. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics* 45: 891–898.

HEIDEL, A.J., M.J. CLAUSS, J. KROYMANN, O. SAVOLAINEN, and T. MITCHELL-OLDS. 2006. Natural Variation in MAM Within and Between Populations of *Arabidopsis lyrata* Determines Glucosinolate Phenotype. *Genetics* 173: 1629–1636.

VON HELVERSEN, O., L. WINKLER, and H.J. BESTMANN. 2000. Sulphur-containing "perfumes" attract flower-visiting bats. *Journal of Comparative Physiology A* 186: 143–153.

HIBBERD, J.M., and S. COVSHOFF. 2010. The regulation of gene expression required for C4 photosynthesis. *Annual review of plant biology* 61: 181–207.

HIBBERD, J.M., J.E. SHEEHY, and J.A. LANGDALE. 2008. Using C4 photosynthesis to increase the yield of rice—rationale and feasibility. *Current Opinion in Plant Biology* 11: 228–231.

HOBALLAH, M.E., T. GÜBITZ, J. STUURMAN, L. BROGER, M. BARONE, T. MANDEL, A. DELL'OLIVO, ET AL. 2007. Single Gene–Mediated Shift in Pollinator Attraction in *Petunia*. *The Plant Cell* 19: 779–790.

HOFBERGER, J.A., E. LYONS, P.P. EDGER, J.C. PIRES, and M.E. SCHRANZ. 2013. Whole Genome and Tandem Duplicate Retention Facilitated Glucosinolate Pathway Diversification in the Mustard Family. *Genome Biology and Evolution* 5: 2155–2173.

HOFBERGER, J.A., D.L. NSIBO, F. GOVERS, K. BOUWMEESTER, and M.E. SCHRANZ. 2015. A Complex Interplay of Tandem- and Whole-Genome Duplication Drives Expansion of the L-Type Lectin Receptor Kinase Gene Family in the Brassicaceae. *Genome Biology and Evolution* 7: 720–734.

HOFBERGER, J.A., A.M. RAMIREZ, E. VAN DEN BERGH, X. ZHU, H.J. BOUWMEESTER, R.C. SCHUURINK, and M.E. SCHRANZ. 2015. Large-Scale Evolutionary Analysis of Genes and Supergene Clusters from Terpenoid Modular Pathways Provides Insights into Metabolic Diversification in Flowering Plants. *PLOS ONE* 10: e0128808.

HOFBERGER, J.A., B. ZHOU, H. TANG, J.D. JONES, and M.E. SCHRANZ. 2014. A novel approach for multi-domain and multi-gene family identification provides insights into evolutionary dynamics of disease resistance genes in core eudicot plants. *BMC Genomics* 15: 966.

HOHMANN, N., E.M. WOLF, M.A. LYSAK, and M.A. KOCH. 2015. A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History. *The Plant Cell*tpc.15.00482.

HOLLOWAY, J.D. 1989. The Moths of Borneo Family Noctuidae Trifine Subfamilies Noctuinae Heliothinae Hadeninae Acronictinae Amphipyrinae Agaristinae. *Malayan Nature Journal* 42: 57–226.

HOLTON, T.A., and E.C. CORNISH. 1995. Genetics and Biochemistry of Anthocyanin Biosynthesis. *The Plant Cell* 7: 1071–1083.

D'HONT, A., F. DENOEUD, J.-M. AURY, F.-C. BAURENS, F. CARREEL, O. GARSMEUR, B. NOEL, ET AL. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488: 213–217.

VAN DER HOOFT, J.J.J., M. AKERMI, F.Y. ÜNLÜ, V. MIHALEVA, V.G. ROLDAN, R.J. BINO, R.C.H. DE VOS, and J. VERVOORT. 2012. Structural Annotation and Elucidation of Conjugated Phenolic Compounds in Black, Green, and White Tea Extracts. *Journal of Agricultural and Food Chemistry* 60: 8841–8850.

HUNTER, S., P. JONES, A. MITCHELL, R. APWEILER, T.K. ATTWOOD, A. BATEMAN, T. BERNARD, ET AL. 2012. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic acids research* 40: D306–12.

HU, T.T., P. PATTYN, E.G. BAKKER, J. CAO, J.F. CHENG, R.M. CLARK, N. FAHLGREN, ET AL. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics* 43: 476–481.

ILTIS, H.H., and T.S. COCHRANE. 2007. Studies in the Cleomaceae V: A New Genus and Ten New Combinations for the Flora of North America. *Novon* 17: 447–451.

ILTIS, H.H., J.C. HALL, T.S. COCHRANE, and K.J. SYTSMA. 2011. Studies in the Cleomaceae I. On the Separate Recognition of Capparaceae, Cleomaceae, and Brassicaceae. *Annals of the Missouri Botanical Garden* 98: 28–36.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE. 2014. Climate Change 2014–Impacts, Adaptation and Vulnerability: Regional Aspects. Cambridge University Press.

JABBOUR, F., S. NADOT, and C. DAMERVAL. 2009. Evolution of floral symmetry: a state of the art. *Comptes Rendus Biologies* 332: 219–231.

JAILLON, O., J.M. AURY, B. NOEL, A. POLICRITI, C. CLEPET, A. CASAGRANDE, N. CHOISNE, ET AL. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: 463–U5.

JIAO, Y., N.J. WICKETT, S. AYYAMPALAYAM, A.S. CHANDERBALI, L. LANDHERR, P.E. RALPH, L.P. TOMSHO, ET AL. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

JOHNSON, S.D., M.E. GRIFFITHS, C.I. PETER, and M.J. LAWES. 2009. Pollinators, "mustard oil" volatiles, and fruit production in flowers of the dioecious tree *Drypetes natalensis* (Putranjivaceae). *American Journal of Botany* 96: 2080–2086.

JORDHEIM, M., Ø.M. ANDERSEN, C. NOZZOLILLO, and V.T. AMIGUET. 2009. Acylated anthocyanins in inflorescence of spider flower (*Cleome hassleriana*). *Phytochemistry* 70: 740–745.

KAGALE, S., S.J. ROBINSON, J. NIXON, R. XIAO, T. HUEBERT, J. CONDIE, D. KESSLER, ET AL. 2014. Polyploid Evolution of the Brassicaceae during the Cenozoic Era. *The Plant Cell* 26: 2777–2791.

KAJALA, K., N.J. BROWN, B.P. WILLIAMS, P. BORRILL, L.E. TAYLOR, and J.M. HIBBERD. 2012. Multiple Arabidopsis genes primed for recruitment into C4 photosynthesis. *The Plant Journal* 69: 47–56.

KANEHISA, M., and S. GOTO. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* 28: 27–30.

KATOH, K., and M.C. FRITH. 2012. Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics* 28: 3144–3146.

KATOH, K., K. MISAWA, K. KUMA, and T. MIYATA. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066.

KELLER, G. 2014. Deccan volcanism, the Chicxulub impact, and the end-Cretaceous mass extinction: Coincidence? Cause and effect? *Geological Society of America Special Papers* 505: SPE505–03.

KENT, W.J. 2002. BLAT—The BLAST-Like Alignment Tool. *Genome Research* 12: 656–664.

KIEŁBASA, S.M., R. WAN, K. SATO, P. HORTON, and M.C. FRITH. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Research* 21: 487–493.

KLIEBENSTEIN, D.J., J. KROYMANN, P. BROWN, A. FIGUTH, D. PEDERSEN, J. GERSHENZON, and T. MITCHELL-OLDS. 2001. Genetic Control of Natural Variation in Arabidopsis Glucosinolate Accumulation. *Plant Physiology* 126: 811–825.

KNILL, T., J. SCHUSTER, M. REICHELT, J. GERSHENZON, and S. BINDER. 2008. Arabidopsis Branched-Chain Aminotransferase 3 Functions in Both Amino Acid and Glucosinolate Biosynthesis. *Plant Physiology* 146: 1028–1039.

KOENIG, D., and D. WEIGEL. 2015. Beyond the thale: comparative genomics and genetics of Arabidopsis relatives. *Nature Reviews. Genetics* 16: 285–298.

KOES, R., W. VERWEIJ, and F. QUATTROCCHIO. 2005. Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. *Trends in Plant Science* 10: 236–242.

KOORNNEEF, M., and D. MEINKE. 2010. The development of Arabidopsis as a model plant. *Plant Journal* 61: 909–921.

KOTEYEVA, N.K., E.V. VOZNESENSKAYA, E.H. ROALSON, and G.E. EDWARDS. 2011. Diversity in forms of C4 in the genus *Cleome* (Cleomaceae). *Annals of Botany* 107: 269–283.

KRAMER, E.M., L. HOLAPPA, B. GOULD, M.A. JARAMILLO, D. SETNIKOV, and P.M. SANTIAGO. 2007. Elaboration of B gene function to include the identity of novel floral organs in the lower eudicot *Aquilegia*. *Plant Cell* 19: 750–66.

KÜLAHOGLU, C., A.K. DENTON, M. SOMMER, J. MAß, S. SCHLIESKY, T.J. WROBEL, B. BERCKMANS, ET AL. 2014. Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules between Two Cleomaceae C3 and C4 Plant Species. *The Plant Cell* 26: 3243–3260.

KU, M.S., J. WU, Z. DAI, R.A. SCOTT, C. CHU, and G.E. EDWARDS. 1991. Photosynthetic and photorespiratory characteristics of Flaveria species. *Plant Physiology* 96: 518–528.

LADD, D.L. 1983. Genetics of flower color in spider flower, *Cleome hasslerana* Chod. Thesis. Kansas State University. Available at: http://krex.k-state.edu/dspace/handle/2097/12969 [Accessed August 25, 2016].

LADIZINSKY, G. 1998. Plant evolution under domestication. Kluwer Academic Publishers, Dordrecht; New York.

LANDRY, J.-F., and P.D. HEBERT. 2013. *Plutella australiana* (Lepidoptera, Plutellidae), an overlooked diamondback moth revealed by DNA barcodes. *ZooKeys*43–63.

LANGMEAD, B. 2010. Aligning short sequencing reads with Bowtie. *Current Protocols in Bioinformatics* Chapter 11: Unit 11 7.

LAZZERI, L., L.M. MANICI, O. LEONI, and S. PALMIERI. 1998. Soil-Borne Phytopathogenic Fungi Control by <i>Cleome Hassleriana <i>Green Manure. *Acta Horticulturae*53–62.

LI, F., G. FAN, C. LU, G. XIAO, C. ZOU, R.J. KOHEL, Z. MA, ET AL. 2015. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nature Biotechnology* 33: 524–530.

LI, H., and R. DURBIN. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25: 1754–1760.

LI, H., P. VIKRAM, R.P. SINGH, A. KILIAN, J. CARLING, J. SONG, J.A. BURGUENO-FERREIRA, ET AL. 2015. A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. *BMC Genomics* 16: 216.

LI, L., C.J. STOECKERT, and D.S. ROOS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 13: 2178–89.

LI, R.Q., H.M. ZHU, J. RUAN, W.B. QIAN, X.D. FANG, Z.B. SHI, Y.R. LI, ET AL. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* 20: 265–272.

LITT, A., and E.M. KRAMER. 2010. The ABC model and the diversification of floral organ identity. *Seminars in Cell & Developmental Biology* 21: 129–137.

LI, W.H., C.I. WU, and C.C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* 2: 150–174.

LI, X.N., N. RAMCHIARY, V. DHANDAPANI, S.R. CHOI, Y. HUR, I.S. NOU, M.K. YOON, and Y.P. LIM. 2013. Quantitative Trait Loci Mapping in *Brassica rapa* Revealed the Structural and Functional Conservation of Genetic Loci Governing Morphological and Yield Component Traits in the A, B, and C Subgenomes of Brassica Species. *DNA Research* 20: 1–16.

LOHAUS, R., and Y. VAN DE PEER. 2016. Of dups and dinos: evolution at the K/Pg boundary. *Current Opinion in Plant Biology* 30: 62–69.

LOWE, T.M., and S.R. EDDY. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* 25: 955–964.

LYONS, E., and M. FREELING. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal* 53: 661–673.

LYSAK, M.A., T. MANDÁKOVÁ, and M.E. SCHRANZ. 2016. Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Current Opinion in Plant Biology* 30: 108–115.

MACHADO, I., A. CRISTINA LOPES, A. VALENTINA LEITE, and C. VIRGÍNIADE BRITO NEVES. 2006. *Cleome spinosa* (Capparaceae): polygamodioecy and pollination by bats in urban and Caatinga areas, northeastern Brazil. *Botanische Jahrbücher für Systematik, Pflanzengeschichte und Pflanzengeographie* 127: 69–82.

MAJOROS, W.H., M. PERTEA, and S.L. SALZBERG. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20: 2878–2879.

MALACARNE, G., M. PERAZZOLLI, A. CESTARO, L. STERCK, P. FONTANA, Y. VAN DE PEER, R. VIOLA, ET AL. 2012. Deconstruction of the (Paleo)Polyploid Grapevine Genome Based on the Analysis of Transposition Events Involving NBS Resistance Genes C. A. Ouzounis [ed.],. *PLoS ONE* 7: e29762.

MARQUARD, R.D., and R. STEINBACK. 2009. A Model Plant for a Biology Curriculum: Spider Flower (Cleome hasslerana L.). *The American Biology Teacher* 71: 235–244.

MARSHALL, D.M., R. MUHAIDAT, N.J. BROWN, Z. LIU, S. STANLEY, H. GRIFFITHS, R.F. SAGE, and J.M. HIBBERD. 2007. *Cleome*, a genus closely related to Arabidopsis, contains species spanning a developmental progression from C3 to C4 photosynthesis. *The Plant Journal* 51: 886–896.

MARSHALL, O.J. 2004. PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics* 20: 2471–2472.

MARTENS, S., A. PREUSS, and U. MATERN. 2010. Multifunctional flavonoid dioxygenases: flavonol and anthocyanin biosynthesis in *Arabidopsis thaliana* L. *Phytochemistry* 71: 1040–1049.

MARTIS, M.M., R. ZHOU, G. HASENEYER, T. SCHMUTZER, J. VRÁNA, M. KUBALÁKOVÁ, S. KÖNIG, ET AL. 2013. Reticulate evolution of the rye genome. *The Plant Cell* 25: 3685–3698.

MATSUURA, H., S. TAKEISHI, N. KIATOKA, C. SATO, K. SUEDA, C. MASUTA, and K. NABETA. 2012. Transportation of de novo synthesized jasmonoyl isoleucine in tomato. *Phytochemistry* 83: 25–33.

MAYROSE, I., S.H. ZHAN, C.J. ROTHFELS, N. ARRIGO, M.S. BARKER, L.H. RIESEBERG, and S.P. OTTO. 2015. Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al. (2014). *New Phytologist* 206: 27–35.

MAYROSE, I., S.H. ZHAN, C.J. ROTHFELS, K. MAGNUSON-FORD, M.S. BARKER, L.H. RIESEBERG, and S.P. OTTO. 2011. Recently Formed Polyploid Plants Diversify at Lower Rates. *Science*1207205.

MCKOWN, A.D., and N.G. DENGLER. 2014. Vein patterning and evolution in C4 plants. *Botany* 88: 775–786.

MCMILLAN, L.E.M., and A.C.R. MARTIN. 2008. Automatically extracting functionally equivalent proteins from SwissProt. *BMC bioinformatics* 9: .

MEINKE, D.W., J.M. CHERRY, C. DEAN, S.D. ROUNSLEY, and M. KOORNNEEF. 1998. *Arabidopsis thaliana*: A Model Plant for Genome Analysis. *Science* 282: 662–682.

MING, R., S. HOU, Y. FENG, Q. YU, A. DIONNE-LAPORTE, J.H. SAW, P. SENIN, ET AL. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452: 991–996.

MITCHELL-OLDS, T., and J. SCHMITT. 2006. Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature* 441: 947–952.

MONDRAGON-PALOMINO, M., and G. THEISSEN. 2009. Why are orchid flowers so diverse? Reduction of evolutionary constraints by paralogues of class B floral homeotic genes. *Annals of Botany* 104: 583–594.

MONSON, R.K. 2003. Gene Duplication, Neofunctionalization, and the Evolution of C4 Photosynthesis. *International Journal of Plant Sciences* 164: S43–S54.

MONSON, R.K.S. 1999. The Origins of C4 Genes and Evolutionary Pattern in the C4 Metabolic Phenotype. *In* C4 Plant Biology, 377–410. Academic Press, San Diego.

MURAT, F., Y. VAN DE PEER, and J. SALSE. 2012. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biology and Evolution* 4: 917–928.

NASRALLAH, M.E., P. LIU, S. SHERMAN-BROYLES, N.A. BOGGS, and J.B. NASRALLAH. 2004. Natural variation in expression of self-incompatibility in Arabidopsis thaliana: implications for the evolution of selfing. *Proceedings of the National Academy of Sciences of the United States of America* 101: 16070–16074.

NAWROCKI, E.P., D.L. KOLBE, and S.R. EDDY. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25: 1335–1337.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3: 418–426.

NOZZOLILLO, C., V.T. AMIGUET, A.C. BILY, C.S. HARRIS, A. SALEEM, Ø.M. ANDERSEN, and M. JORDHEIM. 2010. Novel aspects of the flowers and floral pigmentation of two Cleome species (Cleomaceae), C. hassleriana and C. serrulata. *Biochemical Systematics and Ecology* 38: 361–369.

VAN OEVEREN, J., M. DE RUITER, T. JESSE, H. VAN DER POEL, J.F. TANG, F. YALCIN, A. JANSSEN, ET AL. 2011. Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome research* 21: 618–625.

OHASHI, K., T.T. MAKINO, and K. ARIKAWA. 2015. Floral colour change in the eyes of pollinators: testing possible constraints and correlated evolution. *Functional Ecology* 29: 1144–1155.

OHNO, S. 1973. Ancient Linkage Groups and Frozen Accidents. *Nature* 244: 259–262.

OHNO, S. 1970. The Creation of a New Gene from a Redundant Duplicate of an Old Gene. *In* Evolution by Gene Duplication, 71–82. Springer Berlin Heidelberg, Heidelberg, Germany.

OHNO, S., U. WOLF, and N.B. ATKIN. 1968. Evolution from fish to mammals by gene duplication. *Hereditas* 59: 169–187.

PASTUGLIA, M., R. SWARUP, A. ROCHER, P. SAINDRENAN, D. ROBY, C. DUMAS, and J.M. COCK. 2002. Comparison of the expression patterns of two small gene families of S gene family receptor kinase genes during the defence response in *Brassica oleracea* and *Arabidopsis thaliana*. *Gene* 282: 215–225.

PATCHELL, M.J., M.C. BOLTON, P. MANKOWSKI, and J.C. HALL. 2011a. Comparative Floral Development in Cleomaceae Reveals Two Distinct Pathways Leading to Monosymmetry. *International Journal of Plant Sciences* 172: 352–365.

PATCHELL, M.J., M.C. BOLTON, P. MANKOWSKI, and J.C. HALL. 2011b. Comparative Floral Development in Cleomaceae Reveals Two Distinct Pathways Leading to Monosymmetry. *International Journal of Plant Sciences* 172: 352–365.

PECINKA, A., W. FANG, M. REHMSMEIER, A.A. LEVY, and O. MITTELSTEN SCHEID. 2011. Polyploidization increases meiotic recombination frequency in Arabidopsis. *BMC Biology* 9: 24.

VAN DE PEER, Y., S. MAERE, and A. MEYER. 2009. OPINION The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* 10: 725–732.

PFALZ, M., H. VOGEL, and J. KROYMANN. 2009. The Gene Controlling the Indole Glucosinolate Modifier1 Quantitative Trait Locus Alters Indole Glucosinolate Structures and Aphid Resistance in Arabidopsis. *The Plant Cell* 21: 985–999.

PIRES, J.C., R. WING, and D. WEIGEL. 2013. Brassicales Map Alignment Project (BMAP). Available at: http://www.brassica.info/resource/sequencing/bmap.php [Accessed January 29, 2017].

PRESTON, J.C., and L.C. HILEMAN. 2012. Parallel evolution of TCP and B-class genes in Commelinaceae flower bilateral symmetry. *Evodevo* 3: 6.

PROOST, S., J. FOSTIER, D. DE WITTE, B. DHOEDT, P. DEMEESTER, Y. VAN DE PEER, and K. VANDEPOELE. 2012. i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic acids research* 40: e11.

PUNTA, M., P.C. COGGILL, R.Y. EBERHARDT, J. MISTRY, J. TATE, C. BOURSNELL, N. PANG, ET AL. 2012. The Pfam protein families database. *Nucleic acids research* 40: D290–301.

PUTNAM, N.H., M. SRIVASTAVA, U. HELLSTEN, B. DIRKS, J. CHAPMAN, A. SALAMOV, A. TERRY, ET AL. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317: 86–94.

QUATTROCCHIO, F., A. BAUDRY, L. LEPINIEC, and E. GROTEWOLD. 2006. The Regulation of Flavonoid Biosynthesis. *In* E. Grotewold [ed.], The Science of Flavonoids, 97–122. Springer New York. Available at: http://link.springer.com/chapter/10.1007/978-0-387-28822-2_4 [Accessed August 18, 2016].

RAHMAN, M., and P. MCCLEAN. 2013. Genetic Analysis on Flowering Time and Root System in *Brassica napus L. Crop Science* 53: 141–147.

RAMSAY, N.A., and B.J. GLOVER. 2005. MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. *Trends in Plant Science* 10: 63–70.

RENNY-BYFIELD, S., L. GONG, J.P. GALLAGHER, and J.F. WENDEL. 2015. Persistence of Subgenomes in Paleopolyploid Cotton after 60 My of Evolution. *Molecular Biology and Evolution* 32: 1063–1071.

RENWICK, J. A. A., and K. LOPEZ. 1999. Experience-based food consumption by larvae of Pieris rapae: addiction to glucosinolates? *Entomologia Experimentalis et Applicata* 91: 51–58.

RIACH, A.C., M.V.L. PERERA, H.V. FLORANCE, S.D. PENFIELD, and J.K. HILL. 2015. Analysis of plant leaf metabolites reveals no common response to insect herbivory by Pieris rapae in three related host-plant species. *Journal of Experimental Botany* 66: 2547–2556.

RIZZON, C., L. PONGER, and B.S. GAUT. 2006. Striking Similarities in the Genomic Distribution of Tandemly Arrayed Genes in Arabidopsis and Rice. *PLoS Computational Biology* 2: . Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1557586/ [Accessed August 29, 2016].

RODMAN, J.E., K.G. KAROL, R.A. PRICE, and K.J. SYTSMA. 1996. Molecules, Morphology, and Dahlgren's Expanded Order Capparales. *Systematic Botany* 21: 289–307.

RONQUIST, F., M. TESLENKO, P. VAN DER MARK, D.L. AYRES, A. DARLING, S. HOHNA, B. LARGET, ET AL. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61: 539–42.

ROSIN, F.M., and E.M. KRAMER. 2009. Old dogs, new tricks: Regulatory evolution in conserved genetic modules leads to novel morphologies in plants. *Developmental Biology* 332: 25–35.

SAEED, A.I., V. SHAROV, J. WHITE, J. LI, W. LIANG, N. BHAGABATI, J. BRAISTED, ET AL. 2003. TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* 34: 374–378.

SAGE, R.F. 2004. The evolution of C4 photosynthesis. *New Phytologist* 161: 341–370.

SAGE, R.F., P.-A. CHRISTIN, and E.J. EDWARDS. 2011. The C4 plant lineages of planet Earth. *Journal of Experimental Botany* 62: 3155–3169.

SÁNCHEZ-ACEBO, L. 2005. A Phylogenetic Study of the New World Cleome (Brassicaceae, Cleomoideae). *Annals of the Missouri Botanical Garden* 92: 179–201.

SAWADA, Y., A. KUWAHARA, M. NAGANO, T. NARISAWA, A. SAKATA, K. SAITO, and M. YOKOTA HIRAI. 2009. Omics-Based Approaches to Methionine Side Chain Elongation in Arabidopsis: Characterization of the Genes Encoding Methylthioalkylmalate Isomerase and Methylthioalkylmalate Dehydrogenase. *Plant and Cell Physiology* 50: 1181–1190.

SCHIESTL, F.P., and S.D. JOHNSON. 2013. Pollinator-mediated evolution of floral signals. *Trends in Ecology & Evolution* 28: 307–315.

SCHMID, M., T.S. DAVISON, S.R. HENZ, U.J. PAPE, M. DEMAR, M. VINGRON, B. SCHOLKOPF, ET AL. 2005. A gene expression map of Arabidopsis thaliana development. *Nature Genetics* 37: 501–506.

SCHMITTGEN, T.D., and K.J. LIVAK. 2008. Analyzing real-time PCR data by the comparative C(T) method. *Nature Protocols* 3: 1101–1108.

SCHNABLE, P.S., and N.M. SPRINGER. 2013. Progress Toward Understanding Heterosis in Crop Plants. *http://dx.doi.org/10.1146/annurev-arplant-042110-103827*. Available at: http://www.annualreviews.org/doi/abs/10.1146/annurev-arplant-042110-103827 [Accessed January 16, 2017].

SCHRANZ, M.E., S. MOHAMMADIN, and P.P. EDGER. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Current Opinion in Plant Biology* 15: 147–153.

SCHRANZ, M.E., P.P. EDGER, J.C. PIRES, N.M. VAN DAM, and C.W. WHEAT. 2011. Comparative Genomics in the Brassicales. *In* Genetics, Genomics and Breeding of Oilseed Brassicas, Genetics, Genomics and Breeding of Crop Plants, 206–218. CRC Press, Boca Raton, Florida, USA. Available at: http://dx.doi.org/10.1201/b11406-12 [Accessed March 31, 2016].

SCHRANZ, M.E., M.A. LYSAK, and T. MITCHELL-OLDS. 2006. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends in Plant Science* 11: 535–542.

SCHRANZ, M.E., and T. MITCHELL-OLDS. 2006. Independent Ancient Polyploidy Events in the Sister Families Brassicaceae and Cleomaceae. *The Plant Cell* 18: 1152–1165.

SCHUSTER, J., T. KNILL, M. REICHELT, J. GERSHENZON, and S. BINDER. 2006. BRANCHED-CHAIN AMINOTRANSFERASE4 Is Part of the Chain Elongation Pathway in the Biosynthesis of Methionine-Derived Glucosinolates in Arabidopsis. *The Plant Cell* 18: 2664–2679.

SCHWINN, K., J. VENAIL, Y. SHANG, S. MACKAY, V. ALM, E. BUTELLI, R. OYAMA, ET AL. 2006. A Small Family of MYB-Regulatory Genes Controls Floral Pigmentation Intensity and Patterning in the Genus Antirrhinum. *The Plant Cell* 18: 831–851.

SHARMA, B., L. YANT, S.A. HODGES, and E.M. KRAMER. 2014. Understanding the development and evolution of novel floral form in Aquilegia. *Current Opinion in Plant Biology* 17: 22–27.

SHUTTLEWORTH, A., and S.D. JOHNSON. 2012. The Hemipepsis wasp-pollination system in South Africa: a comparative analysis of trait convergence in a highly specialized plant guild. *Botanical Journal of the Linnean Society* 168: 278–299.

SLATER, G.S., and E. BIRNEY. 2005. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* 6: 31.

SLEWINSKI, T.L., A.A. ANDERSON, S. PRICE, J.R. WITHEE, K. GALLAGHER, and R. TURGEON. 2014. Short-Root1 Plays a Role in the Development of Vascular Tissue and Kranz Anatomy in Maize Leaves. *Molecular Plant* 7: 1388–1392.

SLEWINSKI, T.L., A.A. ANDERSON, C. ZHANG, and R. TURGEON. 2012. Scarecrow Plays a Role in Establishing Kranz Anatomy in Maize Leaves. *Plant and Cell Physiology* 53: 2030–2037.

SMACZNIAK, C., R.G. IMMINK, J.M. MUINO, R. BLANVILLAIN, M. BUSSCHER, J. BUSSCHER-LANGE, Q.D. DINH, ET AL. 2012. Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. *Proceedings of the National Academy of Sciences of the United States of America* 109: 1560–1565.

SOBEL, J.M., and M.A. STREISFELD. 2013. Flower color as a model system for studies of plant evo-devo. *Frontiers in Plant Science* 4: 321.

SODERLUND, C., I. LONGDEN, and R. MOTT. 1997. FPC: a system for building contigs from restriction fingerprinted clones. *Computer applications in the biosciences: CABIOS* 13: 523–535.

SOLTIS, D.E., M.C. SEGOVIA-SALCEDO, I. JORDON-THADEN, L. MAJURE, N.M. MILES, E.V. MAVRODIEV, W. MEI, ET AL. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). *New Phytologist* 202: 1105–1117.

SOLTIS, D.E., and P.S. SOLTIS. 1999. Polyploidy: recurrent formation and genome evolution. *Trends in Ecology & Evolution* 14: 348–352.

SØNDERBY, I.E., F. GEU-FLORES, and B.A. HALKIER. 2010. Biosynthesis of glucosinolates – gene discovery and beyond. *Trends in Plant Science* 15: 283–290.

SPELT, C., F. QUATTROCCHIO, J.N.M. MOL, and R. KOES. 2000. anthocyanin1 of Petunia Encodes a Basic Helix-Loop-Helix Protein That Directly Activates Transcription of Structural Anthocyanin Genes. *The Plant Cell* 12: 1619–1631.

SPILLANE, C., and P.C. MCKEOWN. 2014. The Gene Balance Hypothesis: Dosage Effects in Plants - Springer. *In* Methods in Molecular Biology, Humana Press. Available at: http://link.springer.com/protocol/10.1007%2F978-1-62703-773-0_2 [Accessed October 5, 2015].

STANKE, M., and B. MORGENSTERN. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* 33: W465–W467.

STEPHENS, S.G. 1947. Cytogenetics of *Gossypium* and the Problem of the Origin of New World Cottons. *In* M. Demerec [ed.], Advances in Genetics, 431–442. Academic Press. Available at: http://www.sciencedirect.com/science/article/pii/S0065266008604915 [Accessed August 30, 2016].

STEVENS, P.F. 2001a. Angiosperm Phylogeny Website. *Angiosperm Phylogeny Website*. Available at: http://www.mobot.org/MOBOT/Research/APweb/ [Accessed August 19, 2016].

STEVENS, P.F. 2001b. Angiosperm Phylogeny Website, Version 12. Available at: http://www.mobot.org/MOBOT/research/APweb/ [Accessed October 5, 2015].

STOUT, A.B. 1923. Alternation of Sexes and Intermittent Production of Fruit in the Spider Flower (*Cleome spinosa*). *American Journal of Botany* 10: 57–66.

STREISFELD, M.A., and M.D. RAUSHER. 2009. Altered trans-Regulatory Control of Gene Expression in Multiple Anthocyanin Genes Contributes to Adaptive Flower Color Evolution in Mimulus aurantiacus. *Molecular Biology and Evolution* 26: 433–444.

SUN, J.Y., I.E. SØNDERBY, B.A. HALKIER, G. JANDER, and M. DE VOS. 2010. Non-Volatile Intact Indole Glucosinolates are Host Recognition Cues for Ovipositing *Plutella xylostella*. *Journal of Chemical Ecology* 35: 1427–1436.

SUYAMA, M., D. TORRENTS, and P. BORK. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* 34: W609–W612.

TANG, H., E. LYONS, B. PEDERSEN, J.C. SCHNABLE, A.H. PATERSON, and M. FREELING. 2011. Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* 12: 102.

THE BRASSICA RAPA GENOME SEQUENCING PROJECT CONSORTIUM, X. WANG, H. WANG, J. WANG, R. SUN, J. WU, S. LIU, ET AL. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics* 43: 1035–1039.

THE INTERNATIONAL WHEAT GENOME SEQUENCING CONSORTIUM (IWGSC). 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345: 1251788.

THEIßEN, G. 2001. Development of floral organ identity: stories from the MADS house. *Current Opinion in Plant Biology* 4: 75–85.

THE TOMATO GENOME CONSORTIUM. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485: 635–641.

THOMAS, B.C., B. PEDERSEN, and M. FREELING. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome research* 16: 934–946.

TRAPNELL, C., L. PACHTER, and S.L. SALZBERG. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–11.

TRAPNELL, C., A. ROBERTS, L. GOFF, G. PERTEA, D. KIM, D.R. KELLEY, H. PIMENTEL, ET AL. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7: 562–78.

TSAI, W.C., C.S. KUOH, M.H. CHUANG, W.H. CHEN, and H.H. CHEN. 2004. Four DEF-Like MADS box genes displayed distinct floral morphogenetic roles in *Phaldenopsis* orchid. *Plant and Cell Physiology* 45: 831–844.

TSAI, W.C., Z.J. PAN, Y.Y. HSIAO, M.F. JENG, T.F. WU, W.H. CHEN, and H.H. CHEN. 2008. Interactions of B-class complex proteins involved in tepal development in *Phalaenopsis* orchid. *Plant and Cell Physiology* 49: 814–824.

TUSKAN, G.A., S. DIFAZIO, S. JANSSON, J. BOHLMANN, I. GRIGORIEV, U. HELLSTEN, N. PUTNAM, ET AL. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596–1604.

VANNESTE, K., S. MAERE, and Y. VAN DE PEER. 2014. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 369: .

VEKEMANS, D., S. PROOST, K. VANNESTE, H. COENEN, T. VIAENE, P. RUELENS, S. MAERE, ET AL. 2012. Gamma Paleohexaploidy in the Stem Lineage of Core Eudicots: Significance for MADS-Box Gene and Species Diversification. *Molecular Biology and Evolution* 29: 3793–3806.

VELASCO, R., A. ZHARKIKH, J. AFFOURTIT, A. DHINGRA, A. CESTARO, A. KALYANARAMAN, P. FONTANA, ET AL. 2010. The genome of the domesticated apple (Malus × domestica Borkh.). *Nature Genetics* 42: 833–839.

VIAENE, T., D. VEKEMANS, V.F. IRISH, A. GEERAERTS, S. HUYSMANS, S. JANSSENS, E. SMETS, and K. GEUTEN. 2009. Pistillata-Duplications as a Mode for Floral Diversification in (Basal) Asterids. *Molecular Biology and Evolution* 26: 2627–2645.

VISION, T.J., D.G. BROWN, and S.D. TANKSLEY. 2000. The Origins of Genomic Duplications in *Arabidopsis*. *Science* 290: 2114–2117.

DE VOS, R.C.H., S. MOCO, A. LOMMEN, J.J.B. KEURENTJES, R.J. BINO, and R.D. HALL. 2007. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols* 2: 778–791.

VOZNESENSKAYA, E.V., N.K. KOTEYEVA, S.D. CHUONG, A.N. IVANOVA, J. BARROCA, L.A. CRAVEN, and G.E. EDWARDS. 2007. Physiological, anatomical and biochemical characterisation of photosynthetic types in genus Cleome (Cleomaceae). *Functional Plant Biology* 34: 247–267.

WAGNER, A. 2000. The role of population size, pleiotropy and fitness effects of mutations in the evolution of overlapping gene functions. *Genetics* 154: 1389–1401.

WANG, P., S. KELLY, J.P. FOURACRE, and J.A. LANGDALE. 2013. Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 Kranz anatomy. *The Plant Journal* 75: 656–670.

WANG, W., Y. WU, Y. LI, J. XIE, Z. ZHANG, Z. DENG, Y. ZHANG, ET AL. 2010. A large insert *Thellungiella halophila* BIBAC library for genomics and identification of stress tolerance genes. *Plant Molecular Biology* 72: 91–99.

WANG, X., U. GOWIK, H. TANG, J.E. BOWERS, P. WESTHOFF, and A.H. PATERSON. 2009. Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome Biol* 10: R68.

WANG, Y., A. BRÄUTIGAM, A.P.M. WEBER, and X.-G. ZHU. 2014. Three distinct biochemical subtypes of C4 photosynthesis? A modelling analysis. *Journal of Experimental Botany* 65: 3567–3578.

WEISS, M.R. 1995. Floral Color Change: A Widespread Functional Convergence. *American Journal of Botany* 82: 167–185.

WEISS, M.R. 1991. Floral colour changes as cues for pollinators. *Nature* 354: 227–229.

WEISS, M.R., and B.B. LAMONT. 1997. Floral Color Change and Insect Pollination: A Dynamic Relationship. *Israel Journal of Plant Sciences* 45: 185–199.

WEISS-SCHNEEWEISS, H., K. EMADZADE, T.-S. JANG, and G.M. SCHNEEWEISS. 2013. Evolutionary Consequences, Constraints and Potential of Polyploidy in Plants. *Cytogenetic and Genome Research* 140: 137–150.

WHITTALL, J.B., C. VOELCKEL, D.J. KLIEBENSTEIN, and S.A. HODGES. 2006. Convergence, constraint and the role of gene expression during adaptive radiation: floral anthocyanins in Aquilegia. *Molecular Ecology* 15: 4645–4657.

WILLIAMS, B.P., S. AUBRY, and J.M. HIBBERD. 2014. Molecular evolution of genes recruited into C4 photosynthesis. *Trends in Plant Science* 17: 213–220.

WOLFE, K.H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* 2: 333–341.

WOLFE, K.H., and D.C. SHIELDS. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387: 708–713.

WOODHOUSE, M.R., H. TANG, and M. FREELING. 2011. Different Gene Families in Arabidopsis thaliana Transposed in Different Epochs and at Different Frequencies throughout the Rosids. *The Plant Cell* 23: 4241–4253.

WRZACZEK, M., M. BROSCHE, J. SALOJARVI, S. KANGASJARVI, N. IDANHEIMO, S. MERSMANN, S. ROBATZEK, ET AL. 2010. Transcriptional regulation of the CRK/DUF26 group of receptor-like protein kinases by ozone and plant hormones in Arabidopsis. *BMC Plant Biol* 10: 95.

XU, W., C. DUBOS, and L. LEPINIEC. 2015. Transcriptional control of flavonoid biosynthesis by MYB–bHLH–WDR complexes. *Trends in Plant Science* 20: 176–185.

XU, Z., and H. WANG. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* 35: W265–W268.

YANG, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24: 1586–1591.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences: CABIOS* 13: 555–556.

YANG, Z., and J.P. BIELAWSKI. 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution* 15: 496–503.

YANG, Z., and R. NIELSEN. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* 17: 32–43.

YANG, Z., and R. NIELSEN. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* 46: 409–418.

ZDOBNOV, E.M., and R. APWEILER. 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17: 847–848.

ZHANG, F., A. GONZALEZ, M. ZHAO, C.T. PAYNE, and A. LLOYD. 2003. A network of redundant bHLH proteins functions in all TTG1-dependent pathways of Arabidopsis. *Development* 130: 4859–4869.

ZHANG, X., L. WANG, Y. YUAN, D. TIAN, and S. YANG. 2011. Rapid copy number expansion and recent recruitment of domains in S-receptor kinase-like genes contribute to the origin of self-incompatibility. *The FEBS Journal* 278: 4323–37.

ZHU, X.-G., S.P. LONG, and D.R. ORT. 2010. Improving photosynthetic efficiency for greater yield. *Annual review of plant biology* 61: 235–261.

ACKNOWLEDGEMENTS

Congratulations reader, you made it to the end. This is the bit where I thank the various people without who this thesis could have never been a reality. This has been a six year journey that has taught me so many things about science, career and life but most of all, persistence.

First and foremost, I want to thank my supervisor and promotor (and co-promotor) Eric Schranz. Both of us had no idea what we were getting into when I first stepped out of the elevator at the UvA six years ago. Even before that when I was in your Bachelor's course Eric during my 3rd year, I distinctly remember thinking: "Genome duplications? When does that ever happen? How important could those possibly be?" Little did I know. I want to thank you for all the guidance both scientifically and unscientifically, the many songs, puns, funny videos and cultural references which you have so graciously shared with me. You have built many bridges for me and without them I could have never made it to where I am now. I hope you, Paige, Esme and Graham will stay happy and healthy for a long time.

Next I need to thank my co-PhD, Johannes Hofberger. Jo, you have shown me a scientist's life need never be boring, as long as efficiency is maintained. Thank you for all the advice, discussion and comments and I hope to share marriage stories over a beer soon. Servus!

The groups I've been a part of have been amazing. Thomas, Yorike, Melis, Patrick, Gerard, Harold en Ludek from the IBED have helped me getting up and running into the next gear that is a PhD when coming from an MSc. Our Aethionema sampling trip in Turkey has been something I will never forget. Then my other group in Wageningen, Biosystematics. Setareh, who shared both groups with me, thank you for all your help and comments. I have full confidence in your amazing abilities as an Aethionema expert and I hope our works will be even more fruitful. Frank, thank you for all your work in the lab and the greenhouse. Your amazing Cleome pictures haven't made it to a journal cover (yet) but they made it to this thesis cover for which I am eternally grateful. Phuong, Tao, Rens, Wilma, Zhen, Nynke, Heo, Ronald, Lars, Freek, Sara, thank you for making us Amsterdam newbies feel welcome in Wageningen. The Biosystematics group is special and has taught me so many things that I never thought I would master.

Als laatste bedank ik mijn familie. Pap en mam, heel erg bedankt voor al jullie steun en onwijkbare geloof in mij. Het is voor mij en ons een enorme steun geweest dat jullie altijd klaar staan met raad en daad. Ik hoop dat jullie trots zijn op deze scriptie omdat hij ook voor een beetje van jullie is. Tom, bedankt voor alle steun en gezelligheid; ik hoop dat ik over een jaar of 6 op jouw PhD verdediging mag zijn. Alle familie van den Bergh en Sigmond, die altijd klaar staan voor wat hulp en een spelletje (met wellicht een borreltje erbij). Mijn geweldige en uitgebreide schoonfamilie, van Weele en Funnekotter en hoe zij mij hebben verwelkomd.

Als allerlaatste, twee mensen die het licht zijn van mijn leven. Eerst mijn allerliefste dochter, Florentine, nu nog zo klein en nu al zo slim en mooi. Ik hoop dat je samen met ons de allergelukkigste persoon op aarde word en als wij daar ook maar een klein beetje bij kunnen helpen dan doen we dat met liefde. Ook jij hebt bijgedragen aan het boekje dat hier voor je neus ligt en ik hoop dat als je dit leest je altijd zult denken aan je Papa die voor altijd van je houdt, wat er ook gebeurt. Rosanna, mijn liefde, mijn vrouw, mijn alles. Zonder jou had mijn leven er nooit zo ongelofelijk mooi uit kunnen zien. Jij bent mijn lichtbron in donkere dagen en ik hoop dat we voor altijd samen gelukkig zullen zijn. Ik weet dat ik dit van veel mensen hierboven zeg, maar zonder jou had dit boekje er NOOIT gekomen; Jij hebt me motivatie gegeven die ik anders niet had kunnen vinden. Ik hoop dat ons leven vol feest, schoonheid en geluk zal zitten en dat de onvermijdelijke donderwolken snel voorbij waaien. En anders slaan we ons er wel weer doorheen, het is tot nu immers ook gelukt. Ik kan niet wachten tot we getrouwd zijn, en ik hou van je.

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to: **Erik van den Bergh**
Date: **17 May 2017**
Group: **Laboratory of Biosystematics**
University: **Wageningen University & Research**

1) Start-up phase	*date*
► **First presentation of your project**	
Title: Patterns of postpaleopolyploidy in plants and the potential for photosynthetic processes	May 02, 2011
► **Writing or rewriting a project proposal**	
► **Writing a review or book chapter**	
► **MSc courses**	
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*1.5 credits**

2) Scientific Exposure	*date*
► **EPS PhD student days**	
EPS PhD student days 'Get2Gether', Soest, NL	Jan 29-30, 2015
► **EPS theme symposia**	
EPS theme 4 'Genome plasticity', Wageningen, NL	Dec 09 2011
EPS theme 4 'Genome plasticity', Nijmegen, NL	Dec 07 2012
EPS theme 4 'Genome plasticity', Wageningen, NL	Dec 13 2013
EPS theme 4 'Genome plasticity', Wageningen, NL	Dec 03 2014
► **Lunteren days and other National Platforms**	
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	Apr 04-05, 2011
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	Apr 02-03, 2012
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	Apr 22-23, 2013
► **Seminars (series), workshops and symposia**	
Ecogenomics day 2011	Jun 16, 2011
Plant Genome Evolution, Amsterdam, NL	Sep 04-06, 2011
EPS mini symposium plant breeding	Nov 25 2011
Ecogenomics day 2012	Jun 07, 2012
Biosystematics scientific retreat	Sep 10-11, 2012
Plant Genome Evolution (Amsterdam, The Netherlands)	Sep 08-10, 2013
► **Seminar plus**	
► **International symposia and congresses**	
Conference Comparative & Regulatory Genomics in Plants, Ghent, Belgium	Apr 11-12, 2011
Plant Animal Genome XXII, San Diego, USA	Jan 10-15 2014
Plant Genomics Congress, London, UK	May 12-13, 2014
► **Presentations**	
Talk: EPS theme 4 day: G. gynandra and T. hassleriana photosynthesis evolution	Dec 07, 2012
Talk: Plant Science Meeting	Jun 22 2012
Poster: Ancestral genome and work with A. arabicum and T.hassleriana PAG	Jan 13 2014

Talk: Ancestral genome and work with A. arabicum and T.hassleriana, London	May 13 2014
Talk: Ancestral genome and glucosinolate evolution, Theme 4 day	Dec 3 2014
Talk: Tarenaya and Aethionema: genome evolution of Brassicaceae from an outgroup perspective	May 21 2013
► **IAB interview**	
Meeting with a member of the International Advisory Board of EPS	Sep 29, 2013
► **Excursions**	
Trip to Anatholia for sampling of A. arabicum	May 14-22, 2011
Excursion to PICB Shanghai for Mapping of Cleome with BGI	Feb 11- Mar 09 2012
Visit to Tucson and Berkeley with collaborators	May 18–31, 2013
Subtotal Scientific Exposure	*19.5 credits**

3) In-Depth Studies	*date*
► **EPS courses or other PhD courses**	
Workshop Comparative & Regulatory Genomics in Plants	Apr 13-15, 2011
Current Trends in Phylogenetics	Oct 22-26 2012
NBIC Pattern Recognition	Jan 21-25, 2013
► **Journal club**	
Weekly lierature discussion club organized by Katja Peijnenberg	2011-2012
Literature discussion club Biosystematics	2013-2015
► **Individual research training**	
Visit to INRA Clermont Ferrand, France, mini internship with J. Salse	Apr 01-18, 2014
► **Individual research training**	
Subtotal In-Depth Studies	*9.9 credits**

4) Personal development	*date*
► **Skill training courses**	
Get acquinted with Supevising PhD students, Management skills and Career planning (organizer Career Fair for High School students), Communication skills; Earlham Institute, UK	2014-2016
► **Organisation of PhD students day, course or conference**	
Assisting in organization of Plant Genome Evolution	Sep 04-06, 2011
Assisting in organization of Plant Genome Evolution	Sep 08-10, 2013
► **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*5.5 credits**

TOTAL NUMBER OF CREDIT POINTS*	**36.4**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

** A credit represents a normative study load of 28 hours of study.*

The research described in thesis was financially supported by the Netherlands Organisation for Scientific Research.

Financial support from University of Amsterdam and Wageningen University is greatly appreciated.

Cover photograph by Frank Becker, frank.becker@wur.nl

Printed by Digiforce